What Does the Punched Holes Task Measure?

BY

ALLISON J. JAEGER

B.A., University of Illinois at Chicago, 2007
M.A., University of Illinois at Chicago, 2012

THESIS

Submitted as partial fulfillment of the requirements
for the degree of Doctor of Philosophy in Psychology
in the Graduate College of the
University of Illinois at Chicago, 2015

Chicago, Illinois

Defense Committee:

Jennifer Wiley, Chair
James Pellegrino
Katherine Zinsser
Mike Stieff, Chemistry
Thomas Moher, Computer Science

DEDICATION

I would like to dedicate this thesis to my husband Chris and my mother Phyllis. Without your unending love, support and patience none of this would have been possible.

ACKNOWLEDGEMENTS

First, I must thank my advisor, Jennifer Wiley. Throughout my graduate career Jenny has always had my best interests in mind. She worked hard to make sure her students were supported both intellectually and emotionally, while at the same time pushing us to think more deeply. I feel so blessed to have had such a wonderful mentor and friend. I would also like to thank Jim Pellegrino and Mike Stieff for making time to discuss the development of my work, ask the really tough questions, and provide me with advice on career development. To Tom Moher, thank you for bringing your applied perspective to my work. Your passion for creating tools for the classroom is something I plan to take with me in the future. I would also like to thank Kate Zinsser for being willing to take a leap into Cognitive and bring with it her applied perspective as well. Furthermore, I must express thanks to the others members of my lab and my cohort for their endless support and feedback, and to Chelsea Perschon for her assistance in the completion of this work. Finally, I could not have made it through graduate school without week night meetings at Haymarket with my friends Andrew Jarosz, Katie McCarthy, and Spencer Campbell.

TABLE OF CONTENTS

CHAPTER **PAGE**

TABLE OF CONTENTS (continued)

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

AGFI	Adjusted Goodness of Fit Index
AIC	Akaike Information Criterion
ANOVA	Analysis of Variance
BIC	Bayesian Information Criterion
CC	Cube Comparison
CFA	Confirmatory Factor Analysis
CFI	Bentler's Comparative Fit Index
DAT	Differential Aptitude Test
DAT: SR	Differential Aptitude Test: Spatial Relations Subtest
EF	Executive Function
FB	Form Board Test
gF	General Fluid Intelligence
GFI	Goodness of Fit Index
M	Mean
MIRT	Multivariate Item Response Theory
N	Number in group
ns	Non-significant
PCA	Principle Components Analysis
PH	Punched Holes Task
r	Pearson Correlation Coefficient
RMSEA	Root Mean Square Error of Approximation
SD	Standard Deviation
STEM	Science, Technology, Engineering and Math
STM	Short-term Memory

LIST OF ABBREVIATIONS (continued)

t	Test Statistic Based on Gasser's Student Distribution
VZ	Spatial Visualization
WM	Working Memory
WMC	Working Memory Capacity
X^2	Chi-squared Test

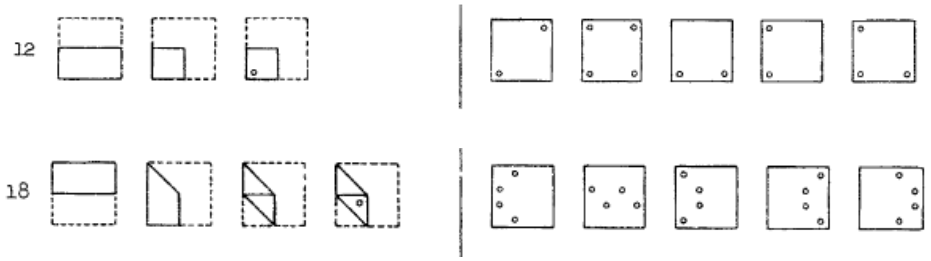
Summary

Visualization and other spatial skills are thought to be important for representation of abstract, complex, dynamic, and spatial relationships. One commonly used measure of spatial visualization (VZ) is the Punched Holes Test (PH) (Ekstrom, French, Harman, & Dermen, 1976). Previous work shows that the PH task is psychometrically reliable, and is related strongly to measures of executive functioning and working memory capacity (WMC; Kane, Hambrick, & Conway, 2005; Marshalek, Lohman, & Snow, 1983). The goal of this study was to more clearly determine what the PH task is actually measuring from a cognitive perspective and specifically attempt to demonstrate if distinct roles can be seen for VZ versus WMC. Pilot work demonstrated that WMC was more predictive of performance on items with multiple basic folds, but not items with atypical folds suggesting these item types may be testing different aptitudes. The current study investigated the role of VZ and WMC in four subtypes of items using a latent variable approach: basic three-fold, basic four-fold, atypical two-fold and atypical three-fold. The results indicated that VZ and WMC constructs were highly related to each other, and both were significant predictors of PH performance. However, performance on all item types was more strongly predicted by VZ, leaving little unique variance to be explained by WMC. At the same time, there was some evidence to suggest that the basic fold items relied more on WMC than the atypical fold items. Additionally, benefits for high VZ and high WMC participants were seen on the basic fold items, but not on the atypical fold items indicating that these two major item categories are differential in nature and may be indicative of different underlying abilities. Directions for future work are described based on these initial findings.

1. Introduction

Spatial visualization is a construct that has been studied for nearly 100 years and relates to one's ability to mentally manipulate 2-dimensional and 3-dimensional objects and figures. Being able to think and reason about spatial information has long been considered an important and unique ability that is of great importance for success in many everyday tasks. Because of this there have been many tasks developed to measure spatial visualization. One such task is the Punched Holes Task (PH) from the Kit of Factor-Referenced Cognitive Tests (Ekstrom et al., 1976).

As shown in the two examples below, in the PH test participants are presented with a series of illustrations on the left side of each problem representing a piece of square paper going through a succession of folds and then a hole being punched through the folded layers of paper:



The task of the participant is to determine which of the five possible patterns of holes presented on the right side of the problem will result from the sequence of folds and punch location presented on the left side. Each answer choice shows a square piece of paper with a pattern of holes. The participant must select the answer choice that depicts the correct number and pattern of holes when the paper is completely unfolded. Problems in the test vary such that they either involve one, two or three folds. Similarly, the punch location can vary such that on some items it goes through all layers of the folded paper at that point, whereas on other items it may only go through a single layer. The complete test is presented in Appendix A.

Typically the PH task is administered in a speeded fashion. There are two sets, each containing 10 items. Participants are given 3 minutes to complete each set. The original method for scoring performance on this task was to take the total number marked correctly minus one-fourth of a point for every answer marked incorrectly. However, it is also commonly scored by simply taking the total number correct or the proportion correct (e.g., Mayer & Sims, 1994; Salthouse, Mitchell, Skovronek, & Babcock, 1989; Sanchez, 2012). This alternative scoring method is useful and preferred in many cases because it reduces the typical male advantage found on spatial tasks by taking into account the fact that women work more slowly and therefore answer fewer items overall (Goldstein, Haldane, & Mitchell, 1990).

1.1 Cognitive Processes Involved in VZ

Interest in understanding individual differences in spatial ability began to emerge in the early 20th century. It was at this time that psychometric evidence was found supporting the idea that a spatial factor of intelligence may exist apart from verbal and numerical factors and that performance on tasks that tap this factor are predictive of STEM success and aptitude (e.g., McFarlane, 1925; Thurstone, 1938;). Eventually research on spatial aptitude veered away from predictive validity studies and turned towards describing the underlying structure of the spatial factor through factor analysis (Cronbach, 1970; Guilford, 1967; Horn & Cattell, 1966). Out of this work it was suggested that spatial ability was not a unitary construct, but instead was composed of several correlated subfactors (Carroll, 1993; Lohman, 1979; McGee, 1979). These included spatial orientation (the ability to imagine how a stimulus or array of stimuli would appear if one's body was reoriented), spatial relations (the ability to rapidly and accurately engage in mental rotation, with an emphasis on speed) and spatial visualization (less clearly

described, but defined by tasks that were unspeeded and involved more complex transformations).

This early psychometric and factor analytic work was beneficial for establishing the existence of a spatial factor and several subfactors, however these approaches did not lend themselves to determining the cognitive processes underlying performance on these tasks. The distinct cognitive processes or mechanisms required by VZ tasks such as PH began to be investigated by looking at models of spatial problem solving based in the information processing tradition. Early models of spatial information processing were described by Cooper and Shepard (1973) as well as by Kosslyn (1981). Cooper and Shepard (1973) put forth a four-stage processing model. In their model the first stage of processing requires the encoding of the stimuli, creating a representation of the stimuli, and then the storage of that representation in working memory. This stage is important because it lays the groundwork for determining what the task goals are, what information is relevant, and what constraints exist. The second stage of processing involves transforming the stimuli, which in the case of the PH task, would first involve folding the mental image, then updating that mental image to include the hole punch, and then finally unfolding the mental image while keeping track of the newly added hole punch. These processes involve storing and retrieving intermediate products, but also require maintaining the representation of the problem elements. Following this transformation stage, a comparison must be made in order to determine which target and comparison figures match. Again, in order to successfully compare the stimuli, an accurate representation must be maintained in memory. Lastly, this comparison stage must be followed by an answer selection.

Kosslyn's (1981) theory regarding the cognitive processes involved in spatial task performance was conceived more generally as a theory of mental imagery. His theory describes

how information is represented in and accessed from visual mental images. The two major components of the model include a visual buffer in which quasi-pictorial, spatial images are supported and a deep representation that is derived from actual visual input or from information stored in long-term memory. Kosslyn argued that the visual buffer is innately determined and of limited capacity therefore, information represented in this space is selectively activated rather than everything being activated at once. The surface images in the visual buffer represent physical information about the objects including extent, brightness, and texture. The purpose of deep processing is to map information stored in long-term memory onto the surface image. Kosslyn suggests that images are constructed from information stored in long-term memory and that this information can be both perceptual and conceptual in nature. Perceptual information includes representations stored from previous perceptual experiences whereas more abstract information (sometimes verbal labels or schemas) can be used and accessed from long-term memory. Kosslyn lays out several processes that may be involved in mental imagery including a Regenerate process which refreshes or reactivates images to keep them active. There are also several processes related to the transformation of visual representations including Rotate, Scan, and Translate. There are also processes for inspecting and classifying patterns including the Find process.

Out of these process models, the cognitive components approach to understanding individual differences in spatial task performance was developed (Pellegrino & Glaser, 1979). Cognitive components analyses were initially conducted to examine sources of individual differences in tasks representing the spatial relations factor (e.g., Carpenter & Just, 1986; Mumaw, Pellegrino, Kail, & Carter, 1984; Pellegrino & Glaser, 1979). Results from these process analyses suggested that individual differences in spatial relations task performance are a

function of the speed and accuracy at which specific mental processes are executed. On simple tasks, processing speed was the major determinant of performance however, as tasks became more complex by using unfamiliar stimuli or stimuli that was 3-dimensional, accuracy became a larger determinant of performance.

Process analyses of VZ tasks are less prevalent in the literature because the tasks are more diverse and cannot be characterized by a singular process model (Pellegrino, Alderton, & Shute, 1984). For these types of tasks both speed and accuracy seem to be important, however, due to the complex nature of VZ tasks, accuracy has been shown to be a greater contributor to performance. Mumaw and Pellegrino (1984) conducted a cognitive components analysis on the Paper Form Board Test. They systematically varied the items on three dimensions including the number elements in the item (2 to 6 elements), the manipulation of the elements (the most simple being holistic matching and the most difficult being rotation and displacement), and whether or not the completed puzzle and the elements matched. The latency data showed that performance was a function of the number of stimulus items and the complexity of the item manipulations such that as the number of items increased or the manipulations became more complex, latency increased as well. Results for accuracy showed that items where the puzzle and all items matched (or all items did not match) were solved more accurately than items where the puzzle only mismatched with a single item. They also compared the performance of high ability and low ability participants and found that low ability participants were less accurate at detecting mismatches and also demonstrated longer latencies. Mumaw and Pellegrino (1984) argue that these data indicate that high ability participants are able to construct higher quality mental representations and are also better able to transform those representations and preserve that information over time.

Shepard and Feng (1972) also examined the cognitive components underlying VZ, but focused on the Surface Development task in which participants need to determine what a flat, unfolded cube would look like if it were folded. The items in their analysis varied in terms of the number of folds required to bring two edges together as well as the number of surfaces that had to be carried along with each fold. Their results showed that decision times increased linearly as a function of both dimensions of complexity. Alderton and Pellegrino performed a similar analysis on another variant of the surface development task and found that performance differences were associated with accuracy, but not speed suggesting that high ability individuals could solve items with more complex fold patterns (Pellegrino et al., 1984).

Out of these process analyses it was suggested that successful performance on VZ tasks is a function of several cognitive capacities, which map on to those laid out in the information processing models of Kosslyn (1981) and Cooper and Shepard (1973). The first important cognitive process is the ability to create mental representations of unfamiliar visual stimuli. These representations need to be stable so that they can be operated on with minimal degradation. Based on the cognitive components literature, participants who are higher in spatial skills are often faster at creating and comparing the representations of unfamiliar stimuli (Pellegrino et al., 1984), a cognitive process that could be representative of spatial visualization ability.

Another important cognitive process involved in VZ task performance is related to the quality of the mental representations created, specifically how much information they can contain. High quality representations become more important as tasks require more complex manipulations and have several interrelated elements. Further, if stimulus representation occurs in a visual buffer system as suggested by Kosslyn (1981) then ability differences may be a

function of individual differences in the capacity of the working memory system. The cognitive components work on VZ task performance outlined above implies a role for visualization and WMC; however, the unique contributions of these constructs have never been directly tested.

1.2 Punched Holes as a Measure of Spatial Visualization

The PH task is a very commonly used measure in the psychometric literature and is based on a task of the same name originally developed by Thurstone (1938). Historically, and currently, this task is used to measure spatial visualization. Since the 1940's it has consistently fallen into the spatial visualization factor (Carroll, 1993; DeFries et al., 1974; Eliot & Smith, 1983; French, 1951; French, Ekstrom, & Price, 1963; Guilford & Lacey, 1947; Lohman, 1979; McGee, 1979; Michael, Guilford, Fruchter, & Zimmerman, 1957; Thurstone, 1938), which again is broadly defined as representing the ability to manipulate or transform complex spatial patterns. Ekstrom et al. (1976) suggest that the spatial visualization factor includes tasks that require figures or objects to be mentally restructured into components. Overall, the PH task has been shown to be a reliable test with estimates of reliability ranging between .75 and .89 (Kane et al., 2004; Kozhevnikov & Hegarty, 2001; Miyake, Friedman, Rettinger, Shah, & Hegarty, 2001; Salthouse, Babcock, Skovronek, Mitchell, & Palmon, 1990). It is also highly correlated with other spatial measures including cube comparison, hidden figures, paper form board, and surface development (Kozhevnikov & Hegarty, 2001; Marshalek et al., 1983).

The PH task has also been used as a predictor task across a number of different literatures. The PH task has been used to predict performance in or aptitude for science, technology, engineering or math related careers. For example, performance on spatial measures was used to predict success in spatially-related industrial jobs (Ghiselli, 1973), course grades in mechanical drawing, shop, art, mathematics and physics (McGee, 1979), and in training courses

for pilots and air crew (Guilford & Lacey, 1947). Lord (1987) showed that science majors have higher mental folding scores than non-science majors. Similarly, Siemankowski and MacKnight (1971) looked at PH scores of students across several majors and found that science, math, and arts majors out-performed non-science majors. Differences in PH scores have also been related to performance in specific science domains including chemistry (Carter, LaRussa, & Bodner, 1987; Staver & Jacks, 1988) and geosciences (Black, 2005). In this literature the PH task is argued to be predictive of these various careers and domains because it is said to be a measure of one's ability to mentally represent and transform shapes and relations among and between entities.

1.3 Punched Holes as a Measure of Executive Function

The PH task has also been used in research on working memory capacity and executive functioning. In some cases it has been used directly as a measure of WMC (e.g., Bichsel & Roskos-Ewoldsen, 1999; Miller & Bichsel, 2004). Other literature has attempted to determine the role of WMC or executive functioning in performance on the task and has argued that performance is primarily accounted for by the central executive component of the WM system (Miyake et al., 2001; Salthouse et al., 1989). The central executive is argued to be a supervisory system that is responsible for attentional control and regulating the flow of information stored within the two subsystems when performing cognitively demanding tasks (Baddeley & Hitch, 1974).

Salthouse et al. (1989) conducted a study in which they were interested in understanding the role of WMC in age related-decline on cognitive tasks by looking at performance on the PH task. They predicted that as the number of folds in a PH item increased, a greater age-related performance decrement would occur because older adults have fewer WMC resources and items

with more folds require more WMC. Additionally, they predicted that controlling for WMC would attenuate the age effects on the PH task. After controlling for computation span, they found that age went from explaining 28% of the variance in PH performance to only 16% of the variance. Although indirect, these results suggest that a major contributor to performance on the PH task is WMC. The authors further interpret their results to suggest that decreased accuracy on the PH task is attributable to an inability or failure to preserve relevant information. While this study does not use the PH task as a measure of WMC, it does indicate a substantial role for WMC in accurate performance.

Miyake et al. (2001) were interested in examining the relationship between WMC, executive functioning, and spatial abilities using a latent variable approach. Participants completed two WMC measures (Letter Rotation and Dot Matrix), two short-term memory (STM) measures (forward Corsi Blocks and Dot Memory), and two executive functioning (EF) measures (Tower of Hanoi and Random Number Generation). The results of a confirmatory factor analysis (CFA) indicated that the WMC and STM measures were better represented as a singular factor (STM-WMC factor), while the EF measures represented a separate factor (EF factor). Additionally, participants completed two spatial visualization measures (PH and DAT Space Relations which is similar to the Surface Development task), two spatial relations measures (Card Rotation and Flags), and two perceptual speed measures (Identical Pictures and Hidden Patterns), which CFA indicated all represented their respective constructs. The primary goal of their study was to assess the relationship that each spatial factor had with the two WM-related constructs. They used structural equation modeling to simultaneously estimate the contribution of STM-WMC and EF in predicting the three spatial factors.

The results of their individual differences study indicated that EF, or the central executive component of the WM system, contributed to 91% of the variance in performance on the VZ tasks, 83% on the spatial relations tasks, and 43% on the perceptual speed tasks. On the other hand, STM-WMC did not significantly contribute to VZ or spatial relations, but did contribute 38% on the perceptual speed tasks. Miyake et al. (2001) argue that the large path coefficients demonstrate that executive functioning contributes greatly (and almost exclusively in the case of VZ) to performance on the spatial tests.

While both of the previously mentioned studies indicate a substantial role for WMC in performance on the PH task, and in the case of Miyake et al. (2001) an almost exclusive one, neither provides definitive evidence for separate roles of VZ and WMC. In the case of Salthouse et al. (1989), baseline spatial ability is never measured and the unique effects of WMC and VZ on PH performance are never directly tested. The results do indicate that WMC is related to performance, but they do not provide evidence as to how or when it matters or if there is another construct that may explain performance more clearly. In the case of Miyake et al., there are several issues that bring into question the validity of their factors and their ability to accurately determine a unique role for EF or WMC. First, all of their factors are only represented by two tasks, which they admit in their discussion is not optimal. Second, when looking at the individual factor loadings for each task, some are quite low, in particular number generation onto the EF factor. This means that the EF factor is primarily representing a single task rather than a set of tasks. Third, there is a high correlation between the EF factor and the STM-WMC factor (.71) which is problematic because they could be representing nearly identical constructs which is why there is nothing left for STM-WMC to explain after EF is accounted for.

The majority of studies using PH, as well as the studies outlined above, have used an individual difference approach, concerned with who does well on the PH task or which aptitudes PH task performance is predictive of. In these studies, PH performance on the test as a whole is assumed to represent a singular construct. However, with the PH task being used to sometimes represent a VZ construct other times being suggested to represent a WM construct, the question arises if this task really does represent a singular construct at all or if certain aspects of performance, or different types of items, may represent separate constructs. Therefore, the goal of this dissertation was to more clearly determine if there are separate roles for WMC and VZ on the PH task using an individual items analysis approach.

1.4 An Item-based Approach to Individual Differences in PH

The present research was concerned with understanding whether different types of items on the PH task may represent separate constructs that are predicted by different individual differences in ability. This item-based approach is analogous to that used in previous research on the Ravens Advanced Progressive Matrices (Wiley, Jarosz, Cushen, & Colflesh, 2011). Although previous work suggests that VZ and WMC are both predictive of PH task performance, the unique contributions of WMC and VZ have never been identified. Thus, the goal of this study was to more clearly determine if there are separate roles for WMC and VZ using an individual items analysis approach.

To begin to analyze what cognitive processes contribute to performance on this task it was important to first examine the types of problems presented in the task and the dimensions that seem to underlie task difficulty and errors. Although there is no cognitive components analysis of performance on the PH task, Kyllonen, Lohman, and Snow (1984) did characterize several major dimensions of difficulty. These dimensions included the number of folds in an

item and the presence of asymmetric or obscured folds. Although the items do seem to vary along these dimensions of difficulty, they do not vary in a systematic way making it difficult to test the unique contributions of each of these dimensions on PH performance. Thus, a series of pilot studies were done to explore possible differences among items, and to design new item sets that would allow for systematic tests of the hypotheses. Two previous studies have developed new PH items, however for different purposes than the current study. In the study by Kyllonen et al. (1984) the goal was to see if training specific problem-solving strategies would improve performance on the PH task. In this study, the goal was not to assess what underlying constructs different item types may be measuring, but rather to see how training affected performance on different item types. Salthouse et al. (1989) also developed new PH items, but did so in order to assess the impact of age-related processing decline on complex cognitive task performance.

1.5 Pilot Studies

In an initial pilot study, a sample of participants completed the PH task while thinking aloud and being eyetracked. They also completed a battery of WMC tasks. The results of this pilot work helped to delineate different types of items in the PH task. Specifically, there appeared to be one set of items whose difficulty could be linked to maintaining an increasing number of folds and another set whose difficulty could be linked to dealing with atypical folding patterns. When the entire shape is folded to fully overlap at all edges and corners, an item was categorized as basic. For example, if a square were to be folded from the top edge to the bottom edge creating two equal size rectangles, that item would represent a basic fold. An additional defining feature of these items is that the resulting dot pattern that is created is always symmetrical or linear. Atypical items were characterized by folds that did not create two equal shapes, for example by only folding a portion of the shape or a single layer to overlap rather than

the entire shape and the resulting hole patterns were not symmetrical or did not create straight lines. An additional characteristic of some of these items were that they included folds that were occluded or made invisible by later folds. An interesting result was that performance on the multiple basic fold items was related to WMC whereas performance on the atypical fold items was not (Jaeger, Jarosz, & Wiley, 2014). Further, coding of the think aloud protocols indicated that participants used more strategies when attempting to solve the atypical fold items compared to the multiple basic fold items. Interestingly, the eyetracking data showed that participants toggled less between the answer bank and problem space when the problems were atypical, suggesting the use of a visualization-based strategy (Perschon, Jaeger, & Wiley, 2015a). An interesting question related to these results was whether performance on these atypical items may be more directly related to visualization skills. Further analysis of the eyetracking data indicated that low VZ participants (as measured by a mental rotation task) toggled more than high VZ participants. There was also a non-significant trend for low VZ individuals to gesture more during problem solving (Perschon, Jaeger, & Wiley, 2015b). Together these results suggest that visualization may be particularly important for the atypical items and that low VZ participants are less effective at visualizing and therefore needed to toggle and gesture more to support their memory.

While this pilot study provided some preliminary evidence to suggest the presence of two types of items, which may tap into different cognitive processes, the original PH task did not systematically vary the items along these dimensions making these hypotheses difficult to test. Specifically, the established PH test includes only a few examples of the atypically folded items (only problems 7 and 9 on set 1 and problems 15, 18, and 20 on set 2). Even if both halves of the test are given (which is not always the case) there are still only 5 items in this potentially

interesting subset. Having only a few critical items on a test is not ideal for reliable measurement because it leads to a greater chance that a person's score on that test is due to error. When a measure is made up of more items it reduces the influence of chance factors such as guessing, therefore leading to more accurate measure of the construct of interest (Nunnally, 1978).

Thus, a second pilot study was conducted with the goal of creating more items to represent each of the two kinds of difficulty identified in the original test: multiple folds and asymmetrical or atypical folds. For this study 10 new items were piloted as shown in Appendix B. Five new items had 4 folds (following Salthouse et al., 1989). There were also 5 new items that involved an atypical fold. For the new basic four-fold items, a strong correlation between PH performance and WMC was expected due to the number of folds required by the item. However, if the new atypical items depend more heavily on visualization skills as hypothesized, then they should not be as highly correlated with WMC.

Results of this second pilot study revealed that performance on the new set of 5 basic four-fold items was found to be highly correlated with the basic three-fold items from the original PH task (Table 1), but were more difficult than the original basic three-fold items, $t(80) = 2.62$, $p < .02$ (mean accuracy of .60 versus .70). These new basic four-fold items were also significantly correlated with WMC as measured by a composite of Running Span and Backward Digit Span (Table 1).

Table I

Correlations for New PH Items

	Original 3-Fold	Original Atypical Fold	New 4-Fold	New Atypical Fold
Original 3-Fold	1	-	-	-
Original Atypical Fold	.07	1	-	-
New 4-Fold	.43**	-.08	1	-
New Atypical Fold	.07	.25*	.12	1
WMC Composite	.46**	.20	.35**	.15

** p < .01, * p < .05, † p < .07

In addition, there were also 5 new atypical fold items that were found to be correlated with the subset of 5 atypical fold items from the original PH task. Performance on the new atypical fold items was higher than performance on the original atypical fold items, $t(80) = 2.87$, $p < .01$, but was still not significantly correlated with WMC (see Table 1).

Overall, the results of this second pilot study demonstrated that there is a subset of PH items that require multiple basic folds and show a strong correlation with WMC, whereas there is another subset of items involving atypical folds and show little correlation with WMC. The creation of these new items was important because it set the stage for more clearly testing whether unique contributions can be seen for both WMC and visualization skills on PH performance using these two different kinds of items.

1.6 Overview of Current Study: Testing for Unique Predictors of Basic and Atypical Fold Items

The goal of the current experiment was to determine what the Punched Holes task is actually measuring and to test if there are subsets of items that represent separate constructs. Based on the results of the pilot work, it seems that there are two distinct subsets of difficult items that determine performance on the Punched Holes task. One subset of difficult items

involves multiple folds, while the other involves atypical folds. While the first kind of item seems to depend heavily on WMC, an interesting question was whether the second type of paper folding item could be seen to rely more directly on VZ. Based on pilot findings, the main goal for this experiment was to test whether these atypical fold items could be shown to be better predicted or uniquely predicted by VZ compared to the basic fold items. An additional goal was to test whether the basic fold items could be shown to be better predicted or uniquely predicted by WMC, especially those with a greater number of folds, compared to the atypical fold items.

1.6.1 Logic of Confirmatory Factor Analyses

To investigate these questions, a factor analytic design was employed. The Punched Holes task was divided into four categories based on item type. These categories included basic three-fold items, basic four-fold items, atypical two-fold items, and atypical three-fold items. Participants also completed a battery of VZ tasks and a battery of WMC tasks. It was hypothesized that a model in which the factor loadings for each item type were constrained to load the same onto VZ and WMC would not be as good of a fit as a model in which item types were unconstrained and allowed load differentially onto these factors. More specifically, it was predicted that multiple basic fold items would load more heavily onto the WMC factor, especially those with a great number of folds, and less onto the VZ factor. This prediction reflects the assumption that keeping track of additional information, in this case additional folds, requires more WMC resources. Additionally, it was predicted that the atypical fold items would load more heavily onto the VZ factor and less onto the WMC factor, regardless of the number of folds. This prediction reflects the assumption that being able to successfully solve the atypical items relies more on one's ability to visualize folds that have been obscured from direct view or to visualize asymmetric patterns.

1.6.2 Logic of Accuracy and Latency Analyses

Following the analyses conducted in the cognitive components tradition, an additional goal was to assess differences in solution accuracy and response latency on the four item types as a function of WMC and VZ. Previous work has demonstrated that accuracy and latency increases as problem complexity increases, such as with increased angular disparity in rotation tasks (Cooper & Shepard, 1973; Pellegrino & Kail, 1982). In the case of the PH task, problems become increasingly complex as more folds are added, therefore it was predicted that accuracy and latency would increase as a function of number of folds. Additionally, stimuli familiarity has been shown to impact response latencies such that familiar stimuli are processed more quickly than unfamiliar stimuli (Mumaw et al., 1984). Based on these previous results it was predicted that although both basic and atypical items would show decreasing accuracy and increasing response times as a function of the number of folds, this pattern would be more pronounced for the atypical fold items.

1.6.2.1 Analyses as a Function of VZ.

Also following work in the cognitive components literature, differences in accuracy and latency across problem type were investigated as a function of overall spatial skills. Specifically, participants were split into high spatial and low spatial groups based upon their performance on the battery of spatial tasks. Previous work has shown that low spatial participants are less accurate and have longer overall response times compared to high spatial participants and that the slope of their response times as item complexity increases is more pronounced (Pellegrino & Kail, 1982). Therefore, it was predicted that low spatial individuals would show decreased solution accuracy and increased response latencies compared to high spatial individuals and that the number of folds in an item would have a greater effect on response accuracy and latency for

low spatial individuals. As for differences in response latency between high and low spatial individuals as a function of fold type (basic versus atypical), the predictions were less clear. One possible outcome was that low spatial individuals would show longer response latencies on atypical fold problems than high spatial individuals. This result would suggest a difference in the speed at which a mental representation can be created, compared, or manipulated. Specifically, low spatial individuals may be slower at creating mental images and comparing those to the visual stimulus provided.

1.6.2.2 Analyses as a Function of WMC

Finally, although no work has previously investigated how WMC impacts performance on different types of PH items, several predictions were made. First, because previous research has implicated WMC as being important for performance on the PH task, it was predicted that high WMC individuals would be more accurate than low WMC individuals across all item types, but that this difference would be especially large for the basic items where it is predicted that WMC will play a more prominent role.

As for differences in response latency, it was predicted that for basic fold items, response latencies overall would be longer for low WMC individuals compared to high WMC individuals and the increase in response latency as the number of folds increased would be more pronounced for low WMC individuals. This prediction reflects the idea that WMC contributes to performance by allowing high capacity individuals to keep track of more folds. In terms of latency predictions for performance on the atypical fold problems as a function of WMC, it was predicted that there would be no difference between high and low WMC individuals. Specifically, this prediction is based on the assumption that these items rely less on WMC and

more on VZ compared to multiple, basic fold items, therefore reducing the impact of being low WMC.

Overall, the goal of this study is to examine whether distinct roles can be seen for WMC and VZ on the Punched Holes Task. While previous work from the cognitive components literature has suggested that these two constructs are what primarily contributes to performance on VZ tasks, the hypothesis that the test may be composed of different item types that differentially relate to these constructs is novel, and has not been directly tested. Further, although the cognitive components involved in performance on some measures of VZ have been explored, the Punched Holes task itself has never received such in depth attention despite its wide use in several different literatures.

2. Method

2.1 Participants

The participants were 149 undergraduates from the University of Illinois at Chicago (84 females) who participated to partially fulfill a course requirement. They ranged in age from 18-31 years old (M = 19.26, SD = 1.58). One hundred and three participants reported being bilingual and the mean age at which they reported becoming fluent in English was 6.02 years (SD = 2.80). Further, the participants reported having taken an average of 2.72 science courses (SD = 2.73) and had a mean composite ACT score of 24.29 (SD = 3.37). An additional 47 participants took part, but their data for some of the tasks were not complete or they failed to follow task instructions. Thus, their data was not included in the analyses. Specifically, 36 participants were excluded because they had incomplete PH data due to not completing all 20 items within the 12 minutes allowed. Of the remaining 11 participants, two were excluded for not completing all the spatial visualization tasks and the remaining nine were excluded for having more than 20 errors on the Rotation Span task.

2.2 Measures

All participants completed a battery of seven tasks. These included the Punched Holes test with the four item subtypes, three working memory capacity measures and three spatial visualization measures.

2.2.1 Punched Holes

The original Punched Holes task from the Kit of Factor Referenced Tests (Ekstrom et al., 1976) was altered from its original form and presented via computer. Participants were given the same instructions as used on the original task and also the same initial practice item. Because performance on the one-fold items was near ceiling in pilot testing, participants were also given

those two items during the practice phase. After being presented with the instructions, completing the three practice items and getting feedback on the practice items, participants completed a set of five basic two-fold items taken from the original Punched Holes stimuli. Because performance on these items was high during pilot testing they served as additional practice and participants were given 2 minutes to complete them. Lastly participants completed the set of 20 target items. Five of the items were basic three-fold items (four taken from the original task), five were basic four-fold items, five were atypical two-fold items (four taken from the original task), and five were atypical three-fold items (one taken from the original task). Presentation of these items was randomized. While the original administration of the Punched Holes task only allows for 3 minutes per 10 items (6 minutes in total), more time was needed for the updated version to insure that participants completed all problem types. Therefore participants were allowed 12 minutes to complete the target 20 items. This task took approximately 15 minutes in total to complete. Performance was computed as the total number of items answered correctly in each subtype. This resulted in four scores, each with a possible maximum of 5. See Appendix C for the full list of target items.

2.2.2 WMC tasks

Three standard WMC measures (Automated Running Span, Backwards Digit Span, and Automated Rotation Span) were administered following the procedure of Kane et al. (2004). Examples of each task can be found in Appendix D.

2.2.2.1 Running Span. The Automated Running Span was based on a version used previously by Broadway and Engle (2010). Participants were presented with a series of letters, but were told to only remember the last few letters in the series. Therefore, they were required to keep track of the last letters while forgetting the earlier letters. On each trial, participants saw a

string between 4 and 9 letters on a computer, and when it terminated they were asked to type in the last 3 to 6 letters, with six trials of each size presented randomly. The task took approximately 10 minutes to complete.

2.2.2.2 Backwards Digit Span. The Backwards Digit Span involves remembering a series of numbers, then repeating them back in reverse order (Unsworth & Engle, 2007). In this computerized version, participants saw string of numbers between 2 and 8 numbers in length, and were asked to type back the numbers in reverse order. Participants had 500ms to view each digit in the number string, and 15 seconds to type their answer once the numbers disappeared. Numbers were presented in sets of increasing size, with two trials of each size. This task took approximately 10 minutes to complete.

2.2.2.3 Rotation Span. The Rotation Span task is a computerized task that involves simultaneous memory and processing of visuospatial stimuli (Harrison et al., 2013). For the processing component of this task participants were presented with a normal or mirror reversed *G*, *F*, *J* or *R* rotated at 0°, 45°, 90°, 135°, 180°, 225°, 270°, or 315°. The task requires the participant to mentally rotate the letter then indicate with a mouse click whether that letter was facing the normal direction ("true") or was mirror reversed ("false"). After the rotation judgment, participants saw a short or long arrow pointing in one of eight directions. This rotation-arrow sequence was repeated from two to five times per trial, with three trials of each size presented randomly. After each rotation-arrow trial, a recall screen appeared in which participants needed to click on the arrows on the recall screen in the order in which they were presented. This task took approximately 15 minutes to complete.

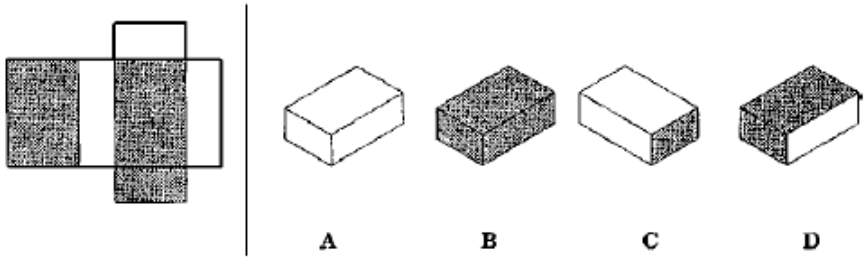
The WMC tasks were scored using standard procedures. The Running Span and Backwards Digit Span tasks were scored by the proportion of items recalled in the correct serial

position (Broadway & Engle, 2010; Unsworth & Engle, 2007). For the Rotation Span task, scores were calculated by summing the number of arrows correctly recalled in the correct order (Foster et al., 2014). Typically, any participants with less than 85 percent accuracy on the processing task are dropped from analyses. For the purposes of the current study any participants with less than 75 percent accuracy were eliminated, which was the case for 17 participants.

2.2.3 VZ tasks

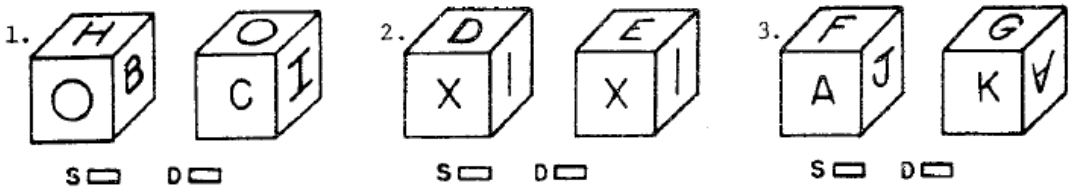
Three VZ tasks were used: The Differential Aptitude Test: Space Relations (DAT: SR) (Bennett, Seashore, & Wesman, 1973), the Cube Comparisons task (CC) (Ekstrom et al., 1976), and the Form Board task (FB) (Ekstrom et al., 1976). The DAT: SR task is similar to the Surface Development Task (Ekstrom et al., 1976) but was preferred because it only required one response per item so its structure and scoring were more similar to the other spatial tasks used in this study. CC was included because it was used previously by Just and Carpenter (1985) to investigate problem solving processes on psychometric tests of spatial ability as they described it as assessing the ability to hold transformed visual images in memory. Finally FB was selected because it is a classic spatial visualization task from the same cognitive test kit at the original Punched Holes test. All three tasks have been demonstrated to load on to the VZ factor and have been used in spatial visualization batteries in the past (Kane et al., 2004; Kozhevnikov, Motes, & Hegarty, 2007; Salthouse et al., 1990).

2.2.3.1 The Differential Aptitude Test: Space Relations. The DAT: SR task (Bennett et al., 1973) presents drawings of solid forms that could be made with paper. The participant's task was to choose the correct 3-dimensional object from four alternatives that would result from folding the given 2-dimensional pattern. See the example below:



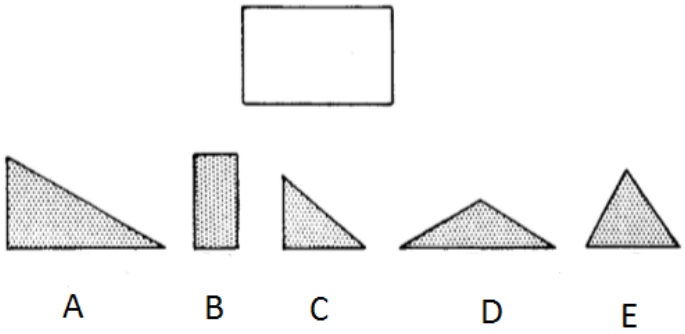
The shaded portions need to correspond to the same shaped portions on the correct 3-D figure. In the traditional administration of this task participants are given 25 minutes to complete 60 items. For this study, following two practice items, participants were given 12.5 minutes to complete 30 items. Presentation of this task was computerized and performance was computed as the total number correct.

2.2.3.2 Cube Comparison. The Cube Comparison task (Ekstrom et al., 1976) is a spatial task in which each item presents two drawings of a cube, as shown in the example below:



Assuming no cube can have two faces alike, the participant was to indicate which items present drawings that can be of the same cube and which present drawing that cannot be of the same cube. This test has two parts that each consist of 21 items. Participants were given 3 minutes to complete each part. Presentation of this task was computerized and was scored by taking the total number of correctly answered items.

2.2.3.3 Form Board Test. The Form Board test (Ekstrom et al., 1976) is a spatial task in which each item presents a set of five geometric two-dimensional shapes that can be put together in some combination to form a two-dimensional geometric figure. See the example below:



Participants were to indicate which two to five of the figures are necessary to form the target figure by selecting each corresponding letter. Participants were given two practice items and feedback on the correct responses for those items. After completing the practice items participants were given 8 minutes to complete 24 target items. There were four different target figures with six items corresponding to each target figure. Presentation of this task was computerized and took approximately 10 minutes to complete. Performance was scored by taking the total number of correctly answered items.

2.3 Procedure

All participants participated in a 2 hour session in which they completed all seven tasks (see Figure 1). First participants completed the updated Punched Holes task. Second, they completed the battery of WM measures starting with Running Span, then Backwards Digit Span, and then Rotation Span. Following the completion of these four tasks participants were allowed a short break. After returning from their break they completed the battery of spatial visualization tasks. First they completed the Space Relations task, followed by the Cube Comparisons task, and lastly the Form Board task. Following the visualization tasks, they completed two additional measures that will not be discussed as part of the dissertation: Series Completion (Kotovsky & Simon, 1973) and Figure Classification (Lohman & Hagen, 2001).

At the end of the session, demographic information was collected from participants. This questionnaire is included in Appendix E. Questions included gender, age, the participant's primary language, and how many years the participant has been speaking English, the number of science courses they have taken as well as self-reported composite ACT scores. This questionnaire also included items assessing the strategies used while completing the Punched Holes task based on the questionnaire developed by Hegarty (2010).

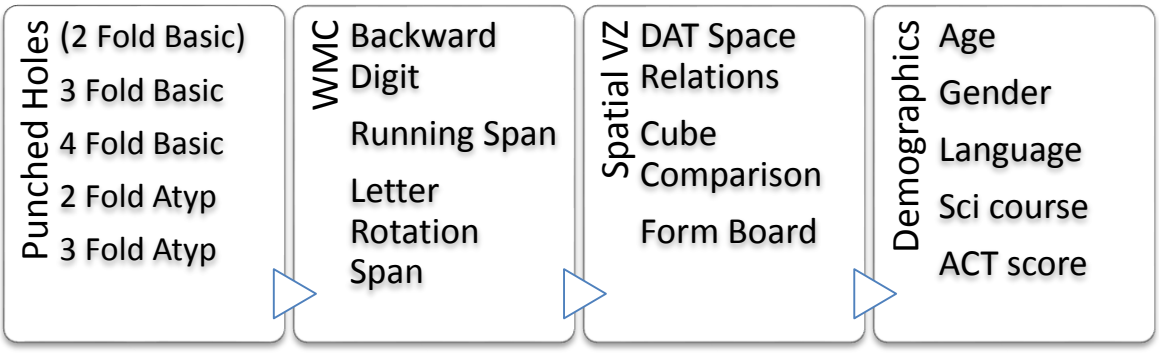


Figure 1. Overview of Procedure

3. Results

3.1 Descriptive Statistics and Preliminary Analyses

Descriptive statistics for the each item subtype on the Punched Holes task are presented in Table 2 including mean, standard deviation, range, skewness, and kurtosis. As shown in Table 2, performance was highest on the basic two-fold items and significantly greater than performance on the basic three-fold items, $t(148) = 5.63$, $p < .001$. To assess accuracy differences as a function of number and type of folds, a 2 (Type of fold: basic, atypical) X 2 (Number of folds: fewer, greater) repeated measures analysis of variance (ANOVA) was conducted. This analysis revealed a significant main effect for both type of fold, $F(1, 148) = 214.79$, $p < .001$, $\eta^2 = .59$, and number of folds, $F(1, 148) = 14.69$, $p < .001$, $\eta^2 = .09$. These main effects indicate that the atypical fold items were more difficult than the basic items, and that the items with more folds were more difficult than the items with fewer folds. These main effects were qualified by a significant interaction between type and number of folds, $F(1, 148) = 5.01$, $p < .03$, $\eta^2 = .03$. To follow up the significant interaction, planned comparisons investigated the effects of number of folds within each item type (basic and atypical). Follow up tests revealed that for basic fold items, participants performance better on items with three folds as compared to four folds, $t(148) = 4.13$, $p < .001$. However, for atypical items, there was no difference in performance between the two-fold items and the three-fold items, $t(148) = 1.11$, *ns*. Additionally, all Punched Holes item subtypes were significantly correlated with each other (see Table 3). Further, there was a trend for the basic fold items to correlate more highly with each other ($r = .55$) and the atypical fold items to correlate more highly with each other ($r = .44$). It is also important to note than although not further analyzed in this study, the basic two-fold items behaved similarly to the basic three and basic four-fold items. Specifically, the basic two-fold

items correlated more strongly with the other basic fold items (r = .49) than the atypical fold items (r = .24). The basic two-fold items also show similar correlations with the WMC and VZ measures as the other PH item types (Table 3).

Descriptive statistics for the WMC and VZ tasks are presented in Table 4. For all measures in this table, higher numbers indicate better performance. In addition to mean, standard deviation and range, Table 4 also summarizes the skewness, kurtosis and reliability estimates. For DAT:SR and FB VZ tasks, reliability estimates were derived from the split-half correlations using odd and even items, adjusted by the Spearman-Brown prophecy formula. For CC, the two sets were correlated for test-retest reliability. As shown in Table 4, reliability for the DAT: SR (α = .86) and FB (α = .83) were good. Test-retest reliability between the two CC item sets was not as good (α = .64). However, the three tasks did correlate well with each other (all rs > .45).

Table II

Descriptive Statistics for the Punched Holes Item Types (N = 149)

Task	*M*	*SD*	Range	Skewness	Kurtosis
PH – 2 Fold Basic	3.85	1.26	0-5	-1.16	.80
PH – 3 Fold Basic	3.21	1.38	0-5	-0.55	-0.60
PH – 4 Fold Basic	2.76	1.45	0-5	-0.17	-0.79
PH – 2 Fold Atypical	1.54	1.26	0-5	0.92	0.62
PH – 3 Fold Atypical	1.43	1.10	0-5	0.92	1.00

Note. For all variables, higher numbers indicate better performance with a maximum possible score of 5. PH = punched holes task

Table III

Intercorrelations Between Dependent Measures (N = 149)

Variable	1	2	3	4	5	6	7	8	9	10	11
1. 2 Fold Basic	---										
2. 3 Fold Basic	.47	---									
3. 4 Fold Basic	.51	.55	---								
4. 2 Fold Atypical	.26	.30	.34	---							
5. 3 Fold Atypical	.22	.26	.28	.44	---						
6. Back Digit	.25	.19	.24	.25	.18	---					
7. Running Span	.10	.20	.19	.14	.06	.52	---				
8. Rotation Span	.03	.12	-.02	.17	.10	.34	.33	---			
9. Space Relations	.41	.54	.48	.43	.34	.38	.29	.25	---		
10. Cube Comp.	.30	.36	.30	.34	.23	.33	.21	.19	.54	---	
11. Form Board	.43	.40	.41	.45	.31	.44	.32	.19	.60	.47	---

Note. Higher scores indicate better performance for all tasks. All correlations greater than or equal to .17 were significant at the .05 level.

In order to compute reliability for the WM tasks, subscores were created. Specifically, for Rotation Span each set size was presented three times so three subscores were created, each including one instance of each set size. For Backwards Digit Span each set size was presented two times therefore two subscores were created. For Running Span, each set size was presented six times so three instances of each set size were combined into two subscores. This yielded three subscores for Rotation Span and two subscores for Running Span and Backwards Digit Span. Cronbach's alpha was then computed using the subscores for each span task. As shown in Table 4, the reliability estimates for Rotation Span ($\alpha = .87$) and Running Span ($\alpha = .78$) were

good, and the estimate for Backwards Digit Span was acceptable ($\alpha = .64$). Although the Cronbach alpha for Backwards Digit Span was not as high as has been found with other working memory tasks, this could likely be the result of computing its reliability with only two subscores, rather than three. The scores on all three of these tasks were within the typical range seen in the literature (Foster et al., 2014). However, Backwards Digit Span and Running Span correlated above .5 with each other, while Rotation Span only correlated at around .34 with each of them. It is also important to note that typically 85 percent accuracy on the processing component of complex span tasks is required, but in the current study a 75 percent accuracy cut-off was used. When using the more strict accuracy cut-off, Rotation Span correlated worse with Backwards Digit Span and Running Span at around .29 with each of them.

Before testing the main hypotheses with CFA, a principle components analysis was conducted with the set of VZ tasks and the set of WMC tasks. The analysis revealed two factors, one factor that the three VZ tasks loaded on to and another that the WMC tasks loaded on to. Both of the factors and their task loadings can be seen in Table 5. Because analyses were proposed using all three WMC measures, they are performed using a factor derived from all three WMC tasks for the main analyses, but follow-up analyses use just the Backward Digit/Running Span composite.

Table IV

Descriptive Statistics for the Independent Measures (N = 149)

Task	*M*	*SD*	Range	Skewness	Kurtosis	Reliability
Running Span	64.34	16.65	20-101	-0.02	-0.38	0.78[b]
Backwards Digit	41.38	7.65	19-60	-0.01	.06	0.64[b]
Rotation Span	24.86	6.93	0-41	-0.60	1.09	0.87[b]
Errors on Rot Span	3.48	3.36	0-14			
Space Relations	14.12	6.37	2-30	0.30	-0.64	0.86[a]
Cube Comparison	26.47	5.36	15-39	0.28	-0.61	0.64[b]
Form Board	8.04	4.55	0-20	0.22	-0.43	0.83[a]

Note. For all variables, higher numbers indicate better performance.
[a] Reliability based on split-half method [b] Reliability based on Cronbach's alpha

Table V

Factor Loadings for VZ and WMC Factors from Principle Components Analysis

VZ Factor		**WMC Factor**	
Task	*Factor Loading*	*Task*	*Factor Loading*
Space Relations	0.83	Backwards Digit Span	0.71
Cube Comparison	0.81	Running Span	0.79
Form Board	0.79	Rotation Span	0.73

3.1.1 Confirmatory Factor Analysis for WMC and VZ. Confirmatory Factor Analyses (CFA) were performed with the AMOS program (Arbuckle, 2013) which uses maximum-likelihood estimation to derive the specified parameters based on the covariance matrix. Following the recommendation of Hu and Bentler (1998), the fit of each model was evaluated with multiple indices. Commonly used indices χ^2 and χ^2/df were used, as well as Bentler Comparative Fit Index (CFI), the root-mean square error of approximation (RMSEA), the goodness of fit index (GFI), and the adjusted goodness of fit index (AGFI).

The most common fit index is the χ^2 statistic. This statistic measures the "badness of fit" or whether the model significantly deviates from the data. For this measure, smaller values indicate no difference between the predicted and the observed covariance, suggesting a good model fit. Because χ^2 is correlated with sample size, many researchers advocate for using the χ^2/df statistic in which values less than 2 indicate a good fit. For CFI, higher values indicate better model fit because it quantifies the extent to which the model fits better than a baseline model. For RMSEA, the recommendation is that values of .05 or lower indicate a good fit. For the GFI and AGFI, the recommendation is that values of .90 or greater indicate a good fit.

A CFA was conducted to specify the extent to which the WMC tasks loaded onto a single WMC factor and the VZ tasks loaded onto a single VZ factor. The estimated model is illustrated in Figure 2. In the figure, circles represent theoretical factors and boxes are exogenous variables. The numbers above the straight, single-headed arrows are the standardized factor loadings. The number next to the curved, double-headed arrow represents the estimated intercorrelation between the two latent variables. Values of fit indices for the two-factor model suggest that it fits the data. This model produced a non-significant $\chi^2(8) = 4.66$, $p = .79$ and $\chi^2/df = 0.58$, indicating that the model did not significantly deviate from the data. The CFI was 1.00, well above the

commonly used criterion of .90 for a good fit, and the RMSEA was 0.00, also constituting a good fit. Additionally, the GFI was .99 and the AGFI was .97. These results align with those of the principle components analysis and suggest that the three VZ tasks load nicely onto a single VZ factor and the three WMC tasks load nicely onto a single WMC factor.

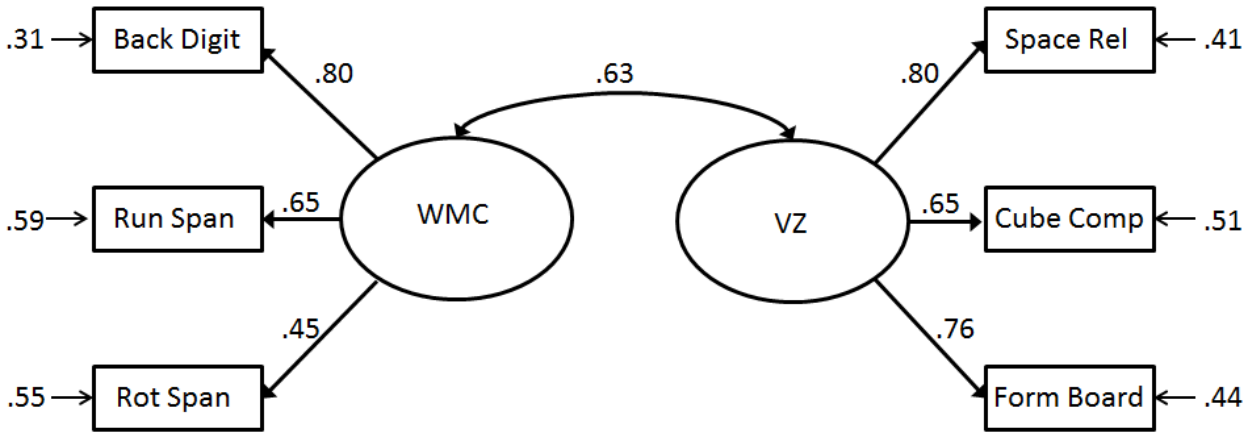


Figure 2. The results of the confirmatory factor analysis for the estimated two-factor model. All loadings and correlations are significant at the .05 level.

3.2 Main Analyses: Do Multiple- and Atypical-fold Items Load Equally onto WMC and VZ?

The main goal of this study was to assess the extent to which performance on the four item subtypes on the Punched Holes task depends on two separate constructs, WMC and VZ. Of particular interest was the hypothesis that the basic-fold items would load more strongly onto the WMC factor than the atypical items with fold occlusions, and that the atypical items would load more strongly onto the VZ factor than the basic items. Further, it was hypothesized that basic

items with a larger number of folds would more strongly load onto the WMC factor than the basic items with fewer folds, while the atypical items would load strongly onto the VZ factor regardless of the number of folds. A series of CFAs were performed to address these issues.

The first step was to estimate the model in which all four item types are constrained to have the same factor loadings on VZ and the same factor loadings on WMC. It is important to note that this does not assume the loadings are the same *between* WMC and VZ, but rather that all items will have the same loadings with each construct. This model (Model A) represents the unstated assumption in the literature that all Punched Holes items represent the same underlying construct. Specifically, it is assumed that all items and should have the same factor loadings onto WMC and VZ. In the current study the hypothesis was that this constrained model would not be as good of a fit compared to an unconstrained model in which the items were allowed to load onto WMC and VZ freely.

Initially, Model A was run without allowing the VZ and WMC latent variables to co-vary, however this resulted in a significant chi-square which indicated a poor model fit, $\chi^2(29) = 58.32$, $p = .001$. Modification indices suggested allowing these two variables to correlate so in all models these two latent variables were allowed to co-vary. Additionally, Model A was run without allowing any of the error terms to co-vary. However, this also resulted in a significant chi-square, $\chi^2(36) = 53.76$, $p = .03$, indicating a poor fit. Modification indices suggested that the error between the basic three-fold items and the basic four-fold items should be allowed to co-vary, as well as the error between the atypical two-fold items and the atypical three-fold items. Because it makes theoretical sense that the unexplained variance in these item types would be related, the model was re-run with these error terms being allowed to co-vary. This resulted in a

model with much more adequate fit indices and these error terms were allowed to co-vary in all models following.

The estimated Model A can be seen in Figure 3, and again, the numbers above the straight, single-headed arrows are the standardized factor loadings. As shown in the figure, the intercorrelation between WMC and VZ was quite high at .62 and was significantly greater than zero. The fit indices for this model were all quite good, as summarized in Table 6 (Model A). In addition to a non-significant chi square, $\chi^2(34) = 31.09$, $p = .61$, and a small χ^2/df value of .91, the CFI value of 1.00 was above the criterion value of .90, the RMSEA value of .00 was below the criterion value of .05, and the GFI and AGFI values (.96 and .94 respectively) were above the criterion value of .90.

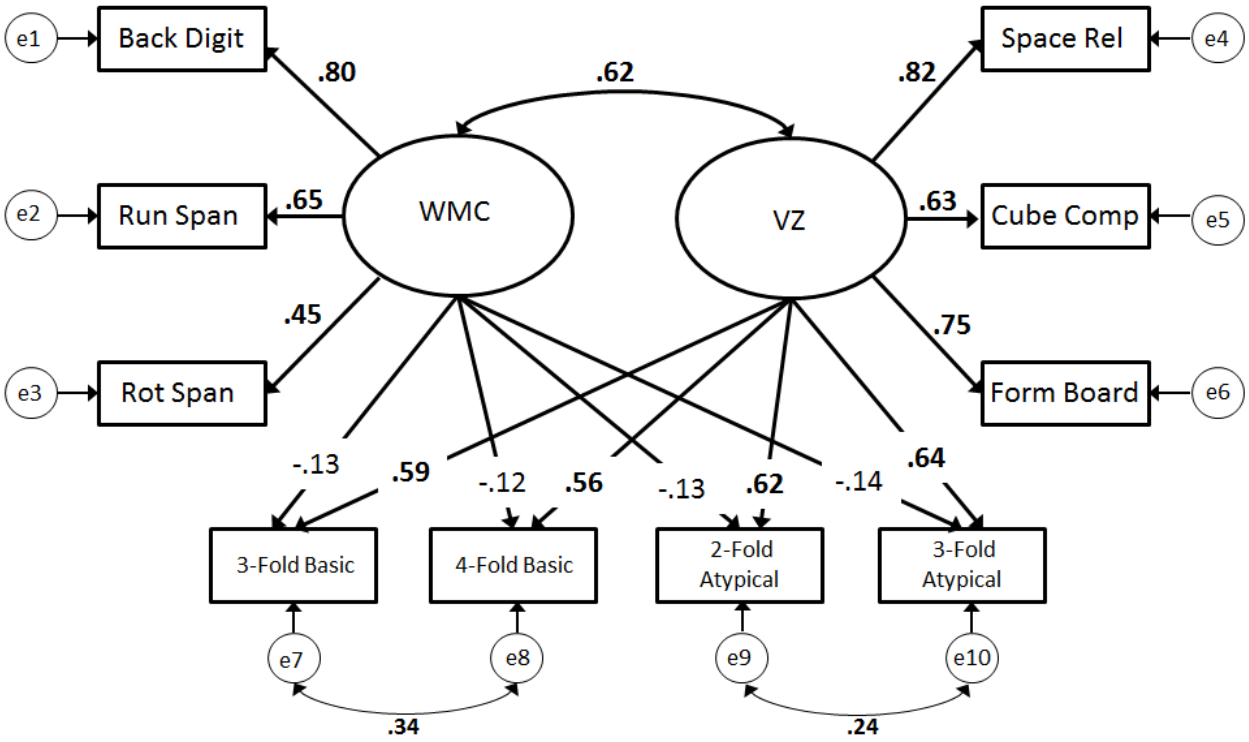


Figure 3. Model A with constrained factor loadings. Values represent the standardized regression weights. All boldface loadings and correlations are significant at the .05 level.

After ascertaining that the constrained model did in fact provide a good fit for the data, the next step was to evaluate the fit of the unconstrained model. Again, it was hypothesized that the model in which the Punched Holes items were allowed to load freely with WMC and VZ would be a better fit than the constrained model. The estimated model can be seen in Figure 4. Again, the fit indices for this model were good and are summarized in Table 6 (Model B). In addition to a non-significant chi square, $\chi^2(28) = 20.92$, $p = .83$, and a small χ^2/df value of .75, the CFI value of 1.00 was above the criterion value of .90, the RMSEA value of .00 was below the criterion value of .05, and the GFI and AGFI values (.97 and .95 respectively) were above the criterion value of .90.

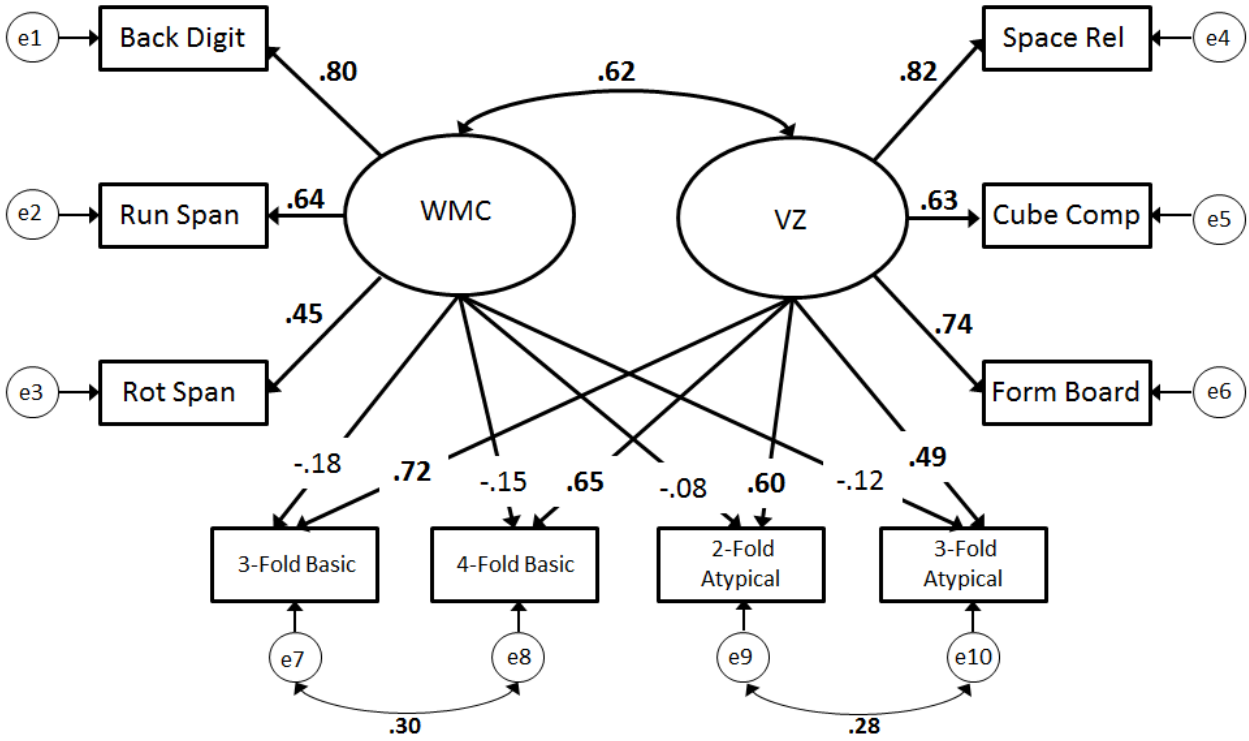


Figure 4. Model B with unconstrained factor loadings. Values represent the standardized regression weights. All boldface loadings and correlations are significant at the .05 level.

To address the hypothesis that the unconstrained model should provide a better fit for the data than the constrained model a chi-square difference test comparing the fit of the two models was conducted. The chi-square difference test indicated that there was no significant difference between the models, $\chi^2(6) = 10.17$, $p = .12$. To further address which model was a better fit for the data, the Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC) were compared across models. Both of these measures represent comparative measures of fit in which lower values indicate a better fit. Contrary to hypotheses, the constrained Model A had a better fit (AIC = 73.09, BIC = 136.18) than the unconstrained Model B (AIC = 74.92, BIC = 156.02). These results suggest that there are not four distinct item types and that all of the item types load similarly onto the VZ and WMC factors. Further, all of the items have negative factors loadings with WMC that do not significantly differ from zero suggesting that WMC is not significantly contributing to performance after the contribution of VZ is accounted for.

Table VI

Fit Indices for the Confirmatory Factor Analysis Models with Full Data (N = 149)

Model	*df*	χ^2	χ^2/df	CFI	RMSEA	GFI	AGFI	AIC	BIC
A (Fig. 3, constrained)	34	31.09	0.91	1.00	0.00	0.96	0.94	73.09	136.18
B (Fig. 4, unconstrained)	28	20.92	0.75	1.00	0.00	0.97	0.95	74.92	156.02
C (Fig. 5, basic vs. atypical)	32	25.50	0.80	1.00	0.00	0.97	0.94	71.50	140.59
D (Fig. 6, fewer vs. more fold)	32	27.81	0.87	1.00	0.00	0.97	0.94	73.82	142.91

Although the previous models did not provide evidence for the existence of four unique item types that differentially load onto VZ and WMC, it could be that there are only two categories of items, those with atypical or occluded folds and those without, and that these two

categories of items have differential factor loadings. A model in which the factor loadings were constrained so that the basic fold items loaded differently onto WMC and VZ than the atypical fold items was evaluated. The estimated model (Model C) can be seen in Figure 5. Again, the fit indices for this model were good and are summarized in Table 6 (Model C). There was a non-significant chi square, $\chi^2(32) = 25.50$, $p = .76$, and a small χ^2/df value of .80. The CFI value of 1.00 was above the criterion value of .90, the RMSEA value of .00 was below the criterion value of .05, and the GFI and AGFI values (.97 and .94 respectively) were above the criterion value of .90.

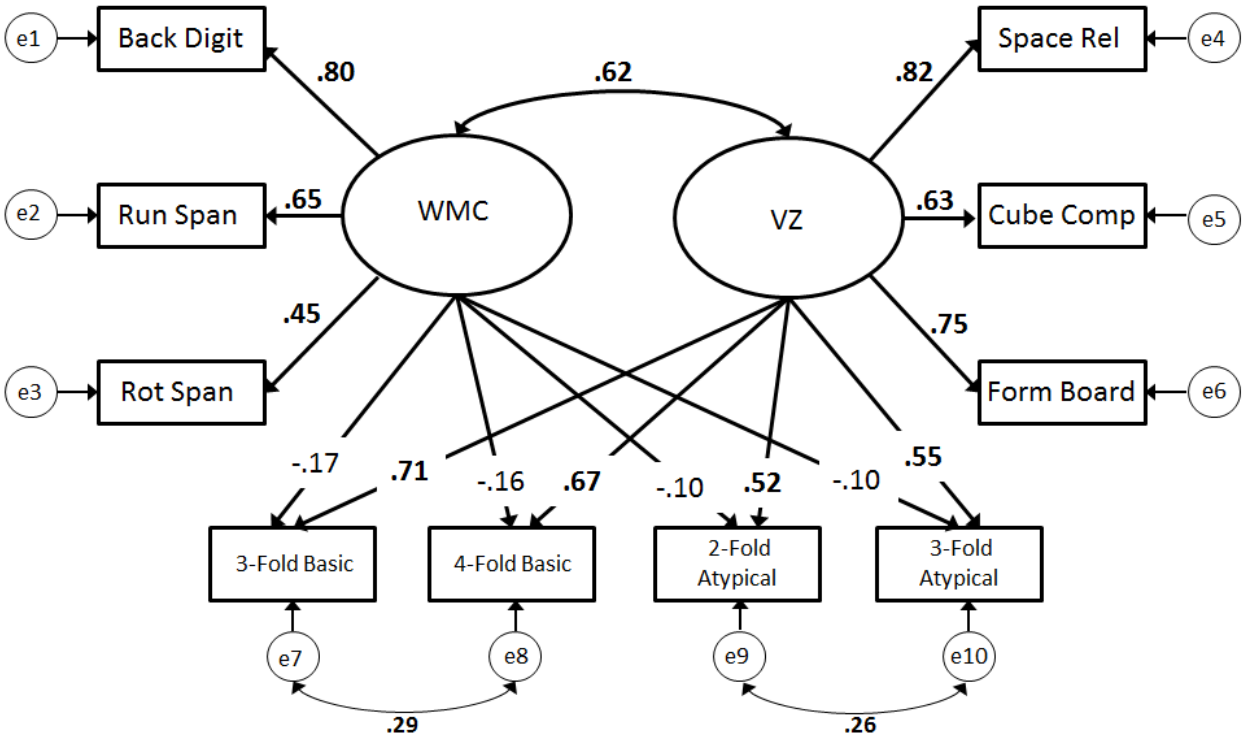


Figure 5. Model C with factor loadings for basic items constrained to be equal and factor loadings for atypical items constrained to be equal. Values represent the standardized regression weights. All boldface loadings and correlations are significant at the .05 level.

An alternate two item category model was also evaluated. In this model (Model D), items with fewer folds were compared to items with more folds. Specifically, the basic three-fold items and atypical two-fold items were constrained to have the same factor loadings and the basic four-fold and atypical three-fold items were constrained to have the same factor loadings. The estimated Model D can be seen in Figure 6. As shown in Table 6, with a non-significant chi square, $\chi^2(32) = 27.82$, $p = .68$, and a small χ^2/df value of .87, the fit indices were good.

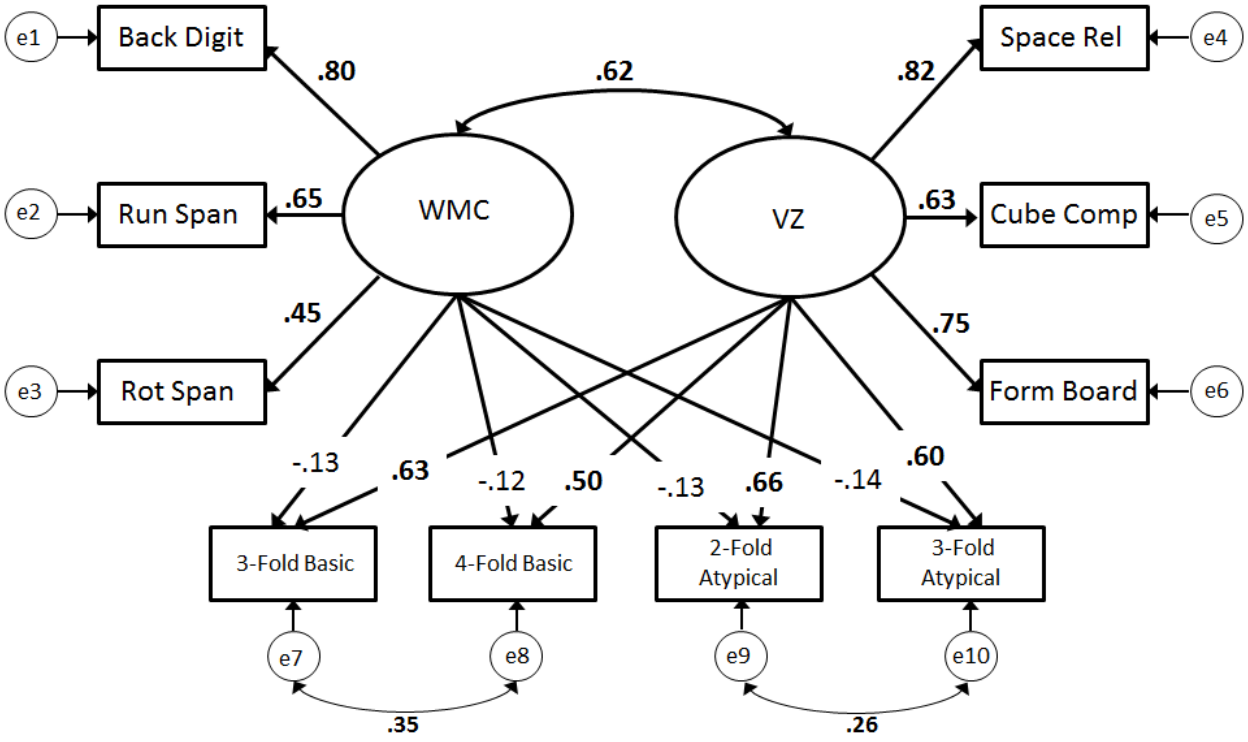


Figure 6. Model D with factor loadings for fewer fold items constrained to be equal and factor loadings for greater fold items constrained to be equal. Values represent the standardized regression weights. All boldface loadings and correlations are significant at the .05 level.

To more directly compare the fit of models C and D, a chi-square difference test comparing their fits was conducted. The chi-square difference test indicated that there was no significant difference between the models, $\chi^2(1) = 2.31$, $p = .87$. To further address which model was a better fit for the data, the AIC and BIC were compared across models, again for these measures lower values indicate a better fit. Model C (basic folds versus atypical folds) had a better fit (AIC = 71.50, BIC = 140.59) than Model D (fewer folds versus more folds) (AIC = 73.82, BIC = 142.91). These results suggest that Punched Holes items with atypical or occluded folds may represent a different construct than items without atypical or occluded folds, and that the number of folds in an item does not impact factor loadings.

An additional chi-squared difference test was performed in order to compare Model C to Model A. No difference was found between Model C and Model A, $\chi^2(2) = 5.59$, $p = .94$. Because of this lack of difference in model fit as indicated by the chi-squared test, other measures of model fit were considered. As shown in Table 6, Model C had the lowest AIC (71.50) and also the lowest χ^2/df at .80. These results provide further evidence that there may be two item subtypes on the Punched Holes task, those with basic folds and those with atypical or occluded folds.

One important issue to consider when thinking about the differences between Punched Holes items is their relative difficulty levels. As briefly described earlier, the atypical fold items (collapsed across number of folds, $M = .30$, $SD = .20$) were significantly more difficult than the basic fold items (collapsed across number of folds, $M = .60$, $SD = 25$), $t(148) = 14.66$, $p < .001$. Because of these large differences in performance on the item subtypes it was important to try running the CFA models on a smaller set of items that were more closely matched on difficulty. A subset of five basic items and five atypical items were selected based on average performance

(see Table 7 for item means and standard deviations). Although still significantly more difficult, performance on the reduced set of atypical fold items (M = .42, SD = .25) was more similar to performance on the reduced set of basic items (M = .50, SD = .27), $t(148) = 3.44$, $p < .001$.

Table VII

Punched Holes Items Means and Standard Deviations

	Basic				Atypical			
	3 Folds		4 Folds		2 Folds		3 Folds	
Item Number	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
1	.72	.45	.54*	.50	.46*	.50	.28	.45
2	.75	.44	.38*	.49	.44*	.50	.46*	.50
3	.71	.46	.68	.47	.17	.37	.14	.35
4	.56*	.50	.62	.49	.37*	.48	.19	.39
5	.48*	.50	.54*	.50	.11	.31	.36*	.48

Note. Items marked with an asterisk were included in the matched difficulty models.

With these reduced item sets, Models A and B were run again (A_1 and B_1). In Model A_1 basic fold items and atypical fold items were constrained to have the same factor loadings on WMC and VZ, while in Model B_1 the two item types were allowed to load freely onto the WMC and VZ factors. The estimated model can be seen in Figure 7. The fit indices for this reduced model were all quite good, as summarized in Table 8 (Model A_1). In addition to a non-significant chi square, $\chi^2(19) = 12.34$, $p = .87$, and a small χ^2/df value of .65, the CFI value of 1.00 was above the criterion value of .90, the RMSEA value of .00 was below the criterion value of .05, and the GFI and AGFI values (.98 and .96 respectively) were above the criterion value of .90.

Once it was determined that Model A_1 did in fact provide a good fit for the data, the next step was to evaluate the fit of the model when none of the loadings were constrained. The estimated model can be seen in Figure 8. Again, the fit indices for this model were good and are summarized in Table 8 (Model B_1). In addition to a non-significant chi square, $\chi^2(17) = 11.77$, $p = .81$, and a small χ^2/df value of .69, the CFI value of 1.00 was above the criterion value of .90, the RMSEA value of .00 was below the criterion value of .05, and the GFI and AGFI values (.98 and .96 respectively) were above the criterion value of .90.

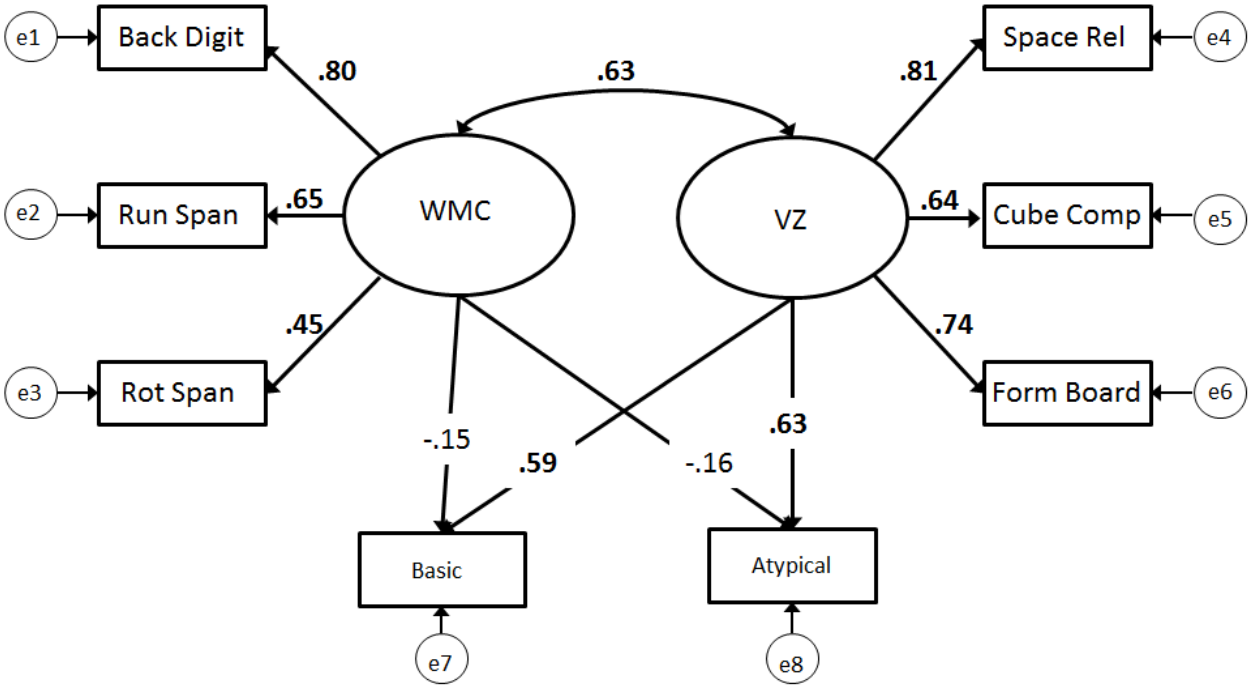


Figure 7. Reduced data Model A_1 with constrained factor loadings. Values represent the standardized regression weights. All boldface loadings and correlations are significant at the .05 level.

To again address the hypothesis that the unconstrained model should provide a better fit for the data than the constrained model, a chi-square difference test comparing the fit of Models A_1 and B_1 was conducted. The chi-square difference test indicated that there was no significant difference between the models, $\chi^2(2) = 0.57$, $p = .25$. To further address which model was a better fit for the data, the AIC and BIC were compared across models. Contrary to hypotheses and similar to the results obtained for the full data, the constrained Model A_1 had a better fit (AIC = 46.34, BIC = 97.41) than the unconstrained Model B_1 (AIC = 49.77, BIC = 106.85). These results suggest that there is no difference between atypical and basic items in terms of how they load onto the VZ and WMC factors. Additionally, both item sets again have negative factor loadings with WMC that do not significantly differ from zero suggesting that WMC is not significantly contributing to performance after the contribution of VZ is accounted for.

In sum, the results from the Confirmatory Factor analyses indicate that spatial visualization overwhelmingly accounts for performance on all four types of Punched Holes items. Although distinct factor loading differences were not found for each individual item subtype, there was some indication as demonstrated in Model C, that there are two major item subtypes, those with basic folds and those with atypical folds. Overall, this pattern of results suggests that the Punched Holes test is in fact best represented a measure of VZ as opposed to WMC or executive control.

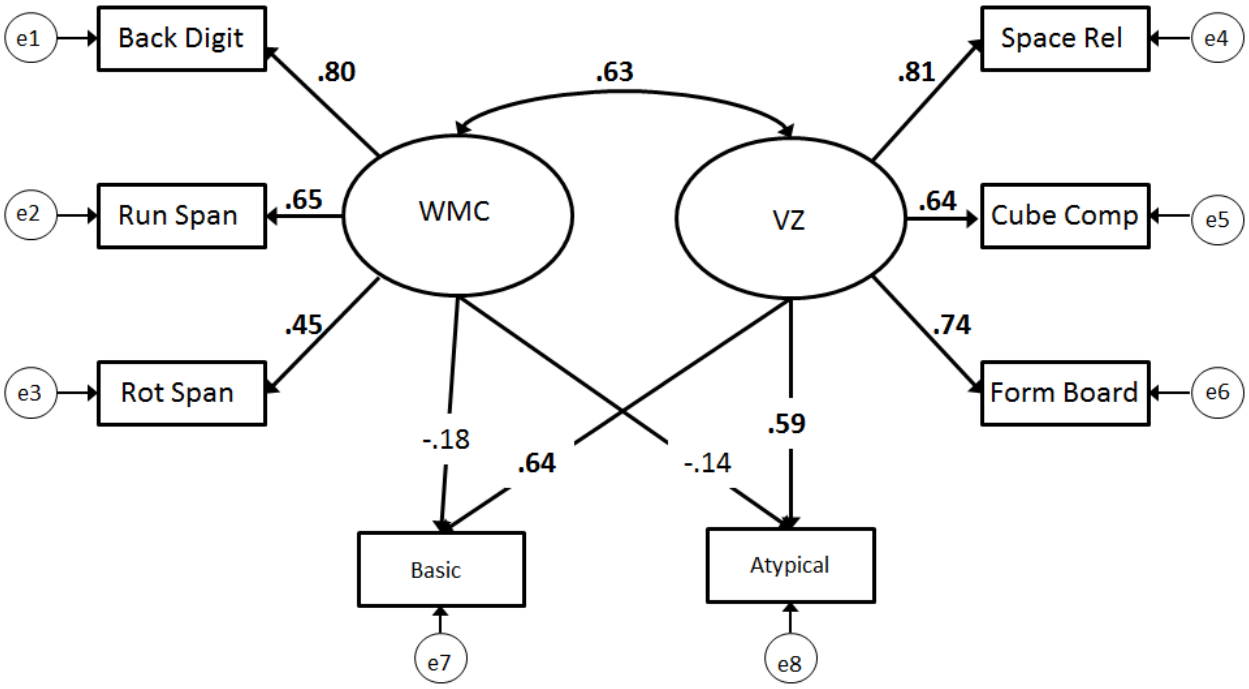


Figure 8. Reduced data Model B_1 with unconstrained factor loadings. Values represent the standardized regression weights. All boldface loadings and correlations are significant at the .05 level.

Table VIII

Fit Indices for the Confirmatory Factor Analysis Models with Matched Data (N = 149)

Model	*df*	χ^2	χ^2/df	CFI	RMSEA	GFI	AGFI	AIC	BIC
A_1 (Fig. 6, constrained)	19	12.34	0.65	1.00	0.00	0.98	0.96	46.34	97.41
B_1 (Fig. 7, unconstrained)	17	11.77	0.69	1.00	0.00	0.98	0.96	49.77	106.85

3.2.1 Follow-up WMC-only Model. In all of the above models, VZ is clearly a very strong predictor of performance on all four PH item types. It seems likely that because of this, the unique effects of WMC on different item types were not able to be shown. In order to look at the role of WMC on the individual item types as originally proposed, it was assessed separately. For this model only the WMC factor was included and items were allowed to load freely onto the factor. This model resulted in a significant chi square test indicating the model did not fit the data well, $\chi^2(14) = 29.46$, $p < .01$. Despite this, the main goal of running this model was to assess whether basic fold items would load more heavily onto WMC than the atypical items. As can be seen in Figure 9, the contribution of WMC was statistically significant for all four item types, but importantly, was the lowest for the atypical three-fold items (.22) and similar for the basic three-fold, basic four-fold and atypical two-fold items (.30, .31, and .33 respectively). These path coefficients suggest that WMC contributes least to the three-fold atypical items. Although the chi-square test was significant, the CFI value of .92 was above the criterion value of .90, and the GFI and AGFI values (.95 and .90 respectively) were at or above the criterion value of .90. However, the RMSEA value of .08 did not fall below the .05 criterion level. The results of this WMC-only model indicate that the three-fold atypical items loaded less onto WMC that the other three item types and that being high in WMC contributes less to success on these most difficult items.

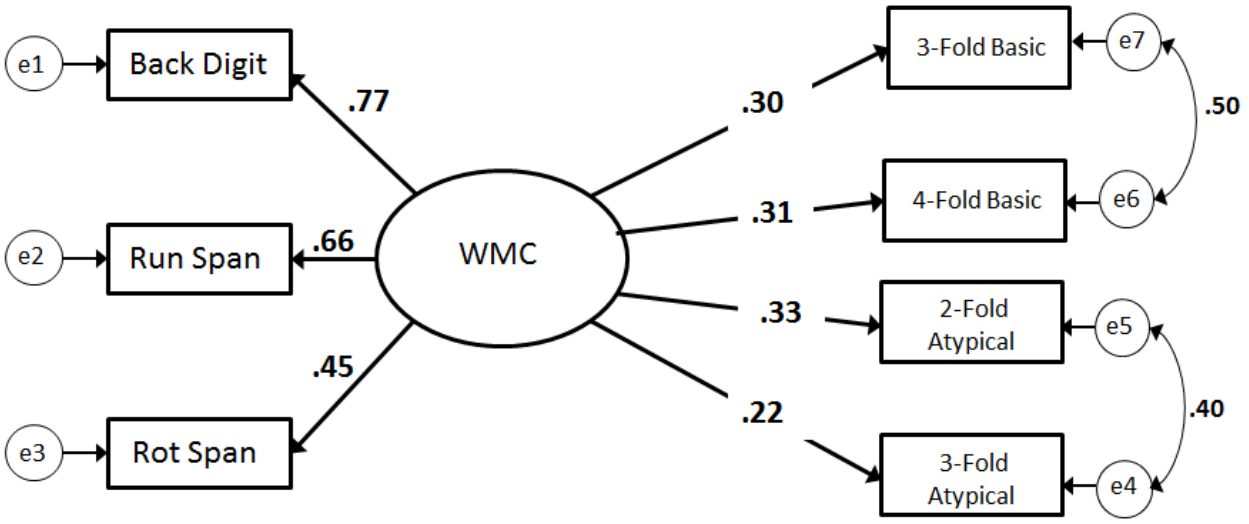


Figure 9. Working memory only model. Values represent the standardized regression weights. All loadings and correlations were significant at the .05 level.

3.3 Item Difficulty and Correct Solution Time

In order to further assess how item difficulty as a function of number of folds and fold occlusion affected performance, overall item response latencies were investigated. As shown in Figure 10, for correct items, as the number of folds in the item increased, so did response latency. A 2 (Fold type: basic, atypical) X 2 (Number of folds: fewer, more) repeated measures ANOVA was conducted to assess the effect of number of folds and type of fold on correct solution time. This analysis revealed a main effect for fold type, $F(1, 148) = 25.61$, $p < .001$, $\eta^2 = .22$, and number of folds, $F(1, 148) = 8.05$, $p < .01$, $\eta^2 = .08$. The interaction between these factors was not significant, $F(1, 148) = 1.26$, *ns*. Differences in response latency on item types were also computed using overall reaction times on all trials and showed similar patterns[i]. Hypotheses were supported and the results indicated that correct solution times were longer for atypical items compared to basic items and that correct solution times were also longer for

greater fold items compared to fewer fold items. It is also important to note that response latencies on all item types in this task were longer than what has been found on other complex spatial visualization tasks (Cooper & Shepard, 1973; Mumaw et al., 1984).

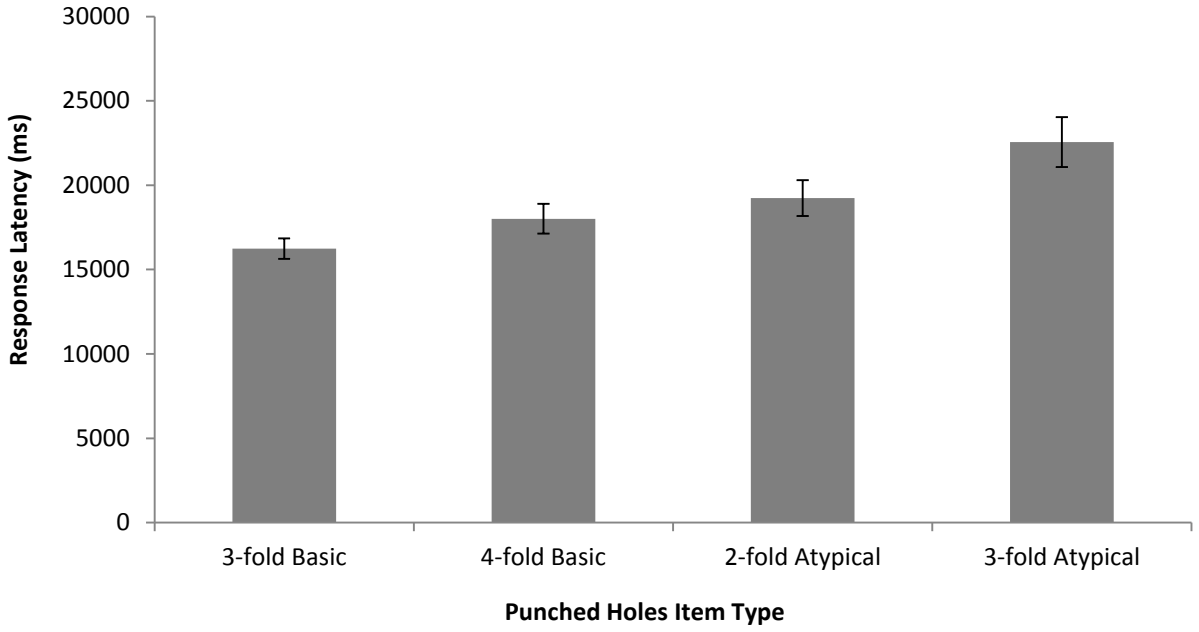


Figure 10. Mean response latencies on correct trials as a function of Punched Holes item type. Error bars represent standard errors.

3.3.1 Analyses as a Function of VZ. Following Cooper and Shepard (1973), solution accuracy and response latencies for the four item types were further analyzed as a function of VZ ability. Participants were split into high and low VZ based on their VZ factor scores which were comprised of performance on the Space Relations task, the Form Board task, and the Cube Comparison task. Participants with factor scores greater than zero were categorized as high VZ and participants with factor scores lower than zero were categorized as low VZ. This resulted in 68 high VZ participants and 81 low VZ participants.

3.3.1.1 Accuracy by VZ. A 2 (Fold type: basic, atypical) X 2 (Number of folds: fewer, more) X 2 (VZ: high, low) repeated measures ANOVA was conducted to assess the effect of fold number, fold type, and spatial visualization ability on participants' solution accuracy (see Figure 11). The results from the ANOVA showed a main effect for type of folds such that performance was more accurate for basic fold items as compared to atypical fold items, $F(1, 147) = 232.27$, $p < .001$, $\eta^2 = .61$. There was also a main effect for number of folds such that performance on fewer fold items was more accurate than performance on greater fold items, $F(1, 147) = 15.27$, $p < .001$, $\eta^2 = .09$. There was also a main effect for VZ such that high VZ participants were more accurate than low VZ participants, $F(1, 147) = 59.78$, $p < .001$, $\eta^2 = .29$.

The interaction between VZ and type of fold was significant, $F(1, 147) = 8.83$, $p < .01$, $\eta^2 = .06$. Specifically, high VZ participants were more accurate than low VZ participants on basic fold items, $t(147) = 7.50$, $p < .001$, and on the atypical fold items, $t(147) = 4.63$, $p < .001$, but the difference in performance between basic and atypical items was greater for the high VZ participants with their performance being especially high on the basic items, $t(147) = 2.97$, $p < .01$. The two-way interaction between VZ and number of folds was not significant, $F(1, 147) = 1.03$, *ns*. There was an interaction between number and type of fold, $F(1, 147) = 4.59$, $p < .05$, $\eta^2 = .03$. Specifically, participants were more accurate on fewer fold items when the items were basic, $t(148) = 4.13$, $p < .001$, but there was no difference between fewer and greater fold items when they were atypical, $t(148) = 1.11$, *ns*. The three-way interaction between number of folds, type of folds, and spatial visualization ability was not significant, $F < 1$.

These results indicate that high VZ participants were more accurate than the low VZ participants and again that more folds and atypical folds were more difficult than fewer folds and basic folds. Being high VZ seemed to contribute most to being able to perform the basic fold

items. This result is consistent with the generally higher weighting of VZ on basic than atypical-fold items in the above models.

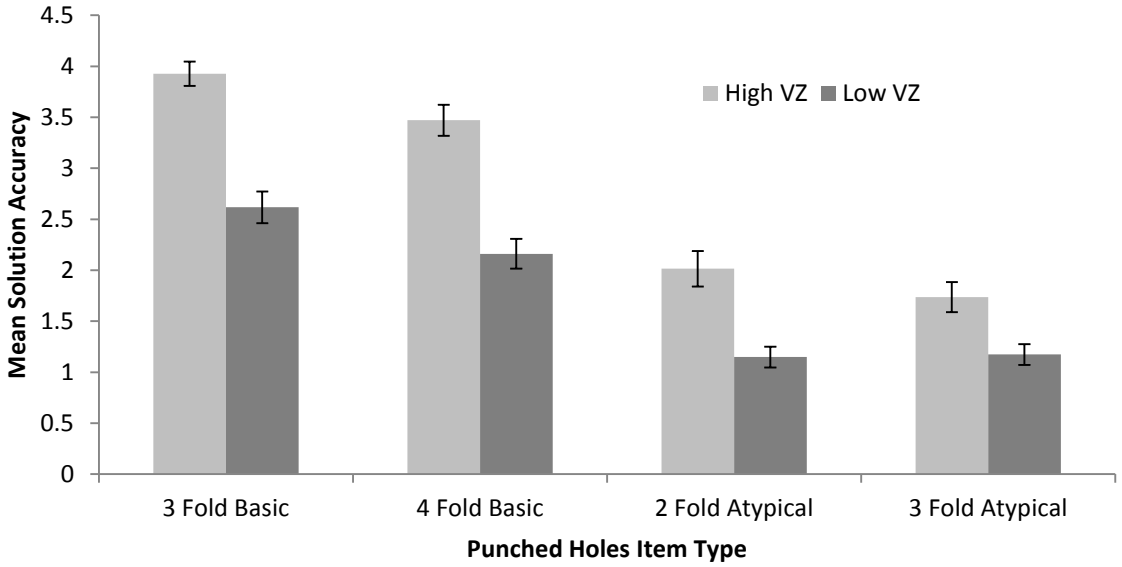


Figure 11. Mean solution accuracy as a function Punched Holes item type and spatial visualization. Error bars represent standard error of the mean.

3.3.1.2 Latency by VZ. Because high VZ participants correctly solved more items of all four types, average response latencies for correct items only would have been based on unequal sample sizes. For example, a mean reaction time for a low VZ participant could have been based on only a single correct item, whereas a high VZ participant's average reaction time could have been based on reaction times for five correct items. To be able to include reaction time data for incorrect trials, reaction times for incorrect items were set to the maximum correct reaction time seen for each individual item. This resulted in reaction time scores that took into account how quickly a participant was able to select the correct response. If they were unable to select the correct response, they received the longest correct reaction time for that item.

A 2 (Fold type: basic, atypical) X 2 (Number of folds: fewer, more) X 2 (VZ: high, low) repeated measures ANOVA was conducted to assess the effect of fold number, fold type, and spatial visualization ability on participants' response latencies (see Figure 12). The results from the ANOVA showed a main effect for type of folds such that performance was faster for basic fold items as compared to atypical fold items, $F(1, 147) = 320.27$, $p < .001$, $\eta^2 = .69$. There was also a main effect for number of folds such that performance on fewer fold items was faster than performance on greater fold items, $F(1, 147) = 87.86$, $p < .001$, $\eta^2 = .37$. There was also a main effect for VZ such that high VZ participants responded more quickly than low VZ participants, $F(1, 147) = 20.04$, $p < .001$, $\eta^2 = .12$.

All three of the two-way interactions were also significant. The interaction between VZ and type of fold was significant, $F(1, 147) = 14.53$, $p < .001$, $\eta^2 = .09$. Specifically, high VZ participants had shorter response latencies than low VZ participants on basic fold items, $t(147) = 5.20$, $p < .001$, but there was no difference between the high and low VZ participants response latencies on the atypical fold items, $t < 1$. The two-way interaction between VZ and number of folds was also significant, $F(1, 147) = 7.83$, $p < .01$, $\eta^2 = .05$. High VZ participants had shorter response latencies than the low VZ participants on the fewer fold items, $t(147) = 5.56$, $p < .001$, but there was no difference in response latencies between the high and low VZ participants on the greater fold items, $t(147) = 1.71$, *ns*. There was also an interaction between number and type of fold, $F(1, 147) = 5.04$, $p < .05$, $\eta^2 = .03$. Specifically, participants showed shorter response latencies for fewer fold items, both when the items were basic, $t(148) = 8.22$, $p < .001$, and atypical, $t(148) = 5.10$, $p < .001$, but this difference was greater for the basic fold items, $t(148) = 2.37$, $p < .02$. The three-way interaction between number of folds, type of folds, and spatial visualization ability was not significant, $F(1, 147) = 1.94$, *ns*. These analyses were also

conducted for overall response latency on all items including incorrect trials and revealed similar patterns [ii].

Taken together, these results indicate that overall high VZ participants completed the items more quickly than the low VZ participants and that more folds and atypical folds took longer than fewer folds and basic folds. Again, being high VZ seemed to contribute most to being able to perform the fewer fold and basic fold items. This result is consistent with the generally higher weighting of VZ on basic than atypical-fold items in the above models.

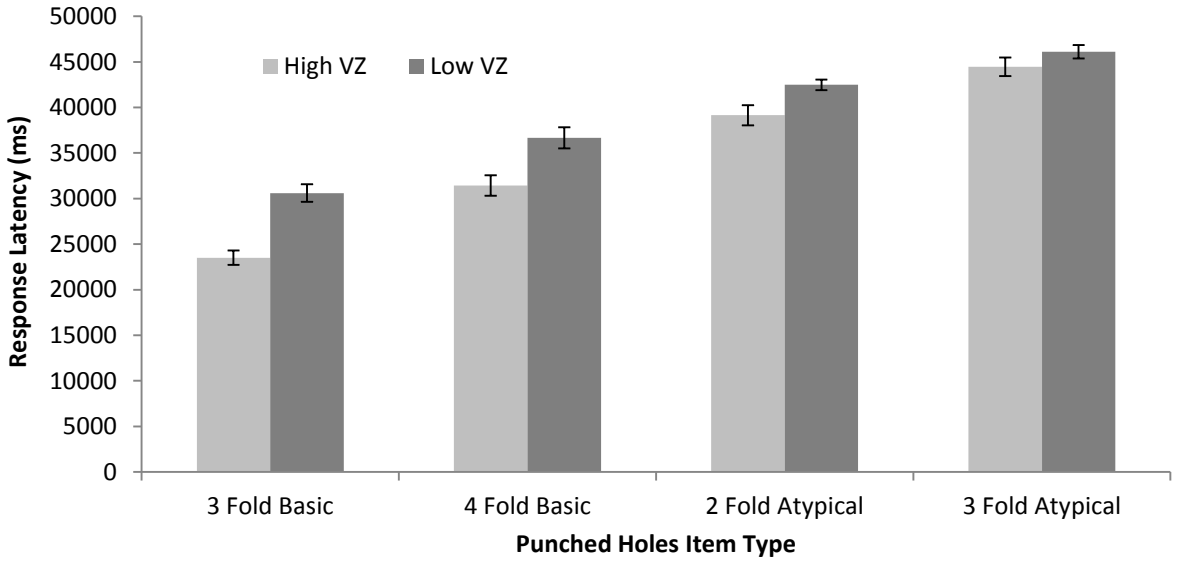


Figure 12. Mean response latencies as a function of Punched Holes item type and spatial visualization ability. Error bars represent standard errors.

3.3.2 Analyses as a Function of WMC. Next, solution accuracy and response latency as a function of WMC were analyzed. Participants were split into high and low WMC based on their composite WMC scores. Composite scores were comprised of performance on the Running Span task and the Backwards Digit Span task. Rotation span scores were not included in the

composite WMC measure for this analysis because, as shown in Table 3 presented earlier, Rotation Span did not correlate with the Punched Holes items in the same manner as the other two span tasks. Participants with a composite score above the mean were categorized as high WMC and participants with a composite score below the mean were categorized as low WMC. This resulted in 70 high WMC participants and 79 low WMC participants.

3.3.2.1 Accuracy by WMC. To assess the effect of fold number, fold type, and working memory capacity on participants' solution accuracy a 2 (Fold type: basic, atypical) X 2 (Number of folds: fewer, more) X 2 (WMC: high, low) repeated measures ANOVA was conducted (see Figure 13). The results from the ANOVA again showed a main effect for type of folds such that performance was more accurate for basic fold items as compared to atypical fold items, $F(1, 147) = 215.71$, $p < .001$, $\eta^2 = .60$. There was again also a main effect for number of folds with performance on fewer fold items being more accurate than performance on greater fold items, $F(1, 147) = 14.99$, $p < .001$, $\eta^2 = .09$. There was also a main effect for WMC indicating that high WMC participants were more accurate than low WMC participants, $F(1, 147) = 8.38$, $p < .01$, $\eta^2 = .05$.

The interaction between number and type of fold was again significant, $F(1, 147) = 4.58$, $p < .05$, $\eta^2 = .03$, and there was no interaction between WMC and type of fold or between WMC and number of folds, Fs < 1.

There was a marginal three-way interaction between number of folds, type of folds, and WMC, $F(1, 147) = 3.45$, $p = .07$, $\eta^2 = .02$. To follow up the three-way interaction, two additional repeated-measures ANOVAs were conducted, one looking at solution accuracy on basic three-fold items and atypical two-fold items as a function of WMC and the other assessing solution accuracy on basic four-fold items and atypical three fold items as a function of WMC. When

looking at basic three-fold items and atypical two-fold items there was a significant main effect for item type indicating that solution accuracy was greater for the basic three-fold items compared to atypical two-fold items, $F(1, 147) = 167.57$, $p < .001$, $\eta^2 = .53$. There was also a significant main effect for WMC indicating that high WMC individuals were more accurate than low WMC individuals, $F(1, 147) = 8.48$, $p < .01$, $\eta^2 = .06$. There was no significant interaction between WMC and item type, $F < 1$.

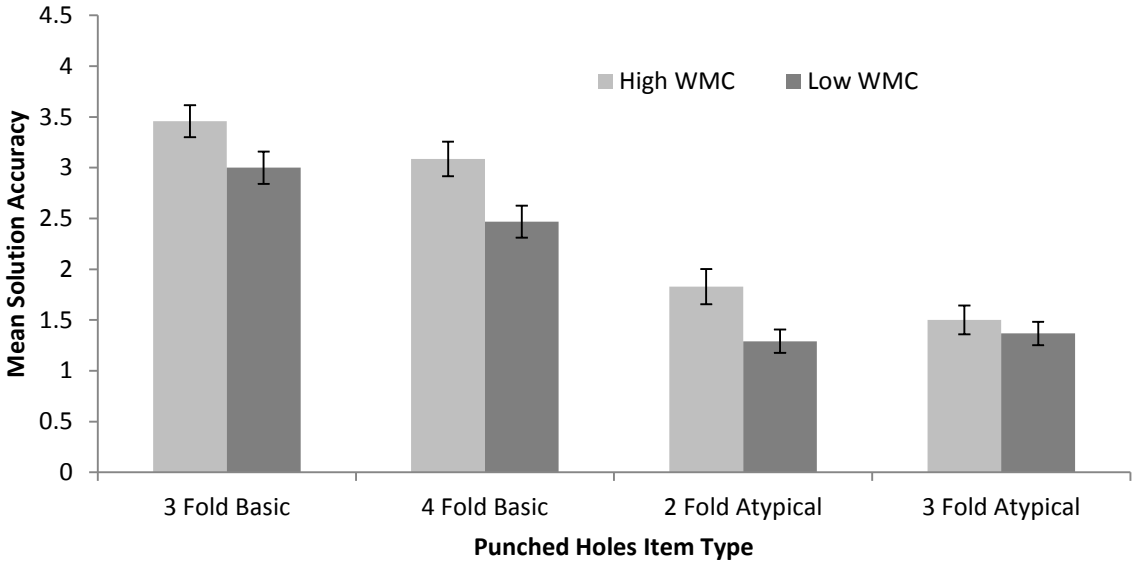


Figure 13. Mean solution accuracy as a function of Punched Holes item type and working memory capacity. Error bars represent standard errors.

When looking at the basic four-fold items and the atypical three-fold items as a function of WMC, the main effect for item type was significant such solution accuracy was greater for basic four-fold items as compared to atypical three-fold items, $F(1, 147) = 113.13$, $p < .001$, $\eta^2 = .44$. There was also a main effect for WMC again indicating that high WMC individuals were more accurate than low WMC individuals, $F(1, 147) = 5.14$, $p < .03$, $\eta^2 = .03$. There was also a

marginal two-way interaction between item type and WMC, $F(1, 147) = 3.68$, $p = .06$, $\eta^2 = .02$. This interaction was driven by the fact that high WMC individuals were more accurate than low WMC individuals on the basic four-fold items, $t(147) = 2.65$, $p < .01$, but did not differ on the atypical three-fold items, $t < 1$.

These results indicate that high WMC participants were more accurate than the low WMC participants and again that more folds and atypical folds were more difficult than fewer folds and basic folds. As was the case with VZ, being high WMC seemed to contribute most to being able to perform the basic fold items. This result is consistent with the higher weighting of WMC on the basic items and two-fold atypical items than the three-fold atypical items in the WMC-only model.

3.3.2.2 Latency by WMC. As in the VZ latency analyses, the adjusted response latency measure was used. For this measure, reaction times for incorrect items were set to the maximum correct reaction time seen for each individual item. A 2 (Fold type: basic, atypical) X 2 (Number of folds: fewer, more) X 2 (WMC: high, low) repeated measures ANOVA was conducted to assess the effect of fold number fold type and working memory capacity on participants' response latencies (see Figure 14). The results from the ANOVA again showed a main effect for type of folds such that participants were faster on basic fold items as compared to atypical fold items, $F(1, 147) = 293.94$, $p < .001$, $\eta^2 = .67$. There was again also a main effect for number of folds such that performance on fewer fold items was faster than performance on greater fold items, $F(1, 147) = 83.83$, $p < .001$, $\eta^2 = .36$. There was also a main effect for WMC such that high WMC participants performance faster than low WMC participants, $F(1, 147) = 4.68$, $p < .05$, $\eta^2 = .03$.

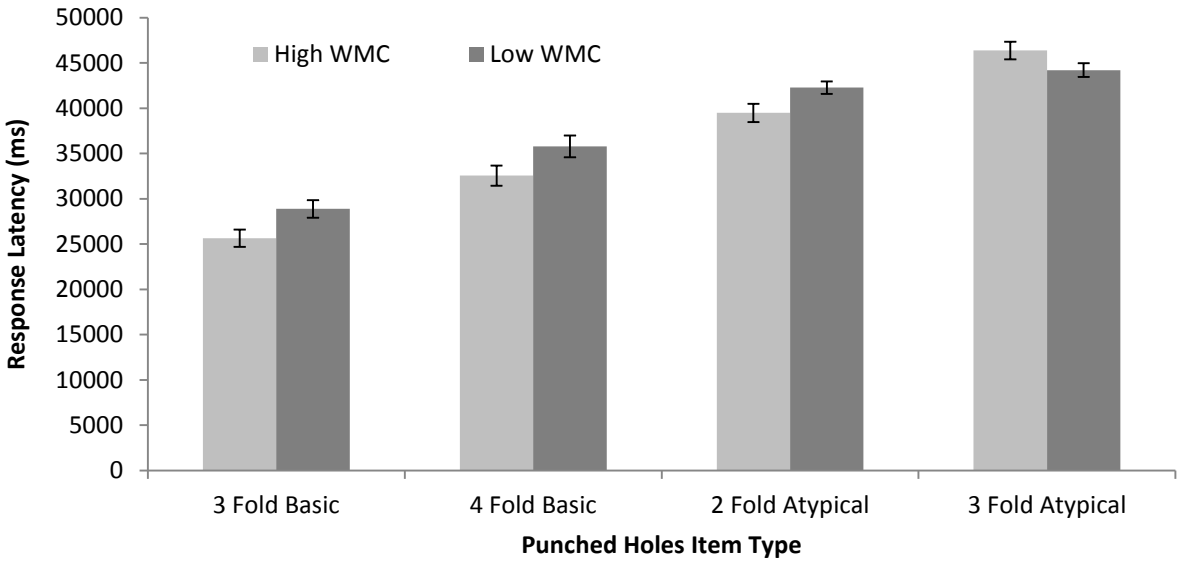


Figure 14. Mean response latencies as a function of Punched Holes item type and working memory capacity. Error bars represent standard errors.

All three of the two-way interactions were also significant. The significant interaction between WMC and type of folds, $F(1, 147) = 4.12$, $p < .05$, $\eta^2 = .03$, was driven by high WMC participants having shorter response latencies on basic fold items as compared to low WMC participants, $t(147) = 2.57$, $p < .02$, but having equal response latencies on the atypical fold items, $t < 1$. The interaction between WMC and number of folds was also significant, $F(1, 147) = 4.03$, $p < .05$, $\eta^2 = .03$. Specifically, high WMC participants had shorter response latencies than low WMC individuals on the fewer fold items, $t(147) = 3.02$, $p < .01$, but there was no difference in response latency between the high and low WMC participants on the greater fold items, $t < 1$. Also, the interaction between number and type of fold was significant, $F(1, 147) = 5.11$, $p < .05$, $\eta^2 = .03$, again indicating that for both typical and atypical folds, participants had shorter response latencies when the items had fewer folds, but this difference was greater for the basic fold items.

All of the previous effects were qualified by a significant three-way interaction between number of folds, type of folds, and working memory capacity, $F(1, 147) = 4.98$, $p < .05$, $\eta^2 = .03$. To follow up the three-way interaction, two additional repeated-measures ANOVAs were conducted, one looking at response latency on basic three-fold items and atypical two-fold items as a function of WMC and the other assessing response latency on basic four-fold items and atypical three fold items as a function of WMC. When looking at basic three-fold items and atypical two-fold items there was a significant main effect for item type indicating that response latencies were shorter for the basic three-fold items compared to atypical two-fold items, $F(1, 147) = 283.34$, $p < .001$, $\eta^2 = .66$. There was also a significant main effect for WMC indicating that high WMC individuals responded more quickly than low WMC individuals, $F(1, 147) = 9.10$, $p < .01$, $\eta^2 = .06$. There was however no significant two-way interaction suggesting that high WMC individuals had shorter response latencies compared to low WMC individuals on both basic three-fold items and atypical two-fold items, $F < 1$.

When looking at the basic four-fold items and the atypical three-fold items as a function of WMC, the main effect for item type was significant such that response latencies were shorter for basic four-fold items as compared to atypical three-fold items, $F(1, 147) = 123.41$, $p < .001$, $\eta^2 = .46$. There was however no main effect for WMC, $F < 1$. These effects were qualified by a significant two-way interaction between item type and WMC, $F(1, 147) = 7.29$, $p < .01$, $\eta^2 = .05$. This interaction was driven by the fact that high WMC individuals had shorter response latencies than low WMC individuals on the basic four-fold items, $t(147) = 1.96$, $p < .05$, but did not differ on the atypical three-fold items, $t(147) = 1.77$, *ns*. These analyses were also conducted for overall response latency on all items including incorrect trials and revealed that high WMC individuals had shorter response latencies on all item types [iii].

These results suggest that for basic items, higher WMC individuals were able to keep track of increasing folds more efficiently than low WMC individuals. However, for atypical items there was no difference in response latency between high and low WMC individuals suggesting that being able to solve these especially challenging items does not rely on WMC. This result is consistent with the trends that were seen in the WMC-only model above.

3.4 Strategy Questionnaire Analyses

Some prior work on the PH task has indicated that the strategies participants use when attempting to solve items contributes to their solution accuracy (Hegarty, 2010). In order to address any effects of strategy use on solution accuracy, results of the strategy questionnaire were analyzed. Overall, 85 percent of the participants reported using more than a single strategy (M = 2.92, SD = 1.21) and the number of strategies used was correlated with their score on all item types (rs > .21, ps < .05) except for the atypical two-fold items. The pattern that using more strategies was generally positively related to performance replicates the results of Hegarty (2010). Further, as shown in Table 9, nearly 84 percent of the participants reported that they imagined folding the paper and punching it in their head, which was the most commonly used strategy. This being the most commonly used strategy also replicates the results reported by Hegarty.

The only strategy that was significantly correlated with performance on the PH items was the *work backwards strategy* in which participants started from the last fold and worked backwards to determine what the unfolded paper would look like. To assess if the use of this strategy significantly impacted performance on the PH items a 2 (Fold type: basic, atypical) X 2 (Number of folds: fewer, more) X 2 (Strategy: use, non-use) repeated measures ANOVA was conducted (see Figure 15). This analysis revealed a main effect for number of folds, $F(1, 147)$ =

206.27, $p < .001$,$\eta^2 = .58$, and type of folds, $F(1, 147) = 13.39$, $p < .001$,$\eta^2 = .08$, and an interaction between those variables, $F(1, 147) = 4.47$, $p < .05$,$\eta^2 = .03$. Additionally, there was a main effect for strategy-use such that participants who reported using the *work backwards strategy* performed better on all four item types than participants who did not use that strategy, $F(1, 147) = 12.90$, $p < .001$,$\eta^2 = .08$. The interactions between number of folds and strategy use, or between type of folds and strategy use were not significant and there was no three-way interaction, Fs < 1.40. The use of this strategy was also significantly correlated with VZ ($r = .28$) and WMC ($r = .19$). No other strategy was correlated with VZ, but the strategy of counting the number of folds in order to determine the final number of holes was significantly correlated with WMC ($r = .21$).

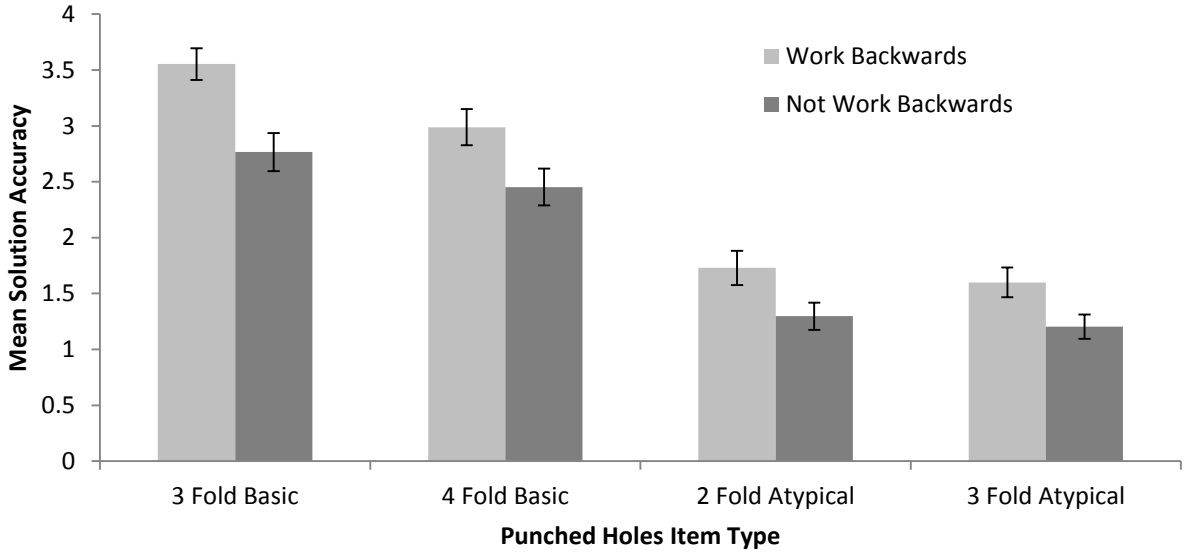


Figure 15. Mean solution accuracy as a function of item type and strategy use. Error bars represent standard errors.

Overall, these results suggest that most participants attempted to use more than one strategy and that using more strategies was related to better performance overall. However, only one strategy, working backwards, was related to better performance across all four item types and was positively related to VZ and WMC.

Table IX

Reported Strategy-Use and Correlations with PH Scores, WMC, and VZ

Strategy	N	3-Basic	4-Basic	2-Atyp	3-Atyp	VZ	WMC
Imagined folding the paper and punching the hole in my mind	125 (83.9%)	.03	.03	-.06	-.10	-.03	.06
Started at the last step and worked backward	85 (57.0%)	**.28**	**.18**	**.17**	**.18**	**.28**	**.19**
Figured out where one hole would be and eliminated answer choices that did not have a hole in that location	84 (56.4%)	-.05	.01	-.15	.01	-.07	-.02
Figured out how many folds/sheets of paper were punched through and how many holes there would be	68 (45.6%)	.13	.10	.12	-.01	.10	**.21**
Used the number of holes/folds to eliminate some of the answer choices	67 (45.0%)	.05	.05	.04	.14	.05	.06

Note. Bolded correlations are significant at the .05 level.

4. Discussion

The goal of this study was to examine performance on four subtypes of Punched Holes items and their relationship with spatial visualization and working memory capacity at the level of latent variables. Specifically, the main goal was to assess the potentially differential roles that VZ and WMC play in performance on the different PH item subtypes.

In order to address these goals, new items were created for the PH task, and as a first step it was important to determine that performance on the different subtypes varied as expected. Results demonstrated that items with more folds were more difficult than items with fewer folds. Also, items with atypical folds were more difficult than items with basic folds. These results are consistent with those obtained previously by Salthouse et al. (1989) and Kyllonen et al. (1984).

Second it was important to determine that the tasks chosen to represent each latent variable in fact did so. Three tasks were chosen to represent each latent construct and results demonstrated that the VZ tasks were all significantly correlated with each other. The three WMC tasks were also significantly correlated with each other, although relations were lower with Rotation Span. However, using both PCA and CFA, the selected tasks were shown to load well onto their respective factor. These correlations and factor loadings replicated those previously found in the literature (Kane et al., 2004; Kozhevnikov, Motes, & Hegarty, 2007; Salthouse et al., 1990), however the correlations for Rotation Span in the current study were lower than what has been found in previous work (about .50; Foster et al., 2014).

Once it was determined that two latent constructs could be created, WMC and VZ, the next goal was to assess the role that these constructs played in performance on the PH item subtypes. It was hypothesized that differential roles for WMC and VZ would be seen for the different item types, specifically that VZ would contribute more to performance on the atypical

items and that WMC would contribute more to performance on the basic items, especially those with a greater number of folds. Overall, although the original Punched Holes task has been used across the literature to represent various constructs including spatial visualization, executive functioning, working memory capacity and general reasoning ability, the results of this study demonstrated that performance on all of the item subtypes was overwhelmingly accounted for by VZ and not WMC. The CFAs suggested that there were not four distinct item types and that all of the item types had similar factor loadings on VZ and WMC. Further, the loadings with WMC did not differ from zero and suggested that WMC did not significantly contribute to performance after the unique contribution of VZ was accounted for. Also in line with these findings is the larger effect size for the main effect of VZ compared to the effect size for the main effect of WMC in the accuracy and latency ANOVA analyses. These results support the idea that the PH task represents one's ability to manipulate or transform images or spatial patterns rather than representing more general attentional control or information regulation.

These results bring into question the validity of the conclusion reached by Miyake et al. (2001) that executive functioning contributes almost exclusively to the variance in spatial visualization performance. Unfortunately Miyake et al. did not test for unique effects of spatial skills against executive functioning in their design so direct comparison between the current results and their results cannot be made. Additionally, directly comparing the results of the current study to those of Miyake et al. is not appropriate because the tasks used to represent each factor differ between the two studies. For example, their factor representing WMC was actually composed of a combination of STM and WMC tasks whereas the factor in the current study was strictly WMC tasks. Additionally, Miyake et al. used a design that distinguished between different spatial subtypes rather than a single spatial factor. However, the results of the current

study do align with those of Miyake et al. because both were unable to clearly distinguish a separate role for WMC in performance on VZ tasks, but in particular for the current study, performance on any PH items.

There are several potential reasons for why the current approach may have been unable to demonstrate a unique role for WMC. First, as expected based on previous work looking at VZ and WMC, there was a strong correlation between the two factors. However, the correlation found in this study (r = .63) was quite high compared to what has been found in previous latent variable analyses (r = .24; Kane et al., 2004). Although a relationship between the two latent variables was expected, the unusually high correlation in this study could have contributed to the inability to see a unique effect of WMC. Similarly to Miyake et al. (2001), this high correlation between factors could indicate that they are in fact not representing two truly separate constructs. Further, all of the VZ tasks chosen were those that fall in the center of the psychometric bull's-eye and are highly correlated with WMC (Marshalek et al., 1984). Creating a more general spatial factor with other measures such as spatial orientation and mental rotation, as well as using spatial measures that are less correlated with WMC, might have allowed for more unique contributions of WMC to be seen.

A second reason may have been the tasks chosen to represent the WMC factor. Although the three measures were significantly correlated with each other, the Rotation Span task did not load on to the factor as well as Backwards Digit and Running Span. The high number of errors on Rotation Span also suggests that it may have introduced noise into the measurement of WMC. This weaker task loading could have reduced the overall predictive power of the WMC factor, while at the same time increasing the shared variance among WMC and VZ constructs due to its spatial nature. Alternatively, including several additional tasks that could be used to compute

separate verbal and spatial WMC factors would have allowed for tests of buffer-specific contributions.

Third, the order of the tasks during the study could have contributed to the inability to find unique effects of WMC. Specifically, the study took approximately two hours to complete with the set of VZ tasks being administered toward the end of the session when fatigue could have been a factor. This fatigue could have reduced the amount of resources available to participants, particularly affecting those who are low in WMC, impacting the amount effort they could have to put into completing the VZ tasks, which may have reduced the ability to find separate influences for WMC and VZ.

Although the overall models were unable to show distinct roles for WMC beyond VZ on the different item types, there was still some evidence supporting the idea that the basic and atypical items are differential in nature. For example, model C which constrained the factor loadings so the atypical items would load differently than basic items provided nearly as good of a fit as model A in which all the items were constrained to have the same factor loadings. Further, model C provided a better fit than model D which constrained loadings as a function of number of folds. Additionally, when the model was run including only WMC, there was a stronger loading for the basic-fold items. Moreover, although it was hypothesized that the atypical items would load more heavily onto VZ than the basic items, the opposite pattern of results was found in the current investigation, possibly because the VZ and WMC constructs were so highly correlated. Specifically, in the CFA models, although all items loaded heavily onto VZ, the basic items had slightly stronger loadings onto the VZ factor than the atypical items. Taken together, these results provide some indication that, although all items rely heavily on VZ, performance on the basic and atypical items may be indicative of different abilities. One

additional approach that may be useful in future work for demonstrating differences between items is using Multivariate Item Response Theory (MIRT) models. This method can be especially useful when considering a task or set of items that may load on more than one dimension and has been shown to be particularly informative for assessing individual item performance (Osteen, 2010). Future research should consider using this more sensitive approach however for the current study a substantially larger sample size would be required. Some recommendations suggest a minimum sample between 250 to 500 participants, while others recommend samples of at least 2,000 to obtain satisfactory item parameter estimates (Ackerman, 1994).

Some additional evidence for differences between basic-fold and atypical-fold items was also suggested by looking at response latencies as a function of WMC. The high WMC individuals were faster to respond to the basic items than the low WMC individuals, suggesting that the high WMC individuals can more efficiently keep track of increasing basic folds. However, for the atypical items there was no difference in response latency between the high and low WMC individuals. This effect seemed particularly driven by the three-fold atypical items where the high WMC individuals actually showed a pattern of being slower than the low WMC individuals. This lack of effect suggests that WMC is not playing the same role on these atypical items.

Finally, it is important to point out that performance on the atypical items was quite low. This low performance could have made it difficult to see a differential role for WMC and VZ because performance was at chance level. Additional atypical items with performance levels more similar to that of the basic items should be developed. By using items that are more closely matched on difficulty separable roles of VZ and WMC could be more apparent. However, it is

also important to note that these most difficult items were difficult for all participants, not just the low VZ participants. In fact, it is on the basic items where the differences between high and low VZ participants are able to be seen. These results suggest that the atypical items may not require the need to visualize. The lack of a role for VZ in these atypical items suggests there is a need for new tests to assess the skills needed for STEM domains, especially domains like chemistry and geology that require a lot of visualization.

4.1 Directions for Future Studies: Penetrative Thinking and Strategy-based Analyses

Because both the constructs used to represent VZ and WMC in this study were found to contribute less to performance on the atypical items compared to the basic items, future work should look into other spatial skill measures or alternative individual difference measures that may play a role particularly on the atypical items. For example, perhaps the atypical items will load more onto a general inductive reasoning factor. The general fluid intelligence (gF) literature commonly uses the PH task and early work on gF has demonstrated that the PH task lies at the center of the psychometric bull's-eye, the center of which is suggested to be gF (Marshalek et al., 1983). Many studies have found the PH task to be strongly related to other measures of gF and in some cases has led the task to be used as a measure of gF (Kane, Hambrick, & Conway, 2005; Kane et al., 2004; Lohman, 1996). Because of this future work should investigate the unique role of gF on the atypical items.

Similarly, considering the role of specific spatial skills, such as penetrative thinking, may provide an additional potential avenue for separating influences of spatial skills from WMC on the atypical items. Penetrative thinking is a relatively new construct in the spatial literature (Kali & Orion, 1996). According to Kali and Orion, penetrative thinking represents the ability to imagine the hidden or internal structure of a form. In their work on understanding geological

thinking, they administered an open-ended test requiring the spatial perception of geologic structures. The test included three types of items: drawing cross-sections of block diagrams, completing block diagrams that only show a single face, and drawing block diagrams of specific locations on a map. Students' responses to these items, as well as recorded interviews from six students were analyzed. Analysis of performance on the test revealed that there were two types of distinct incorrect responses, penetrative errors and non-penetrative errors. Non-penetrative errors were when students made inaccurate judgments about the internal structure of the block diagrams because they only used external or visible information. On the other hand, penetrative errors were those that reflected some attempt to represent the interior properties of the block diagram by continuing the layers exposed on one side of the block or multiple sides, but without accurately integrating the information from all sides into a single representation.

Analysis of these incorrect answer types indicated that students were consistent in the types of errors they made such that 47% of the students gave all non-penetrative incorrect responses and 39% gave all penetrative incorrect responses (Kali & Orion, 1996). In terms of overall performance on the geological thinking task, students who demonstrated penetrative errors scored higher than students who demonstrated non-penetrative errors. Analysis of the student interviews demonstrated that students who made penetrative errors tried to deduce what the interior of the structure looked like using information from multiple sides, and although their drawings reflected errors, they were able to select the correct cross-section when presented with multiple-choice response options. Students who made non-penetrative errors on the other hand, made no attempt to explain the internal structure of the block, for example saying statements such as, "I can see only what is outside. How can I know what is inside?" Additionally, these

students were less able to identify the correct cross-section when presented in multiple-choice format.

Kali and Orion (1996) argue that there are two important processes involved in solving cross-section problems including the ability to perceive the configuration of layers and the ability to visually penetrate a structure. They further suggest that although these two abilities seem to fall within the category of spatial visualization described in the factor analytic literature, many of the commonly used spatial visualization tasks involve mental manipulations of the external surfaces of objects rather than the internal parts. The atypical items from the Punched Holes task may better represent the manipulation of internal parts and can be seen to share several features with penetrative thinking. Specifically, the atypical items require imagining a fold that is hidden or occluded and also require thinking about how many layers the punched hole actually goes through. Because of this task similarity, including a penetrative thinking measure in the future could be another possible approach for separating spatial skills from more general WM resources.

Future work on the new Punched Holes items should consider the relation between solution strategies, VZ, and WMC. Prior work on the original Punched Holes task has demonstrated that there are a variety of strategies available when solving, some of which rely directly on visualization and others which are more analytic and rely less on visualization (Hegarty, 2010). The results of the current study did demonstrate that participants use multiple strategies when solving and that using more strategies was related to better performance. Additionally, it was found that participants who reported using one strategy in particular, working backwards, performed better on all PH item types and that the use of this strategy was correlated with VZ and WMC. These results begin to demonstrate a role for strategy-use on the

PH task, but they do not provide any information of how often each strategy was used, which items they were used on, or how well different participants were able to use them. Future work should collect online information through eyetracking and think aloud to investigate these questions.

While the current study did not collect online information about participants' problem solving process, pilot testing provided some indication that there may be differences in strategies or how effectively they can be used as a function of VZ (Perschon, Jaeger, & Wiley, 2015). Specifically, it was found using eyetracking that low spatial individuals toggled more between the problems and answer bank and also produced more gestures during problem solving. These results may indicate that these participants are using a visualization strategy, but that they may not be effective at mentally animating and therefore must rely on the answer options and gesturing to support their memory. One specific empirical question that is related to strategy use is whether the instructions participants received before starting the Punched Holes task overly determine the use of a visualization strategy. The standard instruction explicitly tells participants to *imagine* the folding and unfolding of pieces of paper, which could be biasing individuals towards a visualization strategy. Future work could implement an instruction that biases towards an analytic strategy and see if that changes the role that WMC and VZ play in performance. In particular, perhaps an instruction that is more analytic in nature could allow for a greater role of WMC to be seen.

4.2 Conclusions

Overall, the main results of this study suggest that the Punched Holes task is best represented as a measure of spatial visualization. This is an important result to convey because it implies that research which has attempted to use the Punched Holes task to represent a different

construct should be evaluated more critically. At the same time, there was some suggestion that items on the Punched Holes task could be meaningfully categorized into two main subtypes, those that contain basic folds and those that contain atypical folds. Although both item subtypes relied heavily on VZ, the atypical items appeared to be less reliant on WMC for successful solution.

Understanding the cognitive constructs that contribute to performance on these items is important because it will provide deeper insight into the abilities that are being measured by spatial thinking assessments. By being able to more clearly demonstrate what aspects of spatial task performance are reliant upon WMC resources and which are not, better understandings of these constructs can be developed. Further, being able to differentiate between these constructs and what kinds of tasks they are important for could eventually help to shed light on how and why different spatial skills may be linked to success in STEM fields.

5. References

Ackerman, T. A. (1994). Using multidimensional item response theory to understand what items and tests are measuring. *Applied Measurement in Education*, *7*, 255-278.

Arbuckle, J. L. (2013). IBM SPSS Amos 22 user's guide. *Crawfordville, FL: Amos Development Corporation*.

Baddeley, A. D., & Hitch, G. J. (1974). Working memory. In G. H. Bower (Ed.), *The psychology of learning and motivation: Advances in research and theory* (Vol. 8, pp. 47–89). New York: Academic Press.

Bennett, G. K., Seashore, H. G., & Wesman, A. G. (1973). *Differential aptitude tests: Administrator's handbook*. New York: The Psychological Corporation.

Bichsel, J., & Roskos-Ewoldsen, B. (1999). Imaginal discovery, working memory, and intelligence. *Journal of Mental Imagery*, *23*, 17 – 34.

Black, A. A. (2005). Spatial ability and earth science conceptual understanding. *Journal of Geoscience Education*, *53*, 402-414.

Broadway, J. M., & Engle, R. W. (2010). Validating running memory span: Measurement of working memory capacity and links with fluid intelligence. *Behavior Research Methods, 42*, 563-570.

Carpenter, P. A., & Just, M. A. (1986). Spatial ability: An information processing approach to psychometrics. In R. J. Stemberg (Ed.), *Advances in the psychology of human intelligence* (Vol. 3, pp. 221-252). Hillsdale, NJ: Erlbaum.

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

Carter, C. S., LaRussa, M. A., & Bodner, G. M. (1987). A study of two measures of spatial

ability as predictors of success in different levels of general chemistry. *Journal of Research in Science Teaching*, *24*, 645-657. doi:10.1002/tea.3660240705

Cooper, L. A., & Shepard, R. N. (1973). Chronometric studies of the rotation of mental images. In W. G. Chase (Ed.), *Visual information processing* (pp. 75-176). New York: Academic Press.

Cronbach, L. J. (1970). *Essentials of psychological testing* (3rd ed.). New York: Harper & Row.

DeFries, J. C., Vandenberg, S. G., McClearn, G. E., Kuse, A. R., Wilson, J. R., Ashton, G. C., & Johnson, R. C. (1974). Near identity of cognitive structure in two ethnic groups. *Science*, *183*, 338-339.

Educational Testing Service, French, J. W., Ekstrom, R. B., & Price, L. A. (1963). *Kit of reference tests for cognitive factors*. Educational Testing Service.

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

Eliot, J., & Smith, I. M. (1983). *An international directory of spatial tests*. Cengage Learning Emea.

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*, 226 - 236.

French, J. W. (1951). The description of aptitude and achievement tests in terms of rotated factors. *Psychometric Monographs*, 5.

Ghiselli, E. E. (1973). The validity of aptitude tests in personnel selection. *Personnel Psychology*, *26*, 461-477.

Goldstein, D., Haldane, D., & Mitchell, C. (1990). Sex differences in visual-spatial ability: The role of performance factors. *Memory & Cognition, 18*, 546-550.

Guilford, J. P. (1967). *The nature of human intelligence*. New York: McGraw Hill.

Guilford, J. P., & Lacey, J. I. (1947). Printed classification tests. *Army Air Force Aviation Psychology Program Research Reports, No. 5*. Washington, DC: US Government Printing Office.

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.

Hegarty, M. (2010). Components of spatial intelligence. In B. H. Ross (Ed.), *The Psychology of Learning and Motivation*, Vol. 52. (pp. 265–297). San Diego: Academic Press.

Horn, J. L., & Cattell, R. B. (1966). Refinement and test of the theory of fluid and crystallized general intelligences. *Journal of Educational Psychology*, *57*, 253 - 270.

Hu, L. T., & Bentler, P. M. (1998). Fit indices in covariance structure modeling: Sensitivity to under-parameterized model misspecification. *Psychological Methods*, *3*, 424 - 453.

Jaeger, A. J., Jarosz, A. F., & Wiley, J. (2014). *Know when to hold 'em, Know when to fold 'em: WMC and spatial reasoning*. Poster presented at the 55th Annual Meeting of the Psychonomic Society, Long Beach, CA.

Just, M. A., & Carpenter, P. A. (1985). Cognitive coordinate systems: Accounts of mental rotation and individual differences in spatial ability. *Psychological Review*, *92*, 137– 171.

Kali, Y., & Orion, N. (1996). Spatial abilities of high-school students in the perception of geologic structures. *Journal of Research in Science Teaching*, *33*, 369-391.

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

Kane, M. J., Hambrick, D. Z., Tuholski, S. W., Wilhelm, O., Payne, T. W., & Engle, R. W. (2004). The generality of working memory capacity: A latent variable approach to verbal and visuo-spatial memory span and reasoning. *Journal of Experimental Psychology: General, 133,* 189–217.

Kosslyn, S. M. (1981). The medium and the message in mental imagery: A theory. *Psychological Review, 88,* 46 – 66.

Kotovsky, K., & Simon, H. A. (1973). Empirical tests of a theory of human acquisition of concepts for sequential patterns. *Cognitive Psychology, 4*, 399-424.

Kozhevnikov, M., & Hegarty, M. (2001). A dissociation between object manipulation spatial ability and spatial orientation ability. *Memory & Cognition, 29*, 745-756.

Kozhevnikov, M., Motes, M. A., & Hegarty, M. (2007). Spatial visualization in physics problem solving. *Cognitive Science, 31*, 549-579.

Kyllonen, P. C., Lohman, D. F., & Snow, R. E. (1984). Effects of aptitudes, strategy training, and task facets on spatial task performance. *Journal of Educational Psychology, 76*, 130-145.

Lohman, D. F. (1979). *Spatial ability: Review and re-analysis of the correlational literature*. Stanford University Technical Report No. 8.

Lohman, D. F. (1996). Spatial ability and g. In I. Dennis & P. Tapsfield (Eds.) *Human abilities: Their nature and measurement* (pp. 97-116). Hillsdale, NJ: Erlbaum.

Lohman, D. F., & Hagen, E. P. (2001*). Cognitive Abilities Test: Form 7, Level H*. Riverside Publishing.

Lord, T. R. (1987). A look at spatial abilities in undergraduate women science majors. *Journal of Research in Science Teaching*, *24*, 757-767.

Marshalek, B., Lohman, D. F., & Snow, R. E. (1983). The complexity continuum in the radex and hierarchical models of intelligence. *Intelligence*, *7*, 107–127.

Mayer, R. E., & Sims, V. K. (1994). For whom is a picture worth a thousand words? Extensions of a dual-coding theory of multimedia learning. *Journal of educational psychology, 86,* 389 – 401.

McFarlane, M. (1925). A study of practical ability. *British Journal of Psychology*, Monograph Supplement No. 8.

McGee, M. G. (1979). Human spatial abilities: Psychometric studies and environmental, genetic, hormonal, and neurological influences. *Psychological Bulletin*, *86*, 889–918.

Michael, W. B., Guilford, J. P., Fruchter, B., & Zimmerman, W. S. (1957). The description of spatial-visualization abilities. *Educational and Psychological Measurement*, *17*, 185 – 199.

Miller, H., & Bichsel, J. (2004). Anxiety, working memory, gender, and math performance. *Personality and Individual Differences*, *37*, 591-606.

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

Mumaw, R. J., & Pellegrino, J. W. (1984). Individual differences in complex spatial processing: *Journal of Educational Psychology*, *76*, 920-939.

Mumaw, R. J., Pellegrino, J. W., Kail, R. V., & Carter, P. (1984). Different slopes for different folks: Process analysis of spatial aptitude. *Memory & Cognition*, *12*, 515–521.

Nunnally, J.C. (1978). *Psychometric Theory.* New York: McGraw-Hill.

Osteen, P. (2010). An introduction to using multidimensional item response theory to assess latent factor structures. *Journal of the Society for Social Work and Research, 1*, 66-82.

Pellegrino, J. W., Alderton, D. L., & Shute, V. J. (1984). Understanding spatial ability. *Educational Psychologist*, *19*, 239 – 253.

Pellegrino, J. W., & Glaser, R. (1979). Cognitive correlates and components in the analysis of individual differences. *Intelligence, 3*, 187-215.

Pellegrino, J. W., & Kail, R. (1982). Process analyses of spatial aptitude. *Advances in the Psychology of Human Intelligence*, *1*, 311-365.

Perschon, C. R., & Jaeger, A. J. (2015a) *Understanding the role of strategy-use on the paper-folding task.* Poster presented at the UIC Undergraduate Research Symposium, Chicago, IL.

Perschon, C., Jaeger, A. J, & Wiley, J. (2015b). *Bringing strategies into the fold: Strategy-use and gesture on the Paper Folding task.* Poster presented at the 87th Annual Meeting of the Midwestern Psychological Association, Chicago, IL.

Salthouse, T. A., Babcock, R. L., Skovronek, E., Mitchell, D. R., & Palmon, R. (1990). Age and experience effects in spatial visualization. *Developmental Psychology*, *26*, 128 - 136.

Salthouse, T. A., Mitchell, D. R., Skovronek, E., & Babcock, R. L. (1989). Effects of adult age and working memory on reasoning and spatial abilities. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *15*, 507 - 516.

Sanchez, C. A. (2012). Enhancing visuospatial performance through video game training to increase learning in visuospatial science domains. *Psychonomic Bulletin & Review*, *19*, 58-65.

Shepard, R. N., & Feng, C. (1972). A chronometric study of mental paper folding. *Cognitive Psychology, 3,* 228-243.

Siemankowski, F. T., & MacKnight, F. C. (1971). Spatial cognitions, a success prognosticator in college science courses. *Journal of College Science Teaching*, *1*, 56-59.

Staver, J. R., & Jacks, T. (1988). The influence of cognitive reasoning level, cognitive restructuring ability, disembedding ability, working memory capacity, and prior knowledge on students' performance on balancing equations by inspection. *Journal of Research in Science Teaching*, *25*, 763-775.

Thurstone, L. L. (1938). *Primary mental abilities.* Chicago: University of Chicago Press.

Unsworth, N., & Engle, R. W. (2007). On the division of short-term and working memory: An examination of simple and complex span and their relation to higher order abilities. *Psychological Bulletin, 133*, 1038-1066.

Wiley, J., Jarosz, A.F., Cushen, P. J., & Colflesh, G. J. H. (2011). New rule use drives the relation between working memory capacity and Raven's advanced progressive matrices. *Journal of Experimental Psychology: Learning, Memory and Cognition, 37*, 256-263.

6. FOOTNOTES

i. A 2 (Fold type: basic, atypical) X 2 (Number of folds: fewer, more) repeated measures ANOVA was conducted to assess the effect of number of folds and type of fold on overall solution time including incorrect trials. This analysis revealed a main effect for fold type, $F(1, 148) = p < .001$, $\eta^2 = .77$, and number of folds, $F(1, 148) = 303.30$, $p < .001$, $\eta^2 = .67$. The interaction between these factors was also significant, $F(1, 148) = 126.79$, $p < .001$, $\eta^2 = .46$. The difference in overall response latency was greater between the two ($M = 47046.36$, $SD = 9942.86$) and three-fold atypical ($M = 71304.12$, $SD = 13979.22$) than between the three ($M = 33241.25$, $SD = 12675.99$) and four-fold basic ($M = 38592.99$, $SD = 13071.50$), $t(148) = 6.09$, $p < .001$.

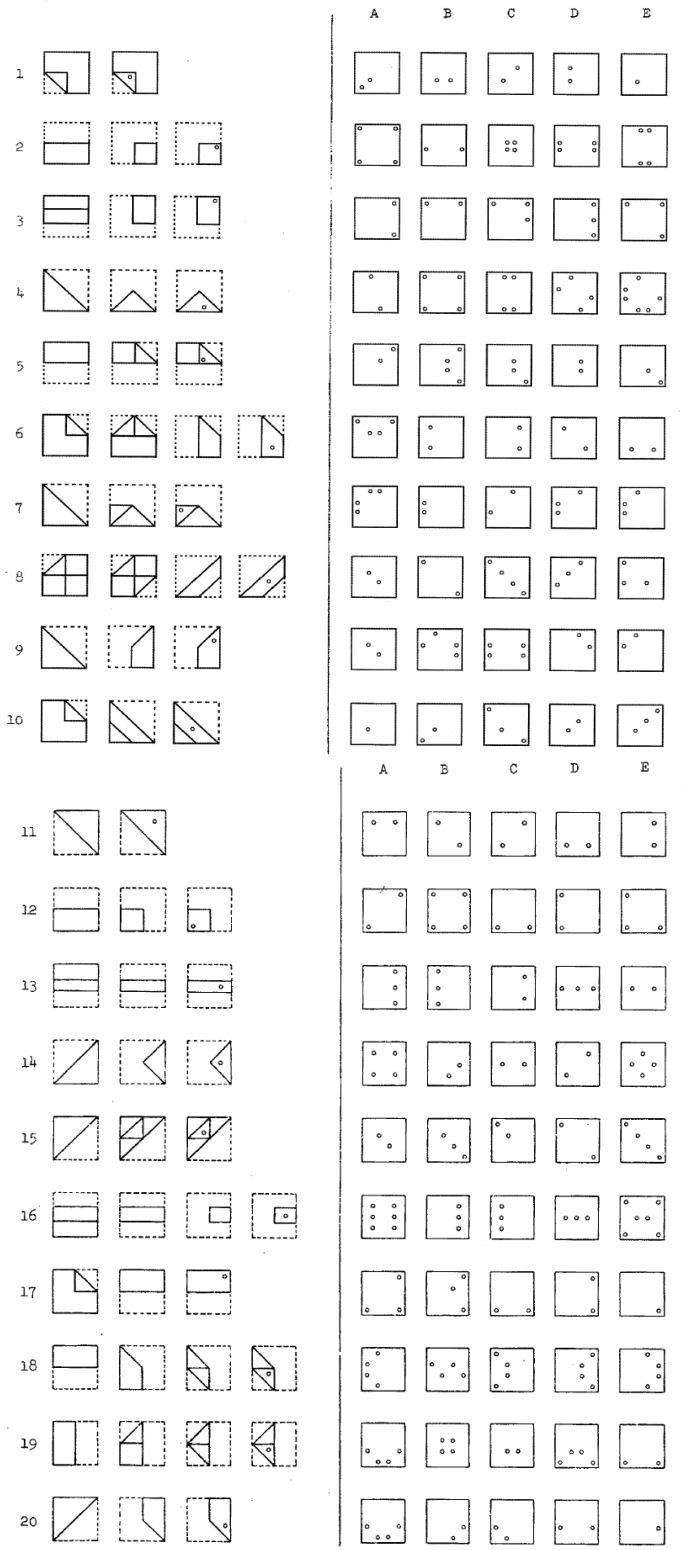
ii. A 2 (Fold type: basic, atypical) X 2 (Number of folds: fewer, more) X 2 (VZ: high, low) repeated measures ANOVA was conducted to assess the effect of fold number, fold type, and spatial visualization ability on participants' overall response latencies. There was a main effect for type of folds, $F(1, 147) = 530.20$, $p < .001$, $\eta^2 = .78$, a main effect for number of folds, $F(1, 147) = 303.54$, $p < .001$, $\eta^2 = .67$, and a main effect for VZ, $F(1, 147) = 28.05$, $p < .001$, $\eta^2 = .16$. The interaction between type and number of folds was significant, $F(1, 147) = 125.23$, $p < .001$, $\eta^2 = .46$, as was the interaction between VZ and type of fold, $F(1, 147) = 6.49$, $p < .02$, $\eta^2 = .04$. The two-way interaction between VZ and number of folds was not significant, $F < 1$, and there was no three-way interaction between number of folds, type of folds, and spatial visualization ability, $F < 1$.

iii. A 2 (Fold type: basic, atypical) X 2 (Number of folds: fewer, more) X 2 (WMC: high, low) repeated measures ANOVA was conducted to assess the effect of fold number, fold type, and working memory capacity on participants' overall response latencies. There was a main

effect for type of folds, $F(1, 147) = 509.53$, $p < .001$, $\eta^2 = .78$, a main effect for number of folds, $F(1, 147) = 304.66$, $p < .001$, $\eta^2 = .68$, and a main effect for WMC, $F(1, 147) = 4.84$, $p < .02$, $\eta^2 = .03$. The interaction between type and number of folds was significant, $F(1, 147) = 129.82$, $p < .001$, $\eta^2 = .47$. There was no interaction between WMC and type of fold, $F(1, 147) = 1.74$, *ns*, and no interaction between WMC and number of folds, $F(1, 147) = 1.12$, *ns*. There was also no three-way interaction between number of folds, type of folds, and working memory capacity, $F(1, 147) = 2.53$, *ns*.

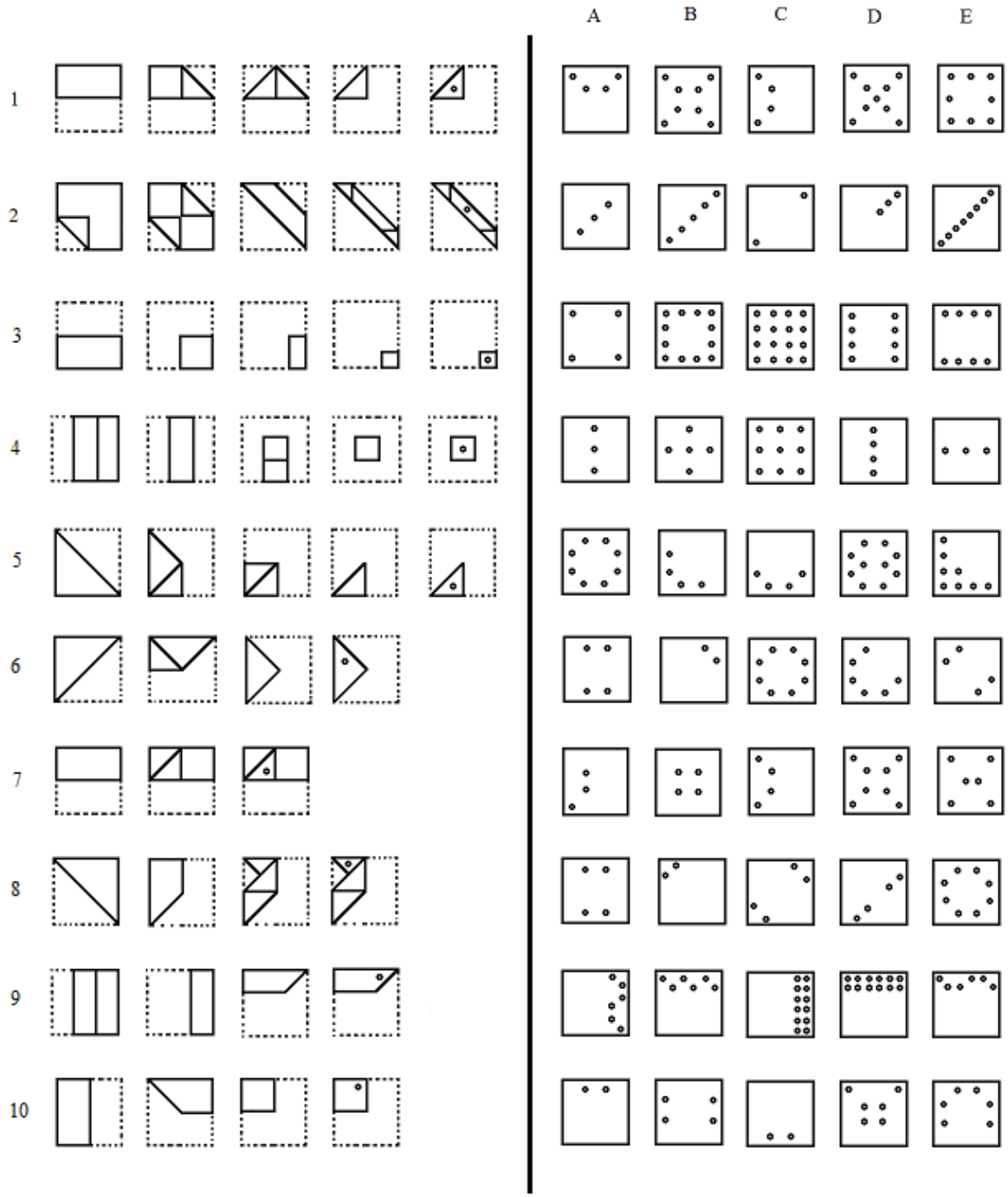
7. APPENDICES

Appendix A



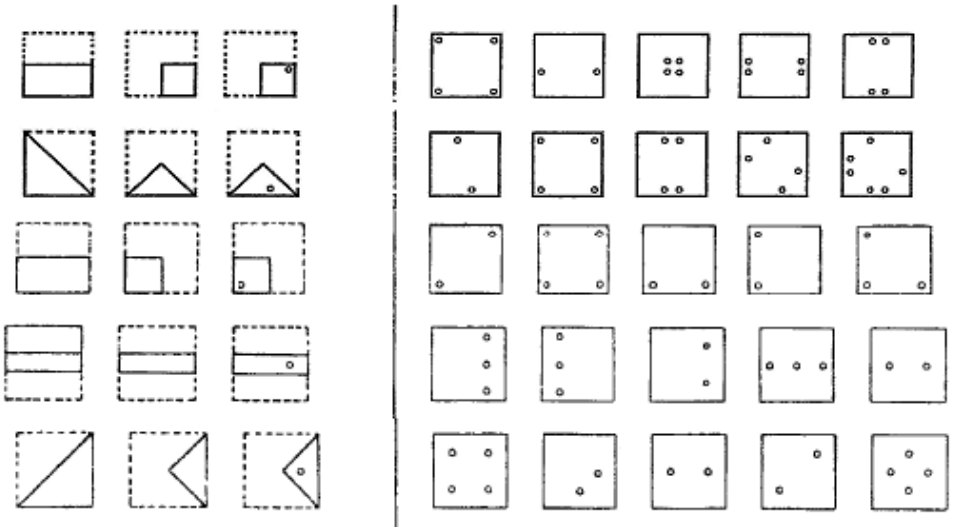
Appendix B

Items 1 through 5 are typical four-fold items and 6 through 10 are atypical fold items

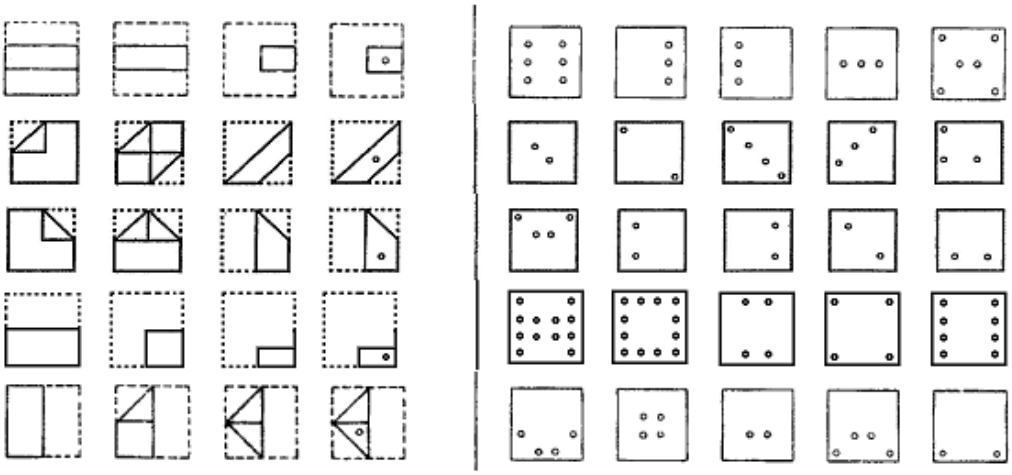


Appendix C

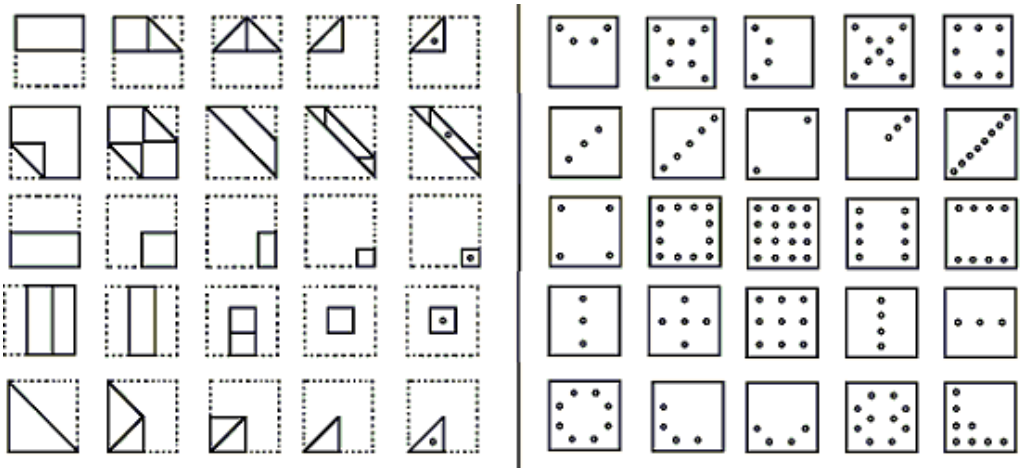
Two Fold Basic



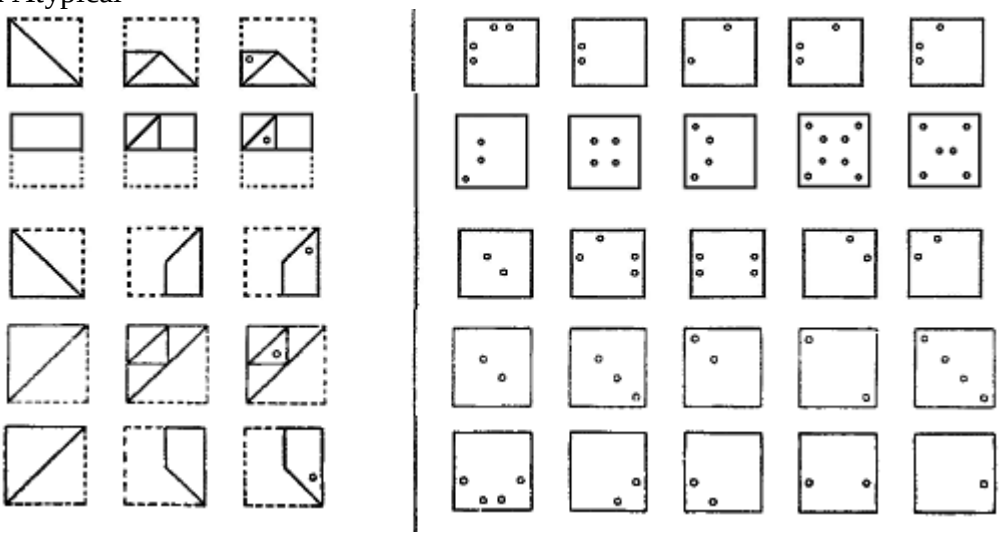
Three Fold Basic



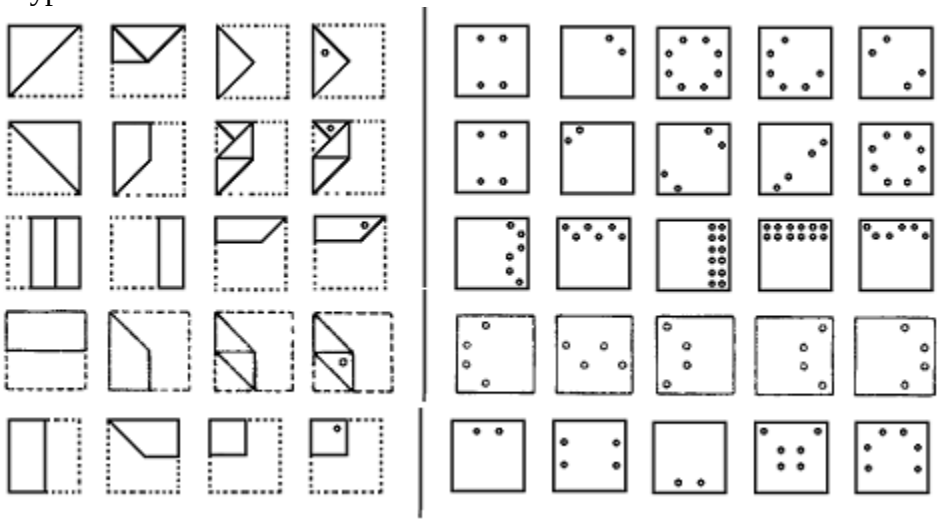
Four Fold Basic



Two Fold Atypical



Three Fold Atypical



Appendix D

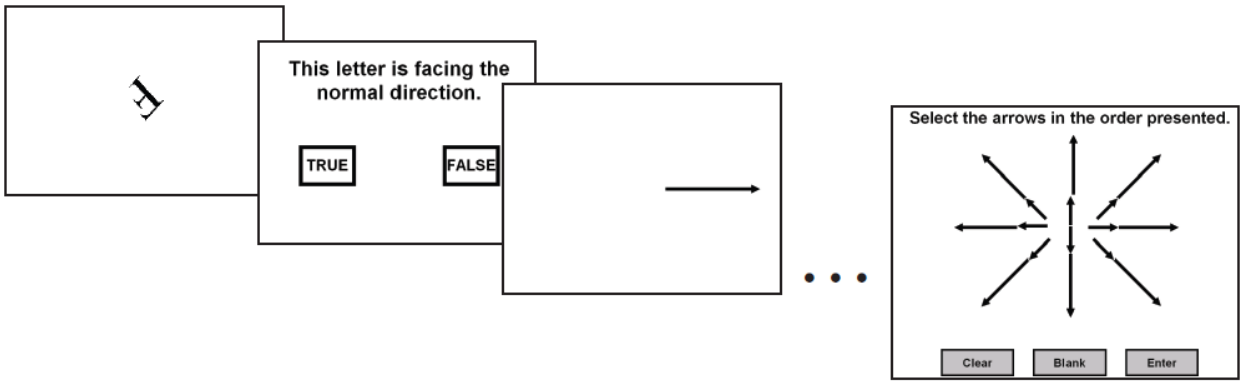
Automated Running Span (Broadway & Engle, 2010):
For the automated running span, participants see a string of letters one at a time, with the instructions to remember only the last few letters. In the example below, the instructions would be to remember the last 4 letters only. Participants would type their response after the last letter disappeared from the screen.

Q F S J K R

Backwards Digit Span (Unsworth & Engle, 2007):
For the backwards digit span, participants are asked to type in the numbers they just saw, in reverse order. Numbers appear one at a time.

3 2 1 4 6

Rotation Span (Harrison et al., 2013):
Subjects are shown a rotated letter then are asked to indicate whether that letter is facing the normal direction or mirror. Next they are presented with an arrow pointing in one of eight directions. At the end participants must indicate the order in which the arrows were presented.



Appendix E

1. What is your gender? (Circle one) M F
2. What is your age?_____
3. **Year in School**: Freshman Sophomore Junior Senior
4. What is your intended major:
5. Are you bilingual (or multilingual)? (circle one) YES NO
6. What is/are your native language(s)?
7. If English is not your first language, **AT WHAT AGE** did you start speaking English fluently?
8. Please fill in any ACT scores (0-36) to the best of your memory:

   ACT COMPOSITE _____ ACT READING _____ ACT ENGLISH_____
   ACT MATH _____ ACT SCIENCE _____

9. If you took the SAT please fill in your scores:

   SAT MATH _____ SAT READING_____ SAT WRITING_____

10. How many college courses (if any) have you taken in science and what were they?

Think back to the first task you completed where you had to imagine the folding and unfolding of pieces of paper and then determine where holes punches would be located when the paper was completely unfolded.

When you were completing this task what kind of techniques were you using to try and solve each problem?

Please select all that you used.

[ ] I imagined folding the paper, punching the hole, and unfolding the paper in my mind

[ ] I started at the last step shown and worked backward to unfold the paper and see where the holes would be

[ ] First, I figured out where one of the holes would be and then eliminated answer choices that did not have a hole in that location

[ ] I figured out how many folds/sheets of paper were punched through; I figured out how many holes there would be in the end

[ ] I used the number of holes/folds to eliminate some of the answer choices

[ ] Other – please describe:

_____
_____
_____
_____

8. HUMAN SUBJECTS COMMITTEE PROTOCAL APPROVAL

This research was approved by the University of Illinois Human Subjects Institutional Review Board under protocol 2001–0489.

9. CURRICULUM VITAE

Allison J. Jaeger

University of Illinois at Chicago
Department of Psychology
1007 W. Harrison St M/C 285
Chicago, IL 60607
Email: ajaege1@uic.edu

Education:

PhD expected Summer 2015, University of Illinois at Chicago
Major: Cognitive Psychology Minor: Learning Technologies
M.A. in Cognitive Psychology, May 2012, University of Illinois at Chicago
B.A. in Psychology (applied track), December 2007, University of Illinois at Chicago

Research interests:

The role of spatial skills in science learning
Text comprehension and metacomprehension accuracy
Learning in science
Multimedia learning
Educational technology

Teaching interests:

Cognition and Instruction
Quantitative and Qualitative Methods
Statistics
Reading Comprehension
Human Learning and Development

Publications:

Jaeger, A. J., & Wiley, J. (in press). Reading an analogy can cause the illusion of comprehension. *Discourse Processes*. DOI: 10.1080/0163853X.2015.1026679

Jaeger, A. J., & Wiley, J. (2014). Do illustrations help or harm metacomprehension accuracy? *Learning & Instruction*, *34*, 58-73. DOI:10.1016/j.learninstruc.2014.08.002

Sanchez, C. A., & **Jaeger, A. J.** (2014). If it's hard to read, it changes how long you do it: Reading time as an explanation for perceptual fluency effects on judgment. *Psychonomic Bulletin & Review*, *22*, 1-6. DOI: 10.3758/s13423-014-0658-6

Wiley, J., Sanchez, C. A., & **Jaeger, A. J.** (2014). The individual differences in working memory capacity principle in multimedia learning. In R. Mayer (Ed.), *The Cambridge*

Handbook of Multimedia Learning (2nd ed., pp. 598-619). New York: Cambridge University Press.

Wiley, J., Ash, I. K., Sanchez, C. A., & **Jaeger, A. J.** (2011). Clarifying readers' goals for learning from expository science texts. In M. McCrudden, J. Magliano, and G. Schraw, (Eds.) *Text relevance and learning from text* (pp. 353-374). Greenwich, CT: Information Age Publishing.

Moher, T., Gnoli, A., **Jaeger, A. J.**, Wiley, J., & Lopez-Silva, B. (2011). Embodied learning for embedded spaces. *Proceedings of the Ninth International Computer-Supported Collaborative Learning Conference, 3,* 1085-1086.

Moher, T., Wiley, J., **Jaeger, A. J.**, Lopez-Silva, B., Novellis, F., & Kilb, D. (2010). Spatial and temporal embedding for science inquiry: An empirical study of student learning. *Proceedings of the International Conference of the Learning Sciences, 1,* 826-833.

Jaeger, A. J., & Wiley, J. (2010). Seductive images and the metacomprehension of science texts [Abstract]. In S. Ohlsson & R. Catrambone (Eds.), *Proceedings of the 32nd Annual Conference of the Cognitive Science Society* (p. 561). Austin, TX: Cognitive Science Society.

Manuscripts in preparation:

Jaeger, A. J., & Wiley, J. (in preparation). What does the Punched Holes test measure?

Jaeger, A. J., Wiley, J., Griffin, T. D., & Britt, M. A. (in preparation). Learning science from multiple documents: Understanding causal chains.

Jaeger, A. J., Wiley, J., & Moher, T. (in preparation). The role of spatial ability in learning from whole-class simulations in science.

Taylor, A., **Jaeger, A. J.**, & Wiley, J. (in preparation). When analogies help: The role of interleaving and spatial skills.

Presentations:

Jaeger, A. J., Griffin, T. D., Britt, M. A., & Wiley, J. (2015, July). *Making connections: Improving student learning about climate change.* Paper to be presented at the 25th Annual Meeting of the Society for Text & Discourse, Minneapolis, MN.

Danielson, R. W., Sinatra, G. M., **Jaeger, A. J.**, & Wiley, J. (2015, July). *Improving learning from refutation-texts with analogies and graphs.* Paper to be presented at the 25th Annual Meeting of the Society for Text & Discourse, Minneapolis, MN.

Perschon, C., **Jaeger, A. J.**, & Wiley, J. (2015, April). *Bringing strategies into the fold: Strategy-use and gesture on the Paper Folding task*. Poster presented at the 87th Annual Meeting of the Midwestern Psychological Association, Chicago, IL.

Perschon, C., **Jaeger, A. J.**, & Wiley, J. (2015, April). *Using eyetracking and think aloud to investigate strategy use on the Paper Folding task.* Poster presented at the University of Illinois at Chicago Student Research Forum, Chicago, IL.

Jaeger, A. J., Jarosz, A. F., & Wiley, J. (2014, November). *Know when to hold 'em, Know when to fold 'em: WMC and spatial reasoning*. Poster presented at the 55th Annual Meeting of the Psychonomic Society, Long Beach, CA.

Jaeger, A. J., & Wiley, J. (2014, August). *Learning from multiple documents: More than just reading between the lines.* Paper presented at the 24th Annual Meeting of the Society for Text & Discourse, Chicago, IL.

Taylor, A., **Jaeger, A. J.,** & Wiley, J. (2014, August). *Analogies and metacomprehension of expository texts.* Paper presented at the 24th Annual Meeting of the Society for Text & Discourse. Chicago, IL.

Jaeger, A. J., Taylor, A. R., & Wiley, J. (2014, May). *Improving metacomprehension accuracy for science text in middle school.* Poster presented at the 86th Annual Meeting of the Midwestern Psychological Association, Chicago, IL.

Wiley, J., Taylor, A., **Jaeger, A. J.**, Griffin, T. D., & Britt, M. A. (2014, April). *Supporting inquiry learning from multiple documents in history and science.* Paper presented at the American Educational Research Association, Philadelphia, PA.

Jaeger, A. J., Griffin, T. D., Britt, M. A., & Wiley, J. (2013, July). *Learning science from multiple documents: "We don't normally do this in science class"*. Paper presented at the 23rd Annual Meeting of the Society for Text & Discourse, Valencia, Spain.

Jaeger, A. J., Sanchez, C. A., & Wiley, J. (2013, July). *How reading affects believing: Perceptually non-fluent text presentations and changes in online processing.* Paper presented at the 23rd Annual Meeting of the Society for Text & Discourse, Valencia, Spain.

Jaeger, A. J., & Sanchez, C. A. (2013, May). *Judgments affected by perceptual fluency are mediated by reading patterns*. Poster presented at the 85th Annual Meeting of the Midwestern Psychological Association, Chicago, IL.

Griffin, T. D., **Jaeger, A. J.**, Jarosz, A. F., Thiede, K. W., & Wiley, J. (2012, July). *Improving metacomprehension in an undergraduate course in research methods.* Poster presented at the 22nd Annual Meeting of the Society for Text & Discourse, Montreal, Canada.

Jaeger, A. J., & Wiley, J. (2012, May). *Pay no attention to that image! Effects of image use instructions on metacomprehension of illustrated text*. Paper presented at the 84th Annual Meeting of the Midwestern Psychological Association, Chicago, IL.

Jaeger, A. J., & Wiley, J. (2011, November). *Reducing illusions of comprehension from illustrated text*. Poster presented at the 52nd Annual Meeting of the Psychonomic Society, Seattle, WA.

Moher, T., Gnoli, A., **Jaeger, A. J.**, Wiley, J., & Lopez-Silva, B. (2011, July). *Embodied learning for embedded spaces*. Symposium paper presented at the Ninth International Computer-Supported Collaborative Learning Conference, Hong Kong.

Jaeger, A. J., & Wiley, J. (2011, July). *Using explanation to improve metacomprehension of illustrated science texts*. Paper presented at 21st Annual Meeting of the Society for Text & Discourse, Poitiers, France.

Jaeger, A. J., Wiley, J., Moher, T., Lopez-Silva, B. A., & Novellis, F. (2011, April). *Embedded simulations support learning in geoscience*. Poster presented in a structured poster session at the American Educational Research Association, New Orleans, LA.

Jaeger, A. J., Wiley, J., & Griffin, T. (2010, November). *Metacomprehension accuracy from illustrated text*. Poster presented at the 51st Annual Meeting of the Psychonomic Society, St. Louis, MO.

Ricks, T. R., **Jaeger, A. J.**, & Wiley, J. (2010, November). *The impact of knowledge and mnemonic chunking strategies in eliminating the FAN effect.* Poster presented at the 51st Annual Meeting of the Psychonomic Society, St. Louis, MO.

Jaeger, A. J., & Wiley, J. (2010, July). *The effects of images on comprehension and metacomprehension of science texts***.** Poster presented at the Society for Text & Discourse, Chicago, IL.

Moher, T., Wiley, J., **Jaeger, A. J**., Lopez-Silva, B., Novellis, F., & Kilb, D. (2010, June). *Spatial and temporal embedding for science inquiry: An empirical study of student learning*. Paper presented at the International Conference of the Learning Sciences, Chicago, IL.

Jaeger, A. J., Moher, T., Wiley, J., Malcolm, P., Lopez-Silva, B. A., Gnoli, A. & Brown, J. (2009, April). *WallCology: Using embedded phenomena to motivate learning about dynamic ecosystems*. Paper presented at the American Educational Research Association, San Diego, CA.

Ricks, T., **Jaeger, A. J.**, & Wiley, J. (2008, May). *The effect of baseball knowledge on the FAN effect*. Poster presented at the Association for Psychological Sciences, Chicago, IL.

Academic Experience:

Graduate Research Assistant, August 2012 – present

IES Funded Project: Reading for Understanding Across Grades 6 through 12: Evidence-based Argumentation for Disciplinary Learning: PI Susan Goldman, (Co-PIs Thomas Griffin, James Pellegrino, Jennifer Wiley)

Responsibilities: Experimental design, data collection and analysis, supervision of research assistants, literature review and manuscript preparation

Graduate Research Assistant, August 2011 – July 2012

IES Funded Project: Improving Metacomprehension and Self-Regulated Learning from Scientific Texts: Joint PIs Thomas Griffin, Keith Thiede, Jennifer Wiley

Responsibilities: Experimental design, data collection and analysis, supervision of research assistants, literature review and manuscript preparation

Graduate Research Assistant, August 2009 – August 2011

NSF Funded Project: Supporting Whole-class Science Investigations with Spatial Simulations: Tom Moher PI and Jennifer Wiley Co-PI

Responsibilities: Lesson and assessment design, data collection, coding and analysis, manuscript preparation

Research Specialist, January 2008 – August 2009

NSF Funded Project: Supporting Whole-class Science Investigations with Spatial Simulations: Tom Moher PI and Jennifer Wiley Co-PI

Responsibilities: Lesson and assessment design, data collection, coding and analysis, manuscript preparation and conference presentations

Teaching Experience:

Instructor (Spring 2014, Fall 2014)

Psychology 353: Laboratory in Cognition and Learning

The main goal of this course is to introduce students to basic research in each of the central areas of cognitive psychology and learning. Students review papers and participate in studies related to each of the major areas of cognition and learning as well as develop, collect data, and analyze the results of their own study. The goal is for students to leave this course with strong writing skills, a good understanding of research in cognitive psychology, and the ability the critically evaluate research.

Practicum for Teaching in Psychology, August 2013 – May 2014

Supervisor: Edward Sargis, Ph. D.

A training course for teaching undergraduate psychology courses

Responsibilities: develop and receive feedback on lesson plans and exams during the first half of the course, teach and be observed by supervisor during second half of course

Graduate Teaching Assistant

Psychology 242: Research Methods (Fall 2010, Fall 2013)

Psychology 353: Lab in Cognition and Learning (Spring 2013, Spring 2015)

Cognitive Division Assistant, Fall 2011 – Spring 2012
Responsibilities: Updating Cognitive Division webpage, organizing weekly brown bag meetings, organizing Cognitive Division visiting day and recruitment

Undergraduate Mentoring:

Michelle Evans (F09, S10) Metacomprehension of Illustrated Science Texts

Samantha Hicks (S10, F11) Science Investigations with Spatial Simulations

Rick Leonard (S10) Science Investigations with Spatial Simulations

Melissa Pasierb (S12) Comprehension and Metacomprehension of Science Texts

Stephanie Blakeslee (S12, F12) Comprehension & Metacomprehension of Science Text

Tegan Michl (F12, S13) Learning from Multiple Documents

Brad Wilson (F12, S13, F13) Perceptual Fluency and Reading Behavior

Recipient of LASURI award (2014 – 2015)

Recipient of Nancy Hirschberg Research Grant (2014)

Ishan Patel (F12, S13) Learning from Multiple Documents

Recipient of Undergraduate Research Experience (URE) award

Alia Mohammad (F13, S14) Learning from Multiple Documents

Raluca Birza (F13) Learning from Multiple Documents

Recipient of Undergraduate Research Experience (URE) award

Amy Thakkar (S14, Sum14) Learning from Multiple Documents

Nathaniel Kelley (S14, Sum 14) Working Memory Capacity and Performance on Spatial Tasks

Chelsea Perschon (S14, F14, S15) WM Capacity and Performance on Spatial Tasks

Recipient of Undergraduate Research Experience (URE) award

Roy Salas (F14, S15) WM Capacity and Performance on Spatial Tasks

Jennifer Chun (S15) WM Capacity and Performance on Spatial Tasks

Himani Kumar (S15) WM Capacity and Performance on Spatial Tasks

Professional Organization Activities:

Student Member, Society for Text and Discourse
Student Member, American Educational Research Association
Student Member, Cognitive Science Society
Student Member, Midwestern Psychological Association
Student Member, American Psychological Association

Student Member of Organizing Committee, 20th Annual Meeting of the Society for

Text and Discourse, Chicago, August 2010
Conference Reviewer, 21st Annual Meeting of the Society for Text and Discourse, Poitiers, France, July 2011
Conference Reviewer, 22nd Annual Meeting of the Society for Text and Discourse, Montreal, Canada, July 2012
Conference Reviewer, 23rd Annual Meeting of the Society for Text and Discourse, Valencia, Spain, July 2013
Conference Reviewer, 24th Annual Meeting of the Society for Text and Discourse, Chicago, July 2013
Student Member of Organizing Committee, 24th Annual Meeting of the Society for Text and Discourse, Chicago, August 2014

Ad Hoc Reviewer: British Journal of Educational Psychology; Journal of Experimental Psychology: Learning, Memory, and Cognition; Discourse Processes

Honors and Awards:

Student Research Award, 2nd prize, UIC Undergraduate Research Symposium, 2008
National Science Foundation Travel Grant, July 2011
Inducted Psi Chi International Honor Society, Spring 2011
Michael J. Piorkowski Award for Intellectual Curiosity and Love of People, April 2015

Professional References:

Jennifer Wiley, Ph.D.
Professor of Psychology
University of Illinois at Chicago
1007 W. Harrison St. (M/C 285)
Chicago, Il 60607
Phone: (312) 355-2501
Email: jwiley@uic.edu

James W. Pellegrino, Ph. D.
Distinguished Professor, Psychology and Education and Co-Director Learning Sciences Research Institute
University of Illinois at Chicago
1007 W. Harrison St. (M/C 285)
Chicago, IL 60607
Phone: (312) 413-2320
Email: pellegjw@uic.edu

Mike Stieff, Ph. D.
Associate Professor, Chemistry and Learning Sciences
University of Illinois at Chicago
845 W. Taylor St. (M/C 111)
Chicago, IL 60607
Phone: (312) 996-4348
Email: mstieff@uic.edu

Tom Moher, Ph. D.
Emeritus Professor
Department of Computer Science
University of Illinois at Chicago
851 S. Morgan (M/C 152)
Chicago, Il 60607
Phone: (312) 996-4562
Email: moher@uic.edu