Behavior Research Methods (2019) 51:2546–2558
https://doi.org/10.3758/s13428-018-1104-x

Do complex span and content-embedded working memory tasks predict unique variance in inductive reasoning?

Amanda Zamary[1] · Katherine A. Rawson[1] · Christopher A. Was[1]

Published online: 20 August 2018

Abstract

Complex span and content-embedded tasks are two kinds of tasks that are designed to measure maintenance and processing in the working memory system. However, a key functional difference between these task types is that complex span tasks require the maintenance of information that is not relevant to the processing task, whereas content-embedded tasks require the maintenance of task-relevant information. The purpose of the present research was to test the hypothesis that more unique variance in inductive reasoning would be explained by content-embedded tasks than by complex span tasks, given that inductive reasoning requires reasoners to maintain and manipulate task-relevant information in order to arrive to a solution. A total of 384 participants completed three complex span tasks, three content-embedded tasks, and three inductive reasoning tasks. The primary structural equation model explained 51% of the variance in inductive reasoning; 45% of the variance in inductive reasoning was uniquely predicted by the content-embedded latent factor, 6% of the variance was predicted by shared variance between the content-embedded and complex span latent factors, and less than 1% was uniquely predicted by the complex span latent factor. These outcomes provide a novel extension to the small but growing literature showing an advantage of using content-embedded rather than complex span tasks for predicting higher-level cognition.

Keywords Working memory · Inductive reasoning · Complex span tasks · Content-embedded tasks

People use inductive reasoning to make inferences and solve problems on a daily basis. Inductive reasoning involves reasoning from the particular to the general; it is an explicit process that involves the discovery of common relationships among stimulus elements via the formation and testing of hypotheses within a stimulus set (Carroll, 1993; Ekstrom, French, Harman, & Dermen, 1976; Johnson-Laird, 2013; Klauer & Phye, 2008; Klauer, Willmes, & Phye, 2002). During inductive reasoning, multiple elements and/or relations between elements are attended to and manipulated in order to derive a solution (Johnson-Laird, 2013; Klauer & Phye, 2008). Reasoners can adopt various strategies during inductive reasoning, such as systematically comparing the stimulus elements and the relations between elements; a more heuristic approach, in which the problem is examined globally and plausible hypotheses are generated and tested; or iterative combinations of more than one strategy (Klauer & Phye, 2008). Regardless of the strategy used, maintaining task-relevant information during processing is important for inductive reasoning to take place (Cowan, 1988; Johnson-Laird, 2013; Oberauer, 2002; Oberauer, Süß, Wilhelm, & Sander, 2007; Sternberg, 1986; Sternberg & Gardner, 1983). Inductive reasoning theories assume that the maintenance of information during processing is achieved by the working memory system (e.g., Johnson-Laird, 2013; Sternberg, 1986; Sternberg & Gardner, 1983).[1]

Maintenance (achieved by the working memory system) is important for inductive reasoning because task-relevant information must be maintained and combined in order to derive a solution (Sternberg, 1986). Consistent with the assumption that maintenance is important for inductive reasoning, a wealth of research has shown strong positive relationships

✉ Amanda Zamary
azamary@kent.edu

[1] Department of Psychological Sciences, Kent State University, P.O. Box 5190, Kent, OH 44242-0001, USA

[1] In a recent review of the working memory literature, Cowan (2017) discusses commonly adopted definitions of working memory and recommends that researchers explicitly state the definition they are adopting to improve conceptual clarity in the field. Following this call for specificity, we adopt what Cowan refers to as the generic working memory definition (Cowan, 1988, 2017), which states that working memory is "the ensemble of components of the mind that hold a limited amount of information temporarily in a heightened state of availability for use in ongoing information processing" (Cowan, 2017, p. 1163).

between reasoning and working memory (e.g., Ackerman, Beier, & Boyle, 2002, 2005; Conway, Cowan, Bunting, Therriault, & Minkoff, 2002; Kane, Hambrick, & Conway, 2005; Oberauer et al., 2007; Unsworth & Engle, 2005). Importantly, most evidence indicates that reasoning (including inductive reasoning) and working memory are highly related but clearly separable constructs (Ackerman et al., 2005; Kyllonen & Kell, 2017; Oberauer, Schulze, Wilhelm, & Süß, 2005; but see Kyllonen & Christal, 1990).

Many of the studies investigating the relationship between working memory and reasoning have examined general reasoning, which can encompass several different subtypes of reasoning (e.g., deductive reasoning, inductive reasoning, and analogical reasoning). Thus, the tasks used in much of this research were not limited to inductive reasoning tasks (see Ackerman et al., 2005, for a comprehensive meta-analysis), although most studies have included at least one inductive reasoning task. For instance, one widely used task within this area of research is the Raven's Progressive Matrices (Raven, Court, & Raven, 1977), a version of which was used in the present research. On each trial of this task, reasoners see a 3×3 matrix in which eight cells contain figures that differ in shape composition, shading, and/or size, and the ninth cell is left empty. The reasoner is given eight additional figures and is asked to identify which figure correctly completes the matrix (on the basis of one or more unspecified rules that determine the relationships between the figures in the matrix). The two other tasks used to measure inductive reasoning in the present research (discussed in greater detail below) are the letter sets and locations tasks from the Kit of Reference Tests for Cognitive Factors (Ekstrom et al., 1976; see also Carroll, 1993; Foster et al., 2015; Harrison et al., 2013; Was, Dunlosky, Bailey, & Rawson, 2012).

Working memory is most frequently measured using *complex span tasks* (Conway et al., 2005; Kane et al., 2004; Shipstead, Harrison, & Engle, 2016). Complex span tasks are a type of working memory task that involves both storage (i.e., maintenance) and processing demands (e.g., Cowan, 2017; Daneman & Carpenter, 1980). Important for present purposes, a critical feature of complex span tasks is that the information being maintained is independent from the information being processed. To illustrate, consider the reading span task (Conway et al., 2005; Kane et al., 2004). In this task, participants are presented with sentences one at a time and are asked to identify whether the sentence makes sense (i.e., the processing component of the task). After each sentence, participants are shown a word to remember for later recall (i.e., the maintenance component of the task). After each block of sentences, participants are asked to recall the to-be-remembered words in serial order. Although participants are told to complete both components as accurately as possible, working memory is measured as performance on the maintenance component. By this measure, working memory reflects the ability to maintain information that is irrelevant to the information being processed in the working memory system.

Another way that researchers have measured working memory is through *content-embedded tasks* (Ackerman et al., 2002; Kyllonen & Cristal, 1990; Was, Rawson, Bailey, & Dunlosky, 2011; Woltz, 1988). Similar to complex span tasks, content-embedded tasks also involve both maintenance and processing demands. In contrast to complex span tasks, the information being maintained for output is the same information that is being processed. To illustrate, consider the ABCD task. On each trial, participants are shown three pieces of information, one at a time, that specify the ordering of the same four letters (ABCD). The first piece of information states the ordering of the letters A and B (e.g., "B comes before A"). The second piece of information states the ordering of the letters C and D (e.g., "D comes after C"). The third piece of information states the ordering of the two sets of letters (e.g., Set 1 comes after Set 2). The participant is then asked to indicate the correct solution (in this case, CDBA). Note that the information being processed (i.e., the ordering of letters and sets of letters) is the same information that is being maintained for output (e.g., CDBA). This measure of working memory reflects the ability to maintain and process *task-relevant* information in the working memory system. This task characteristic differs from complex span tasks (in which the information being maintained is *task-irrelevant*).

Although content-embedded (e.g., Kyllonen & Christal, 1990) and complex span (e.g., Engle, Tuholski, Laughlin, & Conway, 1999) tasks both correlate with measures of reasoning, the vast majority of prior research has used complex span tasks to measure working memory. Given the assumption that working memory is important for inductive reasoning because of its role in maintaining task-relevant information to derive a solution (e.g., Sternberg, 1986), we hypothesize that working memory tasks that emphasize the maintenance and processing of the same information (i.e., content-embedded tasks) would predict more variance in inductive reasoning than do tasks that emphasize the maintenance of task-irrelevant information (i.e., complex span tasks). Importantly, no prior research has simultaneously investigated the predictive power of complex span and content-embedded tasks in inductive reasoning.

Although no prior research has investigated how well these kinds of working memory tasks predict inductive reasoning, prior research has investigated how well these kinds of tasks predict other complex cognitive processes. Was, Rawson, Bailey, and Dunlosky (2011) investigated the extent to which these two task types predicted reading comprehension. Similar to the argument proposed in the present research, Was et al. (2011) hypothesized that reading comprehension would be predicted better by content-embedded than by complex span tasks, because reading comprehension requires the maintenance of task-relevant information. As hypothesized,

reading comprehension was predicted better by content-embedded than by complex span tasks.

Although this finding provides indirect evidence supporting our hypothesis that inductive reasoning will be predicted better by content-embedded than by complex span tasks, reading comprehension and inductive reasoning are distinct constructs that may differentially rely on other cognitive processes. For instance, reading comprehension loads onto a crystalized intelligence factor, whereas inductive reasoning loads onto a fluid intelligence factor (Carroll, 1993). Thus, the extent to which inductive reasoning is predicted better by content-embedded than by complex span tasks remains an open question awaiting direct empirical investigation.

The purpose of the present research was to test the hypothesis that more unique variance in inductive reasoning would be explained by content-embedded than by complex span tasks. Given that both tasks are designed to measure maintenance and processing in the working memory system, we predicted that content-embedded and complex span tasks would share some variance explaining inductive reasoning performance. However, given that content-embedded tasks measure the maintenance and processing of the same information (which is central in inductive reasoning tasks), whereas complex span tasks measure the maintenance of task-irrelevant information, we predicted that content-embedded tasks would also predict unique variance in inductive reasoning performance.

Method

Participants

Participants were recruited from the Psychology Department's participant pool and received course credit for participation. The full sample included 384 students from a large Midwestern university (68% female; 70% white, 14% black, 5% Asian, 4% First Nations, 2% Hispanic or Latino, 1% native Hawaiian or Pacific Islander); 36% were in their first year of college (M = 2.2, SE = 0.1), and 30% were psychology majors. The mean age of participants was 19.9 years (SE = 0.1), and the sample size was determined by rule of thumb for conducting large individual differences studies (n = around 300). We oversampled in order to account for attrition and noncompliance. Most importantly, we did not analyze the data until the full sample was collected.

Materials and procedure

Complex span tasks The complex span tasks used in the present research were versions of the span tasks described in Kane et al. (2004). Each trial of the reading span task (RSPAN) included a set of sentences. The set size ranged from two to six sentences. Sentences were presented individually, and participants were asked to read each sentence silently and then to click a button to indicate whether the sentence made sense (e.g., "Mr. Owens left the lawnmower in the lemon"). Across all trials, half of the sentences made sense, and half did not. If participants did not respond within 4 s, the computer automatically moved them forward. After each sentence, participants were presented with an unrelated word (e.g., *eagle*) for 1 s that they were asked to remember for later recall. At the end of the sentence set, participants were prompted to recall the words in the order in which they had been presented. Participants completed 15 trials, with one trial of each set size in each of three blocks. Trials were presented in a fixed random order within each block.

Each trial of the operation span task (OSPAN) included a set of mathematical expressions. The set size ranged from two to five mathematical expressions. Mathematical expressions were presented individually, and participants were asked to read each expression silently and then to click a button to indicate whether it was correct (e.g., "Is $(4 \times 2) + 5 = 10$?"). Across all trials, half of the expressions were correct, and half were not. If participants did not respond within 4 s, the computer automatically moved them forward. After each expression, participants were presented with a word (e.g., *phone*) for 1 s that they were asked to remember for later recall. At the end of each set of mathematical expressions, participants were prompted to recall the words in the order in which they had been presented. Participants completed 12 trials, with one trial of each set size in each of three blocks. Trials were presented in a fixed random order within each block.

Each trial of the counting span task (CSPAN) included a set of arrays. The set size ranged from two to six arrays; each array was presented individually for as much time as the participant needed. However, participants were told to complete each array as quickly as possible. Each array was composed of a random assortment of squares and circles, including three to nine dark blue circles, a varying number of light blue circles, and a varying number of dark blue squares (the arrays were the same across participants). Participants were asked to count the dark blue circles in each array, clicking on each one as it was counted. A checkmark appeared on the circle to show the participant that that circle had been counted. After they finished counting the dark blue circles in the array, a new array appeared. Participants were asked to remember the number of dark blue circles in each array for later recall. At the end of the array set, participants were prompted to recall the numbers in the order in which they had been presented. Participants completed 15 trials, with one trial of each set size in each of three blocks. Trials were presented in a fixed random order within each block.

The scores on all complex span tasks were computed using partial-credit load scoring (see Conway et al., 2005, for discussion). Additionally, we used serial recall scoring;

participants only received credit for items recalled in their correct ordinal position. Furthermore, participants were only given credit for items that were spelled entirely correctly on the RSPAN and OSPAN, due to ambiguity concerning whether misspellings reflected semantic or typographical errors (e.g., *bean* or *beat* for the target word *bear*). The scores on all complex span tasks were entered into the model as percentages correct.

Content-embedded tasks The content-embedded tasks used in the present research were versions of content-embedded tasks that had been used as measures of working memory in previous research (e.g., Ackerman et al., 2002; Kyllonen & Christal, 1990; Was et al., 2011; Was & Woltz, 2007; Woltz, 1988). The stimuli for all content-embedded tasks are available online at https://osf.io/gcav6/.

On each trial of the ABCD task, participants were required to process three pieces of information to determine the ordering of four letters (A, B, C, and D). First, participants were given the ordering of the letters A and B (e.g., "B comes before A"). Participants clicked a button to replace the first statement with one giving the ordering of the letters C and D (e.g., "D comes after C"). Participants again clicked a button to replace the second statement with one giving the ordering of the two pairs of letters (e.g., "Set 1 comes after Set 2"). Participants clicked a button to advance to the next screen, which showed the eight possible orderings of A, B, C, and D. Participants were asked to select the correct answer, and then the cycle repeated for the next trial. All screens on each trial were self-paced; however, participants were told to respond as quickly as possible. Participants completed 23 trials; the letter and set orderings varied by trial and were presented in the same fixed, random order across participants.

On each trial of the alphabet task, participants were asked to transform sets of letters. Participants were presented with one or two nonadjacent letters from the alphabet with a transformation direction and number (e.g., "T forward 3"; "OZ backward 2"; the answers are W and MX, respectively). Once participants had solved the transformation, they clicked a button to advance to the next screen, which included eight response options. Participants had up to 5 s to select the correct answer; if they did not select an answer, they were automatically moved forward and the trial was counted as incorrect. Participants completed 12 trials in each of two blocks (each block contained both one- and two-letter trials). Letters and transformations varied by trial and were presented in the same fixed, random order across participants.

On each trial of the digit task, participants were asked to answer one or two questions about a string of numbers. Participants were presented with six single-digit numbers for 2 s each (e.g., "5, 8, 1, 4, 9, 8"). After the presentation of the digit string, participants were asked one or two questions about the number string (e.g., "How many even numbers were there?," "What is the smaller of the middle two numbers?"). If the trial involved two questions, the questions were presented individually. All answers were numeric, and participants answered by typing in the correct answer. This phase of the task was self-paced, but participants were asked to answer as quickly and accurately as possible. Participants completed a block of 12 single-question trials and then a block of 12 double-question trials. The questions varied by trial and were presented in a fixed, random order across participants.

The scores on all content-embedded tasks were computed as the number of correct responses per minute, and all participants included in the final analyses completed all trials in all content-embedded tasks. For the digit task, minutes were computed as the time spent on the response screen (given that digit presentation times were fixed). Prior research using content-embedded tasks indicates that meaningful individual differences are captured in both speed and accuracy on these tasks (see Vandierendonck, 2017; Was & Woltz, 2007).

Inductive reasoning tasks The scores on all three inductive reasoning tasks were computed as percentages correct. We used the short form of Raven's Advanced Progressive Matrices (RAPM; Raven, 1962, Set II), used by Stanovich and Cunningham (1992). In brief, Stanovich and Cunningham dropped 18 of the least and most difficult items, given the frequent floor and ceiling effects in college students on these items. On each trial, participants saw a 3×3 matrix, with the first eight cells containing figures differing in shape composition, shading, and size. Eight additional figures were presented below the matrix. Participants were asked to click on the figure that correctly completed the pattern in the matrix. Participants could complete up to 18 trials and were given up to 12 min to complete the task. Trials were presented in ascending order, from least to most difficult.

On each trial of the locations task (Carroll, 1993; Ekstrom et al., 1976), participants were asked to extract a pattern from an array of Xs and dashes (see Fig. 1a for a sample trial). Each array included four rows, and each row contained sets of dashes with an X inserted within one of the sets. The placement of the X in each row was determined by an unstated rule (e.g., in Fig. 1a, the rule is to place the X in the second set of dashes in the position $n + 1$ from the previous row). Below the array, participants were presented with a fifth line that included a set of dashes with the numbers 1 through 5 dispersed in five locations. Participants were asked to figure out the rule and then to select the number that indicated where the X should be placed, given the rule (e.g., in Fig. 1a, the answer is 3). Participants were instructed that the task goal was to get a high score on the test, but to skip a problem if they were unsure of the answer, because they would be penalized for answers that were incorrect. Participants could complete up to 14 trials in each of two blocks and had up to 5 min to spend on each block of trials. If a participant skipped one or more

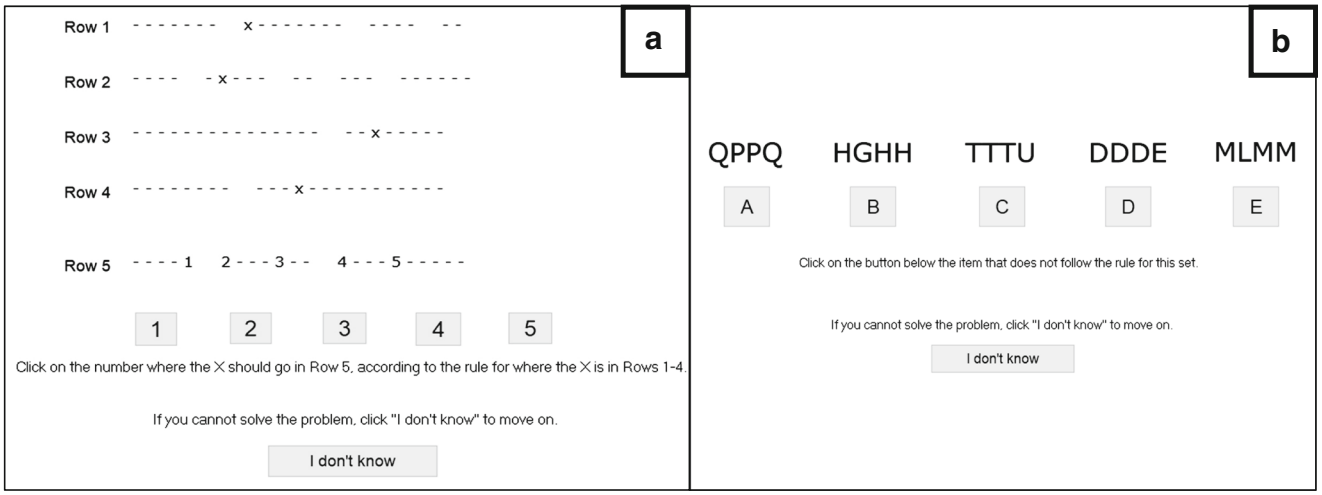


Fig. 1 Sample trials from the locations and letter sets tasks. (a) Locations task. (b) Letter sets task

trials and had time left within the 5-min block, the skipped trials were presented again until either the participant had selected an answer or the 5 min were up.

In each trial of the letter sets task (Carroll, 1993; Ekstrom et al., 1976), participants received five sets of four letters (see Fig. 1b for a sample trial). A rule determined the composition of four of the sets of letters, and one set did not follow the rule (e.g., in Fig. 1b, the rule is three copies of one letter plus one copy of a different letter, and the letter set QPPQ does not follow this rule). Participants were asked to figure out the rule and then to click on the set of letters that did not follow the rule (e.g., in Fig. 1b, the answer is A). Participants were instructed that the task goal was to get a high score on the test, but to skip a problem if they were unsure of their answer, because they would be penalized for answers that were incorrect. Participants could complete up to 15 trials in each of two blocks and had up to 5 min to spend on each block of trials. If a participant skipped one or more trials and had time left within the 5-min block, the skipped trials were presented again until either the participant had selected an answer or the 5 min were up.

The data reported were collected as a part of a larger individual differences study. Participants completed additional tasks that are not relevant for the present purposes and will not be reported in this article.[2] The entire study involved four sessions across a two-week period. Participants did not complete more than one task for any given latent factor during the same session (Session 1: alphabet and locations; Session 2: ABCD, OSPAN, and RAPM; Session 3: RSPAN and letter sets; Session 4: digit and CSPAN).

Results

Prior to conducting the analyses, we examined the data for attrition and evidence of noncompliance. Participants were excluded from analyses if they had more than one missing value from a single latent factor, either due to attrition (n = 36) or due to computer error (n = 1). Of the remaining 347 participants, 13 were excluded from the analysis, given evidence of noncompliance on more than one measured variable on a single latent factor [i.e., for RSPAN and OSPAN tasks, the participant did not respond to more than 60% of the processing trials; for the alphabet task, participant spent less than 90 s on the entire task (including the instructions); for the locations and letter sets tasks, participant spent less than 60 s on the first block (including the instructions) and/or 30 s on the second block; for the RAPM, participant spent less than 120 s on the entire task (including the instructions and practice problems)]. Instead of excluding participants who showed noncompliance on a single measure, we treated that single measured variable as missing data. In total, 34 participants had missing data but were still included in the analysis, given that they were missing no more than one measured variable per latent factor (n = 14 due to attrition, n = 19 due to noncompliance, n = 1 due to a lost data file). No more than 4% of the data were missing for each measured variable. Less than 2% of the data were missing from the entire set of data.

The final sample included 334 participants. Given that we used structural equation modeling and that the parameter estimates were derived using maximum likelihood, a minimum of five cases per parameter estimate is recommended (Mueller & Hancock, 2010). Our sample size well exceeded the minimum requirement for the model to be tested (i.e., 21

[2] For purposes of full disclosure (see Simmons, Nelson, & Simonsohn, 2011), the other tasks from which data were collected included example-based learning tasks and final comprehension tests, as well as individual differences measures of verbal ability and spelling. These data have not yet been analyzed and will be reported elsewhere. For purposes of clarification, we also note that the data reported here are new data from a different sample of participants than were used in the study by Was et al. (2011). The data reported in this article are available online at https://osf.io/4wqp7/.

estimated parameters, with 16 cases per parameter). All analyses were conducted in MPlus version 7.31 (Muthén & Muthén, 2015).[3] The values for missing data were estimated using full-information maximum likelihood.

Preliminary analyses

To ensure that participants were complying with the task instructions and engaging in both the storage and processing tasks for the RSPAN and OSPAN, we checked performance on the processing component of each task. High performance on the processing components of these tasks suggested that participants were complying with the task instructions (RSPAN: $M = 84\%$, $SD = 12$; OSPAN: $M = 77\%$, $SD = 15$). Performance on both components of the RSPAN and OSPAN was similar to the performance found in previous research (e.g., Lewandowsky, Oberauer, Yang, & Ecker, 2010).

Table 1 includes summary statistics, zero-order correlations, and reliability estimates for the measured variables. Importantly, the three measures composing each latent factor correlated highly with each other. [4] We also screened for univariate normality and multivariate normality. Concerning univariate normality, the skewness statistics on each measured variable were all smaller than 1.6, and the kurtosis statistics were all smaller than 3.2, meeting the assumption of univariate normality for the use of maximum likelihood. Concerning multivariate normality, Mardia's measures of multivariate skewness and kurtosis were significant ($z = 652.57$, $p < .001$, and $z = 13.05$, $p < .001$, respectively), indicating multivariate nonnormality. To ensure that multivariate nonnormality did not affect the qualitative pattern of findings, we also calculated estimates for all primary models using 500 bootstrap samples. The parameter estimates were similar following bootstrapping, and the 95% confidence intervals for the standard errors of the regression coefficients indicated that significant parameter estimates were not affected by the bootstrap sampling.

Structural equation modeling

Primary model Both the complex span and content-embedded latent factors were expected to predict inductive reasoning; accordingly, the primary model included paths for both of these directional effects. Additionally, given that we predicted that both complex span and content-embedded tasks measure some of the same facets of the working memory system, the complex span and content-embedded latent factors were expected to correlate with one another. Accordingly, the model included a path for this nondirectional effect. Fig. 2 depicts the hypothesized model with standardized path coefficients and estimated factor correlations.

Concerning model fit, the chi-square test of model fit indicated that the model did not fit the data well (see Table 2, Model 1). However, two limitations of the chi-square test of model fit include (1) that it assumes multivariate normality, and even slight deviations from the specified model may produce large chi-square values, and (2) that it is overly strict when the sample size is large (Bentler & Bonett, 1980; McIntosh, 2006). Given that the multivariate normality assumption was not met and that the sample size was large, other model fit indices were more appropriate. Importantly, all other model fit indices indicated that the model fit the data well (see Table 2, Model 1). All measured variables significantly loaded onto their respective latent factor, and the latent factors were strongly correlated with one another.

All model relationship statistics are reported using standardized estimates. As predicted, the complex span and content-embedded latent factors were strongly correlated ($r = .75$, $p < .001$). Of primary interest, we predicted that the content-embedded latent factor would strongly predict inductive reasoning, given that content-embedded tasks involve the maintenance of task-relevant information. Indeed, the content-embedded latent factor uniquely predicted inductive reasoning [$\beta = .67$, $SE = .13$, $p < .001$; 95% CI: (.42, .93)]. Interestingly, the complex span latent factor did not uniquely predict inductive reasoning [$\beta = .06$, $SE = .13$, $p = .65$; 95% CI: (– .21, .32)]. In total, the model predicted 51% of the variance in inductive reasoning: 45% of the variance was uniquely explained by the content-embedded factor, 6% was explained by overlapping variance between the latent factors, and less than 1% (0.004%) was uniquely explained by the complex span factor.

This statement was verified by testing the two models presented in Fig. 3. In the first model, we cross-loaded content-embedded tasks onto the complex span latent factor (see Fig. 3a). When tasks are loaded in this way, the content-embedded factor only reflects variance unique to content-embedded tasks, whereas the complex span factor reflects both variance unique to complex span tasks and overlapping variance between the two factors. In this model, the complex span factor explained 31% of the variance in inductive reasoning ($\beta = .56$). More importantly, the content-embedded factor still predicted 21% of the variance above and beyond the variance explained by the complex span factor ($\beta = .46$). This finding indicates that the content-embedded latent factor still uniquely predicted a substantial amount of variance in inductive reasoning, even when overlapping variance between the factors was allotted to the complex span factor ($p < .001$).

[3] The primary analyses were also conducted in AMOS version 22, and the values were almost identical to those found in Mplus (Arbuckle, 2013).

[4] The one exception was the locations task. The zero-order correlations between locations and the other inductive reasoning tasks were somewhat weaker than expected. To foreshadow, this task significantly loads onto the inductive reasoning factor, although the loading was numerically weaker than expected. All three of the inductive reasoning tasks have been used to compose a single latent factor in previous research in which the locations task loaded more strongly (Was et al., 2012). Most importantly, neither of the working memory factors would be differentially disadvantaged by this loading, given that the task was part of the inductive reasoning factor.

Table 1 Means, standard deviations, and correlations of the nine measured variables

Variable	*M*	*SD*	1	2	3	4	5	6	7	8	9
1. RSPAN	64	19	**(.86)**	.76	.58	.46	.35	.54	.30	.23	.40
2. OSPAN	81	16	.63	**(.79)**	.57	.50	.46	.52	.28	.27	.35
3. CSPAN	80	17	.46	.46	**(.87)**	.42	.26	.49	.35	.28	.35
4. ABCD	3	1	.40	.42	.36	**(.89)**	.53	.59	.39	.34	.48
5. Alphabet	4	1	.29	.37	.22	.45	**(.81)**	.52	.18	.24	.45
6. Digit	8	3	.47	.43	.43	.52	.45	**(.87)**	.30	.31	.42
7. RAPM	34	17	.23	.21	.28	.31	.14	.23	**(.72)**	.24	.56
8. Locations	38	14	.17	.19	.21	.26	.17	.23	.15	**(.64)**	.38
9. Letter sets	55	15	.34	.28	.30	.41	.37	.36	.42	.28	**(.83)**

Complex span (Variables 1–3) and inductive reasoning scores (Variables 7–9) are out of 100%. Content-embedded scores (Variables 4–6) are number correct per minute. All *p*s < .008. Internal reliability estimates were computed using Cronbach's alpha and are presented on the diagonal (bolded and in parentheses). The observed correlations are presented below the diagonal. The correlations corrected for attenuation are presented above the diagonal

In the second model, we cross-loaded complex span tasks onto the content-embedded latent factor (see Fig. 3b). When tasks are loaded in this way, the complex span factor only reflects variance unique to complex span tasks, whereas the content-embedded factor reflects both variance unique to content-embedded tasks and overlapping variance between the two factors. In this model, the content-embedded factor explained 52% of the variance in inductive reasoning (β = .72). Importantly, the complex span factor now predicted less than 1% of the variance above and beyond the variance explained by the content-embedded factor (β = .02), indicating that the complex span latent factor does not predict variance in inductive reasoning when overlapping variance between the factors was allotted to the content-embedded factor (p = .80). Collectively, the models in Fig. 3 support our hypothesis that more unique variance in inductive reasoning is explained by a content-embedded latent factor than by a complex span latent factor.

To provide further evidence that inductive reasoning is predicted better by the content-embedded latent factor, we conducted an additional set of models. First, we conducted a model that included only the complex span latent factor and inductive reasoning. Consistent with prior research (e.g., Conway et al., 2002; Kane et al., 2005; Kane et al., 2004; Was et al., 2012), the complex span latent factor significantly predicted inductive reasoning (b = .23, SE = .05, p < .001; r = .56, SE = .07). Next, we conducted another model in which we added the content-embedded latent factor back into the model and constrained the relationship between the complex span latent factor and the inductive reasoning latent factor to be

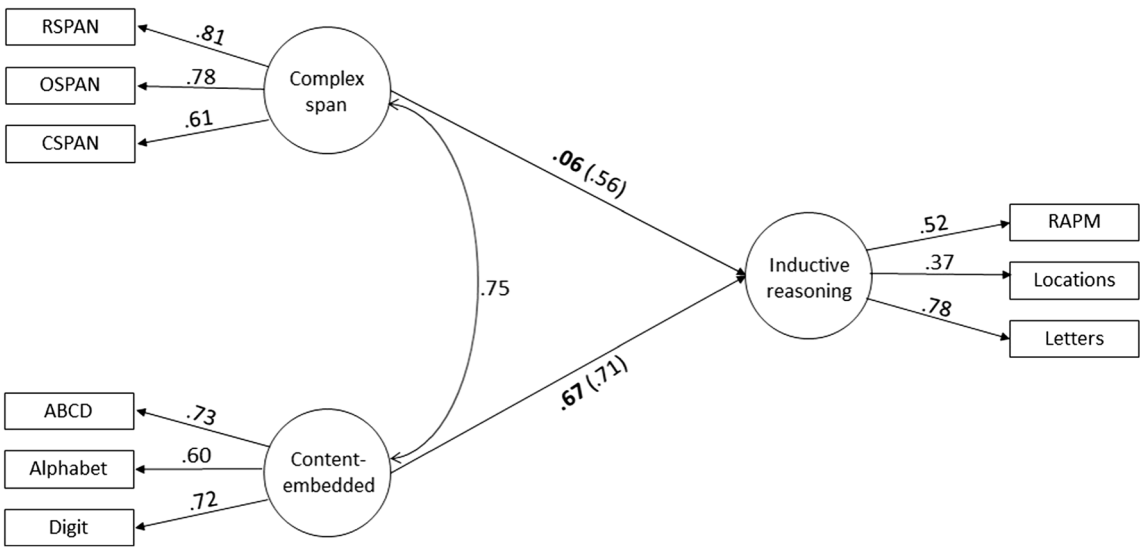
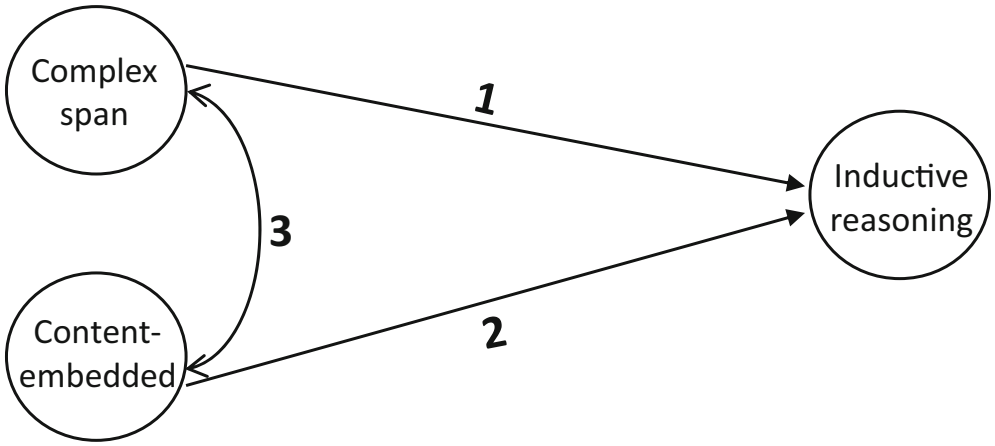


Fig. 2 Hypothesized model displaying standardized parameter estimates (error variances are not displayed in the figure). Estimated factor correlations are shown in parentheses (these values indicate zero-order correlations between latent factors). Standardized path coefficients are shown in bold type (these values indicate relationships between the latent factors when all latent factors are in the model)

Table 2 Model fit indices and path coefficients for the primary model (Model 1) and the alternative models (Models 2–7)



Model Number	Path 1	Path 2	Path 3	RMSEA	CFI	AIC	χ^2 Test of Model Fit
1	**.06**	**.67***	**.75***	**.05 (.03, .08), p = .42**	**.97**	**19,732**	**$\chi^2(24) = 45.55$, p = .005**
2	.10	.63*	.86*	.05 (.03, .07), p = .46	.98	19,380	$\chi^2(24) = 44.30$, p = .007
3	– .15	.90*	.81*	.09 (.07, .11), p = .002	.92	19,644	$\chi^2(24) = 81.96$, p < .001
4a	– .16	.83*	.83*	.06 (.04, .08), p = .17	.97	16,135	$\chi^2(24) = 54.95$, p < .001
4b	– .29	.95*	.88*	.06 (.04, .08), p = .24	.97	15,817	$\chi^2(24) = 51.61$, p < .001
5	– .17	.88*	.80*	.07 (.05, .09), p = .10	.96	19,746	$\chi^2(24) = 59.16$, p < .001
6	.02	.70*	.73*	.05 (.02, .08), p = .53	.98	17,111	$\chi^2(17) = 29.49$, p = .03
7	n/a	n/a	n/a	.09 (.08, .11), p < .001	.91	19,783	$\chi^2(26) = 100.79$, p < .001

The values for Paths 1, 2, and 3 are listed as standardized beta coefficients. The preferred values for the fit indices are as follow: Root Mean Square Error of Approximation (RMSEA): between .05 and .00; Comparative Fit Index (CFI): between .95 and 1.00; Akaike Information Criterion (AIC): smaller values; χ^2 test of model fit: smaller values and non-significant (see Mueller & Hancock, 2010, for additional information on model fit indices). Numbers listed in parentheses in the RMSEA column reflect 90% confidence intervals. * Path is statistically significant at p < .05

equal to the unstandardized parameter estimate when the content-embedded latent factor was not in the model (b = .23). A chi-square difference test between the freely estimated primary model and this fixed parameter model indicated that the primary model fit the data better [$\Delta\chi^2(1) = 11$, p < .01].

Taken together, the findings across these models indicated that inductive reasoning is predicted better by the content-embedded latent factor than by the complex span latent factor.

Alternative scoring methods and models Although we attributed the primary results to functional differences between content-embedded and complex span tasks, we conducted a series of models to rule out alternative, artifactual explanations for why inductive reasoning was predicted better by the content-embedded latent factor compared to the complex span latent factor. To facilitate comparison between the models, the basic outcomes and model fit statistics for all models are provided in Table 2.

First, although complex span tasks have traditionally been scored as percentages correct on the maintenance portion of the task, meaningful individual differences may also be found in performance on the processing portion of the task. If performance on both the maintenance and processing portions of the complex span tasks were taken into account, the complex span factor might be a better predictor of inductive reasoning. To ensure that the same results would hold when processing performance was taken into account in complex span task scores, we ran a model in which complex span task scores included performance on both the maintenance and processing portions of the tasks.[5] This model yielded the same qualitative pattern of results as Model 1 (see Table 2, Model 2).

Another plausible reason for why inductive reasoning was better predicted by the content-embedded latent factor could be because scores on the content-embedded tasks

[5] RSPAN and OSPAN scores were computed as averages between the percentage correct on the maintenance task and the percentage correct on the processing task. Note that an integrated measure for the CSPAN would be redundant with maintenance scores alone, given that participants were required to perform at 100% on the processing task (i.e., by clicking all of the dark blue circles prior to moving forward to the next screen). Therefore, in models involving a combined maintenance-and-processing score for complex span tasks, CSPAN scores are the same as in Model 1.

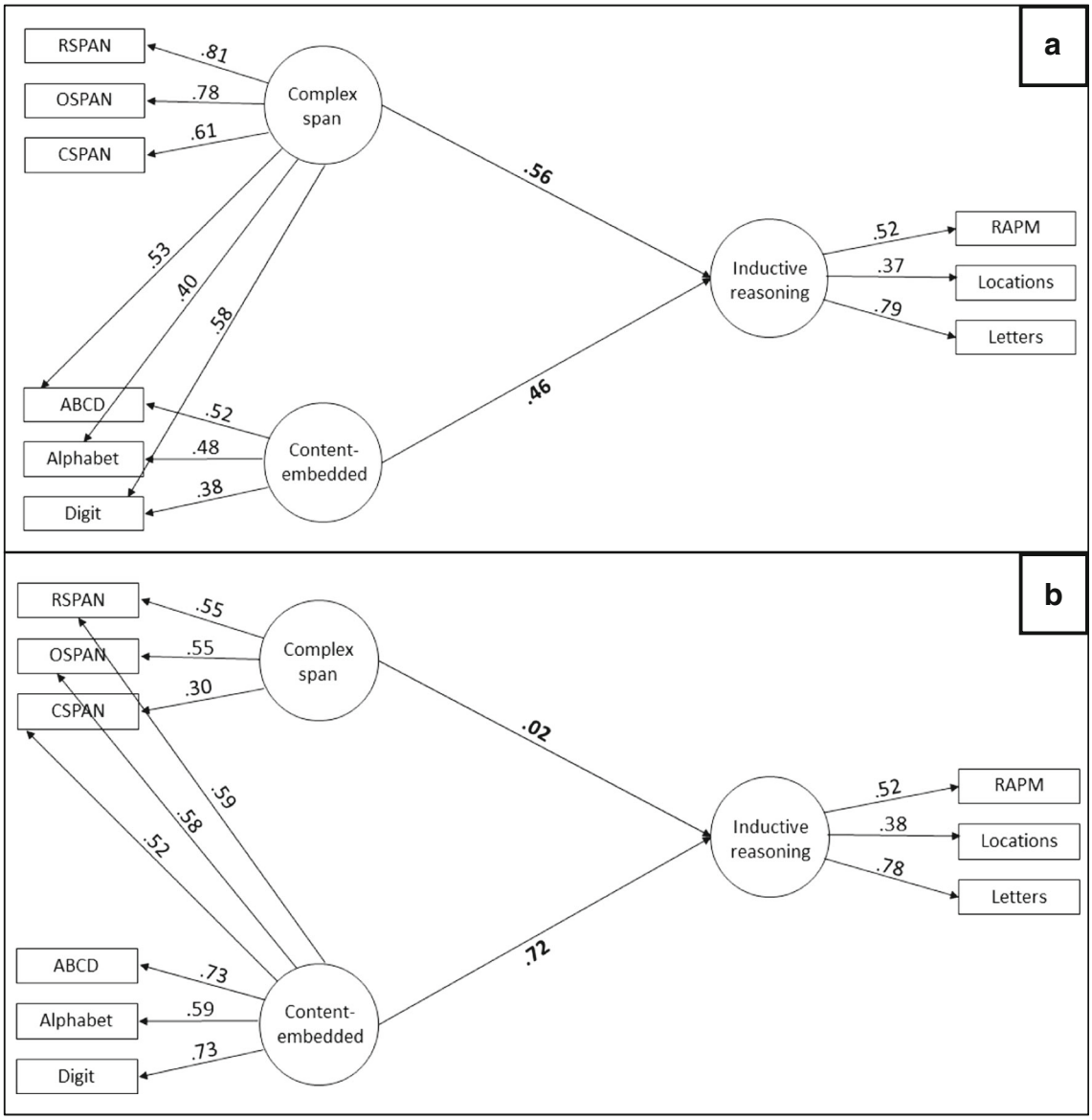


Fig. 3 Models with cross-loaded tasks. Standardized path coefficients are shown in bold type (error variances are not included)

included some variation in processing speed (as they were computed as the number correct per minute), whereas scores on the complex span tasks did not (as they were computed as percentages correct). To ensure that the inclusion of processing speed was not the reason why inductive reasoning was predicted better by the content-embedded latent factor, we conducted a series of three models (i.e., Models 3, 4a, and 4b).

In Model 3, we scored complex span tasks as percentages correct on the maintenance portion of the task and content-embedded tasks also as percentages correct (i.e., processing speed is not taken into account in both factors). In Models 4a and 4b, we computed complex span task scores as percentages correct per minute and the used number correct per minute scores for the content-embedded tasks (i.e., taking processing speed into account in both factors). The percentage correct portion of the scores in Model 4a was based only on performance on the maintenance portion of the tasks, whereas the percentage correct scores in Model 4b were based on performance in both the maintenance and processing portions of the tasks. Models 3, 4a, and 4b all yielded the same qualitative pattern of results as Model 1 (see Table 2), suggesting that processing speed is not the reason why inductive reasoning was predicted better by the content-embedded latent factor than by the complex span latent factor.

Models 5 and 6 concerned the use of the CSPAN as part of the complex span latent factor. In brief, we used the same three complex span tasks as in the earlier study by Was et al. (2011), given that these three tasks are commonly used together to create latent complex span factors (for a discussion, see Was et al., 2011). Although the CSPAN did not load strongly onto

the complex span factor in Was et al. (2011), we retained the CSPAN in the present study in order to allow for the possibility that the weaker factor loading in their study was spurious (particularly given that CSPAN performance was near ceiling (92%) in that study). The CSPAN loaded onto the complex span latent factor more strongly in the present study than in Was et al. (2011; .61 as compared to .47, respectively). However, Was et al. (2011) found better model fit when the CSPAN was loaded onto the content-embedded rather than the complex span factor. They also conducted a follow-up model in which they removed the CSPAN from the model altogether, but they found the same qualitative pattern of results as in their primary model. We conducted these same models with the present data set. The model fit was not improved by loading the CSPAN onto the content-embedded latent factor (Model 5). However, the model fit was improved by removing the CSPAN from the model altogether (Model 6). One possible reason why the model fit could have been negatively impacted by having the CSPAN in the model is that the processing trials in the CSPAN were self-paced (in contrast to those in the RSPAN and OSPAN). This self-pacing may have increased the extent to which individual differences in strategy use contributed to task performance (for relevant discussion, see Friedman & Miyake, 2004, and Lewandowksy et al., 2010). Most importantly, both Models 5 and 6 revealed the same qualitative pattern of results as Model 1 (see Table 2).

Finally, we also compared our hypothesized model to a model in which all working memory tasks were loaded onto a single working memory factor, to ensure that this more parsimonious model would not fit the data better. This model is also relevant to the argument that using a heterogeneous battery of tasks is important to reducing task-specific variance and to measuring working memory better at the construct level (e.g., Lewandowky et al., 2010). Although the single working memory factor significantly predicted inductive reasoning (β = .69, $p < .001$), the model fit statistics indicated worse fit in this single-factor model than in our hypothesized model (see Table 2, Model 7). A chi-square difference test comparing the two models suggested that the model was oversimplified when working memory tasks were loaded onto a single factor [$\Delta\chi^2(2) = 55.24$, $p < .01$]. Furthermore, less variance in inductive reasoning was explained by this model than by the primary model (47% vs. 51%).

General discussion

The present research tested the hypothesis that more unique variance in inductive reasoning would be explained by content-embedded than by complex span tasks. To revisit, a key difference between content-embedded tasks and complex span tasks concerns whether the information being maintained in working memory is relevant to the processing task (i.e., in content-embedded tasks) or irrelevant to the processing task (i.e., in complex span tasks). Given that inductive reasoning tasks require the reasoner to maintain and manipulate task-relevant information to derive a solution, we predicted that more unique variance in inductive reasoning would be explained by content-embedded than by complex span tasks. Confirming this prediction, our primary model explained 51% of the variance in inductive reasoning; 45% of the total variance was uniquely explained by the content-embedded factor, whereas only 6% was explained by overlapping variance between the factors, and less than 1% was uniquely explained by the complex span latent factor. Furthermore, we ruled out numerous artifactual reasons that could account for these results, by testing a series of alternative models. In all of the models tested, inductive reasoning was better predicted better by the content-embedded latent factor than by the complex span latent factor.

Most theories of working memory assume that working memory is a multifaceted system (see Miyake & Shah, 1999, for perspectives on the nonunitary nature of working memory), but the number of facets and their independence from one another are still up for debate. Likewise, the working memory literature includes some disagreement as to what processes of the working memory system are reflected in various kinds of tasks proposed to measure the construct. Although the present research was not designed to tease apart the finer-grained processes involved in complex span versus content-embedded tasks, the present outcomes may inform these theoretical issues.

Some theoretical accounts have been forwarded about the processes underlying complex span tasks and the importance of those processes for reasoning. For example, Unsworth and Engle (2007) argued that performance on complex span tasks reflect both maintenance in primary memory and controlled search and retrieval of content from secondary memory. In contrast, although performance on content-embedded tasks also likely reflects maintenance in primary memory, these tasks likely do not reflect controlled search and retrieval from secondary memory to the same extent that complex span tasks do. In complex span tasks, an interpolated processing task forces to-be-remembered items from primary memory to secondary memory (given that primary memory is capacity-limited). In content-embedded tasks, task-relevant information is not displaced from primary memory by an unrelated processing task, and maintenance of the task-relevant information is less likely to exceed the limits of primary memory.

Although differential involvement of controlled search and retrieval from secondary memory in complex span versus content-embedded tasks is plausible, the extent to which this difference may have contributed to the pattern of outcomes observed here is less clear. Unsworth and Engle (2007) argued that controlled search of secondary memory is particularly

important for reasoning. Consistent with Unsworth and Engle's (2007) argument, Mogle, Lovett, Stawski, and Sliwinski (2008) found that reasoning performance on the Raven's Advanced Progressive Matrices was predicted by complex span tasks when controlling for primary memory, but that complex span tasks predicted nothing above and beyond measures of secondary memory. By this account, complex span tasks would have been better than content-embedded tasks at predicting inductive reasoning in the present study, which was clearly not the case. Additionally, other studies have yielded somewhat mixed results concerning the role of secondary memory in reasoning. For example, Unsworth, Brewer, and Spillers (2009) found that maintenance in primary memory and retrieval from secondary memory both uniquely predict reasoning, and findings from Wilhelm, Hildebrandt, and Oberauer (2013) suggested that primary memory is more important for reasoning than secondary memory. If so, inductive reasoning may have been better predicted by content-embedded tasks versus complex span tasks because they more heavily reflect maintenance in primary memory.

Other processes that may be differentially involved in content-embedded and complex span tasks include those involved in updating (i.e., the transformation and replacement of contents in working memory with more accurate or task-relevant information; see Miyake et al., 2000). Some research has suggested that updating itself involves multiple components (Ecker, Lewandowsky, Oberauer, & Chee, 2010). One component of updating that is of particular interest for present purposes involves intentionally disengaging from outdated or incorrect information in working memory (Ecker et al., 2010; Shipstead et al., 2016). Shipstead et al. recently emphasized the importance of disengagement for successful reasoning, given that initial focus on particular stimuli elements, relationships, and hypotheses may be incorrect. Importantly, although previous research shows a strong relationship between updating tasks and complex span tasks (i.e., Schmiedek, Hildebrandt, Lövdén, Wilhelm, & Lindenberger, 2009), Shipstead et al. argued that this strong relationship is largely driven by the other two proposed components of updating (i.e., retrieval and transformation). Furthermore, they argued that complex span tasks do not heavily reflect the disengagement component of updating. In contrast, disengagement may be captured to a greater degree by content-embedded tasks. For example, in the ABCD task, the stimuli and the structure of the instructions are the same on every trial (the letters ABCD, information about the relationship between A and B, the relationship between C and D, and the relationship between set orders). Given that the same elements are used on every trial and only relationships change, intentionally disengaging from temporary relationships between elements at the start of each trial is important to reduce interference. Otherwise, lingering relationships from previous trials may make it difficult to maintain and output the correct solution in the current trial. To the extent that disengagement plays a key role in successful reasoning, the predictive power of content-embedded tasks over complex span tasks in part may reflect greater involvement of disengagement processes.

Another possible explanation for why inductive reasoning was predicted better by the content-embedded latent factor than by the complex span latent factor concerns the extents to which these tasks involve the use of rules. Arguably, both content-embedded and complex span tasks involve rule application. For instance, in the alphabet task, participants are required to apply a transformation rule to a set of letters (e.g., "OZ backward 2"). Similarly, in the OSPAN, participants are required to apply rules of mathematics during the processing portion of the task. With that said, when the complex span task scores are based on performance on the maintenance portion of the task alone, variability in rule application would not be reflected in these scores. Indeed, when performance on the processing portion was taken into account into complex span task scores, the estimated correlation between the complex span and content-embedded latent factors was stronger relative to the models that only included performance on the maintenance portion of the task (see Table 2, Models 2 and 4b). However, inductive reasoning was still predicted better by the content-embedded latent factor than by the complex span latent factor, suggesting that rule application must not be the differentiating factor between these task types. Note that in both of these types of working memory tasks, participants are simply required to apply the provided rule. In contrast, the quintessential feature of inductive reasoning tasks is that leaners must infer the rule themselves before applying. This task feature represents an important functional difference between inductive reasoning tasks and both of these types of working memory tasks.

Importantly, the theoretical discussion here is only speculative—this research was not designed to isolate what processes are differentially tapped by content-embedded and complex span tasks. Nonetheless, the novel findings reported here will be informative for guiding further theoretical work on the component processes involved in these two kinds of working memory task and their involvement in inductive reasoning. These outcomes also provide an important extension to the small but growing literature showing an advantage of using content-embedded tasks versus complex span tasks for predicting higher-level cognition (e.g., reading comprehension; Was et al., 2011). Thus, future research investigating the involvement of working memory in complex cognitive tasks that involve the maintenance and processing of task-relevant information will likely profit from including content-embedded tasks as measures of working memory.

Author note The research reported here was supported by a James S. McDonnell Foundation 21st Century Science Initiative in Bridging Brain, Mind, and Behavior Collaborative Award.

References

Ackerman, P. L., Beier, M. E., & Boyle, M. O. (2002). Individual differences in working memory within a nomological network of cognitive and perceptual speed abilities. *Journal of Experimental Psychology: General*, *131*, 567–589. doi:https://doi.org/10.1037/0096-3445.131.4.567

Ackerman, P. L., Beier, M. E., & Boyle, M. O. (2005). Working memory and intelligence: The same or different constructs? *Psychological Bulletin*, *131*, 30–60. doi:https://doi.org/10.1037/0033-2909.131.1.30

Arbuckle, J. L. (2013). AMOS (Version 22) [Computer software]. Chicago: IBM SPSS.

Bentler, P. M., & Bonett, D. G. (1980). Significance tests and goodness of fit in the analysis of covariance structures. *Psychological Bulletin*, *88*, 588–606. doi:https://doi.org/10.1037/0033-2909.88.3.588

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

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

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*, 769–786. doi:https://doi.org/10.3758/BF03196772

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

Cowan, N. (2017). The many faces of working memory and short-term storage. *Psychonomic Bulletin & Review*, *24*, 1158–1170.

Daneman, M., & Carpenter, P. A. (1980). Individual differences in working memory and reading. *Journal of Verbal Learning and Verbal Behavior*, *19*, 450–466. doi:https://doi.org/10.1016/S0022-5371(80)90312-6

Ecker, U. K. H., Lewandowsky, S., Oberauer, K., & Chee, A. E. (2010). The components of working memory updating: An experimental decomposition and individual differences. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *36*, 170–189.

Ekstrom, R. B., French, J. W., Harman, H. H., & Dermen, D. (1976). Manual for Kit of Factor-Referenced Cognitive Tests. Princeton: Educational Testing Service.

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*, 309–331. doi:https://doi.org/10.1037/0096-3445.128.3.309

Foster, J. L., Shipstead, Z., Harrison, T. L., Hicks, K. L., Redick, T. S., & Engle, R. W. (2015). Shortened complex span tasks can reliably measure working memory capacity. *Memory & Cognition*, *43*, 226–236.

Friedman, N. P., & Miyake, A. (2004). The relations among inhibition and interference control functions: A latent-variable analysis. *Journal of Experimental Psychology: General*, *133*, 101–135. doi:https://doi.org/10.1037/0096-3445.133.1.101

Harrison, T. L., Shipstead, Z., Hicks, K. L., Hambrick, D. Z., Redick, T. S., & Engle, R. W. (2013). Working memory training may increase working memory capacity but not fluid intelligence. *Psychological Science*, *24*, 2409–2419.

Johnson-Laird, P. N. (2013). Inference in mental models. In K. J. Holyoak, & R. G. Morrison (Eds.), The Oxford handbook of thinking and reasoning (pp. 134–154). New York: Oxford University Press.

Kane, M. J., Hambrick, D. Z., & Conway, A. R. A. (2005). Working memory capacity and fluid intelligence are strongly related constructs: Comment on Ackerman, Beier, & Boyle (2005). *Psychological Bulletin*, *131*, 66–71. doi:https://doi.org/10.1037/0033-2909.131.1.66

Kane, M. J., Hamrick, 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*, 189–217. doi:https://doi.org/10.1037/0096-3445.133.2.189

Klauer, K. J., & Phye, G. D. (2008). Inductive reasoning: A training approach. *Review of Educational Research*, *78*, 85–123.

Klauer, K. J., Willmes, K., & Phye, G. D. (2002). Inducing inductive reasoning: Does it transfer to fluid intelligence? *Contemporary Educational Psychology*, *27*, 1–25.

Kyllonen, P. C., & Christal, R. E. (1990). Reasoning ability is (little more than) working-memory capacity?! *Intelligence*, *14*, 389–433. doi:https://doi.org/10.1016/S0160-2896(05)80012-1

Kyllonen, P., & Kell, H. (2017). What is fluid intelligence? Can it be improved? In M. Rosén, K. Y. Hansen, & U. Wolff (Eds.), Cognitive abilities and educational outcomes: A Festschrift in honour of Jan-Eric Gustafsson (pp. 15–38). Switzerland: Springer.

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

McIntosh, C. (2006). Rethinking fit assessment in structural equation modeling: A commentary and elaboration on Barrett (2007). *Personality and Individual Differences*, *42*, 859–867.

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*, 49–100. doi:https://doi.org/10.1006/cogp.1999.0734

Miyake, A., & Shah, P. (Eds.). (1999). Models of working memory: Mechanisms of active maintenance and executive control. New York: Cambridge University Press.

Mogle, J. A., Lovett, B. J., Stawski, R. S., & Sliwinski, M. J. (2008). What's so special about working memory. *Psychological Science*, *19*, 1071–1077.

Mueller, R. O., & Hancock, G. R. (2010). Structural equation modeling. In G. R. Hancock & R. O. Mueller (Eds.), The reviewer's guide to quantitative methods in the social sciences (pp. 371–384). New York: Routledge.

Muthén, L. K., & Muthén, B O. (2015). MPlus (Version 7.31) [Computer software]. Los Angeles: Muthén & Muthén.

Oberauer, K. (2002). Access to information in working memory: Exploring the focus of attention. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *28*, 411–421. doi:https://doi.org/10.1037/0278-7393.28.3.411

Oberauer, K., Schulze, R., Wilhelm, O., & Suß, H. M. (2005). Working memory and intelligence—Their correlation and relation: comment on Ackerman, Beier, and Boyle (2005). *Psychological Bulletin*, *131*, 61–65. doi:https://doi.org/10.1037/0033-2909.131.1.61

Oberauer, K., Süß, H.-M., Wilhelm, O., & Sander, N. (2007). Individual differences in working memory capacity and reasoning ability. In A. R. A. Conway, C. Jarrold, M. J. Kane, A. Miyake, & J. N. Towse

(Eds.), *Variation in working memory* (pp. 49–75). New York: Oxford University Press.

Raven, J. C. (1962). Advanced Progressive Matrices, Set II. London: H. K. Lewis.

Raven, J. C., Court, J. H., & Raven, J. (1977). Raven's Progressive Matrices and Vocabulary Scales. New York, NY: Psychological Corp.

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

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

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science*, *22*, 1359–1366. doi:https://doi.org/10.1177/0956797611417632

Stanovich, K. E., & Cunningham, A. E. (1992). Studying the consequences of literacy within a literate society: The cognitive correlates of print exposure. *Memory & Cognition*, *20*, 51–68.

Sternberg, R. J. (1986). Toward a unified theory of human reasoning. *Intelligence*, *10*, 281–314.

Sternberg, R. J., & Gardner, M. (1983). Unities in inductive reasoning. *Journal of Experimental Psychology: General*, *112*, 80–116.

Unsworth, N., Brewer, G. A., & Spillers, G. J. (2009). There's more to the working memory-fluid intelligence relationship than just secondary memory. *Psychonomic Bulletin & Review*, *16*, 931–937. doi:https://doi.org/10.3758/PBR.16.5.931

Unsworth, N., & Engle, R. W. (2005). Working memory capacity and fluid abilities: Examining the correlation between operation span and raven. *Intelligence*, *33*, 67–81.

Unsworth, N., & Engle, R. W. (2007). The nature of individual differences in working memory capacity: Active maintenance in primary memory and controlled search from secondary memory. *Psychological Review*, *114*, 104–132. doi:https://doi.org/10.1037/0033-295X.114.1.104

Vandierendonck, A. (2017). A comparison of methods to combine speed and accuracy measures of performance: A rejoinder on the binning procedure. *Behavior Research Methods*, *49*, 653–673. doi:https://doi.org/10.3758/s13428-016-0721-5

Was, C. A., Dunlosky, J., Bailey, H., & Rawson, K. A. (2012). The unique contributions of the facilitation of procedural memory and working memory to individual differences in intelligence. *Acta Psychologica*, *139*, 425–433.

Was, C. A., Rawson, K. A., Bailey, H., & Dunlosky, J. (2011). Content-embedded tasks beat complex span for predicting comprehension. *Behavior Research Methods*, *43*, 910–915.

Was, C. A., & Woltz, D. J. (2007). Reexamining the relationship between working memory and comprehension: The role of available long-term memory. *Journal of Memory and Language*, *56*, 86–102. doi:https://doi.org/10.1016/j.jml.2006.07.008

Wilhelm, O., Hildebrandt, A. H., & Oberauer, K. (2013). What is working memory capacity, and how can we measure it? *Frontiers in Psychology*, *4*, 433. doi:https://doi.org/10.3389/fpsyg.2013.00433

Woltz, D. J. (1988). An investigation of the role of working memory in procedural skill acquisition. *Journal of Experimental Psychology: General*, *117*, 319–331.