

**From the holdings of the Department of Special Collections , Princeton University Library,
Princeton, NJ**

These documents are either in the public domain, or can only be used for educational and research purposes (Fair use) as per U.S. Copyright law (text below). By accessing this file, all users agree that their use falls in the public domain or within fair use as defined by the copyright law. We do not charge any permission or usage fees for the publication of images of material in our collections, including those provided by Princeton University Library, via our website, catalog, or directly from staff. For more information regarding copyright, credit, and citation please review the information on the Copyright, Credit and Citation Guidelines page of our website.

<https://library.princeton.edu/special-collections/policies/copyright-credit-and-citation-guidelines>. If you have any questions, including how to order higher-quality images from the collection, contact Special Collections Public Services via our Ask Us! Form. <https://library.princeton.edu/special-collections/ask-us>

Doctoral Dissertations, 2011-2026

Psychology

2017

Decoding 3 Categories of Conventional Wisdom in Pattern Similarity Analyses

Chow, Michael Anthony

Princeton, NJ : Princeton University

<http://arks.princeton.edu/ark:/88435/dsp012n49t431c>

U.S. Copyright law: The copyright law of the United States (Title 17, §108, United States Code) governs the making of photocopies or other reproductions of copyrighted material. Under certain conditions specified in the law, libraries and archives are authorized to furnish a photocopy or other reproduction. One of these specified conditions is that the photocopy or reproduction is not to be used for any purpose other than private study, scholarship, or research. If a user makes a request for, or later uses, a photocopy or reproduction for purposes in excess of fair use that user may be liable for copyright infringement. This institution reserves the right to refuse to accept a copying order if, in its judgment, fulfillment of the order would involve a violation of copyright law. <https://www.copyright.gov/title17/title17.pdf>

Decoding Three Categories of Conventional Wisdom in Pattern Similarity Analyses

Michael Anthony Chow

A DISSERTATION
PRESENTED TO THE FACULTY
OF PRINCETON UNIVERSITY
IN CANDIDACY FOR THE DEGREE
OF DOCTOR OF PHILOSOPHY

RECOMMENDED FOR ACCEPTANCE
BY THE DEPARTMENT OF
PSYCHOLOGY

Advisers: Andrew Conway and Uri Hasson

June 2017

© Copyright by Michael Anthony Chow, 2017. All rights reserved.

Abstract

Many methods in neuroscience characterize shared patterns of activity through the analysis of covariance matrices. Depending on the experimental designs and dimensions of interest, the domains of these analyses are sometimes referred to as functional connectivity, pattern similarity analysis, or inter-subject correlation. Progress in these domains has led to a richer understanding of the brain. However, unexpected properties of commonly employed covariance methods within these domains calls into question conventional wisdom that sometimes motivates analyses. While the common practice of averaging pairwise correlations within- and across-groups of observations may appear straightforward, and comparing these within- and across-group averages is simple to perform, researchers often rely on verbal explanations of what such a contrast might mean.

In this dissertation, we show that these verbal explanations rest on stringent assumptions, and as such, do not hold without qualification. Procedures which selectively average groups prior to correlating, or use elaborate forms of parceling the data and correlating across parcels, rely on the same—often implicit—assumptions. Throughout this paper we demonstrate that these analyses can be viewed as special cases of a latent factor model, in which all of the variances of the latent factors are assumed to be equal. Moreover, this model is not only able to explain such procedures, but has been studied and extended across the social, health, and physical sciences. As a result, we draw connections to analogous issues discussed in the behavioral psychology and psychometric literatures. The first section of this dissertation is empirical research on individual differences in working memory and intelligence, that uses a latent factor model common to individual differences psychology. The second section of this dissertation uses this model to evaluate the three aforementioned procedures in neuroscience.

Such a perspective is valuable because it creates a clear mapping between methodological issues that were (and often continue to be) widespread in covariance analyses from behavioral psychology and their neuroscience counterparts.

Table of Contents

Abstract.....	i
Table of Contents	iii
Acknowledgements.....	vi
Introduction	1
References	6
<u>Section I - The Scope and Control of Attention: Sources of Variance in Working Memory</u>	
Capacity	10
Abstract	10
Introduction	11
Scope of Attention	12
Control of Attention	14
Comparing Scope and Control Tasks	15
The Current Study	17
Method	17
Participants	18
Procedure	18
Visual Array (VA) tasks	18
Change detection.....	18
Color estimation.	19
Operation span.	20
Reading span.	20
Symmetry span.	20
Intelligence tasks.....	20
Raven's Advanced Progressive Matrices, Set II.	21
Cattell's Culture Fair Task.	21
DAT Space Relations Test.	21
ETS Surface Development.....	22
AFQT Reading Comprehension.	22
AFQT Analogies.....	22
Results	22
Data Screening	22
Statistical Procedures	22
Descriptive Statistics	23
Resolution and Number for Visual Arrays	23
One vs. Two Factors of Working Memory Capacity	27
Model of intelligence and g	29
The relationship of VA and CS to g.....	29
Discussion	35

Relationship between VA and CS.....	35
Orthogonality of Number and Resolution	35
Conclusion	40
Section II – Decoding Three Categories of Conventional Wisdom in PSA	48
Overview of Pattern Similarity Analyses	50
Three Common Approaches to Covariance Analysis	55
1. Within and Between Contrasts, Template Regression.	55
2. Averaging Across Items as a Stand-in for Underlying Pattern.	60
3. Splitting Items into Halves Within Group.	61
Linear DAGs, Simpson’s Paradox, and Testing for Different Underlying Patterns	65
Overview of Sections	69
Methods.....	70
Software Tools	70
Restricted Factor Model	70
Example 1: Testing Correlated Factors and the Within- vs Between-group Contrast ..	74
E1.1. Correcting for Unequal Item Reliability Across Groups	77
E.1.2. Simulations Evaluating Bias in Testing for Distinct Groups ($p < 1$)	78
E1.3. Discussion.....	81
Example 2: Relating inter-item and item-total correlation	85
E.2.1. Defining Scenario and Covariance Methods of Interest	86
E.2.2. Inter-item, Item-total, and Total-total Correlation	89
Table 2-1. Within and Between Group Correlations	90
E.2.3. Measuring Item Reliability and Composite Items	93
Inter-Item Correlation.	93
Item-Total Correlation and Composite Items.	93
E.2.4. Latent Pattern Similarity Through Between-group Measures	97
BG Inter-Item Correlation.....	97
BG Item-Total Correlation.	97
Total-Total Correlation.	98
E.2.5. Estimating Pattern Similarity and Consistency from Summary Data	99
Hasson et al. (2009).	99
Regev et al. (2013).	101
E2.6. Discussion.....	104
Example 3: Correlated Residuals and Multiple Groups.....	106
E.3.1. “Decorator Experiment” Data	107
E.3.2. Correlated Residuals	110
E.3.4. Performing Model Selection via Cross-validation	117
E.3.5. Discussion.....	118
General Discussion	121
Representational Similarity Analysis (RSA)	123
Limitations and Future Directions.....	128
Dynamic factor analysis.	128
Spatial factor analysis.	129

Exploratory factor analysis.	129
Factor Models in Neuroscience.	130
Alternative Approaches.	131
Conclusion	131

Acknowledgements

Right out of the gate, let me just say that I am deeply grateful to all those who have guided this research--whether it be through thoughtful critiques, gestures of support, or reminding me to go outside once in a while. You are too many to name.

Thanks to my advisor Andy Conway, who introduced me to an extended family of working memory researchers and the world of psychometrics. I'll never forget the range of problems that passed through lab meetings: sign-language interpreting, working out for working memory's sake, and stereotype threat, to name a few. A big thanks to Uri Hasson, for advising me on applying psychometric ideas to neuroscience analyses. Your guidance and frank feedback were critical for shaping this work. Moreover, thanks to the members of the Conway and Hasson labs. I'm grateful for all of the outings, and brain scanning time we shared, as well as your examples as outstanding researchers.

I'm grateful to my undergraduate advisors, who set me on the right track. Kandi Turley-Ames for waking me to interesting problems in psychology, and giving me my first experiences with research. Dewayne Derryberry, whose statistics lectures and advice continue to bounce around my head regularly.

My first positive experience at Princeton began with Wouter Kool, who housed me during interviews. In the years to come, Shane Blackman, and Kyle Keller, and he provided a much needed social side to the department—progressives, off-the-books psych prom, and more. Definite tip of the cap to you, and to others who made department life so enjoyable—Kaite Yang, Jon Berliner, Nina Rouhani, Laura Bustamante, Jesse Schwab—to name a few. It is at this point that I must acknowledge the sheer recreational delight that came from many summer outings on Mochi Liu's handmade raft, afloat in Lake Carnegie. You are a master of aquatics.

Finally, thank you to my officemate and Conway lab partner in crime, Latasha Holden, for more quality time and sagely advice than is reasonable.

Thank you to my brother and sister, and to their families, for being an abundant source of joy in my life. Thank you to my de facto sister, Davina Hull, who has seen me through thick and thin. Thank you to my partner, Cella Sum, for putting up with my research ramblings and filling my days with joy.

This dissertation is dedicated to my parents, Lou Ann Randall and Dan Chow, for their constant support.

Introduction

Over the past century, psychologists have displayed great interest in covariance-based methods. From Spearman's consideration of how a correlation may be underestimated due to measurement error, to ongoing debate over what may be inferred from the correlation structure between intelligence measures, methodologists in the field (psychometricians) have grappled not only with the issue of developing robust models in order to represent psychological hypotheses from covariance structures. More recently, covariance-based approaches have become foundational to many analysis regimes in functional neuroimaging, including functional connectivity (Greicius, Krasnow, Reiss, & Menon, 2003; van den Heuvel & Hulshoff Pol, 2010), pattern similarity analysis (Haxby, 2001; Kriegeskorte, 2008), and inter-subject correlation (Hasson, 2004). These areas focus on the representation of information or transformational processes through focusing on the spatial and/or temporal structure of neural activity. The insightful use of these methods has led to meaningful inferences on the distributed nature of representation (Haxby, 2001), and the shared temporal dynamics of communication (Stephens, Silbert, & Hasson, 2010). As with any statistical analysis, however, we should be careful with the implementation of these methods, to avoid a mismatch between the appropriate inferences a given analysis allows, and those made by researchers.

Although several neuroscience approaches to covariance analysis--such as pattern similarity analysis and inter-subject correlation--appear distinct enough as to have acquired separate names, they often bear an uncanny resemblance to each other, as well as to classic psychological approaches. Through considering the similarities between standard forms of covariance analysis performed in behavioral psychology and neuroscience, as well as models which may be used to give explicit justification to their use, this dissertation will argue that

methodological pitfalls exposed in behavioral psychology are often prevalent in these neuroscience analyses as well.

In behavioral psychology, an experiment might consist of administering test batteries designed to measure potentially distinct constructs, such as working memory and intelligence. For example, researchers might administer six tests to participants, three for each battery, with an interest in drawing inference from the correlations between tests (Figure 1). Possible inferences may include,

- Are working memory and intelligence the same psychological construct?
- Are specific tests more noisy measurements of their psychological construct?

Following intuition, researchers might reason that if working memory and intelligence are two distinct constructs, then the correlations within a construct (e.g. correlating two tests of working memory) should be greater than those between constructs (e.g. correlating a working memory and intelligence test). Moreover, in order to obtain a single, accurate measure of noisy-ness for each test, they might correlate each test against the average of other within-construct tests. While both of these procedures have been endorsed by psychologists at some point or another, a highly influential paper by Bollen & Lennox (1991) used explicit models to argue that these procedures are often driven by conventional wisdom, and do not hold without qualification.

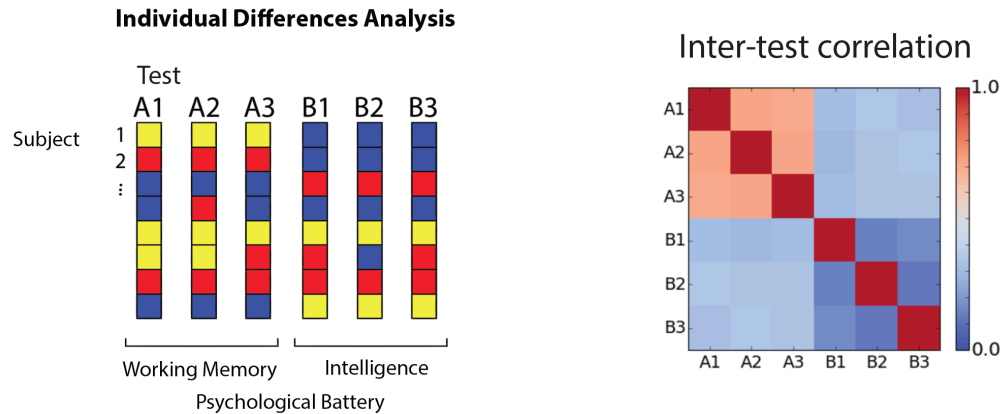


Figure 1. Hypothetical individual differences experiment, in which each subject completes a working memory and intelligence battery, with three tests each (left), and the corresponding correlation matrix between tests.

In the neuroscience analysis regimes of pattern similarity analysis and inter-subject correlation¹, however, it is clear that the two procedures are often the dominant approach taken by researchers. For example, in an experiment using pattern similarity analysis, researchers might employ functional magnetic resonance imaging (fMRI) to measure bold activation in the brain of a participant who is told to attend to either faces or scenes in an image, which are thought to induce distinct psychological face and scene states (Figure 2, left). In this case, when testing for distinct psychological states, a common approach is exactly the contrast of within- and between-state correlations (Aly & Turk-Browne, 2015; e.g. Favila, Chanales, & Kuhl, 2016; Haxby, 2001; Iordan, Greene, Beck, & Fei-Fei, 2015; Kriegeskorte et al., 2008, p. 1133;

¹ Pattern similarity analysis generally refers to relating spatial patterns of activity for a brain region of interest (ROI) across time. For example, calculating the correlation between a spatial pattern of activity during one trial, with another pattern of activity during another trial. Inter-subject correlation often refers to relating patterns of temporal activity for a voxel or region of interest (ROI) across subjects. For example, within an ROI, calculating the timecourse of neural activity while one subject listens to an auditory stimulus to that of another subject listening to the same auditory stimulus.

LaRocque et al., 2013; Op de Beeck, Brants, Baeck, & Wagemans, 2010; Schapiro, Rogers, Cordova, Turk-Browne, & Botvinick, 2013; Wolosin, Zeithamova, & Preston, 2013; Xue et al., 2010). Moreover, a common inter-subject correlation experiment design might show two groups of participants the video narrative, but instruct each group to view it from a different perspective (Figure 2, right). In addition to comparing correlations within and between groups, many studies also favor correlating each subject against the others in their group over using the pairwise correlations between subjects (e.g. Ames, Honey, Chow, Todorov, & Hasson, 2015; Ben-Yakov, Honey, Lerner, & Hasson, 2012; J. Chen et al., 2015; Hasson et al., 2009; Hasson, Yang, Vallines, Heeger, & Rubin, 2008; Honey, Thompson, Lerner, & Hasson, 2012; Regev, Honey, Simony, & Hasson, 2013).

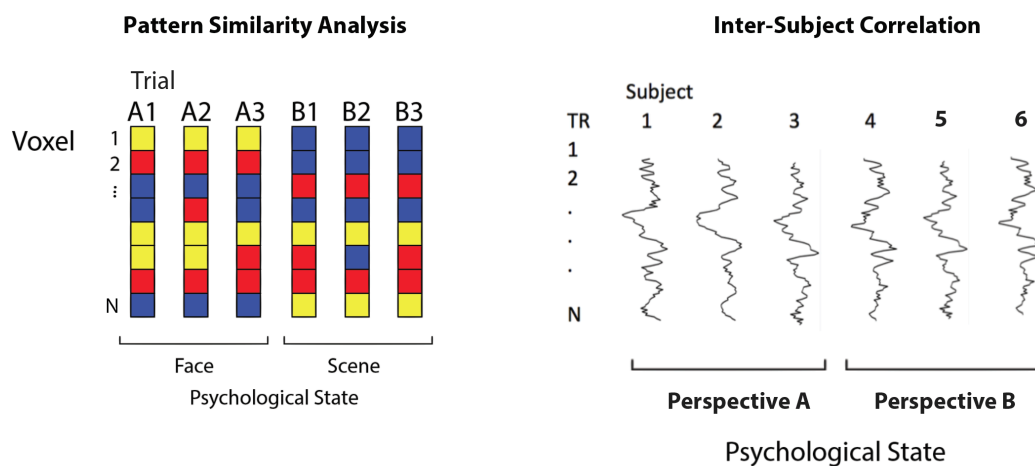


Figure 2. Hypothetical experiments using pattern similarity analysis (left), and inter-subject correlation (right).

This dissertation does not exist to suggest that the approach taken in pattern similarity analysis and inter-subject correlation should reasonably have been avoided. Rather, the occurrence of these approaches in behavioral psychology suggests that simple procedures--such

as contrasting correlations within- and between-groups of items--are highly enticing.

Emphasizing the shared aspects of pattern similarity analyses, inter-subject correlation, and behavioral psychology approaches, expands the methodological tool kit available to researchers in these three domains.

In order to illustrate the classic behavioral psychology perspective, Section I consists of empirical research which demonstrates a common psychometric approach to covariance analysis in behavioral psychology. In this article, structural equation models with latent variables, also called confirmatory factor analyses, are used to examine individual differences in working memory and intelligence measures. Section II demonstrates how this confirmatory factor analysis approach may be applied to pattern similarity analysis and inter-subject correlation of neuroimaging data.

References

- Aly, M., & Turk-Browne, N. B. (2015). Attention Stabilizes Representations in the Human Hippocampus. *Cerebral Cortex*. <http://doi.org/10.1093/cercor/bhv041>
- Ames, D. L., Honey, C. J., Chow, M. A., Todorov, A., & Hasson, U. (2015). Contextual alignment of cognitive and neural dynamics. *Journal of Cognitive Neuroscience*, 27(4), 655–664. http://doi.org/10.1162/jocn_a_00728
- Ben-Yakov, A., Honey, C. J., Lerner, Y., & Hasson, U. (2012). Loss of reliable temporal structure in event-related averaging of naturalistic stimuli. *NeuroImage*, 63(1), 501–506. <http://doi.org/10.1016/j.neuroimage.2012.07.008>
- Bollen, K., & Lennox, R. (1991). Conventional wisdom on measurement: A structural equation perspective. *Psychological Bulletin*, 110(2), 305.
- Chen, J., Honey, C. J., Simony, E., Arcaro, M. J., Norman, K. A., & Hasson, U. (2015). Accessing Real-Life Episodic Information from Minutes versus Hours Earlier Modulates Hippocampal and High-Order Cortical Dynamics. *Cerebral Cortex*, bhv155. <http://doi.org/10.1093/cercor/bhv155>
- Chen, P.-H. C., Chen, J., Yeshurun, Y., Hasson, U., Haxby, J., & Ramadge, P. J. (2015). A reduced-dimension fMRI shared response model. In *Advances in Neural Information Processing Systems* (pp. 460–468).
- Favila, S. E., Chanales, A. J. H., & Kuhl, B. A. (2016). Experience-dependent hippocampal pattern differentiation prevents interference during subsequent learning. *Nat Commun*, 7. Retrieved from <http://dx.doi.org/10.1038/ncomms11066>

- Greicius, M. D., Krasnow, B., Reiss, A. L., & Menon, V. (2003). Functional connectivity in the resting brain: a network analysis of the default mode hypothesis. *Proceedings of the National Academy of Sciences*, 100(1), 253–258.
- Hasson, U. (2004). Intersubject synchronization of cortical activity during natural vision. *Science*, 303(5664), 1634–1640. <http://doi.org/10.1126/science.1089506>
- Hasson, U., Avidan, G., Gelbard, H., Vallines, I., Harel, M., Minshew, N., & Behrmann, M. (2009). Shared and idiosyncratic cortical activation patterns in autism revealed under continuous real-life viewing conditions. *Autism Research*, 2(4), 220–231. <http://doi.org/10.1002/aur.89>
- Hasson, U., Yang, E., Vallines, I., Heeger, D. J., & Rubin, N. (2008). A hierarchy of temporal receptive windows in human cortex. *The Journal of Neuroscience*, 28(10), 2539–2550.
- Haxby, J. V. (2001). Distributed and Overlapping Representations of Faces and Objects in Ventral Temporal Cortex. *Science*, 293(5539), 2425–2430. <http://doi.org/10.1126/science.1063736>
- Honey, C. J., Thompson, C. R., Lerner, Y., & Hasson, U. (2012). Not Lost in Translation: Neural Responses Shared Across Languages. *Journal of Neuroscience*, 32(44), 15277–15283. <http://doi.org/10.1523/JNEUROSCI.1800-12.2012>
- Iordan, M. C., Greene, M. R., Beck, D. M., & Fei-Fei, L. (2015). Basic Level Category Structure Emerges Gradually across Human Ventral Visual Cortex. *Journal of Cognitive Neuroscience*, 27(7), 1427–1446. http://doi.org/10.1162/jocn_a_00790
- Klami, A., Virtanen, S., Leppäaho, E., & Kaski, S. (2015). Group factor analysis. *IEEE Transactions on Neural Networks and Learning Systems*, 26(9), 2136–2147.

- Kriegeskorte, N. (2008). Representational similarity analysis – connecting the branches of systems neuroscience. *Frontiers in Systems Neuroscience*.
<http://doi.org/10.3389/neuro.06.004.2008>
- Kriegeskorte, N., Mur, M., Ruff, D. A., Kiani, R., Bodurka, J., Esteky, H., ... Bandettini, P. A. (2008). Matching Categorical Object Representations in Inferior Temporal Cortex of Man and Monkey. *Neuron*, 60(6), 1126–1141.
<http://doi.org/10.1016/j.neuron.2008.10.043>
- LaRocque, K. F., Smith, M. E., Carr, V. A., Witthoft, N., Grill-Spector, K., & Wagner, A. D. (2013). Global Similarity and Pattern Separation in the Human Medial Temporal Lobe Predict Subsequent Memory. *Journal of Neuroscience*, 33(13), 5466–5474.
<http://doi.org/10.1523/JNEUROSCI.4293-12.2013>
- Murphy, K., Birn, R. M., Handwerker, D. A., Jones, T. B., & Bandettini, P. A. (2009). The impact of global signal regression on resting state correlations: Are anti-correlated networks introduced? *NeuroImage*, 44(3), 893–905.
<http://doi.org/10.1016/j.neuroimage.2008.09.036>
- Op de Beeck, H. P., Brants, M., Baeck, A., & Wagemans, J. (2010). Distributed subordinate specificity for bodies, faces, and buildings in human ventral visual cortex. *NeuroImage*, 49(4), 3414–3425. <http://doi.org/10.1016/j.neuroimage.2009.11.022>
- Regev, M., Honey, C. J., Simony, E., & Hasson, U. (2013). Selective and Invariant Neural Responses to Spoken and Written Narratives. *Journal of Neuroscience*, 33(40), 15978–15988. <http://doi.org/10.1523/JNEUROSCI.1580-13.2013>

- Schapiro, A. C., Rogers, T. T., Cordova, N. I., Turk-Browne, N. B., & Botvinick, M. M. (2013). Neural representations of events arise from temporal community structure. *Nature Neuroscience*, *16*(4), 486–492. <http://doi.org/10.1038/nn.3331>
- Stephens, G. J., Silbert, L. J., & Hasson, U. (2010). Speaker-listener neural coupling underlies successful communication. *Proceedings of the National Academy of Sciences*, *107*(32), 14425–14430. <http://doi.org/10.1073/pnas.1008662107>
- van den Heuvel, M. P., & Hulshoff Pol, H. E. (2010). Exploring the brain network: A review on resting-state fMRI functional connectivity. *European Neuropsychopharmacology*, *20*(8), 519–534. <http://doi.org/10.1016/j.euroneuro.2010.03.008>
- Wolosin, S. M., Zeithamova, D., & Preston, A. R. (2013). Distributed hippocampal patterns that discriminate reward context are associated with enhanced associative binding. *Journal of Experimental Psychology: General*, *142*(4), 1264–1276. <http://doi.org/10.1037/a0033609>
- Xue, G., Dong, Q., Chen, C., Lu, Z., Mumford, J. A., & Poldrack, R. A. (2010). Greater Neural Pattern Similarity Across Repetitions Is Associated with Better Memory. *Science*, *330*(6000), 97–101. <http://doi.org/10.1126/science.1193125>

Section I - The Scope and Control of Attention: Sources of Variance in Working Memory Capacity

Abstract

Working memory capacity is a strong positive predictor of many cognitive abilities, across various domains. The pattern of positive correlations across domains has been interpreted as evidence for a unitary source of inter-individual differences in behavior. However, recent work suggests that there are multiple sources of variance contributing to working memory capacity. The current study ($N = 71$) investigates individual differences in the scope and control of attention, in addition to the number and resolution of items maintained in working memory. Latent variable analyses indicate that the scope and control of attention reflect independent sources of variance and each account for unique variance in general intelligence. Also, estimates of the number of items maintained in working memory are consistent across tasks and related to general intelligence whereas estimates of resolution are task-dependent and not predictive of intelligence. These results provide insight into the structure of working memory, as well as intelligence, and raise new questions about the distinction between number and resolution in visual short-term memory.

Published as:

Chow, M., & Conway, A. R. (2015). The scope and control of attention: Sources of variance in working memory capacity. *Memory & cognition*, 43(3), 325-339.

Introduction

Working memory is a limited capacity system responsible for the active maintenance of information, as well as its retrieval from long term memory. Recent work has emphasized the distinction between two components of working memory: the *scope* and *control* of attention (Cowan et al., 2005). The scope of attention refers to the amount of information that can be actively maintained at a given time, whereas the control of attention refers to the ability to focus on relevant information, and away from irrelevant information.

Working memory is an extremely active area of research in psychology and neuroscience. One motivating factor behind this research effort is the strength of correlations between measures of working memory capacity and measures of higher cognitive function, such as reading comprehension, spatial reasoning, and fluid intelligence (Daneman & Carpenter, 1980; Kane et al., 2004; Kyllonen & Christal, 1990). While research in this area has traditionally focused on the relationship between the control of attention and higher cognitive function, several recent studies have linked the scope of attention to higher cognitive function as well (Cowan et al., 2005; Fukuda, Vogel, Mayr, & Awh, 2010; Shipstead, Redick, Hicks, & Engle, 2012; Unsworth, Fukuda, Awh, & Vogel, 2014). Moreover, the scope of attention can be fractionated into two components: the number of items that can be maintained and the resolution, or acuity, of those items. Importantly, estimates of number, but not resolution, have been found to correlate with measures of spatial reasoning (Fukuda et al., 2010). This apparent dissociation is important because working memory tasks have been accused of lacking discriminant validity, as they tend to correlate with a broad range of abilities (Ackerman, Beier, & Boyle, 2005; but see, Kane, Hambrick, & Conway, 2005). Through examining the scope of attention, the control of attention, and ways in which both facets of working memory are related, this dissociation

within the scope of attention may be leveraged as a powerful tool in understanding the working memory system.

Scope of Attention

An influential theory of working memory is the Embedded Process Model (Cowan, 1999). According to this framework, capacity is constrained by the “focus of attention”, which actively maintains items in memory and insulates them from interference and forgetting. In order to assess the number of items that can be simultaneously maintained in the focus of attention, several tasks have been developed. One of the most common procedures is the “visual array” task. For example, Luck & Vogel (1997) used a visual array change-detection procedure, in which participants were required to maintain an array of simple objects in memory, and then indicate whether a probe array was the same or different from the one seen previously. Accuracy was near ceiling when the number of objects in the array was 3 or less but with 4 or more objects in the array, accuracy declined. It is also important to note that this pattern of accuracy in detecting changes was the same even when objects increased in complexity. Moreover, by extending a simple model of capacity by Pashler (1988), which takes into account guessing rates, Cowan (2001) demonstrated that capacity estimates tend to be around 4 items. Importantly, the fact that many other types of visual array tasks, such as those which display only a single item probe, or require the participant to estimate properties of a display item, produce similar capacity estimates provides strong evidence that the focus of attention can maintain approximately 4 items (Cowan, 2001).

There is also substantial evidence that the number of items held in memory can be distinguished from the resolution, or quality of representation, for those items (Ester, Anderson, Serences, & Awh, 2013; Zhang & Luck, 2008, 2011). Resolution in change detection tasks is

often operationalized by increased similarity between probe and memory items, under the rationale that successful recognition requires more detailed memory of the probed item. Using the change detection paradigm, Xu & Chun (2005) found that the intra-parietal sulcus could be functionally dissociated into two distinct regions, one that tracks the number of items regardless of complexity, and another that tracks the number of relevant features. In addition, Awh, Barton, and Vogel (2007) demonstrated that although measures of number and resolution in a change detection task were reliable across different types of stimuli, they were uncorrelated with each other. Finally, Fukuda et al. (2010) compared number, resolution, and spatial reasoning measures using structural equation modeling, and found that only number and spatial reasoning covaried. The correlation between resolution and spatial reasoning was near zero. In fact, by deriving a 95% confidence interval from their reported standard errors, it is clear that if a non-trivial correlation exists between resolution and spatial reasoning, then likely to be fairly weak (r between $-.05$ and $.25$).

However, unlike the convergent evidence for a fixed capacity limit across a wide range of tasks, evidence for a dissociation between number and resolution comes primarily from studies that employed only one visual array task. While illustrating this dissociation using a single task is an important step, there is a risk that the effect is driven by task-specific factors. Whereas change detection tasks manipulate the degree of change on a few, discrete levels, more recent tasks have required participants to estimate the some property of an item (e.g. color) in a continuous manner (Wilken & Ma, 2004; Zhang & Luck, 2008). Consensus among a range of tasks used to operationalize these constructs would allow for a stronger argument for dissociation. Fortunately, many of the same tasks used to assess capacity limits can be adapted to provide measures of resolution as well.

Control of Attention

While the scope of attention refers to the number and resolution of representations that can be held simultaneously in an active state, the control of attention refers to processes that allow for the maintenance of task-relevant representations in the face of distraction (Kane, Bleckley, Conway, & Engle, 2001), as well as selective retrieval from memory (Unsworth & Engle, 2006a). To this end, the control of attention has largely been operationalized through tasks that interject processing components between memoranda, under the rationale that participants must either maintain representations for memoranda throughout the distracting periods, or retrieve them afterward. Such tasks are typically referred to as complex span tasks.

Complex span task scores correlate with performance on tasks that require the top-down guidance of attention. For example, dichotic listening (Conway, Cowan, Bunting, 2001), Stroop (Kane & Engle, 2003), Ericson flanker (Heitz & Engle, 2007), and anti-saccade tasks (Kane et al., 2001, Unsworth, Schrock, & Engle, 2004) all demonstrate relationships to complex span task performance.

In addition, there is mounting evidence that complex span tasks rely on retrieval from episodic memory (Spillers & Unsworth, 2011; Unsworth & Engle, 2006a, 2006b; for more general overview, see Postle, 2007). One source of evidence comes from the pattern of errors committed on complex span tasks. Consistent with the use of temporal-contextual cues to guide memory search and retrieval, low working memory participants are more likely to omit earlier list items, allow prior list intrusions for the first position recalled, and have a broader transposition gradient (Unsworth & Engle, 2006a). One explanation for this is that they employ noisier memory cues. Further evidence of this is that they show a reduced tendency to recall words in correct temporal order during delayed free recall, suggesting that recalled words serve

more poorly as cues for further recall (Spillers & Unsworth, 2011). Furthermore, memory retrieval during complex span tasks is associated with increased medial temporal lobe activity (Chein, Moore, & Conway, 2011; Faraco et al., 2011). As the medial temporal lobe is thought to be involved with the binding of material in long-term memory (McClelland, McNaughton, & O'Reilly, 1995; O'Reilly, Bhattacharyya, Howard, & Ketz, 2011), these tasks may rely on temporary storage and selective retrieval from episodic memory rather than (or in addition to) active maintenance.

Importantly, convergent evidence that complex span tasks are associated with top-down guidance of attention and retrieval from episodic memory comes from the use of processing components ranging from solving basic arithmetic to making symmetry judgments, and memoranda ranging from letters to spatial locations. Thus, the effects appear to be domain general (Kane et al., 2004).

Comparing Scope and Control Tasks

While much research has focused on either the scope or control of attention individually, there is good reason to believe that they tap largely overlapping constructs. Specifically, (Cowan et al., 2005) (2005) argued that the critical factor for a valid working memory task is its ability to prevent strategic grouping and rehearsal. Under this line of reasoning, visual array tasks prevent rehearsal by using stimuli that are difficult to verbalize, while complex span tasks prevent rehearsal by interjecting a processing task that occupies attention during intervals between memoranda (Cowan, Saults, & Morey, 2006). Consistent with this reasoning, even complex span tasks with a rudimentary processing component correlate with intelligence (Lépine, Barrouillet, & Camos, 2005). Moreover, for children who are too young to rehearse or group items, simple digit span tasks correlate with intelligence as well (Cowan et al., 2005).

By comparing the scope and control of attention to measures of higher cognitive function, Cowan et al. (2005) were able to assess whether each component contributed unique variance to higher cognitive function. The scope of attention was measured using a wide range of tasks, including visual arrays, which contained only a storage component, but were designed to reduce strategic grouping and rehearsal. The control of attention was assessed using two complex span tasks. While the complex span tasks explained more variance in higher cognitive function, this variance was task-specific. That is, the additional variance was particular to the individual tasks. Cowan et al. (2005) concluded that disrupting rehearsal in complex span allows a purer estimate of the number of items in memory, implicating these tasks as measures of the scope of attention.

However, Shipstead & Engle (2012) highlight an issue that arises from claiming the scope of attention is necessary for explaining the relationship between complex span task performance and higher cognitive function. Namely, for short list lengths, in which the amount of information to be retained falls below capacity estimates for most participants, ceiling effects in measuring capacity should reduce the correlation between complex span and high-order cognition. However, the correlation between complex span and general intelligence remains equally strong across a wide range of list lengths (Bailey, Dunlosky, & Kane, 2011; Salthouse & Pink, 2008; N. Unsworth & Engle, 2006b). Through reexamining two datasets that contained a visual array task in addition to complex span and general intelligence measures, they concluded that the scope and control of attention represent distinct but largely overlapping constructs. Moreover, the variance shared between visual array and general intelligence measures was largely explained by complex span. However, one limitation they mention is the use of a single visual array task, which is likely to induce a large degree of task specific variance.

In spite of the strong evidence for overlap between the scope and control of attention, comparisons between the two constructs have yet to address the role of resolution. If resolution is uncorrelated with measures representing the control of attention as well, it represents a fairly astounding finding because over a century of psychometric research suggests that two similar cognitive ability measures should reveal a strong positive correlation, or at the very least, a weak positive correlation. It is extremely rare to find orthogonal manifest variables when testing cognitive ability (Ackerman et al., 2005).

The Current Study

The present study extends recent work (Fukuda et al., 2010; Shipstead et al., 2012) by administering multiple types of visual array tasks to represent the scope of attention and by administering a broad range of reasoning tasks to provide an accurate assessment of general intelligence. By using multiple measures to reflect the number and resolution of items in the scope of attention, we can reduce the proportion of task specific variance in that factor, allowing a stronger test of whether it is distinct from the control of attention, and a more clear view of its shared variance with general intelligence (for an example using many tasks to examine number, but not resolution, see Unsworth et al., 2014). Moreover, assessing and validating resolution from those tasks allows for the opportunity to test whether orthogonality between measures of number and resolution is a general property of visual short-term memory, or a task specific phenomenon. In addition, it allows for novel comparisons between resolution and the control of attention. If resolution is orthogonal to the control of attention as well, then it could serve as a starting point for development of analogous measures in tasks designed to measure the control of attention.

Method

Participants

Participants ($N = 71$) were recruited from Princeton University and the surrounding community. Students recruited from the Psychology Department participated in exchange for course credit ($n = 55$). Students and community members recruited from the University participant pool ($n = 16$) were compensated \$12/h for their participation.

Procedure

Each participant completed a battery of tasks designed to measure working memory capacity and general intelligence. The tasks were completed in two sessions that lasted approximately two hours each. Breaks were allowed between tasks. General intelligence measures were split evenly between the two sessions. Complex span tasks were administered in the first session, while visual array tasks were administered in the second session. All participants completed general intelligence and working memory tasks in the same, alternating order. In the first session, tasks were administered to groups of up to six participants. In the second session, tasks were administered individually.

Visual Array (VA) tasks

Change detection. The change detection procedure was based on a task previously developed by Fukuda et al. (2010). For each trial, participants viewed an array of rectangles and ovals. Each shape contained one of two different internal patterns. Participants viewed each array for 500 ms, and after a blank interval of 1000 ms were shown a single probe shape with a pattern inside. They then responded as to whether the probe shape was identical to the one they had viewed in the visual array at that location. The probe could either be identical (no change), a different shape (big change), or the same shape but with a different internal pattern (small change). Arrays contained either 4 or 8 items.

Within each block, half of the trials were no-change trials. For each change trial, there was a 2/3 chance it would be a big change, and 1/3 chance it would be a small change. 16 trials for each set size were interleaved randomly each block. The total task consisted of 9 blocks, and 4 practice trials at each set size, yielding 296 trials total. From large change conditions, the number of items in memory was estimated using the formula given by Cowan (2001): $k = \text{set size} * (\text{correct_hits} - \text{false_alarms})$, where k is the total items in memory, set size is the size of the initial array, correct hits is the proportion of correct change trials, and false alarms are the proportion of incorrect no-change trials. From small change conditions, the resolution of items in memory was calculated in the same way. This was done for each set size. Note that the same no-change trials were used in calculating number and resolution. This is consistent with estimation methods in previous studies (Awh, Barton, & Vogel, 2007; Fukuda, Vogel, Mayr, & Awh, 2010).

Color estimation. The color estimation procedure was based on a task previously developed by Zhang & Luck, (2008). On each trial, participants viewed an array of 4 or 6 colored squares for 50 ms. Following a 1000 ms interval, the squares were presented again without color, and a single probe square was outlined in bold. Participants estimated the original color of the probe by selecting from a continuous color wheel, which surrounded the array. As in Zhang & Luck (2008), for each set size, maximum likelihood estimation over a mixture of a von Mises and uniform distribution were used to determine the number and resolution of items simultaneously. Number was represented by the weight given to the von Mises distribution, while resolution was estimated by its dispersion parameter, with greater dispersion parameters representing poor resolution.

Complex Span (CS) tasks

Three automated complex span tasks were administered (Unsworth, Heitz, Schrock, & Engle, 2005; see also Redick et al., 2012). On each trial of these tasks, participants alternated between performing a secondary task and viewing a to-be-remembered (TBR) item. After a number of rounds, participants reported as many TBR items as possible, in the same order they were presented, by selecting the items from a grid. A point was awarded for each TBR item recalled in correct serial order, which is often referred to as partial credit load scoring (Conway et al., 2005). Participants were instructed to click a button labeled “blank” to skip a serial position.

Operation span. The operation span (OSPAN) consisted of solving basic arithmetic, followed by a TBR letter. Set sizes ranged from 3 to 7 items. Each set size was completed 3 times.

Reading span. The reading span (RSPAN) consisted of indicating whether a sentence was grammatical, followed by a TBR letter. Set sizes ranged from 3 to 7 items. Each set size was completed 3 times.

Symmetry span. The symmetry span (SSPAN) task consisted of solving for whether an 8x8 grid of black and white squares was symmetrical, followed by a TBR blue square presented on a 4x4 grid. Set sizes ranged from 2 to 6 items. Each set size was completed 3 times.

Intelligence tasks

All measures of intelligence were administered on paper. Before each task, an experimenter reviewed the instructions aloud, went through practice problems, and asked participants whether they had any questions. For each task, one point was awarded for each correct response.

Raven's Advanced Progressive Matrices, Set II. In this task, a 3x3 matrix has 8 images in it, while the bottom-right cell is missing in each problem. Participants indicate which of 8 options would complete the matrix by looking at the relations among the images across the rows and down the columns. After completing the first two problems from Set I as practice, participants were given the 18 odd numbered problems from Set II, and had 15 minutes to complete the task.

Cattell's Culture Fair Task. This task consisted of four subtests. In the first, a row of three boxes and an empty box were shown, and participants chose which option best completes a pattern shown in the first three boxes. In the second, five boxes were shown and participants indicate which two boxes do not belong to the pattern made by the other three boxes. In the third, images fill a 2x2 or 3x3 matrix of boxes, but the bottom-right box was empty. Participants chose, from a number of options, the box that best completed the pattern. Finally, the last test consisted of an image containing a number of shapes and a dot. The goal of this test was to choose from different configurations of shapes, the one in which a dot could be placed in a similar location relative to the shapes (e.g. encompassed by a triangle, but not by a circle). In some instances, the images contained two dots, and participants were to follow a similar rule in choosing responses. The number of items for each subtest were 13, 14, 13, and 10, respectively. Scores for each subtest were summed to create one overall score.

DAT Space Relations Test. For each item, flat shapes were presented along with five printed 3-dimensional depictions of what each shape might look like when folded along its edges. Participants chose the 3-dimensional depiction that could be made from the flat shape. This test consisted of 18 items.

ETS Surface Development. For each item, a flat shape was presented along with an accurate 3-dimensional depiction of that shape folded along its edges and rotated. Lines on the flat shape were numbered, while lines on the folded shape had letters next to them. Participants indicated which numbers on the flat shape corresponded to letters on the folded shape. This test consisted of 5 shapes with 5 items each, yielding 25 items total.

AFQT Reading Comprehension. For each item, participants read short paragraphs and chose from five options the one that best completed each paragraph. This test consisted of 15 items.

AFQT Analogies. For each item, participants were presented with an incomplete analogy (e.g. QUART is to LITER as INCH is to) and chose from five options the one that best completed each analogy. This test consisted of 18 items.

Results

Data Screening

Participants with scores greater than 4 standard deviations on any measure were considered univariate outliers and discarded. From this procedure, 2 participants were removed. In addition, the mvoutlier package in R was used to plot the ordered, robust mahalanobis distance against the quantiles of the Chi-squared distribution. Participants who deviated largely from the expected quantiles were removed and the plot was reconstructed until there appeared to be no substantial curvature. From this procedure, 2 additional participants were discarded as multivariate outliers.

Statistical Procedures

A series of latent variable models (aka confirmatory factor analyses) were conducted. Multiple indices of fit are reported for each model. First, the chi-square statistic measures the

mismatch between the observed and reconstructed covariance matrices. Higher values indicate greater mismatch. However, moderate to large sample sizes may lead to statistically significant differences, even when the discrepancy between covariance matrices is slight. So in addition, we report the root mean square error of approximation (RMSEA), standardized root mean square residual (SRMR), non-normed fit index (NNFI), comparative fit index (CFI), and Akaike's information criterion (AIC). Following the suggestion of Hu and Bentler (1999), we consider the following to be evidence of good fit: $RMSEA < .08$, $SRMR < .08$, $NNFI > .95$, $CFI > .95$. AIC is a measure of parsimony used for model comparisons, with lower AIC indicating a better fit after applying a penalty for free parameters.

All analyses were performed using the R software language (R Core Team, 2012). In addition to the base R libraries, custom libraries were used for general data processing (Wickham, 2007, 2011), outlier detection (Filzmoser & Gschwandtner, 2013), and structural equation models (Rosseel, 2012).

Descriptive Statistics

Descriptive statistics for all measures are provided in Table 1. All meet the acceptable criteria suggested by Kline (2011) for latent variable model analyses (absolute skew < 3 ; kurtosis < 10). The full correlation matrix is provided in Table 2.

Resolution and Number for Visual Arrays

We tested whether measures for number and resolution in the scope of attention load on distinct factors, and whether resolution is task dependent, by contrasting three models. In the first model, all visual array measures loaded on a single factor. In the second model, measures of resolution and number loaded on separate factors. The third model was similar to the second,

except that the resolution factor was split into two, so that each task loaded onto a separate resolution factor.

Fit indices are provided in Table 3. All models failed to meet any of the criteria for a good fit to the data. In order to address whether the poor fit was because of task-specific covariance among measurements, new models were constructed which allowed

Table 1: Descriptive statistics

	mean	sd	min	max	skew	kurtosis
s span	31.15	7.67	13.00	42.00	-0.47	-0.81
r span	62.12	10.48	24.00	75.00	-1.58	3.20
o span	63.54	9.61	36.00	75.00	-1.05	0.61
cd_4_num	2.96	0.57	1.46	3.89	-0.86	-0.03
cd_8_num	3.65	1.45	0.31	6.59	-0.05	-0.92
cd_4_res	1.59	0.75	-0.25	3.52	-0.09	-0.27
cd_8_res	1.37	1.18	-1.38	3.64	-0.04	-0.44
color_4_num	66.96	17.38	21.82	92.30	-0.72	-0.09
color_6_num	47.91	15.14	5.40	78.84	-0.43	0.23
color_4_res	19.97	3.90	8.56	33.15	0.35	1.43
color_6_res	21.99	7.38	5.73	52.99	1.01	3.16
ravens	13.70	2.36	8.00	18.00	-0.27	-0.54
cft	31.12	3.82	21.00	39.00	0.01	-0.38
surfdev	20.43	4.93	6.00	25.00	-1.10	0.33
spacerel	13.40	2.94	6.00	18.00	-0.52	-0.36
readcomp	11.69	2.05	6.00	14.00	-0.74	-0.25
analogy	13.19	2.07	8.00	17.00	-0.40	-0.34

Note. rspan = reading span; sspan = symmetry span; ospan = operation span; cd-4-num = change detection (number), set size 4; cd-8-num = change detection (number), set size 8; cd-4-res = change detection (resolution); cd-8-res = change detection (resolution); color-4-num = color estimation (number), set size 4; color-6-num = color estimation (number), set size 6; color-4-res = color estimation (resolution), set size 4; color-6-res = color estimation (resolution), set size 6; ravens = Raven's Advanced progressive matrices; cft = Cattell's culture-fair test; surfdev = surface development; spacerel = space relations; analogy = analogy battery; readcomp = reading comprehension.

Table 2: Correlations among variables.

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
span	-																
rspan	.50***	-															
ospan	.22	.57***	-														
cd_4_num	.32**	.34**	.13	-													
cd_8_num	.52***	.38**	.24	.66***	-												
cd_4_res	.39***	.20	.19	.45***	.45***	-											
cd_8_res	.17	-.01	.17	.26*	.43***	.58***	-										
color_4_num	.26*	.21	.03	.58***	.43***	.43***	.27*	-									
color_6_num	.15	.13	-.05	.50***	.38**	.38**	.17	.77***	-								
color_4_res	-.25*	-.18	-.11	-.07	.02	-.05	.02	.15	.14	-							
color_6_res	-.16	.00	-.07	.06	.03	-.24	-.23	.12	.41***	.47***	-						
ravens	.43***	.26*	.01	.41***	.61***	.30*	.15	.29*	.34**	-.11	.03	-					
cft	.40***	.23	.06	.34**	.42***	.27*	.17	.22	.08	-.11	-.19	.37**	-				
surfdev	.58***	.46***	.26*	.38**	.51***	.37**	.14	.22	.29*	-.04	.06	.54***	.49***	-			
spacerel	.47***	.37**	.22	.50***	.49***	.39**	.21	.39**	.39**	.11	.06	.58***	.37**	.67***	-		
readcomp	.25*	.35**	.27*	.11	.20	.11	.08	.09	.11	-.02	.01	.22	.19	.38**	.28*	-	
analogy	.22	.38**	.31*	.26*	.31*	.17	.02	.07	.02	-.15	-.08	.40***	.33**	.44***	.35**	.27*	-

Note. rspan = reading span; span = symmetry span; ospan = operation span; cd-4-num = change detection (number), set size 4; cd-8-num = change detection (number), set size 8; cd-4-res = change detection (resolution); cd-8-res = change detection (resolution); color-4-num = color estimation (number), set size 4; color-6-num = color estimation (number), set size 6; color-4-res = color estimation (resolution), set size 4; color-6-res = color estimation (resolution), set size 6; ravens = Raven's Advanced progressive matrices; cft = Cattell's culture-fair test; spacerel = space relations; analogy = analogy battery; readcomp = reading comprehension.

residuals to covary. Specifically, for each task and each measure (either number or resolution), the residuals were allowed to covary across the two set sizes. However, all fit indices for these models also failed to meet the criteria for a good fit.

All three of our theoretically motivated models failed to fit the data. In an attempt to find a model that could fit the data we considered a four-factor model, consisting of two factors for number and two factors for resolution. This model provided a better fit than the other models but still did not satisfy our criteria for adequate fit (see Table 3).

An

Table 3: Fit indices for all visual array tasks

	model	chisq	df	pvalue	rmsea	srmr	nnfi	cfi	aic
1	vstm.1f	108.91	20.00	0.00	0.26	0.15	0.42	0.59	1410.44
2	vstm.2f	87.54	19.00	0.00	0.23	0.15	0.53	0.68	1391.07
3	vstm.3f.2res	58.17	17.00	0.00	0.19	0.09	0.69	0.81	1365.71
4	vstm.4f	44.51	18.00	0.00	0.15	0.09	0.81	0.88	1350.04

Note. rmsea = root mean square error of approximation; srmr = standardised root mean square residual; nnfi = non-normed fit index; cfi = comparative fit index; aic = Akaike's information criterion.

examination of the correlation matrix shows that measures of resolution are correlated within each task (change detection, $r = .58, p < .001$; color estimation, $r = .47, p < .001$). However, the correlations between number and resolution *across* tasks is complicated. For the change detection task the correlations between number and resolution are positive at small and large set sizes. In contrast, for color estimation, the correlations between number and resolution vary as a function of set size. Different patterns of correlation across set sizes within the visual array tasks are not predicted by any of the models, which helps to explain the relatively poor fit of our three theoretically motivated models.

One vs. Two Factors of Working Memory Capacity

In order to assess whether visual array and complex span tasks are better represented by distinct components, we compared a single factor model to a two factor model. In the single factor model, visual array and complex span tasks loaded on the same factor. In the two factor model, visual array tasks loaded on one factor, while complex span tasks loaded on another. Consistent with previous modeling efforts using automated span tasks (Shipstead, Redick, Hicks, & Engle, 2012), residuals for OSPAN and RSPAN were allowed to covary in each of the models.

Fit indices are provided in Table 4. A direct comparison of the models favored the two factor model ($\chi^2 = 9.47$, $p < .01$). All fit statistics for the single factor model failed to meet the criteria listed in the Statistical Procedure section, indicating a poor overall fit to the data. However, all fit statistics for the two factor model met or exceeded the same criteria, indicating a good overall fit. The two factor model is displayed in Figure 1. Consistent with Shipstead et al. (2012), the latent variable for visual array tasks is labeled VA, while the latent variable for complex span tasks is labeled CS.

Table 4: Fit indices for working memory tasks

	model	chisq	df	pvalue	rmsea	srmr	nnfi	cfi	aic
1	1.factor	22.67	12.00	0.03	0.12	0.08	0.90	0.94	1178.05
2	2.factor	13.94	11.00	0.24	0.06	0.06	0.97	0.98	1171.33

Note. rmsea = root mean square error of approximation; srmr = standardised root mean square residual; nnfi = non-normed fit index; cfi = comparative fit index; aic = Akaike's information criterion.

Figure 1

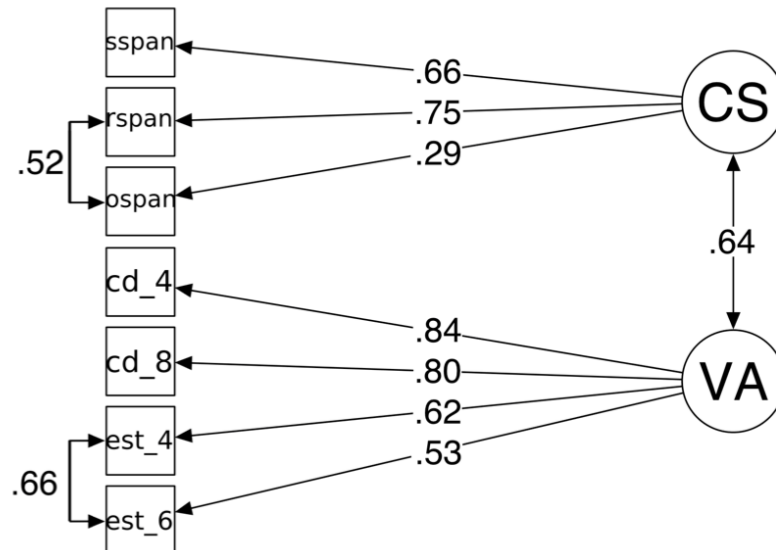


Figure 2. Two factor model for working memory tasks. est-4 = color estimation, set size 4; est-6 = color estimation, set size 6; cd-4 = change detection, set size 4; cd-8 = change detection, set size 8; rspan = reading span, sspan = symmetry span, ospan = operation span.

Model of intelligence and g

The battery of intelligence tests consisted of verbal, spatial, and fluid reasoning tasks.

The battery consisted of only 6 tests but we tapped different domains in order to estimate a general factor of intelligence. Indeed, a one factor model of the 6 intelligence tests proved to be sufficient. The model is illustrated in Figure 2. All fit statistics satisfied the criteria listed in the Statistical Procedure section, indicating a good overall fit of the model.

The relationship of VA and CS to g

The models tested thus far suggest that (a) VA and CS are separate but related factors; and (b) a strong general factor emerged from the battery of intelligence tests we

Figure 2

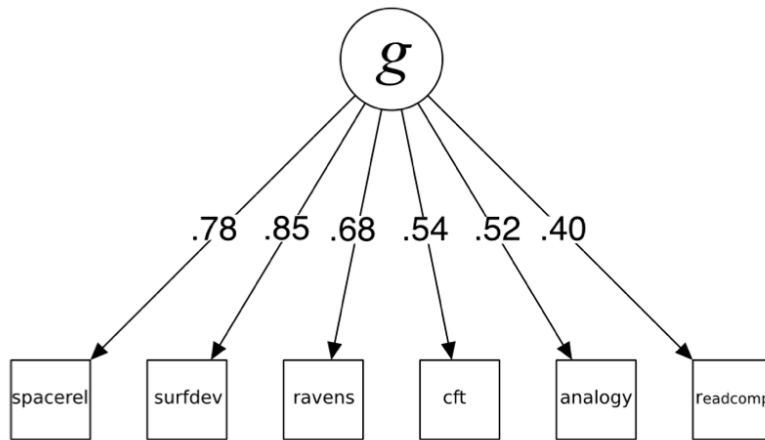


Figure 1. One factor model of intelligence. ravens = Raven's Advanced progressive matrices; cft = Cattell's culture-fair test; spacerel = space relations; analogy = analogy battery; readcomp = reading comprehension.

administered. Based on these results, we tested the hypothesis that VA and CS each account for unique variance in g by computing factor scores for VA, CS, and g (based on the models reported above) and then conducting a multiple regression analysis with g as the outcome variable and VA and CS as the predictors. A summary of the regression analysis is reported in Table 6. Our prediction that VA and CS account for independent sources of variance in g was supported. The regression coefficients for both VA and CS were significant (for both, $p < .05$) and together VA and CS accounted for 85% of the variance in g .

Finally, to be consistent with prior work, we considered models that included all manifest variables from the entire study. Based on Shipstead et al. (2012), we tested three models. The first model predicts that VA is the primary source of variance in g . The second model predicts that CS is the primary source of variance in g . The third model, which we prefer, and is consistent with our regression analysis, assumes that both VA and CS account for unique

variance in g . The fit statistics for these models are provided in Table 5, and SEM parameter estimates for the third model are presented in Figure 3. The second and third model are clearly superior to the first but the distinction between model 2 and model 3 is less clear. While model 2 is more parsimonious, we are hesitant to conclude that CS is essential and VA is irrelevant. The above regression analysis, and previous work, suggests that both CS and VA contribute to variability in general intelligence. The lack of discrimination between models 2 and 3 in the current study may be due to the small sample and/or the limited sample of tasks for both CS and VA.

Table 5: Fit indices for model of intelligence

	model	chisq	df	pvalue	rmsea	srmr	nnfi	cfi	aic
1	sem.va	75.86	60.00	0.08	0.06	0.08	0.94	0.95	2178.06
2	sem.cs	69.48	60.00	0.19	0.05	0.07	0.96	0.97	2171.69
3	sem.both	67.57	59.00	0.21	0.05	0.07	0.97	0.97	2171.77

Note. rmsea = root mean square error of approximation; srmr = standardised root mean square residual; nnfi = non-normed fit index; cfi = comparative fit index; aic = Akaike's information criterion.

Figure 3

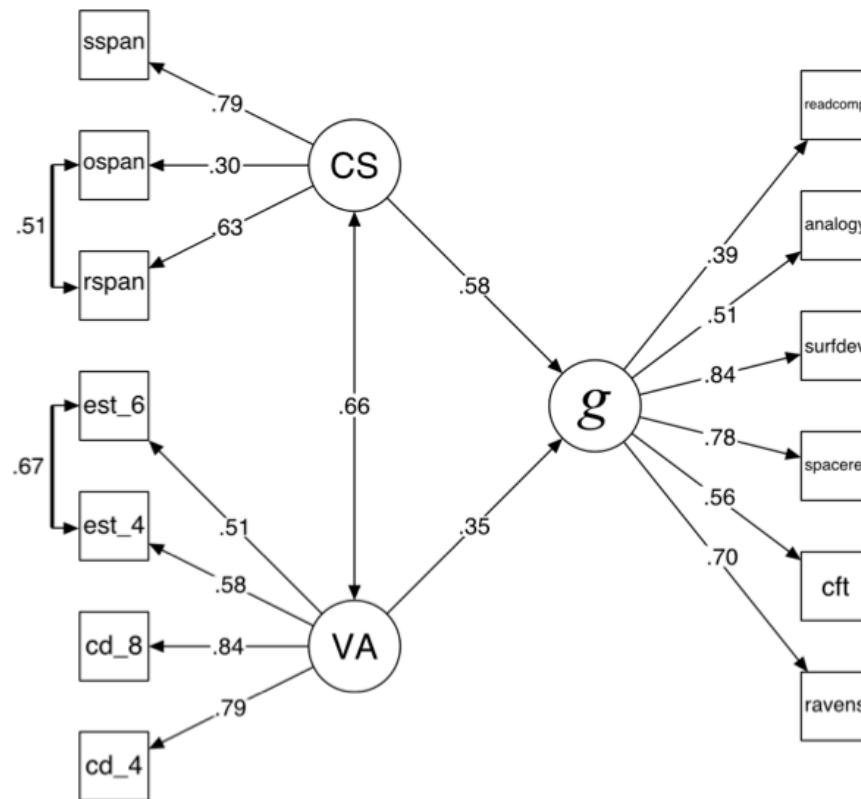


Figure 3. Full model. est = color estimation; cd = change detection; rspan = reading span; sspan = symmetry span; ospan = operation span; ravens = Raven's Advanced progressive matrices; cft = Cattell's culture-fair test; spacerel = space relations; analogy = analogy battery; readcomp = reading comprehension.

In order to ensure that the current findings are not a result of the large number of measurements, relative to number of participants, the same latent variable models were examined, but with measures that came from the same type of visual short-term memory task (either change detection or color estimation) averaged together. This is known as measurement parceling, and is done to both reduce the number of free parameters, as well as increase the test reliability of each resulting measurement (Little, Cunningham, Shahar, & Widaman, 2002). Consistent with advice given by Little et al. ((2002) for reducing subfactors to parcels, in order to gain a general, unidimensional measure of g , each pair of tasks were averaged within the spatial, formal inductive, and verbal reasoning domains. Complex span tasks were not parceled, so that overall there were 8 measurements. Comparisons among the models with measurement parcels were consistent with the original models. See Table 7 for model fits, and Figure 4 for SEM estimates of the model with both CS and VA as regressors.

Table 7: Fit indices for model of intelligence with measure parcels

	model	chisq	df	pvalue	rmsea	srmr	nnfi	cfi	aic
1	sem.va	32.34	17.00	0.01	0.12	0.08	0.87	0.92	1365.10
2	sem.cs	24.95	17.00	0.10	0.08	0.07	0.93	0.96	1357.71
3	sem.both	22.63	16.00	0.12	0.08	0.07	0.94	0.97	1357.39

Note. rmsea = root mean square error of approximation; srmr = standardised root mean square residual; nnfi = non-normed fit index; cfi = comparative fit index; aic = Akaike's information criterion.

Figure 4

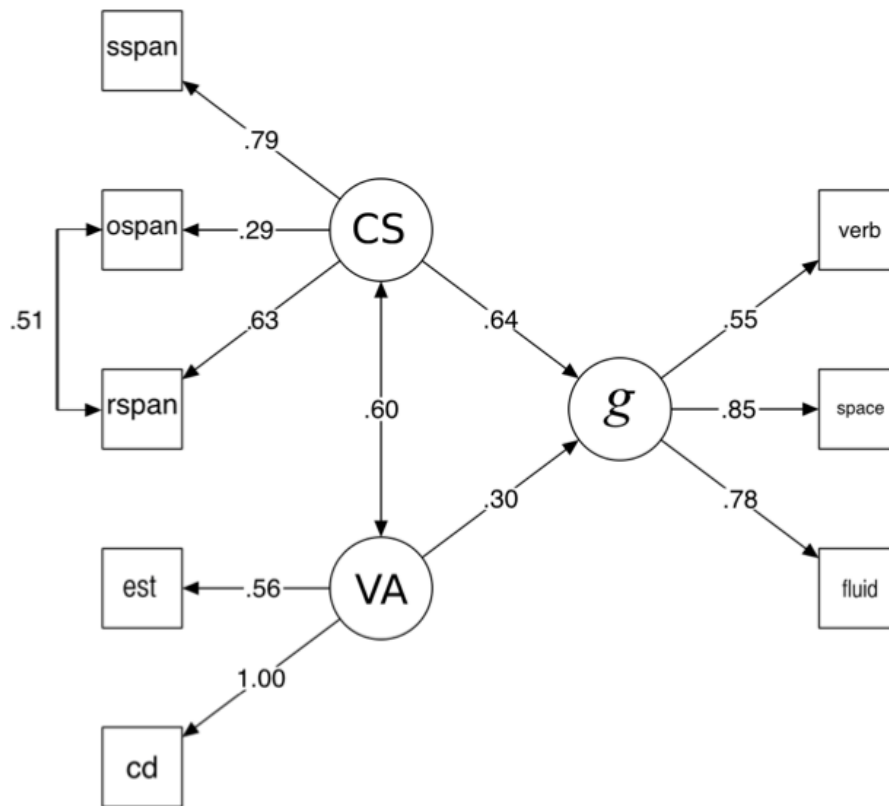


Figure 4. Full model using measure parcels. *est* = color estimation; *cd* = change detection; *rspan* = reading span; *sspan* = symmetry span; *ospan* = operation span; *ravens* = Raven's Advanced progressive matrices; *cft* = Cattell's culture-fair test; *spacerel* = space relations; *analogy* = analogy battery; *readcomp* = reading comprehension.

Discussion

Relationship between VA and CS

The current results provide further support for the argument that visual array tasks and complex span tasks tap distinct but related constructs, replicating and extending a recent studies by Shipstead et al. (2012) and Unsworth et al. (2014). Critically, the use of visual array tasks that are common in research, as well as a broad set of intelligence measures, provides convergent evidence for an important distinction between these two task paradigms. Moreover, the latent correlation between VA and CS observed here was intermediate to those obtained by Shipstead et al. (2012) from two separate data sets ($r = .64$ compared to $r = .61$ and $r = .69$). Though the visual array tasks used here differed from Shipstead et al. (2012) in several critical ways, and intelligence measures from three subdomains were used, these findings replicate the same pattern of results from their two-factor working memory models.

Moreover, the constructs underlying visual array tasks and complex span tasks each account for unique variance in general intelligence. This finding supports recent arguments that there are multiple sources of variance underlying intelligence (and working memory capacity). This result is inconsistent with unitary source models of general cognitive ability. Several lines of converging evidence in the last decade suggest that unitary source models are implausible and should be replaced by multiple source models (Conway & Kovacs, 2013).

Orthogonality of Number and Resolution

Contrary to previous studies (Awh et al., 2007; Fukuda et al., 2010; Scolari, Vogel, & Awh, 2008), which did not find evidence of correlation between measures of number and resolution, our change detection and color estimation tasks showed moderately strong correlations between resolution and number, though the color estimation task did so only at set

size 6. If accurate, these results go beyond suggesting that the orthogonality of number and resolution may be task specific, and question the validity of the number and resolution constructs, as they were measured in these tasks. Given that the correlations appear selective to different set sizes, it appears that the relationship between number and resolution may not be as straightforward as previous accounts have claimed. While previous findings suggest that resolution and number are uncorrelated in change detection tasks, some degree of caution seems warranted. That number and resolution are correlated in the current change detection task is puzzling, as it was constructed using the stimuli and procedures given by Fukuda et al. (2010). However, there were a few differences between their study and the current procedure, which may help explain the conflicting outcomes.

First, unlike (Fukuda et al., 2010), which averaged Cowan's k scores across set sizes, the current study averaged scores across stimuli, before calculating Cowan's k for each set size. Averaging across set sizes is problematic, in that capacity estimates at set sizes that do not exceed a participant's capacity are necessarily biased downward (Rouder, Morey, Morey, & Cowan, 2011). However, calculating Cowan's k in the manner of (Fukuda et al., 2010) still produced significant correlations between number and resolution measures.

Second, although the stimuli for both versions of the task were the same, the exact instructions differed. Thus, it is possible that a large subset of participants misunderstood task instructions and performed poorly in all aspects of the task. This could induce correlation between all measures. However, the number of capacity estimates that were at or below zero, which indicates performance around chance levels, in this task appears comparable to Fukuda et al. (2010). It is worth noting that in both studies a substantial number of participants fell near or below zero for resolution estimates.

A third possibility, which addresses the large number of low or negative resolution estimates, is that mixing small and large change trials in the same blocks produces capacity estimates which may not correctly control for response bias due to guessing. This is because Cowan's k only takes into account response bias in the case where an object (or relevant feature) is retained in memory or lost completely (Rouder et al., 2011). While information about whether a single feature of an object is present in a probe may be sufficient to detect a change from sample to test, both features of the object are necessary to determine whether there was no change. Thus, while the model takes into account only a single guessing rate, participants may employ a guessing rate based on whether they retained only the probed shape in memory, and a separate guessing rate if they retained neither relevant feature in memory. A participant could strategically reduce overall false alarms at the expense of change detection accuracy for small changes by always indicating “no change” if they only know that a large change has not occurred. This might occur, for example, if the participant found the small change component to be too difficult.

Differential strategy use of this nature could artificially reduce or eliminate correlations between number and resolution, while keeping both reliable across set sizes. Moreover, in reanalyzing the change detection data of (Fukuda et al., 2010), a pattern emerges which seems to be consistent with this explanation. Specifically, residuals for the linear models regressing resolution on number appear U-shaped (see Figure 5), with a cluster of participants who scored near zero for resolution, but high for number. It is possible that high capacity participants enhanced number estimates by sacrificing performance on small change trials. Intriguingly, a similar issue arose in an experiment by (Cokely, Kelley, & Gilchrist, 2006), which related complex span performance to a partial list cueing paradigm. Initially, complex span performance

was found to be uncorrelated with a partial list cueing effect. However, the authors discovered that high complex span participants were engaging in a strategy that negated the anticipated effect, creating U-shaped residuals (see Figure 6). Upon controlling for strategy use, a correlation between complex span and the partial list-cuing paradigm emerged.

Figure 5

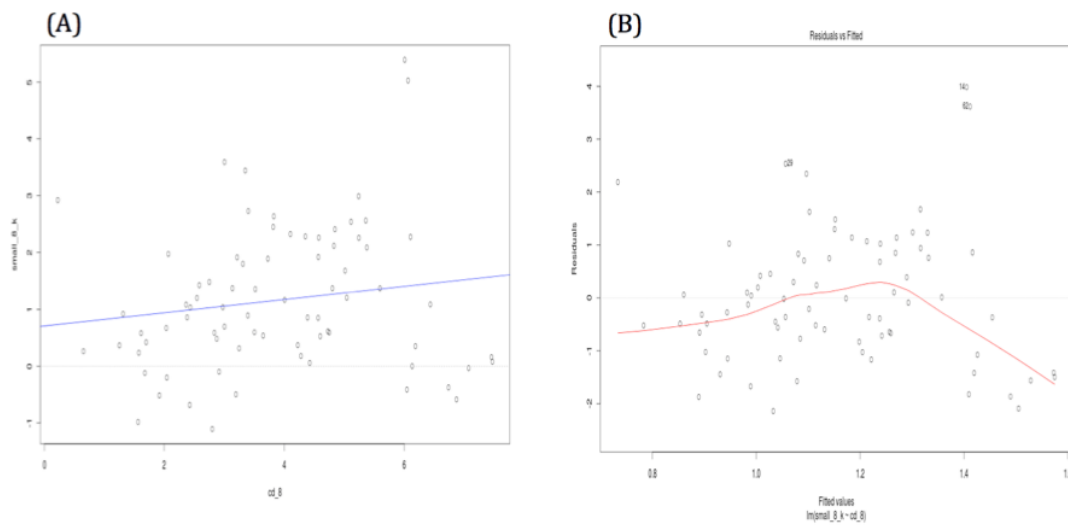


Figure 5. Resolution (small_change.k) regressed on number (cd.8) for set size 8 of a change detection task from Fukuda et al (2010). Panel (A) shows a scatter plot with the best fitting linear model in blue. Panel (B) shows the residuals versus fitted values from the linear model. A clear U-shaped pattern can be observed in the residuals.

However, the finding that resolution estimates in the color estimation task were reliable within task, but correlated neither with number nor resolution in the change detection task raises the possibility that the change detection task used in the current study failed to yield valid measures of resolution. However, this explanation fails to account for the correlation between resolution and number at set size 6 in the color estimation task.

Figure 6

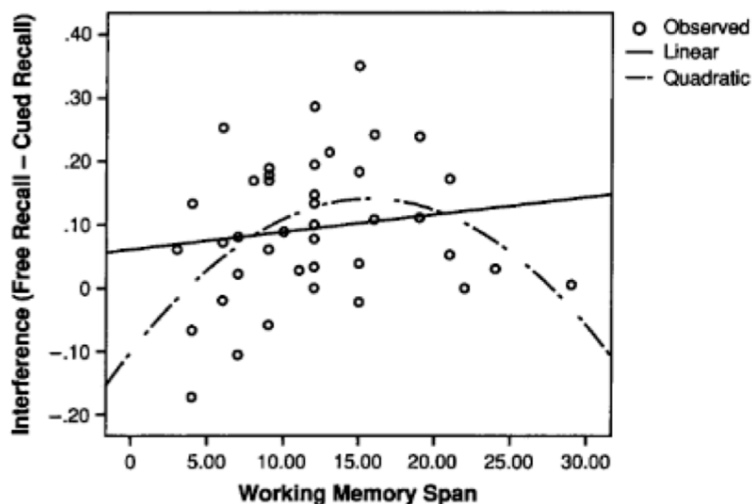


Figure 6. Plot taken from Cokely, Kelley, & Gilchrist (2006) that shows a distinct U-shaped pattern. After questioning participants about their strategy use, they discovered that high performers on complex span tasks had used a strategy which negated the effect of their partial list cueing paradigm. The solid line traces the best fitting linear model, while the dotted line traces the best fitting quadratic model.

Whether the cluster of participants observed in the change detection data of Fukuda et al. (2010) represent a meaningful pattern, or the current study failed to accurately tap into resolution and number estimates is unclear. Further control and replication are necessary to clarify whether resolution and number are correlated. One possibility for controlling the potential contributions of strategy use in the change detection paradigm would be to administer blocks with only one type of change trial in addition to mixed blocks. Blocks with one trial type would allow for an unbiased assessment of resolution and number, as participants would have little incentive to reduce resolution performance. Moreover, performance on these blocks could be compared with mixed blocks, to assess whether participants make strategic trade-offs.

Finally, there is a long-standing debate over whether performance on the change detection and color estimation tasks is better reflected by the slot-based models used in the current study, where items are remembered in an all-or-nothing fashion, or models in which a

continuous memory resource is distributed across items (Bays & Husain, 2008; Cowan & Rouder, 2009; Luck & Vogel, 1997; Rouder et al., 2011; van den Berg, Shin, Chou, George, & Ma, 2012). While previous studies examining the relationship between visual working memory capacity and intelligence capacity (Fukuda et al., 2010; Shipstead et al., 2012; Nash Unsworth et al., 2014) have also assumed the slot-based models are valid, the failure to find support simple latent variable models of resolution across tasks raise the possibility that other accounts may be more useful. This is especially important to note when considering the findings of Unsworth (2014), as they (presumably) use measures where capacity and resolution are often estimated simultaneously using slot-based models (as in Zhang & Luck, 2011), but make no mention of resolution estimates. It would be interesting to see whether they find a common factor underlying resolution estimates across tasks. Another worthwhile avenue to explore would be to explore how these alternative models account for individual differences in visual working memory performance. Evidence in favor of models that distribute a continuous memory resource across items was reviewed recently by Ma, Husain, and Bays (Ma, Husain, & Bays, 2014).

Conclusion

The current study replicates and extends recent findings by (Shipstead et al., 2012), suggesting that the scope and control of attention contribute independent sources of variance to working memory capacity and general intelligence. These findings support multiple source theories of capacity and intelligence and are inconsistent with general ability models (cf., Conway and Kovacs, 2013; Kovacs & Conway, 2014). The current study failed to replicate certain aspects of Fukuda et al. (2010) with respect to the relationship between number and resolution measures derived from visual array tasks. Several possibilities, including differential strategy use across studies, might explain these differences. However, it is clear that more work

is necessary to leverage individual differences in order to assess the validity of number and resolution as conceptualized by slot-based models of visual short-term memory, as well as the claim that resolution and number in the scope of attention are orthogonal.

References

- Ackerman, P. L., Beier, M. E., & Boyle, M. O. (2005). Working memory and intelligence: The same or different constructs? *Psychological Bulletin*, 131(1), 30.
- Awh, E., Barton, B., & Vogel, E. K. (2007). Visual working memory represents a fixed number of items regardless of complexity. *Psychological Science*, 18(7), 622–628.
- Bailey, H., Dunlosky, J., & Kane, M. J. (2011). Contribution of strategy use to performance on complex and simple span tasks. *Memory & Cognition*, 39(3), 447–461.
- Bays, P. M., & Husain, M. (2008). Dynamic shifts of limited working memory resources in human vision. *Science*, 321(5890), 851–854. doi:10.1126/science.1158023
- Chein, J. M., Moore, A. B., & Conway, A. R. A. (2011). Domain-general mechanisms of complex working memory span. *NeuroImage*, 54(1), 550–559.
- Cokely, E. T., Kelley, C. M., & Gilchrist, A. L. (2006). Sources of individual differences in working memory: Contributions of strategy to capacity. *Psychonomic Bulletin & Review*, 13(6), 991–997.
- Cowan, N. (1999). An embedded-processes model of working memory. *Models of Working Memory: Mechanisms of Active Maintenance and Executive Control*, 62–101.
- Cowan, N. (2001). The magical number 4 in short-term memory: A reconsideration of mental storage capacity. *Behavioral and Brain Sciences*, 24(1), 87–185.
- Cowan, N., Elliott, E. M., Sauls, J. S., Morey, C. C., Mattox, S., Hismjatullina, A., & Conway, A. R. . (2005). On the capacity of attention: Its estimation and its role in

- working memory and cognitive aptitudes. *Cognitive Psychology*, 51(1), 42–100.
- Cowan, N., & Rouder, J. N. (2009). Comment on “Dynamic Shifts of Limited Working Memory Resources in Human Vision.” *Science*, 323(5916), 877c.
- Cowan, N., Saults, J. S., & Morey, C. C. (2006). Development of working memory for verbal-spatial associations. *Journal of Memory and Language*, 55(2), 274–289.
- Daneman, M., & Carpenter, P. A. (1980). Individual differences in working memory and reading. *Journal of Verbal Learning and Verbal Behavior*, 19(4), 450–466.
- Ester, E. F., Anderson, D. E., Serences, J. T., & Awh, E. (2013). A neural measure of precision in visual working memory. *Journal of Cognitive Neuroscience*, 25(5), 754–761. doi:10.1162/jocn_a_00357
- Faraco, C. C., Unsworth, N., Langley, J., Terry, D., Li, K., Zhang, D., ... Miller, L. S. (2011). Complex span tasks and hippocampal recruitment during working memory. *NeuroImage*, 55(2), 773–787.
- Fukuda, K., Vogel, E., Mayr, U., & Awh, E. (2010). Quantity, not quality: The relationship between fluid intelligence and working memory capacity. *Psychonomic Bulletin & Review*, 17(5), 673–679.
- Heitz, R. P., & Engle, R. W. (2007). Focusing the spotlight: Individual differences in visual attention control. *Journal of Experimental Psychology: General*, 136(2), 217.
- Kane, M. J., Bleckley, M. K., Conway, A. R. A., & Engle, R. W. (2001). A controlled-attention view of working-memory capacity. *Journal of Experimental Psychology: General*, 130(2), 169–183.
- Kane, M. J., & Engle, R. W. (2003). Working-memory capacity and the control of

- attention: the contributions of goal neglect, response competition, and task set to Stroop interference. *Journal of Experimental Psychology: General; Journal of Experimental Psychology: General*, 132(1), 47.
- Kane, M. J., Hambrick, D. Z., & Conway, A. R. (2005). Working memory capacity and fluid intelligence are strongly related constructs: comment on Ackerman, Beier, and Boyle (2005). *Psychological Bulletin*, 131(1), 66–71.
- Kane, M. J., Hambrick, D. Z., Tuholski, S. W., Wilhelm, O., Payne, T. W., & Engle, R. W. (2004). The generality of working memory capacity: a latent-variable approach to verbal and visuospatial memory span and reasoning. *Journal of Experimental Psychology: General; Journal of Experimental Psychology: General*, 133(2), 189–217.
- Kyllonen, P. C., & Christal, R. E. (1990). Reasoning ability is (little more than) working-memory capacity?!. *Intelligence*, 14(4), 389 – 433.
doi:[http://dx.doi.org/10.1016/S0160-2896\(05\)80012-1](http://dx.doi.org/10.1016/S0160-2896(05)80012-1)
- Lépine, R., Barrouillet, P., & Camos, V. (2005). What makes working memory spans so predictive of high-level cognition? *Psychonomic Bulletin & Review*, 12(1), 165–170.
- Little, T. D., Cunningham, W. A., Shahar, G., & Widaman, K. F. (2002). To parcel or not to parcel: Exploring the question, weighing the merits. *Structural Equation Modeling*, 9(2), 151–173.
- Luck, S. J., & Vogel, E. K. (1997). The capacity of visual working memory for features and conjunctions. *Nature*, 390(6657), 279–280.
- Ma, W. J., Husain, M., & Bays, P. M. (2014). Changing concepts of working memory.

- Nature Neuroscience*, 17(3), 347–356. doi:10.1038/nn.3655
- McClelland, J. L., McNaughton, B., & O'Reilly, R. C. (1995). Why there are complementary learning systems in the hippocampus and neocortex: Insights from the successes and failures of connectionist models of learning and memory. *Psychological Review*, 102(3), 419–457.
- O'Reilly, R. C., Bhattacharyya, R., Howard, M. D., & Ketz, N. (2011). Complementary learning systems. *Cognitive Science*, 1, 1–20. doi:10.1111/j.1551-6709.2011.01214.x
- Pashler, H. (1988). Familiarity and visual change detection. *Attention, Perception, & Psychophysics*, 44(4), 369–378.
- Postle, B. R. (2007). Activated long-term memory? The bases of representation in working memory. *The Cognitive Neuroscience of Working Memory*, 333–349.
- Redick, T. S., & Engle, R. W. (2006). Working memory capacity and attention network test performance. *Applied Cognitive Psychology*, 20(5), 713–721.
- Rouder, J. N., Morey, R. D., Morey, C. C., & Cowan, N. (2011). How to measure working memory capacity in the change detection paradigm. *Psychonomic Bulletin & Review*, 18(2), 324–330.
- Salthouse, T. A., & Pink, J. E. (2008). Why is working memory related to fluid intelligence? *Psychonomic Bulletin & Review*, 15(2), 364–371.
- Scolari, M., Vogel, E. K., & Awh, E. (2008). Perceptual expertise enhances the resolution but not the number of representations in working memory. *Psychonomic Bulletin & Review*, 15(1), 215–222. doi:10.3758/PBR.15.1.215
- Shipstead, Z., Redick, T. S., Hicks, K. L., & Engle, R. W. (2012). The scope and control

- of attention as separate aspects of working memory. *Memory*, 20(6), 608–628.
- Spillers, G. J., & Unsworth, N. (2011). Variation in working memory capacity and temporal–contextual retrieval from episodic memory.
- Unsworth, N., & Engle, R. W. (2006a). A temporal-contextual retrieval account of complex span: An analysis of errors. *Journal of Memory and Language*, 54(3), 346–362.
- Unsworth, N., & Engle, R. W. (2006b). Simple and complex memory spans and their relation to fluid abilities: Evidence from list-length effects. *Journal of Memory and Language*, 54(1), 68–80.
- 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(6), 1038.
- Unsworth, N., Fukuda, K., Awh, E., & Vogel, E. K. (2014). Working memory and fluid intelligence: Capacity, attention control, and secondary memory retrieval. *Cognitive Psychology*, 71, 1–26. doi:10.1016/j.cogpsych.2014.01.003
- Van den Berg, R., Shin, H., Chou, W.-C., George, R., & Ma, W. J. (2012). Variability in encoding precision accounts for visual short-term memory limitations. *Proceedings of the National Academy of Sciences*, 109(22), 8780–8785. doi:10.1073/pnas.1117465109
- Wilken, P., & Ma, W. J. (2004). A detection theory account of change detection. *Journal of Vision*, 4(12), 11–11. doi:10.1167/4.12.11
- Xu, Y., & Chun, M. M. (2005). Dissociable neural mechanisms supporting visual short-term memory for objects. *Nature*, 440(7080), 91–95. doi:10.1038/nature04262

Zhang, W., & Luck, S. J. (2008). Discrete fixed-resolution representations in visual working memory. *Nature*, 453(7192), 233–235. doi:10.1038/nature06860

Zhang, W., & Luck, S. J. (2011). The number and quality of representations in working memory. *Psychological Science*, 22(11), 1434–1441.
doi:10.1177/0956797611417006

Section II – Decoding Three Categories of Conventional Wisdom in PSA

Decoding Three Categories of Conventional Wisdom used in Pattern Similarity Analysis

This section written in collaboration with Janice Chen, Chris Baldassano, and Uri Hasson

Overview of Pattern Similarity Analyses

The use of covariance-based methods is foundational to many analysis regimes in functional neuroimaging, including functional connectivity (Greicius, Krasnow, Reiss, & Menon, 2003; van den Heuvel & Hulshoff Pol, 2010), pattern similarity analysis (Haxby, 2001; Kriegeskorte, 2008), and inter-subject correlation (Hasson, 2004). These areas focus on how information or transformational processes may be represented by the spatial and/or temporal structure of neural activity. The insightful use of these methods has led to meaningful inferences on network dynamics across brain regions (Greicius et al., 2003), the distributed nature of representation (Haxby, 2001), and the shared temporal dynamics of communication (Stephens, Silbert, & Hasson, 2010).

While these analyses have much to offer, they have in some cases been hindered by a mismatch between the appropriate inferences a given analysis allows, and those made by researchers. For example, recently in the area of functional connectivity, a common step called Global Signal Regression was found to constrain correlation matrices in a way that was unanticipated by many researchers (Murphy, Birn, Handwerker, Jones, & Bandettini, 2009). This resulted in many simulation studies and model arguments that questioned basic findings in the field, and offered corrections to existing methods. One such study by Saad et al. (2012) used a simple set of linear models to simulate data with a known structure, and demonstrated that under basic conditions, the use of global signal regression distorts that underlying structure and subsequent analyses. Such demonstrations are important, because they use explicit definitions of structure and representation to evaluate practices and beliefs that may have been justified using verbal arguments.

A special case of Global Signal Regression—sometimes called cocktail blank normalization—was recently criticized as a common preprocessing step in pattern similarity analysis (Diedrichsen, Ridgway, Friston, & Wiestler, 2011; Garrido, Vaziri-Pashkam, Nakayama, & Wilmer, 2013; Walther et al., 2015). By emphasizing how to measure and test the structures that give rise to covariance matrices, researchers in neuroscience provided clear explanations of when such an approach is inappropriate, as well as providing reasonable alternatives (see Diedrichsen et al., 2011). Moreover, while various areas of neuroscience differ in the expected spatiotemporal structure of the data, many of the representations and approaches taken within functional connectivity, pattern similarity analysis, and inter-subject correlation differ in name but all share an underlying interest in a specific covariance matrix. For example, hypothetical experiments and the measures of interest for each area may be as follows (Fig 0-1).

- **Functional connectivity** – relationship of temporal activation across two groups of voxels or ROIs, within a subject.
- **Pattern Similarity Analysis** – relationship of spatial activation (measured for voxels within a ROI) across two groups of trial conditions, within a subject.
- **Inter-subject Correlation** – relationship of temporal activation across two groups of subjects, within a voxel or ROI.

In essence, the names of these measures are loosely based on the dimensions of interest (spatial, temporal, subject). Another set of names for them might be inter-voxel/roi, inter-trial, and inter-subject covariance analysis, respectively (see Figure 0-1). This characterization of the three approaches is overly simple, but will be an essential perspective of this paper. The advancement of this perspective in behavioral psychology has resulted in a simple, but far reaching discussion about the challenges common to diverse research areas. Thus, insightful

methodological critiques have identified issues that span the analyses of covariance between raters, test items, surveys over multiple occasions, and so on (Bollen & Lennox, 1991; Borsboom, 2006; Mellenbergh, 1994; Sijtsma, 2009b). For example, from this perspective, the problem of global signal regression and cocktail blank normalization have also been addressed within the broader psychology literature under the names acquiescence bias (Cronbach, 1942; Savalei & Falk, 2014) and random intercept factor analysis (Maydeu-Olivares & Coffman, 2006).

Regardless of the suffix applied to “inter-“, many simple approaches to covariance analysis—such as averaging correlations, or correlating averages—and their conventional interpretations have been the subject of criticism in several fields, such as ecology (Pollet, Stulp, Henzi, & Barrett, 2015), politics (Bollen & Lennox, 1991), and psychology (Monin & Oppenheimer, 2005). Many problematic interpretations result because these procedures are done without using an explicit model to explain the rationale and assumptions behind such actions (Bollen & Lennox, 1991).

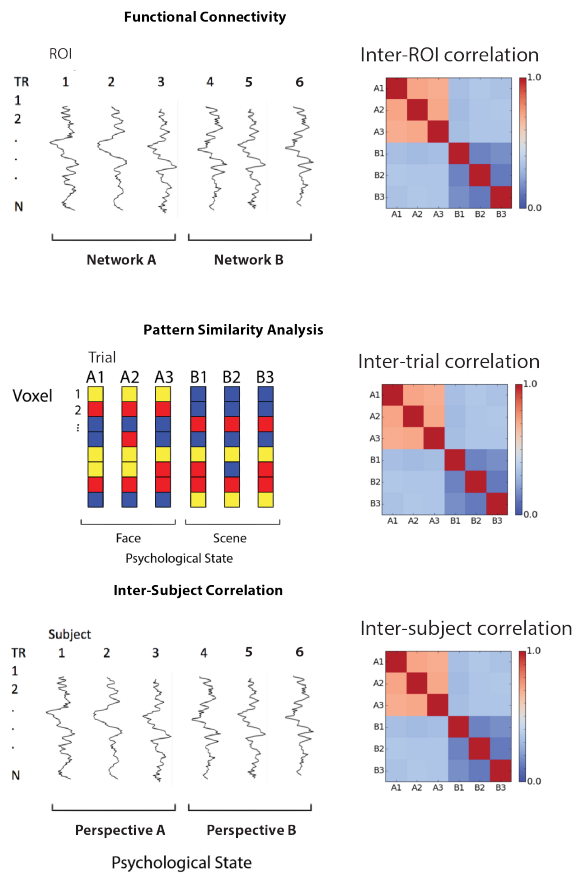
In this paper, we will make the case that in fMRI pattern similarity analyses, common operations conducted over covariance matrices have attempted to measure two properties behind groups of observed patterns: consistency and similarity. Whereas “consistency” has often been measured by the strength of covariances within groups (e.g. Aly & Turk-Browne, 2015; Hasson, 2004; Jordan, Greene, Beck, & Fei-Fei, 2015; Lahnakoski et al., 2014; LaRocque et al., 2013; Xue et al., 2010, experiments 1 and 2), “similarity” has often been measured by the strength of covariances across groups (e.g. Aly & Turk-Browne, 2015; Hasson et al., 2009; Haxby, 2001; Jordan et al., 2015; Regev, Honey, Simony, & Hasson, 2013; Schapiro, Rogers, Cordova, Turk-

Browne, & Botvinick, 2013; Xue et al., 2010, experiment 3). In addition, three common strategies have emerged for making calculations within and across groups:

1. Comparing the average of one block of a correlation matrix to another (inter-subject correlation, Hasson et al., 2009; pattern similarity analysis, Xue et al., 2010).
2. averaging items within a group as a stand-in for a shared group pattern, before performing correlations against this average (inter-subject correlation: Hasson, Yang, Vallines, Heeger, & Rubin, 2008; Honey, Thompson, Lerner, & Hasson, 2012; pattern similarity analysis: Looser, Guntupalli, & Wheatley, 2013; Rice, Watson, Hartley, & Andrews, 2014).
3. splitting each group of items into half (or into k partitions), and performing correlations across them (Haxby, 2001; Nili et al., 2014; Walther et al., 2015).

In the remainder of the introduction, we review the use of the three above strategies in turn, before considering how simple models, such as those used to examine erroneous conclusions regarding covariances (e.g., Simpson's paradox), can illuminate possible weaknesses or rigid assumptions behind these strategies. The remainder of the paper will use a simple latent factor model to robustly characterize consistency and similarity.

Figure 0-1. Covariance analysis across domains



Hypothetical experiments using functional connectivity, pattern similarity analysis, and inter-subject correlation. Each experiment involves rows of observations, columns of “items” that have group labels, and might focus analyses on the “item-item” covariance matrix.

Three Common Approaches to Covariance Analysis

As with functional connectivity, research employing pattern similarity analyses (Haxby, 2001, 2012) and inter-subject correlation (Hasson, 2004; Jääskeläinen et al., 2008) often make heavy use of correlation matrices. In both of these approaches, it is common practice to segment a correlation matrix in blocks, and then perform contrasts between different blocks. To preface, two common block contrasts have emerged to measure the two properties: consistency and similarity. Consistency has been calculated as the mean of within-group correlations, and has sometimes been referred to as coherence (Iordan et al., 2015), stability (Aly & Turk-Browne, 2015), or reliability (Hasson, 2004). Similarity, or rather a lack thereof, sometimes referred to as dissimilarity or distinctness, has been calculated as the difference between averaged within- and between-group correlations (e.g. Haxby, 2001; Kriegeskorte et al., 2008).

In addition, researchers have opted to perform various averaging procedures, such as correlating each item against the average of the others within its group, or calculating correlations between items in one group with averages across items in another group. While these procedures are sometimes regarded as alternatives to the above contrasts, in testing hypotheses regarding consistency or similarity, it is not clear when they may produce differing results. Finally, a variety of splitting procedures have been suggested, in which items are split into halves (or k partitions) within each group, and correlations are done across halves. Below, we detail these three common practices, and their attributed relationships to either consistency or similarity.

1. Within and Between Contrasts, Template Regression. As an example from pattern similarity analysis, suppose that a study is conducted in which participants view either faces or

scenes (see Fig 0-1). In this study, researchers might define “consistency” as the average strength of within-category correlations between neural patterns present during individual face or scene trials. Interpretations for this type of measure are that it indicates that “information content is similar within that particular category” (Jordan et al., 2015, p. 1428), or that neural states are stable when voxel patterns are similar within category (Aly & Turk-Browne, 2015; Ezzyat & Davachi, 2014). By this reasoning, contrasting the average of correlations within one category (e.g. face vs face) against the average of correlations within another (e.g. scene vs scene) tests whether one set of condition evoked pattern is more cohesive, contains more similar information content, or is more stable than another (Fig 0-2A). This type of contrast appears commonly in studies of subsequent memory (e.g. Ezzyat & Davachi, 2014; Karlsson Wirebring et al., 2015; Wolosin, Zeithamova, & Preston, 2013; Wolosin et al., 2013; Xue et al., 2010). For example, Xue et al. (2010) calculate the correlations between voxel patterns across stimulus repetitions, and find that subsequently remembered stimuli had higher mean correlations across repetitions. They conclude that “successful episodic memory encoding occurs when the same neural representations are more precisely reactivated across study episodes, rather than when patterns of activation are more variable across time” (Xue et al., 2010, abstract).

Additionally, researchers might contrast the average of correlations within each category (e.g. face vs face; scene vs scene) against the average of correlations between categories (e.g. face vs scene) as a measure of the “distinctness” of faces from scenes (Fig 0-2B). An interpretation for this procedure was given by Haxby (2012, p. 853):

The idea was straightforward and based on a concept from conventional statistics, namely split-sample cross-validation. If a given stimulus category evoked a **distinct pattern of**

activity, then independent observations of the response to that category should be more similar to each other than to responses to different categories. [emph. added]

Here, it should be noted that distinctness can be interpreted in two ways. First, distinctness may be used to describe the degree to which two patterns are discriminable in any way (e.g. different variances, means, or consistency). This definition is more in line with machine learning and prediction oriented approaches (see Aho, Derryberry, & Peterson, 2014; Breiman, 2001).

Alternatively, distinctness may refer to the degree to which patterns may differ in a specific manner (e.g. reflect some characteristic of interest without regard to “consistency” as described above). While many studies have focused solely on distinctness in the first sense, it seems likely that, when researchers conduct separate analyses of consistency and distinctness (e.g. Aly & Turk-Browne, 2015; Iordan et al., 2015; Xue et al., 2010), there is an implicit assumption that they are measuring these properties in a separable manner.

In inter-subject correlation, similar tests are performed within a voxel, using temporal patterns of activity across subjects (Jääskeläinen et al., 2008; Lahnakoski et al., 2014; Nummenmaa, Saarimäki, et al., 2014; Nummenmaa, Smirnov, et al., 2014). For example, Lahnakoski (2014) had subjects view movie clips either from one of two perspectives, either that of a detective or decorator, which they used to group the subjects. They then contrasted the average correlations within one group against the average correlations within the other, as above². However, for examining distinctness, rather than contrasting the average of within-group correlations with the average of between group correlations, they opted to perform a second-order correlation between the unique, non-diagonal elements of an observed correlation matrix,

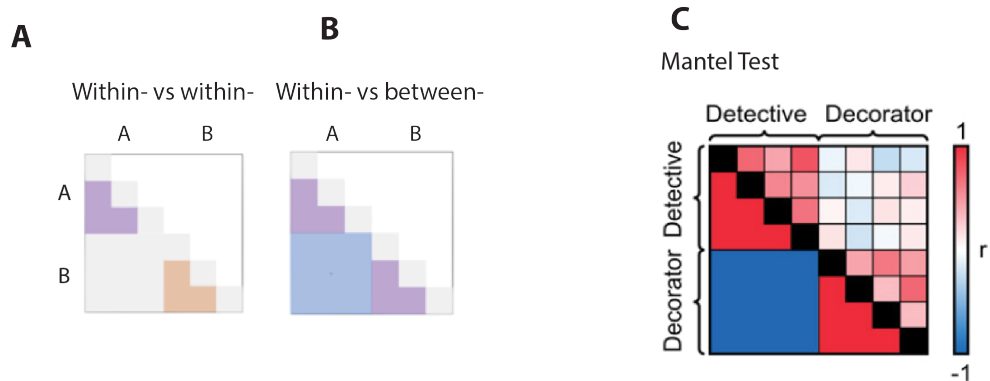
² Note that, in this case, their null hypothesis significance was modified, to take into account that all participants participated in both conditions.

and the corresponding elements of a “template matrix” (Fig 0-2C). They interpret a non-zero result of this procedure as follows:

...viewing the movie from a similar perspective increases the ISC [inter-subject correlation] of brain activity compared with watching it from different perspectives (i.e. the activity is **consistent within a perspective but different across perspectives**). (p. 319)

As we will argue below, there is a need to account for the differences in the consistency within a group, to properly compare distinctness across groups. Such an issue may seem trivial, but causes serious issues even for simple tests. One such case occurs with the standard t-test, used to examine whether two populations have different means, which returns misleading results for two populations with unequal variances. In such a case, explicit assumptions about the population variances is necessary in order to make an inference about means (for more robust alternatives, see Kruschke, 2013; Ruxton, 2006). Moreover, in the same sense that a t-test may be represented as a simple linear model with a “template” that consists of 0’s for the first group, and 1’s for the other, the template correlation procedure used by Lahnakoski (2014) is simply an alternative way of representing the within- vs between-group contrast common to pattern similarity analysis.

Figure 0-2. Comparing within- and between-group correlations



(A) Within- vs within-group contrast of pairwise correlations. (B) Within- vs between-group contrast of pairwise correlations. The average correlation within each group is first calculated, and then the overall average within-group is contrasted with the average between-groups. (C) The Mantel Test performs a similar contrast to the within- vs between-group contrast by forming a vector of all unique pairwise correlations (excluding the diagonal), and then correlating it against a vector with ones corresponding to within-group correlations, and zeros corresponding to between-group correlations.

2. Averaging Across Items as a Stand-in for Underlying Pattern. One variation on the averaging of within- and between-group correlations, found commonly in inter-subject correlation studies, first averages across items within a group to calculate a single, composite pattern / timecourse. For example, Hasson and colleagues often use the term inter-subject correlation to denote correlating each item (i.e. a subject's timecourse) against the average of the others within a group (e.g. Ames, Honey, Chow, Todorov, & Hasson, 2015; Ben-Yakov, Honey, Lerner, & Hasson, 2012; J. Chen et al., 2015; Hasson et al., 2009, 2008; Honey et al., 2012; Regev et al., 2013). In this sense, their measure of consistency for a group is the average one-against-other correlation across all subjects in that group. The intuitive appeal and potential usefulness behind such a procedure is given by Hasson et al., (2009), who found that while viewing a movie, pairwise correlations within a group of autistic subjects were lower (i.e. responses were less consistent across those subjects), and observed that “aggregating the signal across all autistic individuals averaged out some of the uncorrelated and idiosyncratic fluctuations, and revealed a more typical response in all cortical regions that is similar to the signal seen in typical subjects”.

From this perspective, the average timecourse across subjects may serve as an approximation of an unobserved, or latent timecourse that is common to a group (e.g. the non-idiosyncratic fluctuations). Accordingly, the procedure of correlating each subject against this approximation may be viewed as an attempt to regress (or correlate) each subject's timecourse against this latent timecourse. While using the average timecourse from other subjects as a proxy for latent underlying model which driving the responses in each subject has proven valuable, it also has some limitations. For example, using such an average as a stand-in for regression

procedures leads to under- and over-estimation of coefficients, due to unaccounted random error in its estimation (see Bollen & Lennox, 1991).

In this paper, we propose the latent factor models as an effective framework to model inter-subject, inter-trial, and inter-voxel covariance matrices. The use of latent factor models allows us to explicitly define and test assumptions as to the underlying latent responses, to account for different sources of variance within and across groups, as well as providing a platform for testing competitive models hypothesis.

For example, one clear advantage of the latent factor model is the ability to correct for bias due to unaccounted random error that occurs when estimating a latent timecourse via averaging (Bollen & Lennox, 1991). Prior studies used inter-subject correlation, in the sense of correlating against an average, to try to discriminate both the consistency and dissimilarities/similarities (inter-subject correlation, Ames et al., 2015; J. Chen et al., 2015; Hasson et al., 2009; Honey et al., 2012; Lerner, Honey, Katkov, & Hasson, 2014; Regev et al., 2013; pattern similarity analysis, Looser et al., 2013). However, as will be shown below, using these methods may under-specify what is being measured, as they don't explicitly test for differences in variance within and between groups as opposed to differences in the latent factor across the groups. In this work, we show how latent factor models can help us for testing for the source of differences across groups.

3. Splitting Items into Halves Within Group. The latent factor models can also account for additional nuisance variables in the data. Oftentimes, factors that may not be of interest for a particular experiment--such as the temporal proximity of trials in pattern similarity analysis, or using the same subject for multiple conditions in inter-subject correlation--influence observed correlations. For example, voxel patterns obtained from trials close in time may increase the

correlations between those trials, due to the temporal dynamics of BOLD response. As a result, researchers may want to omit pairs of correlations that are affected by these nuisance factors. For example, Haxby (2001)—reflecting on the procedure in a later article—stated that he “made independent observations by dividing the [trials] for each subject into two halves — even-numbered and odd-numbered runs” (Haxby, 2012, p. 853). Here, it is important to note that it is unlikely Haxby is referring to independence in the statistical sense, as that would imply trials have no expected correlation across halves³. Rather, he conveying that a specific source of dependence, such as that caused by temporal proximity, was removed by only correlating across halves (Fig 0-3).

Another example, which may benefit from explicit modelling, comes from Hasson et al. (2009), who showed autistic subjects two viewings of the same movie. In this work, Hasson et al. (2009), argued that differences in the response patterns between the autistic and neurotypical groups while they watch a movie arise from differences in response reliability (i.e. variance) across groups and not from difference in the latent response pattern across the two groups. They concluded this after finding a reduction in reliability of responses within the autistic groups, while they obtained strong correlation of the average responses across subjects.

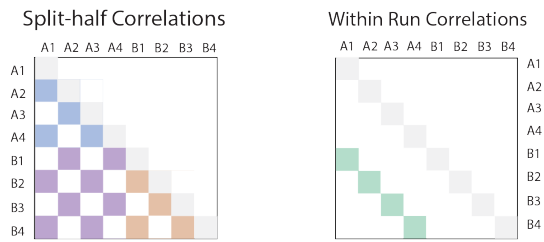
While such splitting procedures are often reasonable, they operate by removing specific pieces of information (e.g. cells of a correlation matrix), prior to performing an analysis. This approach while producing sensible analyses, may benefit from explicit modelling, which explain when such analyses would be appropriate. In the same sense that modelling approaches have

³ While a true covariance of 0 does not imply that two random variables are independent (e.g. their relationship may be U-shaped), independence implies a true covariance of 0 (see Ross, 2014, p. 124, Theorem 4.7.4). However, random variables that covary may also be independent, when conditioned on other variables (as is often the case with residuals in linear models).

been used to consider global signal regression and cocktail-blank normalization, they may also provide a reasonable context in which to consider various splitting procedures. In behavioral psychology, latent factor models provide a general account for procedures such as split-half reliability, which also includes correlating across splits of the data, and a closely related measure named Cronbach's alpha⁴ (Graham, 2006). Moreover, incorporating different sources of covariance into models has received extensive treatment in this field. For example, researchers who administer a battery in which different traits are measured with each of several methods (e.g. behavioral, questionnaire), may be interested in the relationship between traits, but need to account for method-specific covariance (e.g. all behavioral items are especially correlated; see Maas, Lensvelt-Mulders, & Hox, 2009). Incorporating the structure of different sources of covariance into a model, rather than cutting up the data in order to separate them, allows researchers to test hypotheses in a unified manner.

⁴ Specifically, Cronbach's alpha is equivalent to the average split-half reliability across all possible splits of the data (Sijtsma, 2009b).

Figure 0-3. Splitting procedures and run-specific variance



Covariance matrices corresponding to trials from one of two possible conditions (A or B), and numbered by run (1,2,3,4). The matrix on the left shows the correlations used when splitting the data into odd and even runs, and correlating across splits. Blue and Orange cells are those within the A and B conditions, respectively, while Purple cells are those between conditions. The matrix on the right shows the correlations between trials within the same run. These correlations may be especially well correlated, due to temporal factors.

Linear DAGs, Simpson's Paradox, and Testing for Different Underlying Patterns

Interpreting the meaning of correlations in the context of a shared pattern of activity has a long history, ranging from Spearman's (1904) consideration of what the correlation between two variables should be if they share (to differing degrees) the same underlying cause of variation, to the development of classical test theory in psychology (Novick, 1966), to many careful treatments of Simpson's Paradox (for review, see Kievit, Frankenhuis, Waldorp, & Borsboom, 2013; Pearl, 2014). In these cases, the interpretation for measures or calculations can be viewed from the lens of a generative data model, in order to understand their strengths and limitations under different scenarios (examples in functional connectivity, Murphy et al., 2009; Saad et al., 2012; examples in psychology, Bollen & Lennox, 1991; Borsboom, 2006; Sijtsma, 2009b).

For example, suppose researchers set out to measure the effect of dose on health outcomes for a specific medication (Fig 0-4A). An increased dose may look effective at the population level, but have no effect when analyzed within each gender. The model of Figure 0-4A shows a simple generative model for this scenario. Here, a participant's sex causes an increase in dose and health, such that they are correlated to sex⁵ by .8 and .5, respectively. The error terms e_{dose} and e_{health} refer to unsystematic error (noise) for dose and health, respectively. Each square node in this graph can be represented as a linear function of those pointing to it,

$$\text{dose} = .8 * \text{sex} + e_{\text{dose}}$$

$$\text{health} = .5 * \text{sex} + e_{\text{health}}$$

⁵ Sex here is presented as dichotomous, which cannot have a perfect ($r = 1$) true correlation with a normal distribution, but it does not affect this example.

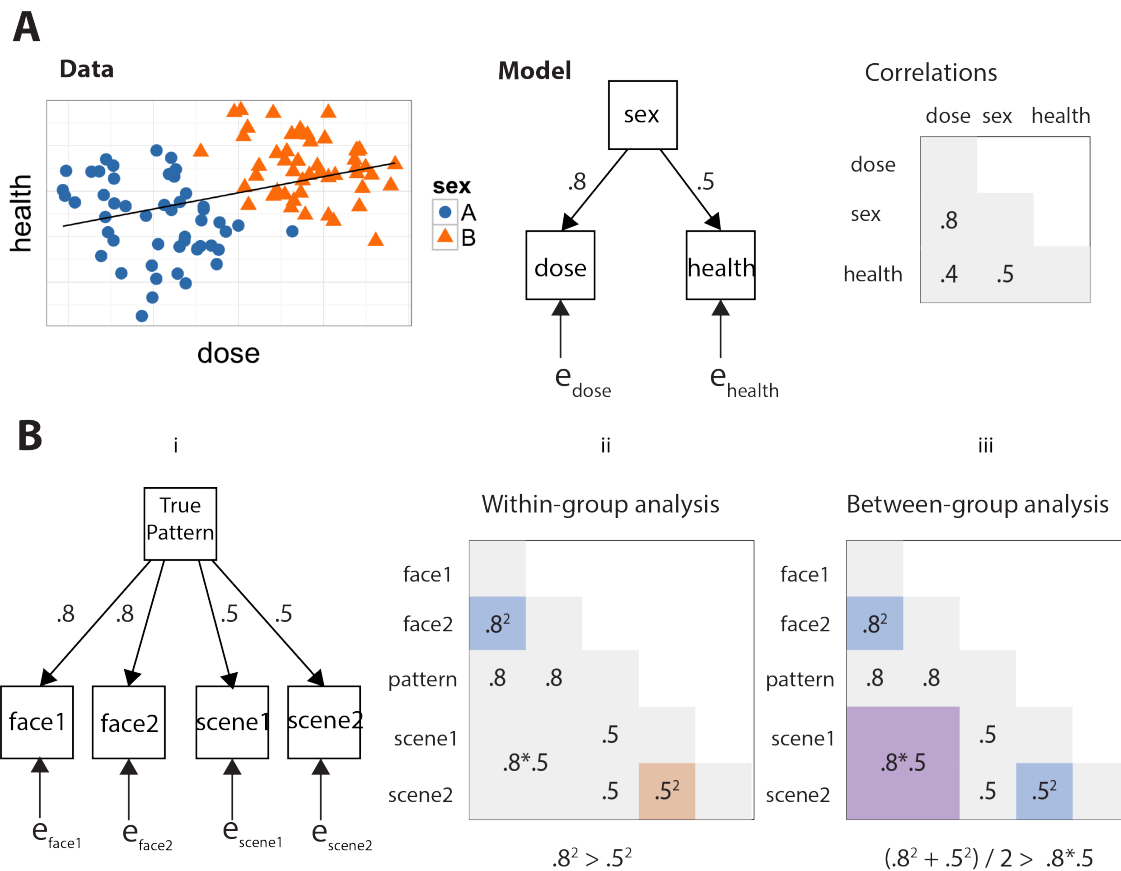
moreover, if all measures are standardized, so that they have a mean of 0 and variance of 1, then the correlation between dose and health is the product of their correlations with sex, $r = .8 * .5 = .4$. At the population level, researchers might see a correlation between dose and health, and wrongfully conclude that increasing dose leads to better health. Worse, the true impact of dose on health may be *negative* and yet still allow a positive correlation at this level of analysis (Simpson's Paradox; see Kievit et al., 2013). Thus, this generative model is useful in illustrating how researchers might see something that interests them (a positive correlation), but not for the reason they might think (dose causes health).

To foreshadow potential difficulties with interpreting the within- vs between-group contrast as indicating whether there are distinct patterns of neural activity between groups, consider the scenario in which the same “true” pattern True-Pattern underlies observed neural patterns elicited by two categories of stimuli, faces and scenes. In addition, suppose individual face trial and scene trial patterns are correlated with True-Pattern at .8 and .5, respectively (Fig 0-4B.i). While Example 1 in this paper will go into greater detail, we consider problems that are apparent by considering the expected correlations these two groups of trials. Accordingly, the two categories have differing levels of *consistency*, reflected by their unequal expected within-group correlations (purple and orange squares in Fig 0-4B.ii). However, although they are not *distinct* (they all share the same underlying pattern True-Pattern, rather than reflecting two distinct underlying patterns), the correlation within the face category is higher than the correlation between the two categories (Fig 0-4B.iii). If this positive within- vs between-group contrast is supposed to indicate that two categories are *distinct*--then swapping in the labels from the previous example, so that the pattern is sex, category A faces are dose, and category B faces are health—we would come to the surprising conclusion that there is a distinct pattern of sex

underlying dose and health. In other words, we would have to conclude from this scenario that some participants were one sex when given the dose, but another sex when health outcomes were measured—a clearly nonsensical inference, given the scenario. What is actually happening in this scenario is that the within- vs between-group contrast is affected *when items in one category are more strongly correlated with the underlying pattern True-Pattern than are items in the other category*. Whether this matters or not depends on what researchers were trying to test. Were they testing for evidence against a single underlying face pattern? Were they testing for evidence against equal strength correlations with (potentially) a single underlying face pattern? Or were they testing distinctness in the anything-to-discriminate-the-two-categories sense?

These problems with conducting the within- vs between-group contrast analysis may be due to issues that arise when dealing with unobserved, or latent variables. That is, we only observe the voxel patterns evoked on each trial, but we may want to know if these trial patterns reflect the influence of category specific patterns. In this paper, we will use a model called *latent factor analysis* which elaborates on the example above to demonstrate that current practices reflect a serious misunderstanding about the relationship between within- and between-group correlations. Along the way we will demonstrate simple ways to remedy some of the more egregious problems. Furthermore, we will use the model to demonstrate that the same problem exists when using averaging procedures across items, or splitting the data into partitions before correlating across partitions. By walking through these problems and the ways in which a latent factor approach might address them, we hope to demonstrate that this model offers a simple but highly extendable solution to these problems.

Figure 0-4. Models and Correlational Structures



(A) From left to right, example data corresponding to the scenario in which sex causes an increase in both dose and health outcome, the graphical depiction specifying the data generating model, the correlation matrix implied by the model's parameters. Note that the variance of all variables was defined to be 1. (B.i) The data generating model for a hypothetical pattern similarity analysis scenario, in which TruePattern generates data for face and scene trials, where face trials are more highly correlated with TruePattern. (B.ii) Model implied correlation matrix with colors corresponding to the two groups used for the within- vs within-group contrast. Underneath, the calculation for the contrast shows that it correctly finds a difference for the two groups. (B.iii) Model implied correlation matrix with colors corresponding to the two groups used for the within- vs between-group contrast. Underneath, the calculation for the contrast shows that, even though the model parameters are known in this scenario, and only a single pattern TruePattern underlies the trials, the within- vs between-contrast is not zero (i.e. is biased as a test of pattern similarity).

Overview of Sections

In this paper we will use a family of models, like the ones shown above, to explicitly operationalize two ways in which underlying patterns may be thought of as distinct between experimental conditions: differences in pattern *consistency* between trials, and *distinct* underlying patterns. Through using these models to evaluate three common beliefs—that distinct patterns of activity lead to stronger within-group correlations than between-group correlations, that averaging sets of items may be used as a stand-in for an underlying pattern, and that parceling the data provides unbiased solutions—it becomes clear that all of these beliefs require qualification. Three examples will be presented in detail. Below, we summarize how each example evaluates one of these beliefs.

- Example 1. The individual calculations in the within- vs between-group contrast may be viewed as estimations of consistency and pattern similarity, and using this contrast as a test of similarity relies on the assumption that the consistencies for all groups are equal. With minor modification this contrast can be made robust to unequal consistency across groups (the case in the pattern example).
- Example 2. A number of procedures not only compare the correlations within and across groups of items, but sum or average sets of items beforehand. This practice may be viewed as an attempt to estimate an underlying pattern for a group. The limitations discussed in Example 1 apply to these procedures, and they introduce additional issues due to unaccounted measurement error in estimating underlying group patterns.
- Example 3. The practice of splitting each group of items into partitions, prior to calculating correlations across partitions, is done to remove unwanted sources of variances, such as increased correlation between items within the same scanning run.

Latent factor models illustrate how to accommodate these idiosyncratic correlations without selectively discarding correlations between items.

Methods

Software Tools

Analyses and simulation were run using R. The libraries dplyr (Wickham & Francois, n.d.) and reshape (Wickham, 2007) were used for data processing and transformation. The library lavaan (Rosseel, 2012) was used to specify, simulate, and fit latent factor models. Dependency management was handled using packrat and docker (see Boettiger, 2015).

Restricted Factor Model

The restricted factor model (RFM; Fig 0-5) consists of two levels: one relating items within a group, and one relating groups to each other. At the first level, latent variables are used to represent shared voxel patterns or temporal activity. We denote each of the latent variables as η_g . They are allowed to covary with one another. We denote the covariance between a given pair of latent variables, i and j , with

$$\text{Corr}(\eta_i, \eta_j) = \rho_{i,j}$$

which means that for N groups of items, there will be $N*(N-1)/2$ pairs of potential correlations of interest at this level, one for each unique pair of groups. In the case of voxel-wise pattern similarity analysis, each latent pattern may be taken to correspond to a different trial condition, such as the category from which stimuli are drawn (e.g. faces or scenes).

At the second level, items are assumed to belong to only one group. For example, each item may correspond to a specific trial. Items within each group, g , are represented as follows:

$$y_p = \lambda_{g,p}\eta_g + \epsilon_p$$

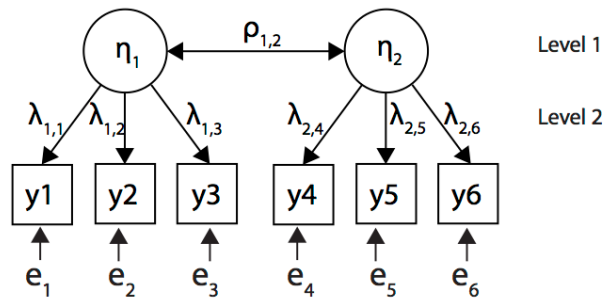
where y_p is activation measurement of item p , η_g is its corresponding latent group pattern, ϵ_p is the measurement error for item p , and $\lambda_{p,g}$ is the effect of the latent signal η_g on item p .

Importantly, we assume that $\text{Corr}(\epsilon_p, \eta_g) = 0$, that $\text{Corr}(\epsilon_i, \epsilon_j) = 0$ when $i \neq j$, and $E[\epsilon_i] = 0$.

That is, all errors are independent of the latent signal, each other, and have a mean of 0. The form of this model is a simple linear equation, similar in form to those used in the previous section, with the only difference being that the value of η_g at each observation (e.g. voxel or time-point) is now unknown. For example, suppose we are considering activation at the first voxel in an experiment. If $\lambda_{p,g} = .5$, ϵ_p is drawn from a standard normal distribution, and $\eta_g = 1$, then this model predicts that y_p will be .5, plus some degree of error. Importantly, even without knowing η_g , we may still examine the correlational structure of the model. Finally, for simplicity and ease of interpretation we standardize all items (y_p) and latent variables to have a mean of 0 and variance of 1. Standardizing these variables allows the convenient property that the factor loadings (λ) are the correlation between a measure y_p and its corresponding latent pattern (Bollen & Lennox, 1991). See Table 0-1 for a summary of model parameters.

This model highlights two ways that groups may differ, each corresponding to one level of the model. On the first level, group factors may be dissimilar to one another (i.e. their correlation may be less than 1), or they may be identical (in which case, their correlation would be 1). On the second level, items within one group may have higher correlations with their underlying group factor, when compared to items within another group. In the next section, we consider the between group analysis from these two perspectives.

Figure 0-5. Restricted Factor Model



Path diagram of the Restricted Factor Model, in which each latent factor loads on only one group, and factors may be correlated with one another.

Table 0-1. Restricted Factor Model components.

<u>Component</u>	<u>Interpretation</u>
η_g	Latent variable pattern for group g
y_p	Observed measure for item p, also written $y_{g,p}$ to denote its group
e_p	error component for item p
$\lambda_{g,p}$	Correlation of item p with its group's latent variable, g
$\rho_{g,h}$	Correlation between latent variable patterns for groups g and h
Pattern Similarity	The correlation between two underlying patterns (e.g. $\rho_{1,2} = 1$ denotes a perfect correlation between group1 and group 2 patterns)
Item Reliability	The proportion of shared variance between an item and the latent variable pattern for its group (e.g. $\lambda_{g,p}^2$)
Consistency	Summary of item reliabilities within a group. Defined in this paper as mean of item reliabilities.
C_g	Unweighted linear composite for group g (e.g. $C_1 = y_1 + y_2 + y_3 + y_4$; $C_2 = y_5 + y_6 + y_7 + y_8$)
R_{C_g}	Reliability of C_g (i.e. the proportion of shared variance between C_g and η_g)

Example 1: Testing Correlated Factors and the Within- vs Between-group Contrast

While many researchers have offered verbal interpretations and justification for the within- vs between-group contrast, the use of the restricted factor model (RFM) offers an explicit context in which these interpretations and justifications can be evaluated. For the sake of simplicity, this section will use an RFM in which all item reliabilities within a given group are assumed to be equal (Fig 1-1). Using the hypothetical face versus scene experiment as an example, this model has three parameters of interest: the item reliability for face trials (λ_F), the item reliability for scene trials (λ_S), and the correlation between the latent face and scene patterns (ρ). Note that, although in this model the underlying patterns are latent rather than observed, its expected item-item correlation matrix may still be derived (see Bollen 1989; Bollen and Lennox). Correlations within a group are the square of each loading (e.g. λ_F^2), while correlations between groups are the product of all three parameters of interest ($\lambda_F * \lambda_S * \rho$).

Simulation of the restricted factor model can evaluate potential bias in the within- vs between-group contrast. One way to do this is to consider what component of the data generative model each calculation in the contrast is trying to measure. The within- vs between-group contrast calculates two terms: the average of all pair-wise correlations between groups, and the average of all pair-wise correlations within groups. When there are an equal number of items within each group⁶, the average correlation between groups is $\lambda_F * \lambda_S * \rho$, and the average within groups is $(\lambda_F^2 + \lambda_S^2) / 2$. Note that when the average within groups are equal, $\lambda_F^2 = \lambda_S^2 = \lambda_{All}^2$,

⁶ It appears common for researchers to take the average pairwise correlation within each group first, in which case, it does not matter if the groups have equal numbers of items. Often, however, the descriptions are not clear on whether this was done first.

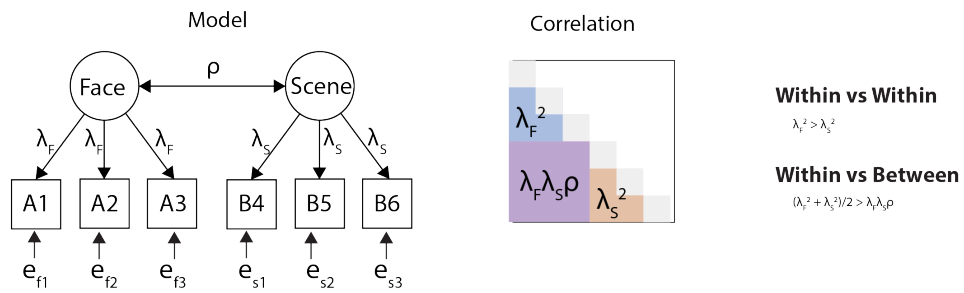
then the within- vs between-group contrast will be positive when $\lambda_{All}^2 * \rho < \lambda_{All}^2$, which is true for $\rho < 1$. In other words, when all item reliabilities are equal across groups, the within- vs between-group contrast is a function of the pattern similarity parameter, ρ . It is expected to be positive for distinct group patterns (e.g. when $\rho < 1$).

However, when the item reliabilities are not equal, so that $\lambda_F^2 \neq \lambda_S^2$, then even when $\rho = 1$, the within- vs between-group contrast is expected to be positive, since

$$\lambda_F * \lambda_S * \rho < \lambda_F * \lambda_S * 1 \leq (\lambda_F^2 + \lambda_S^2) / 2$$

due to the inequality of the arithmetic and geometric mean. Thus, even when there is only one underlying pattern ($\rho = 1$), unequal item reliabilities across groups may lead the between group contrast to be significant. This holds when there are more than two groups (supplement A).

Figure 1-1. Inter-item correlation



Path diagram and model-implied correlation matrix corresponding to a hypothetical face versus scene pattern similarity analysis experiment. The within- vs within-group contrast and within- vs between-group contrast may be expressed in terms of the model-

E1.1. Correcting for Unequal Item Reliability Across Groups

In this sense, the common approach of pooling all within group correlations may be viewed as an estimate of a single item reliability term, under the assumption that all groups have equal item reliabilities. If this assumption is violated, it causes a bias when testing hypotheses about pattern similarity (ρ). However, rather than pooling all within group correlations together, we could take the average within each group separately. Referring to Fig 1-1, it is clear that this will lead to individual estimates of λ_F^2 and λ_S^2 . These can then be used to adjust the between group average in order to estimate ρ . Provided the average within-group correlations are positive, this solution is equivalent to fitting the above restricted factor model using least squares estimation⁷. The corrected procedure would be as follows,

1. Estimate λ_F = square root of the average correlation within Face group
2. Estimate λ_S = square root of the average correlation within Scene group
3. Estimate ρ = average correlation between groups / ($\lambda_F * \lambda_S$)

where the resulting parameters will be an estimate of the restricted factor model. Whereas many analyses have performed NHST on the within- vs between-group contrast, this procedure emphasizes the measurement of quantities of interest directly. Researchers may then perform NHST on the parameters or use a procedure, such as bootstrapping to produce confidence intervals around them. In the following section, we consider how well this approach to correcting

⁷ The model estimates all correlations in each of the three within- or between-group block as a single estimated. Note that the estimate that produces the least squared-error within a block is its mean.

for separate item reliabilities compares to the traditional within- vs between-group contrast, when it comes to testing for distinct group patterns.

E.1.2. Simulations Evaluating Bias in Testing for Distinct Groups ($\rho < 1$)

In order to evaluate the bias in using the within- vs between-group contrast to test against the null hypothesis that group items have distinct underlying patterns, we simulated data using the RFM for the face/scene scenario defined above. For each simulation, the loadings λ were taken from [.25, .3, .45, .65, .85], with 100 observations. The lowest three loadings were calculated from the model-implied correlations used in simulations by Diederichson et al. (2011, p. 1669, “Accessing similarities across conditions”).

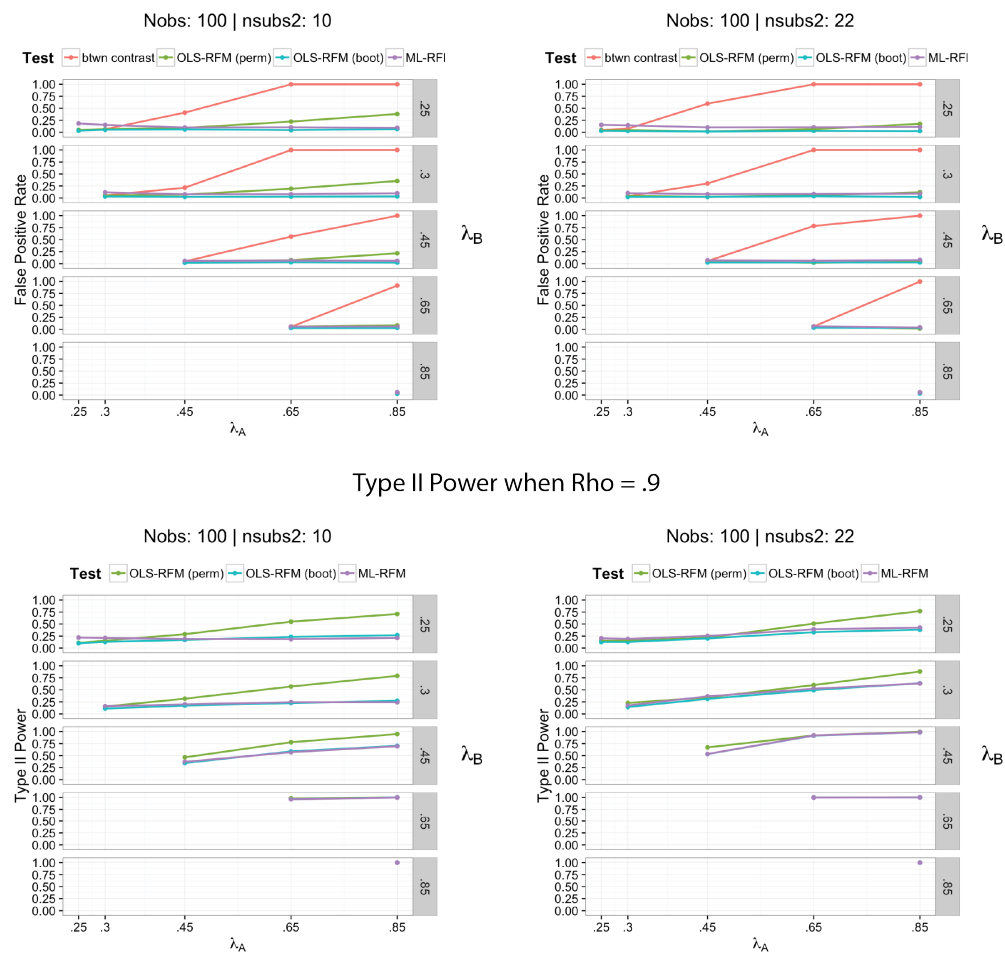
On each simulated data set, we evaluated the traditional within- vs between-group contrast approach, in which a permutation test over item labels is used for NHST. In addition, we compared three more test procedures, in which the RFM is estimated. In the first two procedures (OLS-RFM), the RFM was estimated using the transformation discussed in the previous section, and NHST was conducted either using the same permutation test as for the within- vs between-group contrast or via the non-parametric bootstrap (see Table 1-1). The final procedure estimated the RFM through maximum likelihood, and then used a chi-square difference test to compare a reduced model--which fixed $\rho = 1$ (i.e. same group pattern)--to a model where ρ was a free parameter (also called likelihood-ratio test; see Bollen, 1989, p. 292).

The results are shown in Figure 1-2. The traditional within- vs between-group contrast performed well when item reliabilities were the same between groups, but had high false positive rates when item reliabilities differed between groups. For example, when $\lambda_a = .45$, $\lambda_b = .25$, and $N_{\text{items}} = 10$, the false positive rate is over 35%. While using the OLS-RFM procedure with a permutation had fewer false positives, they were still above the alpha level set in the NHST

($\alpha = .05$) when item reliabilities were unequal. In contrast, the bootstrapped OLS-RFM had reasonable false rejection rates (near .05). One reason for the poor performance of the permutation test compared to the bootstrap, even for the OLS-RFM procedure, is that when item reliabilities differ across groups, permuting group labels creates non-homogeneity of item reliabilities within a group (i.e. items with different levels of reliability may be put into the same group)—a violation of one of the assumptions behind the simple OLS-RFM procedure. In this sense, the process being represented here is that groups share the same underlying pattern *and* the item reliabilities are mixed within groups, which does not correspond to our intended null hypothesis. Finally, the ML-RFM procedure performed more similarly to the OLS-RFM with bootstrap, but in some cases, for example when both loadings were .25, had higher false-positive rates. This occurs, in part, because the chi-square test is an approximate test, that is inaccurate with small numbers of observations (see Lee & Song, 2004).

This power analysis illustrates that the standard within- vs between-group contrast is inadequate to separate evidence for differences in item reliability and evidence of distinct underlying group patterns in pattern similarity analysis. The RFM is beneficial in that it can serve as an explicit representation of belief, which—even if that representation is seen as simplistic—provides a context in which verbal statements of what the contrast is doing become scrutable.

Figure 1-2. Power Analyses for Within vs Between Pattern Similarity



Type II Power when Rho = .9

Power analyses for the between group similarity test. Top: false positive rate when the groups share the same underlying pattern. Bottom: power to detect differences when the groups have distinct patterns (pattern correlation, $r = .9$). *btwn_isc* = the average of within group correlations minus the average of between group correlations, with permuting subject labels. *btwn_lm* = estimating ρ via the linear model, with permuting subject labels. *btwn_lm_boot* = same procedure as *btwn_lm*, except that testing was done by the non-parametric bootstrap of TRs. *cfa* = estimating ρ directly via maximum likelihood on the latent factor model, with significance testing using the chi-square test.

E1.3. Discussion

The RFM confers several advantages in measuring item reliability and group distinctiveness: an account that focuses on measurement, a generative data model, and a rich set of tools for evaluating different models. In the following section we discuss the value of these three points, as well as limitations of the RFM model used here.

First, the model characterizes potential differences between groups as occurring at two levels, the level of item reliability and the level of group patterns. Inadequately accounting for item reliability will affect inferences about the relationship between underlying group patterns. Importantly, the parameters of this model have a straightforward interpretation. Item reliabilities are the correlation of an item with its underlying group pattern, and pairs of latent group patterns are related through their correlation. In the context of the RFM, the practice of averaging all within-group correlations may be viewed as estimating a single item reliability parameter for those group items.

Second, using the RFM as a data generative model suggests what assumptions must be taken in order for certain inferences to follow. For example, claiming that statistical significance for the within- vs between-group contrast is evidence for distinct group patterns requires the unnecessary qualification that item reliabilities are the same across groups. This situation is comparable to the limitations of the classical t-test when used for groups that have unequal variance. The point of bringing up this potential issue is not to claim that researchers have reached incorrect conclusions, but to emphasize that it is not clear which set of conclusions (item reliability or group distinctiveness) they meant to express. To the degree that they were looking for evidence of either outcome, the bias mentioned here is not an issue—but the RFM may serve

as an effective interpretive tool for researchers to be explicit about which outcome(s) they are testing, and what they view their null hypotheses as representing.

Finally, the power analysis demonstrated that either maximum likelihood or least squares may be used to estimate the parameters of the RFM (see Krijnen, 1996; Ximénez, 2009), and that hypothesis tests may be conducted by comparing reduced versions of the model—where parameters are fixed to certain values—to the full version of the model, using a chi-square test. Much research has gone into testing hypotheses using these types of models, as well as frequentist and Bayesian approaches to model fitting. One major advantage of the general RFM is that it can lift the constraint that item reliabilities within a group are equal, whereas other models of pattern similarity require that they be set equal, or their ratios to one another are known beforehand (e.g. item 1 has a factor loading twice as large as item 2; (Allefeld & Haynes, 2014; Diedrichsen et al., 2011)).

However, it should be noted that this demonstration did not take into account spatial covariance, which arises in pattern similarity analysis. This was done to illustrate the simple conditions under which issues arise (e.g. no spatial covariance). While omitting these sources of variance both in simulations and in models occurs sometimes in methods papers on inter-subject correlation (G. Chen et al., 2016; but see Kauppi, Pajula, & Tohka, 2014), and often in methods papers on pattern similarity analysis (Diedrichsen et al., 2011; Diedrichsen, Wiestler, & Ejaz, 2013; Garrido et al., 2013; Kriegeskorte, 2008; Nili et al., 2014), there are useful extensions of latent factor models that may accommodate these types of covariance. These will be discussed in more detail in the General Discussion.

While we have considered simple contrasts done over pairwise correlations here, several procedures, especially within the inter-subject correlation literature, calculate correlations against

various averages. This practice is analogous to having a number of participants complete a survey or cognitive test, and then averaging across items on the test to get a single test score per participant. In a widely cited psychology paper⁸, Bollen and Lennox (1991) demonstrated models under which this approach may or may not be sensible, but also made the point that these correlations are difficult to interpret due to bias introduced by measurement error. In the following section, we revisit these points with emphasis on pattern similarity analysis and inter-subject correlation.

⁸ Google scholar reported 3,181 citations as of 29 June, 2016

Table 1-1. Tests of Distinct Group Patterns used in Power Analysis

Name	Description	Conclusion
Between group contrast w/ permutation	Traditional approach. Calculate whether the average correlation within groups is greater than the average correlation between groups. NHST is conducted by permuting item labels (i.e. whether they belong to the same group or not).	Bias when item reliabilities are unequal
OLS-RFM w/ permutation	Estimate item reliability parameters, and then use in calculation of pattern similarity (see section XXX). NHST is conducted by permuting item labels.	Bias when item reliabilities are unequal.
OLS-RFM w/ non-parametric bootstrap	Estimate as above. NHST is conducted by using the non-parametric bootstrap. Each bootstrap sample is drawn by resampling observations (i.e. sample a random voxel) without replacement. The p-value was defined as the proportion of bootstrap samples in which ρ was less than 1.	Robust to unequal item reliabilities. Produces confidence intervals.
ML-RFM	Estimate item reliability and pattern similarity through maximum likelihood estimation. A chi-squared test between the reduced model where $\rho = 1$ was fixed, and where ρ was a free parameter was used to test for distinct group patterns.	Robust to unequal item reliabilities.

Note. OLS-RFM = ordinary least-squares estimation of reduced factor model; ML-RFM = maximum likelihood estimation of reduced factor model.

Example 2: Relating inter-item and item-total correlation

In addition to examining the pairwise correlation between items, researchers often employ a range of aggregating procedures, such as summing or averaging sets of items of beforehand. While such an approach may seem intuitive, it has led to misunderstandings in several fields (see Bollen & Lennox, 1991; Monin & Oppenheimer, 2005; Pollet et al., 2015), where researchers have chosen to use either the correlation of averages or average of correlations. For example, Monin & Oppenheimer (2005, Study 2) discuss a situation in which multiple raters judge the familiarity of photographs of individuals, for which they had a normed attractiveness measure. They calculated the correlation of attractiveness with either familiarity judgements within each rater, or average familiarity judgements across raters—illustrating that averaging familiarity across raters resulted in a higher correlation. The same use of averaging can be found in the fMRI pattern similarity literature; for example, researchers examining the effects of smoothing on voxelwise response patterns across two levels of stimuli (category and subordinate), used correlations of averages for categories, stating: “because the correlations for the across-category condition were always much higher than the correlations for the [across-subordinate condition, we had to bring the correlations to the same level prior to smoothing to be able to compare the conditions reliably” (Brants, Baeck, Wagemans, & Op de Beeck, 2011, p. 1374). This explanation identifies differences in the magnitude of correlations as a problem, and introduces a solution; but additional clarity may be achieved by exploring the reasoning behind, and effectiveness of, such approaches.

In the ISC (inter-subject correlation calculated across time) literature, Hasson and colleagues often calculate the correlation for each subject against the average of all other subjects within a condition (e.g. Ames et al., 2015; Hasson et al., 2009; Lerner et al., 2011). Other

researchers opt instead for the pairwise correlation between subjects (e.g. Abrams et al., 2013; Herbec, Kauppi, Jola, Tohka, & Pollick, 2015; Kauppi, 2010; Lahnakoski et al., 2014), but both practices are referred to as inter-subject correlation. It is clear that these two approaches are viewed as similar, and we will argue that such a view is often warranted, so that the use of both definitions of inter-subject correlation is likely not problematic for “is this zero”-type null hypothesis significance tests. However, even if the two methods produce similar results on a significance test, we have not addressed the question of effect size; we should also ask how these two forms of inter-subject correlation might measure the same quantity of interest.

While the use of different approaches to correlational analyses is understandable, and the intuition behind them is reasonable, the use of an explicit model helps both to explain these intuitions and to expose potentially unforeseen issues. In this section, we examine the practice of selectively averaging and correlating items from the perspective of a more general RFM, in which all items may have distinct reliabilities. From this perspective, the use of averaging may be viewed as an estimate of a latent pattern. However, error in this estimate introduces a bias, when an average is used as a stand-in for that latent pattern.

E.2.1. Defining Scenario and Covariance Methods of Interest

In this section, we will use a study by Hasson et al. (2009) to illustrate the use and interpretation of different averaging procedures. In this study, autistic and control subjects underwent fMRI scanning while viewing a short movie. In order to illustrate the degree to which subject timecourses were correlated, control subject timecourses (Fig 2-1A) and autistic subject timecourses (Fig 2-1B) within visual cortex (V1+) are shown along with the mean timecourse for each group. On the top-right of each of these figures, the average pairwise correlation between subjects is given. In making a point about consistency, Hasson et al. (2009) note that the average

pairwise correlation is higher within the control group (mean $r = .32$) compared to the autism group (mean $r = .14$; see Fig 2-1D).

In order to examine similarity between groups, they use the same data in a different way: they correlate the mean timecourses for the two groups (Fig 2-1C), noting “that by averaging the time courses within a group the responses become highly correlated across groups.” Finally, the researchers correlate each subject within the autism group against the average of other subjects within the control group, stating that “[c]orrelating the autistic individuals’ response time courses with the average [control] response time courses enabled us to recover more typical response time courses in each individual with autism.” (Hasson et al., 2009, p. 226)

Thus, in the paper above, the average pairwise correlation seems to test the degree to which people within the same group are related (consistency), and correlation between mean group timecourses seems to test the degree to which people across groups are related (similarity). We will argue here that latent factor model provides a unified way to assess the consistency and similarity of responses within and between groups. It is important to emphasize that while the latent model analysis provide a more systematic and principled way to test such hypotheses, our re-assessment of the data agrees and support with their fundamental conclusions: the autism group had lower consistency than the control group, but both groups shared a common signal (See E.2.5). In the following section, we provide clear language to distinguish between the measures above and other variations.

Figure 2-1. Hasson (2009) Autism Experiment

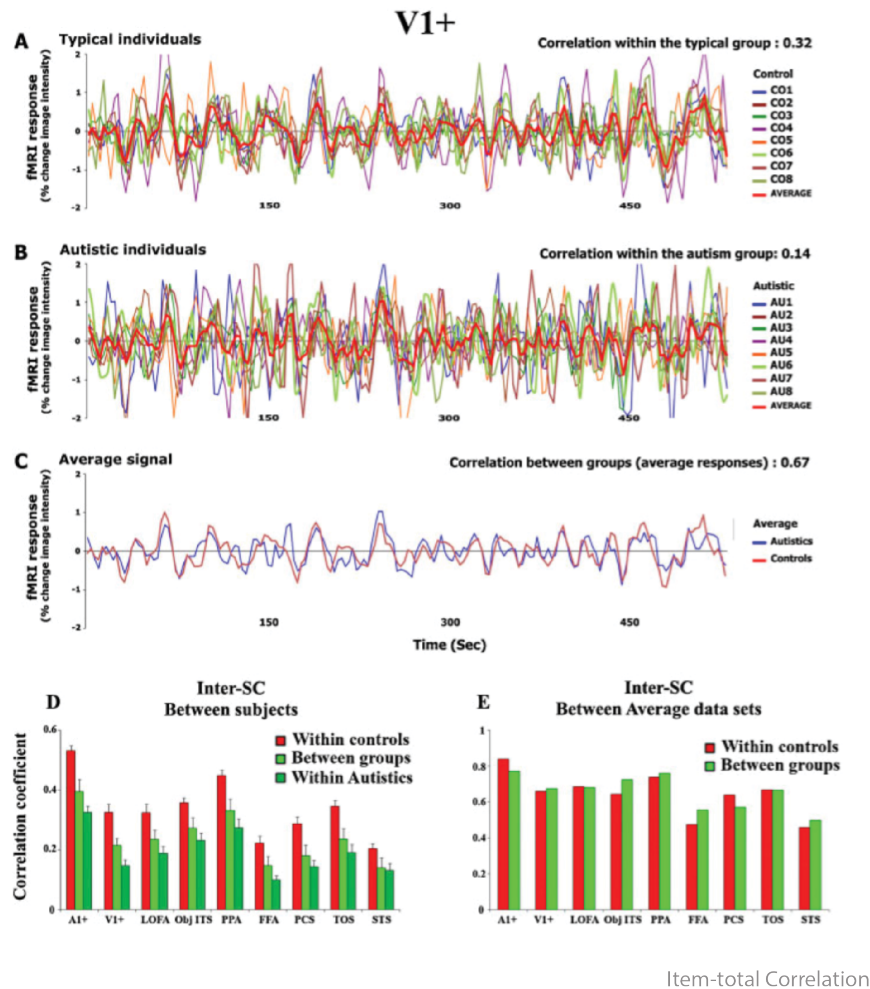


Figure of results from autism study of Hasson et al. (2009). Response timecourses within visual cortex for (A) the typical/control group, and (B) the autistic group. Note that the thick red line on each plot denotes the mean timecourse across subjects within a group. Correlations reported on the top-right of (A) and (B) are the average of pairwise correlations. (C) The averaged timecourses across subjects for each group. (D) Average of pairwise correlations within each group, and between groups. (E) Correlation of average timecourses across groups. Note that this was done by splitting the control group in half, and correlating its average timecourse against the remaining half ("Within controls") or the autistic group ("Between groups")

E.2.2. Inter-item, Item-total, and Total-total Correlation

Within a group, we define “inter-item correlation” as the correlation between two items, and “item-total correlation” as the correlation of an item with the sum or average of the remaining items in its group. For example, correlating $y_{A,1}$ and $y_{A,2}$ from the autism scenario is inter-item correlation, while correlating $y_{A,1}$ against the sum of others, $y_{A,2} + y_{A,3}$, is item-total correlation. We refer to the result of summing items as an “item composite”. These names are drawn from the field of psychometrics (see Furr & Bacharach, 2013), where an item may correspond to a survey or test question. One benefit of this correspondence is that studies employing pattern similarity analysis often include psychological tasks and batteries. In the same sense that a simple test, like a t-test or ANOVA, may be used on both behavioral and neural data collected for a study, keeping terminology consistent allows clear mappings and reusability across neuroscience and other psychological domains.

When these correlations use items belonging to separate groups, we prefix them with BG (between group). For example, correlating $y_{A,1}$ and $y_{B,4}$ is “BG inter-item correlation”, while correlating $y_{A,1}$ and the sum of group B items, $y_{B,4} + y_{B,5} + y_{B,6}$, is “BG item-total correlation”. Finally, when the sums of two groups are correlated, we refer to it as “total-total correlation”. Each definition, its representation under the RFM, and limitations discussed below, are summarized in Table 2-1. In addition, Fig 2-2 illustrates each measure as a function of the items used.

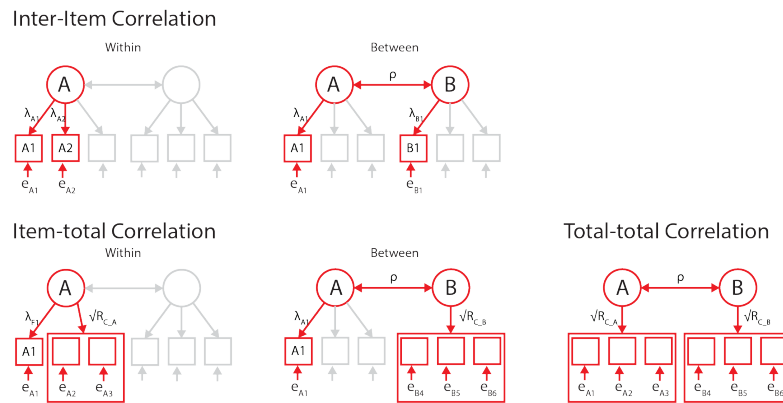
Table 2-1. Within and Between Group Correlations

Measure	Definition	Model Interpretation	Issue
Inter-item correlation	Pairwise correlation	$\text{Corr}(y_{a,1}, y_{a,2}) = \lambda_{a,1}\lambda_{a,2}$	
Item-total correlation	One item correlated against total of own group (C_a)	$\text{Corr}(y_{a,\text{new}}, C_a) = \lambda_{a,\text{new}}\sqrt{R_{Ca}}$	Conflates item reliability with reliability of total
BG inter-item correlation	Pairwise correlation of items in different groups	$\text{Corr}(y_{a,1}, y_{b,5}) = \lambda_{a,1}\lambda_{b,5}\rho_{a,b}$	Conflates latent correlation with item reliability
BG item-total correlation	One subject correlated against total of other group (C_b)	$\text{Corr}(y_{a,1}, C_b) = \lambda_{a,1}\rho_{a,b}\sqrt{R_{Cb}}$	Conflates latent correlation with item reliability, and composite reliability
Total-total correlation	Total for one group (C_a) correlated with total for another group (C_b)	$\text{Corr}(C_a, C_b) = \rho_{a,b}\sqrt{R_{Ca}R_{Cb}}$	Conflates latent correlation with both composite reliabilities

Note. $\sqrt{R_{Cg}}$ is the square-root of the reliability for the sum of items within group g (see section

“Item-Total Correlation and Composite Items”). The symbol $y_{a,1}$ refers to an item within group a, that is the first item overall.

Figure 2-2



graphical representation of the five different analyses. Note that taking the average (i.e. composite) of items is represented by putting a large box around the items. As noted in the section “Item-Total Correlation and Composite Items”, this reflects that under the RFM, the sum across items has a similar form to individual items.

E.2.3. Measuring Item Reliability and Composite Items

Inter-Item Correlation. Examining the correlational structure of the RFM reveals that inter-item correlation and item-total correlation have slightly different properties. For example, under the model the correlation between two items in the same group (i.e. inter-subject correlation) is given by

$$\text{Corr}(y_{g,i}, y_{g,j}) = \lambda_{g,i} \lambda_{g,j},$$

which says the correlation between two items within a group, g , is the product of their correlations with the latent group pattern. Thus, as items become more strongly correlated with the group pattern, inter-subject correlation will increase.

Item-Total Correlation and Composite Items. While the interpretation of inter-item correlation is fairly straightforward, understanding the item-total correlation requires evaluating a new component: the sum across items. This sum is sometimes referred to as an unweighted linear composite, since it does not apply any relative weighting to the items (Bollen & Lennox, 1991; Green & Yang, 2009). Bollen & Lennox (1991, p. 309) express the composite of items within a group, g , as follows:

$$\begin{aligned} C_g &= y_{g,1} + y_{g,2} + y_{g,3} + y_{g,4} \\ &= (\lambda_{g,1} + \lambda_{g,2} + \lambda_{g,3} + \lambda_{g,4})\eta_1 + (\epsilon_1 + \epsilon_2 + \epsilon_3 + \epsilon_4) \\ &= \lambda_c \eta + \epsilon_c, \end{aligned}$$

where $\lambda_c = (\lambda_{g,1} + \lambda_{g,2} + \lambda_{g,3} + \lambda_{g,4})$ and $\epsilon_c = (\epsilon_1 + \epsilon_2 + \epsilon_3 + \epsilon_4)$. The linear composite is similar

in form to each measure ($y_{g,i}$), where λ_C represents the relationship between C_g with its underlying signal, and ϵ_C its noise. Generally, increasing the number of subjects in the composite increases its signal to noise ratio, and thus its reliability (Lord, Novick, & Birnbaum, 1968; Sijtsma, 2009a). Moreover, we can calculate the correlation of the composite with a measure from the same group, that is not used in its calculation, as:

$$\text{Corr}(y_{g,\text{new}}, C_g) = \frac{\lambda_{g,\text{new}}\lambda_{g,c}}{\sqrt{\lambda_{g,c}^2 + \text{Var}(\epsilon_C)}} = \lambda_{g,\text{new}}\sqrt{R_C},$$

where $\sqrt{R_C}$ is the square root of the reliability for the composite (see Table 0-1). From the middle part of the equation, it is clear that if there were zero error variance, then this correlation would simplify to $\lambda_{g,\text{new}}$. However, in practice, the error variance in estimating the latent pattern is not zero, and will reduce this correlation. This means that whereas the inter-item correlation does not depend on the number of items in a group, item-total correlation will tend to underestimate an item's reliability for groups with fewer items (e.g. subjects or trials).

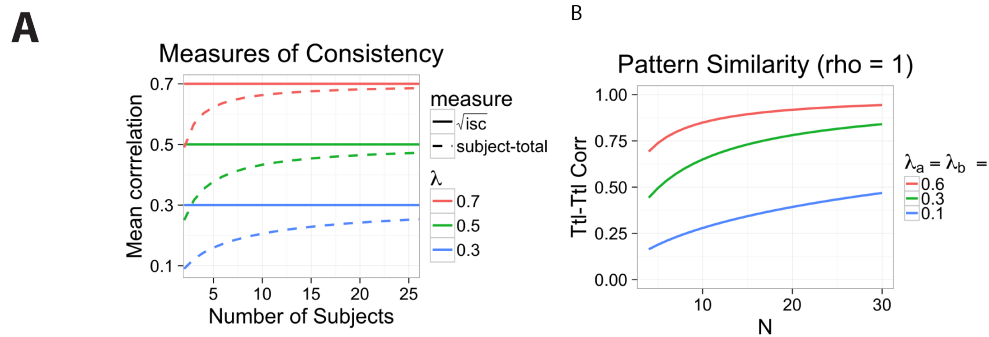
In order to illustrate the degree to which item-total correlation underestimates item reliability, we calculated inter-item and item-total correlation from the formulas above, based on a simplified form of a single latent factor model, in which item reliabilities were equivalent for all subjects, but the value of the item reliability and the number of subjects were manipulated. The results are plotted in Figure 2-3A. Note that in this simple case where item reliabilities are equal within the group, comparing the equations for inter-item and item-total correlation suggests that there is a connection between the methods. Specifically, as items are added to a group, so that the reliability of the composite approaches 1, then the item-total correlation

approaches the square root of the inter-item correlation⁹. However, for studies that use different sized samples with item-total correlation as a measure of consistency, failure to account for composite reliability may lead to spurious findings. In practice, it appears that studies using subject-total correlation to test for different levels of consistency across conditions generally ensure that there are equal (or nearly equal) numbers of subjects in each condition, so this may not be a major concern (e.g. Ames et al., 2015; Lerner et al., 2011; Regev et al., 2013).

One reason researchers may favor item-total correlation is that it yields a single estimate per subject, which may be useful for individual differences studies. Another reason may be that it tends to yield a larger value than inter-item correlation. However, simply obtaining a larger value is not, of itself, a compelling reason to use one over the other (see Marsh et al., 2013).

⁹ If item reliabilities are not equal within a group, but non-negative, then as the composite reliability approaches 1, the square root of the averaged inter-item correlations will be less than the averaged item-total correlations. The proof for this is identical to the one given for the within vs between contrast in the first section.

Figure 2-3. Effects of averaging on consistency and pattern similarity



(A) Bias in the item-total correlation in terms of measuring consistency. Loadings were set as either .3, .5, .7. The solid line denotes the square-root of the average inter-item correlation, which does not depend on the number of items. The dashed line denotes the item-total correlation, which approaches the correct loading as the number of items increases. (B) Total-total correlation bias in terms of measuring pattern similarity. Although pattern similarity in this case is 1, total-total correlation underestimates it both as a function of the loadings (consistency) and number of items.

E.2.4. Latent Pattern Similarity Through Between-group Measures

BG Inter-Item Correlation. Calculating correlations between groups, while a function of item and composite reliability, also incorporates the correlation between latent variables. In the case of BG inter-item correlation, the formula for the correlation between two measures is given by (Bollen & Lennox, 1991):

$$\text{Corr}(y_1, y_5) = \lambda_1 \lambda_5 \rho_{12}.$$

In this formula, a low correlation can come from two sources: low item reliability in one group or a low correlation between latent signals. For this reason, using BG inter-subject correlation to estimate the similarity between two latent patterns should be avoided. One possible exception might be to test for the existence of non-zero correlation between latent patterns, since $\rho_{12} = 0$ implies a BG inter-item correlation of 0.

BG Item-Total Correlation. In the cases of BG item-total and total-total correlation, the reliability of the composites again affect estimates. For the item-total, the correlation is expressed by:

$$\text{Corr}(y_5, C_1) = \frac{\lambda_5 \lambda_C \rho_{12}}{\sqrt{\lambda_C + \text{Var}(\epsilon_C)}} = \lambda_5 \rho_{12} \sqrt{R_C}$$

similar to the within group item-total correlation, as the composite reliability becomes 1 (i.e. as error variance approaches 0), the correlation becomes the product of an item's correlation with its latent signal, only now it is also multiplied by the correlation between latent signals for the two groups. Thus, if two groups have different levels of item reliability, BG item-total correlations for the two groups may differ from within-group item-total correlations, even if they

are generated from the same underlying pattern – an unsatisfactory property for measuring the correlation between the underlying patterns. However, similar to the use of BG inter-item correlation mentioned above, this measure is also expected to be 0 if $\rho_{12} = 0$.

Total-Total Correlation. While the previous two methods allow for testing whether pattern similarity between two underlying signals differs from zero, they also conflate item reliability with pattern similarity. In contrast, the total-total correlation gets a step closer to estimating pattern similarity. Since the square-root of reliability of a linear composite is analogous to the correlation of an item with its latent pattern, the correlation between two composites can be expressed by (see Spearman, 1910):

$$\text{Corr}(C_1, C_2) = \rho_{12} \sqrt{R_{C1} R_{C2}} ,$$

Where R_{C1} is the reliability of C_1 , and likewise for R_{C2} . In the case that there is no error in either composite, their correlation will equal ρ_{12} . Thus, for large samples or those with high consistency, where composite reliabilities tend to be high, the correlation between those composites may be a good estimate of pattern similarity. Moreover, this formula can be rearranged to solve for ρ_{12} directly:

$$\rho_{12} = \frac{\text{Corr}(C1, C2)}{\sqrt{R_{C1} R_{C2}}} .$$

If the reliabilities are known, then the correlation between signals is the correlation between composites, divided by the square root of their reliabilities. Note that this is analogous to the correction in Example 1 for inter-item correlation, with a composite serving to create an item with increased reliability. Bollen & Lennox (1991) derive the formula for computing the reliability of an unweighted composite, as a function of item reliabilities, and Green & Yang

(2009) discuss calculating the reliability of weighted linear composites. In order to illustrate the bias induced by measurement error in the composites, Fig 2-3B shows the expected correlation between two composites for equal-sized groups with the same underlying pattern ($\rho = 1$), as a function of the number of items in each group. In practice, fitting the latent factor model directly takes reliability into account (via the factor loadings) when calculating pattern similarity.

E.2.5. Estimating Pattern Similarity and Consistency from Summary Data

Hasson et al. (2009). These formulas, in addition to clarifying how the RFM conceptualizes many procedures over correlations, allow for an alternate view of various correlations reported in previous studies. For example, we can re-analyze Hasson et al. (2009)'s inter-item correlations, BG inter-item correlations, and total-total correlations. To preview, this re-analysis does not contradict the null hypothesis significance testing done by Hasson et al. (2009), but illustrates two additional pieces of inference that a model--such as the RFM--may provide. First, it provides an explicit relationship between within- and between-group correlations, based on group similarity. Second, it allows us to examine the *degree* to which groups are similar, while controlling for measurement error (e.g. low consistency). Below, we illustrate this by using within-group correlations, and the assumption that two groups share a single underlying signal, to calculate the between-group correlations implied by an RFM.

The sample in Hasson et al. (2009) was composed of 8 subjects in the control group, and 9 in the autism group. They split the control group into half, in order to calculate the total-total correlation between the split-half control groups, as well as between one of the control halves and the autism group. We assumed that the data corresponded to a two factor model with equal loadings within a factor. As in Example 1, we used the square-root of the average inter-subject correlation to estimate the loadings for each group, and used their product to calculate the BG

inter-subject correlation implied if pattern similarity between the two groups was perfect ($\rho = 1$). These results are shown in Figure 2-4 (left). Note that while the average of observed BG inter-subject correlations is intermediate to the average of observed WG inter-subject correlations, it is fairly consistent, as argued in Hasson et al (2009), with a model in which the two groups have perfect pattern similarity (e.g. a single underlying signal).

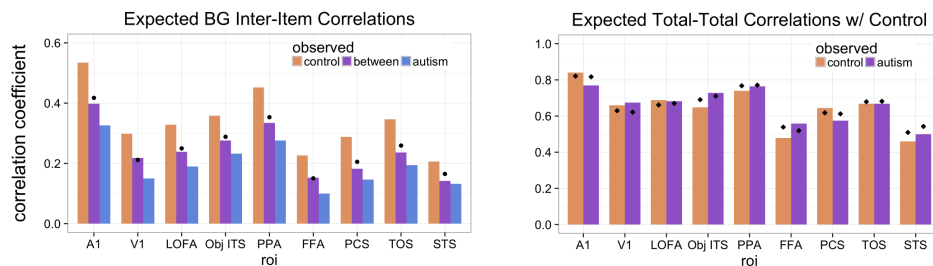
Next we used the same loading estimates, and the assumption that pattern similarity was perfect between the two groups, to consider the total-total correlations between the autism and control groups. Importantly, as discussed above, these correlations are affected by the number of subjects in each group, as the number of subjects affect the reliability of each total (i.e. composite reliability). In order to account for this, we estimated the composite reliability of each group, as a function of its sample size and the estimated loadings for that group, using the Spearman-Brown Prophecy Formula¹⁰ (Spearman, 1910). Finally, since the total-total correlation is a product of the correlation between factors (ρ) and the square-root of the reliabilities for both totals, we substituted in the reliabilities estimated above and fixed $\rho = 1$ to calculate the model--implied total-total correlation when the factors are perfectly correlated. The results are displayed in Figure 2-4 (right). As with the pairwise correlations, total-total correlations for one half of the control group with the other, or with the autism group appear consistent with those implied by the model.

Such an approach is overly simplistic, in that it made use of only the reported summary

¹⁰ The Spearman-Brown Prophecy formula for estimating composite reliability within a group is $\widehat{R}_C = f(\lambda, N) = \frac{N\lambda^2}{1+(N-1)\lambda^2}$, where N is the number of items used in the composite. For example, the average inter-item correlation for the control group in A1 is $\lambda^2 = 0.534$, with N=8 subjects, so we estimated the reliability of the total as $8*(.534) / (1 + (8 - 1)*.534)$.

data, and assumed that pattern similarity was perfect, rather than calculating similarity directly. Nonetheless, it demonstrates that one piece of the data (in this case, pairwise correlations within each group) and a model--such as the RFM--provide a clear link between what may appear to be disparate correlational measures. One approach not considered here is the subject-total correlation. In order to provide a similar illustration on this type of data, we consider the summary data in Regev et al (2013).

Figure 2-4. Hasson et al. (2009) BG inter-item and total-total correlation



A) Plots by Hasson et al (2009), with points representing what the model would expect to see if pattern similarity was 1. Note that each point was calculated using only the within-group correlations reported (bars without points) and sample size, in addition to the assumption that the latent factors were perfectly correlated.

Regev et al. (2013). One experiment that did report item-total correlation values was by Regev et al. (2013). In two conditions of their experiment, subjects attended to the same narrative in either written or spoken form. Importantly, for the written form, the words of the narrative were displayed with the same timing as the audio narrative, so that subjects could reasonably be correlated between conditions. The researchers initially focused on using item-total correlation to test for different consistencies within the two groups, with differences indicating that areas were “modality-selective”, and no evidence of differences indicating that areas were “modality-invariant”. They went on to assess another form of “modality invariance”

by using the BG item-total correlation to test for non-zero pattern similarity. Such a consideration seems reasonable, and a potentially useful addition to their discussion is a measure of effect size—or the *strength* of modality invariance, in terms of pattern similarity.

The researchers assessed the magnitude of pattern similarity by calculating the total-total correlation within two ROIs, and argued that observing values of .47 and .69 meant responses across modalities were highly similar. However, while a useful illustration that the groups are similar to some extent, such an approach does not take into account the dampening effect of measurement error on the total-total correlation. As in the Hasson et al. (2009) example above, we can use the summary measures of various correlations to get a rough estimate for what the data might look like if the two conditions had perfect pattern similarity. Such a consideration is not meant to call into question the basic findings of Regev et al. (2013), but rather to use a RFM to illustrate a type of measurement invariance that was not the focus of their paper.

Note that the Spearman-Brown Prophecy Formula expresses composite reliability for a group as a function of its group loading and the number of items in the group. Accordingly, item-total correlation for a group--the product of its loading and the square-root of composite reliability—is also a function of this (unknown) group factor loading and the (known) number of items used in the composite. We first used numerical approximation¹¹ to estimate this group factor loading from the reported item-total correlations for each condition, and then used each

¹¹ The model-implied item-total correlation for group g , $r_{g\cdot}$, may be expressed as $r_{g\cdot} = \lambda_g f(\lambda_g, N_g - 1)$, where the function f is the Spearman-Brown formula for the reliability of the group's composite, $f(\lambda, N) = \frac{N\lambda^2}{1 + (N-1)\lambda^2}$. It may be readily verified that when the number of items is fixed, and loadings are greater than 0, the model-implied item-total correlation is strictly monotonic. Another way to phrase this is that for a fixed number of items, where loadings are equal across items, increased consistency (higher loadings) leads to greater item-total correlation.

loading to estimate composite reliability from the Spearman-Browne Prophecy Formula, as above. Since the BG item-total correlation is the product of three terms--one group's loading, the other's composite reliability, and pattern similarity (ρ)--we substituted in our above estimates of the first two terms, and fixed ρ to 1, in order to estimate what this correlation should look like with perfect pattern similarity. The result is shown in Figure 2-5. While the average observed BG item-total correlation is intermediate to the observed WG item-total correlation, so too are the model-implied BG item-total correlations.

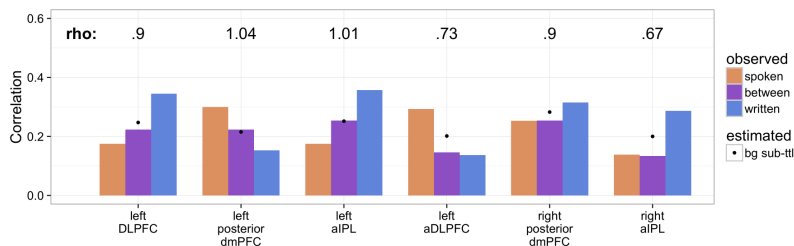
While fixing pattern similarity to 1 was done to calculate the model-implied BG item-total correlation, we also used the observed BG item-total correlation to estimate pattern similarity (ρ). In order to do this, we took the formula for model-implied BG item-total correlation, and substituted in the above estimates of one group's loading and the other's composite reliability, along with their observed BG item-total. In areas that were argued to have differing levels of consistency between conditions, there appears to be a range of magnitudes for pattern similarity (Fig 2-5). For example, left posterior dorsal-medial prefrontal cortex (dmPFC) and left anterior inferior parietal lobe (aIPL) were estimated with essentially perfect similarity estimates, at $\rho = 1.04$ and 1.01 , respectively. In this sense, it is possible that areas like these are modality-invariant, not only in the partial, "any non-zero degree of shared response" sense, but in that the two modality conditions may have the same underlying pattern of temporal activation, regardless of their consistency levels.

While the simple calculation of pattern similarity from summary measures is suggestive, it comes at the cost of simplifying the RFM, and producing only point estimates. More complete inference from fitting an RFM (or alternative models) to the raw data is necessary to fully examine consistency and pattern similarity in these data. The above analysis supports the item-

total correlation approach of Regev et al. (2014) and extends it by providing an explicit model to explain the rationale behind their analysis of consistency and pattern similarity, as well as options for measuring the effect size of pattern similarity.

E2.6. Discussion

Figure 2-5. Regev et al. (2014) BG subject-total correlation



Observed BG subject-total correlations from Regev et al. (2014), with estimates of the BG subject-total correlation using the Spearman-Brown Prophecy Formula to estimate the loading for an individual item. The estimated loading and number of subjects were then used to estimate composite reliability for each group reliability, and fixing rho to 1 (points). In addition, we calculated estimates of rho from the same formula, using the estimate of composite reliability and observed BG item-total correlation (text).

--

The more general form of the RFM, where each item may have a different reliability, clarifies the relationship between item composite and pairwise correlation approaches. The practice of summing or averaging items within a group may be viewed as an attempt to estimate the group's latent pattern. However, since the composite will likely be a noisy measure of the latent pattern, this will underestimate the similarity between latent patterns.

Within a group, correlating an item against the sum of others (item-total correlation) may be viewed as a biased measure of that item's reliability. Between groups, correlating an item against the sum of others may be viewed as reflecting both that item's reliability and the

correlation between group patterns.

While obtaining larger correlations by using averages may be appealing, and in some cases—such as testing whether a group correlation is 0—may yield sensible results, the use of these approaches benefits greatly from an explicit consideration of what they are supposed to measure.

A final set of commonly used procedures, that provide another variation on the theme of running contrasts over correlations, split the data into partitions within each group and then calculate correlations across partitions. In the following section, we consider the rationale for these procedures, as well as how they may be modeled as special cases of the RFM.

Example 3: Correlated Residuals and Multiple Groups

Beginning with Haxby (2001), researchers have used a number of splitting and correlating procedures in analyses of spatial fMRI patterns. For example, Haxby (2001) split their data into odd and even runs, and then correlated each pair of trials across runs (i.e. each even-odd trial pair). This was done to ensure that within-run factors (e.g. temporal dependencies) did not distort correlations (Haxby, 2012)¹². Splitting between even and odd runs has the benefit of removing within-run correlations, but comes at the cost of removing many others, such as the correlations between each pair of runs within a half. Depending on their beliefs about the data, researchers may exclude different correlations—such as those between temporally adjacent trials in pattern similarity analysis, or those within the same subject in inter-subject correlation.

Another motivation for correlating across splits is in order to remove bias due to measurement error (Allefeld & Haynes, 2014; Nili et al., 2014; Walther et al., 2015). One example is given by Walther et al. (2015), who demonstrate that even when there is only a single latent pattern underlying the data, measurement error may cause the Euclidean distance between items to be greater than 0. While it is important to note that estimating the distance between items is not equivalent to estimating the distance between latent patterns, their observation has an important parallel at the level of pattern similarity in latent variable models: estimates of pattern similarity (a correlation) will always be less than or equal to 1. Thus, even when there is a single latent pattern underlying the data (i.e. $\rho = 1$), random error in a two-factor model may cause estimates of ρ to be less than 1, leading to its underestimation on average.

¹² Haxby, 2012 refers to this as ensuring “independent observations”. We take this to mean independent residuals of underlying patterns, since all observations are likely non-independent in a pattern similarity analysis (i.e. share some meaningful correlation).

However, it should be noted that tests like a simple one-way ANOVA run into similar situations, in that the variance between observed group means is virtually always non-zero, even when the means are not truly different from one another. An alternative way to view the dilemma faced by Walther et al. (2015) is to consider it to be an issue of model selection. From this perspective, this issue arises from accepting and fitting a model with two latent factors, when a single factor is more appropriate. Rather than assuming a two-factor model and trying to correct pattern similarity, we need to first evaluate whether a one- or two-factor model is more reasonable.

In this section, we will consider accounting for run- or pair-specific variance. We first illustrate the concept using a real fMRI data set, demonstrate how the model works with multiple (e.g. 20 subjects') covariance matrices, and conclude with an example of performing model selection through cross-validation.

E.3.1. “Decorator Experiment” Data

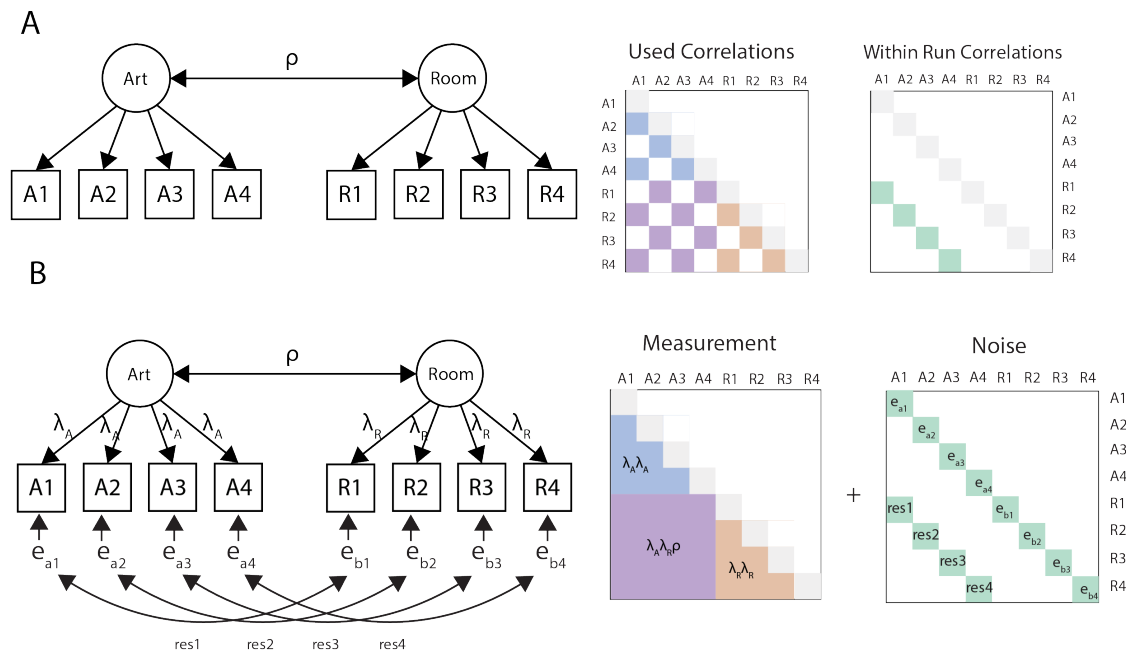
Aly and Turk-Browne (2015) conducted an fMRI experiment in which all participants viewed the same set of stimuli — 20 images of room with art hanging on the wall — but for each image trial were cued to attend only to the room (“room condition”) or art (“art condition”). They derived observed voxel-patterns by using GLM to fit regressors for each of P trials within a voxel. Stacking the regressors for all N voxels within an ROI resulted in a $N \times P$ matrix. Since ROIs were defined specifically for each subject, N was not constant across subjects, but the number of trials, P , was always 160. For simplicity, we reduced these trials into the average pattern for each condition within the four pairs of runs in which the data were collected, resulting in ROIs with a $N \times 8$ ($= 4$ run pairs $\times 2$ conditions) matrix for each subject.¹³ Experiment designs

¹³ As discussed in Example 2, averaging represents the assumption that the items being averaged all load on the same factor(s). The use of averaging for latent factor models has been discussed

such as this are increasingly common in the fMRI literature (e.g. Favila, Chanales, & Kuhl, 2016; Haxby, 2001; Iordan et al., 2015; Kriegeskorte et al., 2008, p. 1133; LaRocque et al., 2013; Op de Beeck, Brants, Baeck, & Wagemans, 2010; Schapiro et al., 2013; Wolosin et al., 2013). A simple RFM of this data is shown in Figure 3-1A. It should be noted that averaging and using the correlations both within and across runs differs from the approach taken by Aly & Turk-Browne (2015), who used only the within run correlations. While their focus within run is perfectly reasonable, looking within and across runs allows us to ask questions about pattern similarity and consistency, while also considering the potential issue of within-run specific variance.

extensively in psychology (see Little, Rhemtulla, Gibson, & Schoemann, 2013; Marsh, L"properties":{"formattedCitation":{"\rtf(s.

Figure 3-1. Splitting Odd vs Even Runs

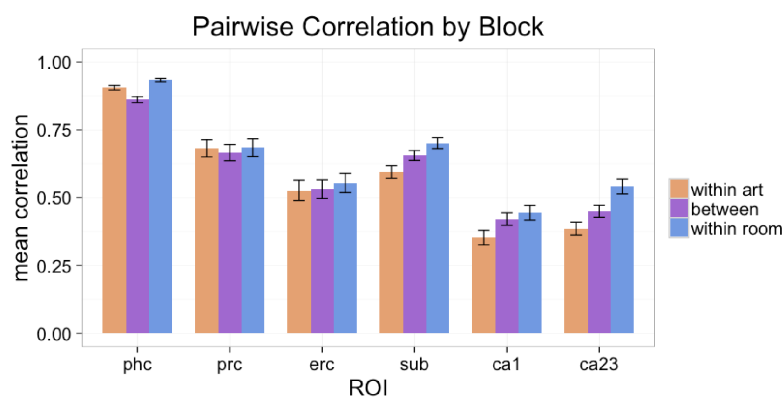


(A) Illustration of splitting between odd and even runs. Removed elements are shown as white spaces. Many elements of the correlation matrix, including within run correlations are removed. (B) Illustration of how a latent factor model may account for within run correlations by correlating residual errors.

In their original analyses, Aly & Turk-Browne (2015) used the within- and between-group contrast to evaluate whether “attention would induce state-dependent activity patterns” (p. 7), and found the contrast for all ROIs shown in Figure 3-2 was significant in a null-hypothesis significance test ($p < .05$ for all ROIs). Such an approach corresponds to the evaluation of pattern similarity emphasized in Examples 1 and 2, as the RFM shown in Figure 3-1A contains state-dependent patterns when the correlation between latent factors is imperfect (i.e. $\rho < 1$). Moreover, they contrasted the within-group correlations for the two states in order to “examine whether stability differed between the art and room states” (p. 7), and found the contrast was significant for all ROIs, with the exception of perirhinal cortex (PRc). In the context of the RFM,

this is a test of the factor loadings, and so is an evaluation of consistency. Importantly, in both the correlations of unaggregated (see Figures 4 and 5 in Aly & Turk-Browne, 2015) and aggregated trials (our Figure 3-2), it appears that potential differences in within-group correlations (i.e. consistency) may affect the use of within- vs between-group contrasts to test pattern similarity (see Example 1).

Figure 3-2



The average correlation within- and between- art and scene trials. As in the original analyses by Aly & Turk-Browne (2015), the correlations within room trials are often larger, and the between-group correlations are sometimes intermediate to the within-group correlations. As in their original figures, standard errors were calculated across subjects. [NOTE: I also have this same plot prepared for their original analysis].

E.3.2. Correlated Residuals

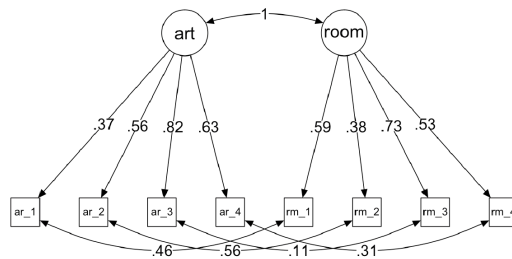
Whereas splitting into even and odd runs operates by removing select covariances, latent factor models may incorporate idiosyncratic covariances between one or more pairs of items through correlating their residual errors. This is illustrated on the example data in Figure 3-1B which shows a RFM with four additional parameters, one for the covariance between residuals of each pair of art and room trials within a run. A major advantage of this approach is that it preserves the covariance matrix.

This approach is illustrated on a single subject from the Decorator Experiment. We fit the

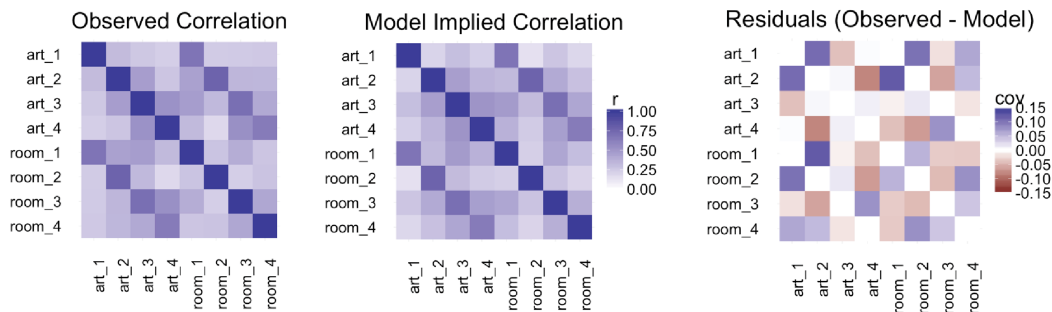
RFM shown in Fig 3-3A, with the exception that loadings were free for all 8 items, rather than having a single loading per group. The estimated covariance matrix is shown alongside the observed covariance matrix in Fig 3-3B. First, there is a clear increase in correlation between items in the same run (e.g. A1 vs R1, A2 vs R2, etc.), that is also incorporated into the model. Second, plotting the difference of the observed and model-implied covariance matrices reveals that the model essentially used the additional residual parameters to “remove” these correlations, by estimating them to be the observed correlation. In this sense, controlling for idiosyncratic correlations is similar to adding intercept terms to linear models, in order to control for unwanted differences in means.

Figure 3-3. Art vs Room Experiment w/ Correlated Residuals

A



B



(A) Fitted model for a single subject. Note that the similarity between latent art and room patterns was not a fixed, but a free parameter, estimated as 1. (B) Observed correlation matrix, model-implied correlation matrix, and the difference between them.

E.3.3.

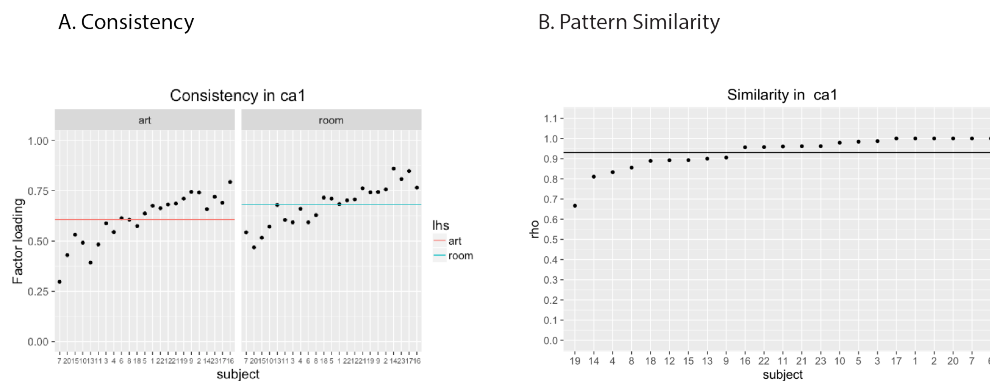
Extending the Model to Multiple Groups

One important aspect of the Decorator Experiment not addressed above is that the researchers did not collect data from only one subject, but 21 subjects. Thus, they were not only interested in examining pattern similarity at the individual level, but at the level of sets of subjects as well. Such issues also arise in fMRI studies that use the GLM approach to model activation within a single voxel, where first level analyses estimate some parameter within subjects—such as calculating a contrast between activation for two trial types--while the second level may combine these parameters across subjects. This may be done by (1) estimating individual parameters for each subject, (2) estimating the same parameters for each subject, or (3) treating parameters across subjects as drawn from a second-level random effect (Mumford & Poldrack, 2007; Poldrack, Mumford, & Nichols, 2011, ch. 6).

The RFM specified above (Fig 3-1B) may be fit in corresponding ways to those of above. First, each subject could be fit individually. For example, the same RFM could be fit for each subject, with the same free parameters for each one. Second, one or more model parameters may be restricted to be equal across the entire set of subjects. For example, the model could be constrained so that the same parameter for the correlation between Art and Room is estimated across all subjects. This is sometimes referred to as parameter or measurement invariance in psychology (see Van De Schoot, Schmidt, De Beuckelaer, Lek, & Zondervan-Zwijnenburg, 2015). When it comes to null hypothesis significance testing, a classic strategy used for pattern similarity analysis is to calculate parameters for each individual, but then use those parameters as the inputs to a separate, second-level analysis. For cases where the null hypothesis is that there is a single group level parameter, then the variance between the individual parameter estimates is treated as sampling error. For example, Diedrichsen et al. (2011) estimated the equivalent of a latent factor model within each subject individually, and then test that the average latent pattern

similarity across subjects is not zero by using a simple t-test. Moreover, Aly and Turk-Browne (2015) calculated the within- vs between-contrast for each subject, and used the same t-test. Finally, much recent work has focused on the use of treating parameters across subjects as random effects. That is, each subject's parameter is taken to be drawn from a random distribution with, for example, some mean and variance. A smaller variance term is similar to the invariant approach, while a larger variance term is similar to the variant approach. This is typically done using Bayesian methods, with the prior over the random-effect affecting the level of flexibility (B. Muthén & Asparouhov, 2012; van de Schoot et al., 2013; Van De Schoot et al., 2015).

Figure 3-4. RFM across all subjects for a single ROI (CA1)



A) Individual estimates of pattern consistency for each subject. Subjects are ordered by the average strength of their consistency estimates (i.e. within-group factor loadings). B) Pattern similarity in CA1 for each subject.

In order to demonstrate the classical approach, we used the first approach to separately estimate parameters for each subject. For the CA1 ROI, Fig 3-4A shows each subject's consistency measure for the Art and Room conditions (e.g. λ_{Art} and λ_{Room} for each subject). Fig 3-4B shows each subject's estimated pattern similarity parameter, ρ . In order to describe the results across subjects, we could simply take the mean across subjects, and use the variance

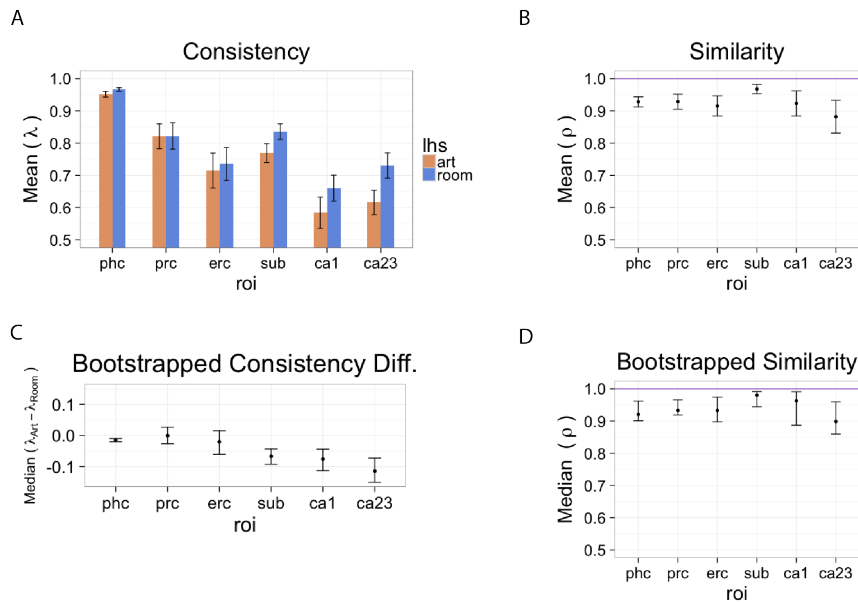
across subjects to calculate 95% confidence intervals by assuming that the individual parameter estimate across subjects are normally distributed (Fig 3-5A,B). However, because ρ may only be equal to or less than 1, when a single latent pattern underlies both trial conditions, error in its measurement can only lead to underestimation. Relatedly, it will not be normally distributed. One possible approach to reducing this underestimation may be using medians rather than means. To illustrate, 95% confidence intervals were derived by using the bootstrap to resample subjects with replacement, and then calculating the median on the parameters of interest for the resampled subjects—either the difference of group consistencies (Fig 3-5C) or pattern similarity (Fig 3-5D). Similar to the original analysis of Aly and Turk-Browne (2015), the bootstrapped confidence intervals of consistency differences includes 0 for perirhinal cortex (PRc); however, unlike the original analysis, this was also the case for entorhinal cortex (ERc). In addition, bootstrapped estimates of pattern similarity appear high (all $\rho > .9$), with the upper bounds of subiculum (SUB) and a sub-region of cornuammonis (CA1) near perfect similarity (both $\rho = .99$).

From the perspective of model selection, the problem of underestimating pattern similarity arises due to using a two-factor model when one-factor would be more appropriate. Rather than trying to change the way ρ is estimated, assuming a two-factor model, an alternative would be to first evaluate whether a one- or two-factor model provides a better representation of the data. A major problem for model comparison is accounting for the complexity of different models. For example, a “saturated” latent factor model is one which has a free parameter for every covariance between unique pairs of items, so perfectly fits the observed covariance. Measures such as Akaike’s Information Criterion (AIC) and Bayesian Information Criterion (BIC) are often used to compare models while accounting for this complexity (see Aho et al.,

2014), but are not clearly defined in this case, as the RFM presented considers voxels to be independent observations, when they will likely be related through spatial dependencies (Le Rest, Pinaud, Monestiez, Chadoeuf, & Bretagnolle, 2014). This problem was considered for a closely related method called path modeling by Bullmore et al. (2000), who addressed the problem of unmodeled temporal structure in the analysis of fMRI timecourses, the calculation of effective degrees of freedom, and fit indices that may also be applied to factor models¹⁴.

¹⁴ Path modeling and latent factor analysis may be viewed as types of structural equation models (Bollen, 1989).

Figure 3-5. Estimating RfM parameters across subjects in pattern similarity analysis



A) Pattern consistency, or the “stability” of the patterns underlying art and room trial activation. Note that the error bars are 95% confidence intervals, calculated by treating the individual subject observations as normally distributed. B) Pattern similarity across underlying art and trial patterns. Note that because valid similarity values range from -1 to 1, even if the underlying patterns are the same, the estimated similarity values may be less than 1, due to random error. C) Bootstrapped estimates of the differences between pattern consistency for art and room trials, with 95% confidence intervals. D) Bootstrapped estimates of the median pattern similarity between subjects, with 95% confidence intervals. This approach is more conservative than using the mean, but still may estimate pattern similarity as less than 1, due to random error.

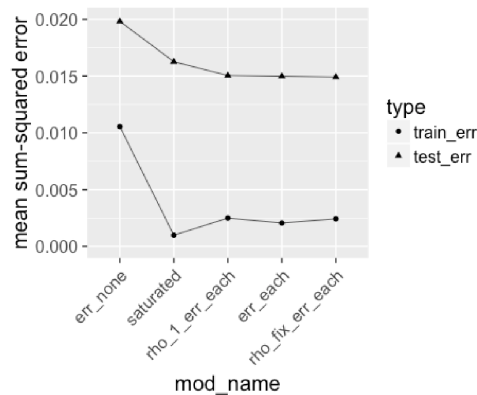
E.3.4. Performing Model Selection via Cross-validation

One weakness of the above analysis is that it chooses a single model specification for the data. In reality, a more empirical approach would be to decide between several possible models. In order to illustrate this approach, we performed leave-two-out cross-validation over the four runs (See model in Fig 3-1B). That is, for each of the six possible split-half combinations, two runs were used for training, and two were held out for testing. We manipulated two model features: residual covariances and pattern similarity. The two within-run residual covariances were either given separate parameters, or kept at 0. In the case where within-run residual covariances were included, the pattern similarity parameter (ρ), was either variant, invariant, or fixed to 1 across subjects. Finally, we included the “saturated” model, which perfectly fits the observed covariance matrix. For each training-test combination, each of these 5 models was fit to the training data, and the model-implied covariance matrix¹⁵ was compared against the observed covariance matrix within the test set, using the mean sum-squared error across the 10 unique elements of the test data covariance matrix. This measure of prediction error was selected because it corresponds to the criteria used to fit the model (unweighted least squares).

The average error in the test and training samples for CA1 is shown in Fig 3-6. Not including within-run residual covariance and the saturated model performed worst in the test group. The best performing model for this ROI fixed ρ to be invariant across subjects, but did not restrict it to 1. In general, fixing ρ to 1 did not outperform models

¹⁵ Note that due to the standardizing of trials beforehand, the covariance matrix was a correlation matrix.

Figure 3-6. Model Selection via Cross-validation in ca1



For CA1, the average of each model's goodness of fit to test and training data across the 6 combinations of leave-two-out splits. Goodness of fit was measured as the mean sum of squares of the difference between the observed and model-implied correlation matrices. Error is lower on training data because each model was fit to minimize this error in that data. The "saturated" model assumes the observed covariance is the best prediction. Unlike models with "err_each" in their name, the "err_none" model does not include additional covariance between items within runs. The component "rho_1", "rho_fix" indicate models in which pattern similarity is fixed to 1, or invariant across subjects, respectively.

that

either left ρ free, or kept ρ invariant across subjects¹⁶. In this sense, the cross-validation procedure led to the selection of two-factor models across ROIS, consistent with the claim by Aly & Turk-Browne (2015) that the task induced category specific activity patterns. Importantly, while it is difficult to interpret the averages of within- and between-group correlations, the selected models characterize the degree to which latent patterns are specific to a category through their correlation.

E.3.5. Discussion

Latent factor models may account for different sources of variance, offer simple extension to multiple covariance matrices, and allow for model selection through prediction. An important benefit of using a model of the data, rather than producing a summary measure, is that

¹⁶ Note that the best fitting model for ERC did not include within-run covariance.

it not only is explicit about the interpretation of such a measure, but whether it should be used for inference in the first place may be treated as an empirical question. For example, simply using a model that should--but fails to--take into account within-run covariance will produce estimates of pattern similarity and consistency that are incorrect, because it is a poor representation of the data. This may also be the case when performing a null hypothesis significance test while contrasting blocks of a correlation matrix, but the misspecification will take the form of inappropriate assumptions—which Example 1 argues may not be clearly formulated in many cases. While we used the mean-squared error to evaluate competing models, other error measures, such as entropy loss, have been considered (Bien & Tibshirani, 2011; Huang, Liu, Pourahmadi, & Liu, 2006; Knafl & Grey, 2007).

In certain instances, performing odd/even splits of the data can be useful for omitting correlations that include additional sources of variance, such as those within the same run. Allowing residuals to covary in the RFM illustrates one possible rationale for such a procedure, and suggests that using holistic models of the data may meet the goals of splitting procedures while keeping the data intact. The model used in this example focused on correlating residual errors within a run, which consisted of only two items. With more than two items in a run, two common approaches exist. The first, sometimes called a Correlated Uniqueness (CU) model, would include free parameters for each residual covariance between items within a run. Note that a two-factor RFM with CU within each group is essentially “ignoring” within-group covariances, and may be viewed as a case of probabilistic canonical correlation analysis (see section 3.2 in Bach & Jordan, 2005). The second, sometimes called Multi-trait Multi-method (MTMM) models, arise in psychology when “traits” of interest, such as cognitive or personality characteristics, are assessed using multiple methods, such as behavioral tasks and questionnaires.

In the two-factor RFM case, the traits would be the art and room conditions, and the methods would be runs. This model fits separate factors for “traits”/conditions—as in the two-factor RFM—but also an additional factor for each method/run. There are many types of MTMM models, but in one common approach, each of the method/run factors are restricted to be uncorrelated with all other factors (multi-trait uncorrelated-methods, Geiser, Bishop, & Lockhart, 2015; see Eid, 2000). Interestingly, a special case of this model, where there is only a single method/run factor for all items, and its loadings are restricted to be equal, is equivalent to the model used by Diedrichsen et al. (2011) for explaining why subtracting the overall mean pattern from each item is often inappropriate (see also Maydeu-Olivares & Coffman, 2006).

Moreover, moving from considering the analysis of a single covariance matrix to many matrices (e.g. 20 subjects) highlights important parallels with the move from a one- to two-level GLM. While much discussion in neuroimaging has considered this issue for GLM, with many articles focusing on the value of random effects, researchers performing pattern similarity analysis have largely ignored the question—by treating first level parameters, such as the result of the within- vs between-group contrast for each subject, as the raw input into second-level analyses. The neglect of this question is understandable, as creating reasonable implementations for second-level pattern similarity analyses may be computationally challenging, and possibly unnecessary in simple cases.

Finally, by modeling many covariance matrices, the relationship between pattern similarity analysis and representational similarity analysis may be clarified. For example, representational similarity analysis might use the observed covariance matrix for each subject as an input to a second-level analysis. At the second level, researchers often calculate some similarity measure between each observed covariance matrix and another “model matrix”, which

represents some hypothesized underlying structure. Accordingly, the goal of comparing the results for different “model matrices” is to select the correct underlying structure for the data, a form of model selection. From the perspective of latent factor analysis, this is analogous to fitting a saturated model for each subject, and then submitting it to a second-level procedure. In contrast to this two-step approach, latent factor analysis uses a hypothesized structure to fit the data directly, which has the benefit of producing an explicit model-implied covariance matrix, rather than a single summary measure. The two-step approach taken by RSA appears useful when researchers only have access to similarity or distance matrices from computational models, and have to discern what structural representation for the data might reasonably correspond to these matrices, in order to distinguish between them. Another benefit to RSA is that, whereas latent factor models may vary in complexity, from containing a single parameter to trivially fitting an observed covariance matrix, each “model matrix” in RSA has essentially the same complexity in terms of being able to over-fit the data. However, when it is not necessary, creating a second-order summary--such as a covariance matrix--to represent a hypothesis, and then trying to work backwards to the structure of the data may be less clear than simply expressing structure at the level of the observed data. As will be considered in more detail in the General Discussion, this point has arisen when considering an analogous practice to RSA in ecology, known as the Mantel Test (see Guillot & Rousset, 2013; Legendre, Fortin, & Borcard, 2015).

General Discussion

Three common practices in fMRI research—the use of simple contrasts of within- and between-group blocks of correlations, averaging groups of items, and splitting items within a

group—reflect various beliefs about the kinds of structure in neural data that may give rise to patterns of covariance. Considering these practices through the lens of a simple restricted factor model (RFM) conveys several advantages. First, it provides a basic account that allows for two realistic, and possible to conflate, ways in which groups of patterns may differ: consistency and distinctiveness. Contrasting the average of within- and between-group correlations may be viewed as estimating this model using ordinary least squares. Second, within the RFM, averaging groups of items may be viewed as an attempt to measure a latent, group pattern. Third, it explicates how splitting items within a group may be viewed as an attempt to control for item-pair-specific or run-specific covariance. Importantly, the RFM's explicit account of measurement error allows it to incorporate the aims of strategies such as averaging, or splitting items into partitions, while also producing a model-implied covariance matrix for the entire data. This is important for common approaches to model selection, such as cross-validation, which allow many models to be empirically tested against one another.

Considering these three practices in the context of the model suggests some useful rules of thumb:

1. The within vs between contrast may yield spurious results when groups of items have different levels of consistency (e.g. one group has a higher average within-group correlation). Correlating an observed correlation matrix against a simple template matrix is equivalent to this contrast.
2. Using the average pattern within a group of items as a stand-in for an underlying group pattern is problematic, due to measurement error in the average pattern.
3. Splitting a group of items into k partitions, then correlating across partitions, does not resolve the above issues.

From a broader perspective, the RFM, and the use of latent factor analysis in general, is beneficial because it connects analyses that may seem distinct (e.g. pattern similarity analysis, inter-subject correlation) with a well-known inter-disciplinary family of methods. Rather than emphasizing the area an analysis is being applied to, much methodological work on latent factor models has focused on the specific functional properties (e.g. spatial, temporal, or spatio-temporal) of the data. This exposes the recurrence of potentially risky practices, such as cocktail-blank normalization (for review, see Garrido et al., 2013; see also, Walther et al., 2015; Diedrichsen et al., 2011), that have been addressed in other areas. An additional benefit is that detailed explanations and extension of latent factor analysis occur in textbooks across the social, health, and physical sciences (Bollen, 1989; Congdon, 2014; Lee, 2007), as well as more generally in machine learning (Barber, 2012); accordingly, many software programs and libraries exist to conduct factor analysis—and systems of linear equations more broadly--such as lavaan (Rosseel, 2012), openMX (Boker et al., 2011), and MPlus (L. K. Muthén & Muthén, 2015).

One approach which is often used as an alternative to, or in conjunction with, the covariance-based analyses above is Representational Similarity Analysis (RSA). In order to give a more complete picture, we consider the relationship between the RFM and RSA in the next section--followed by limitations, extensions, and alternatives to the RFM.

Representational Similarity Analysis (RSA)

RSA is a procedure in which dissimilarity matrices are correlated with each other (e.g., across datasets) or with a template (Kriegeskorte, 2008; Nili et al., 2014). A common RSA procedure is to use the correlations between an observed dissimilarity matrix and those generated

by competing hypotheses as a means of distinguishing between them¹⁷. As noted by Slami et al. (2013), this practice is equivalent to the Mantel Test, a procedure often used in ecology (Mantel, 1967). In ecology, this procedure takes the form of using a matrix of observed genetic distances to distinguish between “cost distances” generated by competing hypotheses. Moreover, both RSA and ecology work have extended this approach to include multiple regression using distance, or template matrices. However, serious concerns have been raised over common methodological claims and applications of this procedure in ecology¹⁸ (Graves, Beier, & Royle, 2013; Guillot & Rousset, 2013; Legendre et al., 2015). For example, a recent survey of the problem stated (Legendre et al., 2015, abstract),

[o]ur main conclusion is that Mantel tests should be restricted to questions that, in the domain of application, only concern dissimilarity matrices, and are not derived from questions that can be formulated as the analysis of the vectors and matrices from which one can compute dissimilarity matrices.

In other words, to the extent that a question can, it should be expressed in terms of the raw data. One of the central themes of this paper is that latent factor models explain how simple covariance analyses that have been expressed in terms of a similarity (e.g. covariance) matrix, may be expressed in terms of the raw data.

However, as in the ecology literature, researchers using RSA are often dealing with

¹⁷ Kriegeskorte and colleagues have suggested replacing correlation with alternative measures of association, such as Spearman’s rho or Kendall’s Tau. Such measures have also been suggested for the Mantel Test (Dietz, 1983).

¹⁸ Note that the NHST with permuted items in Example 1 is equivalent for the within-vs-between contrast, a simple template regression, and the Mantel test with a simple template.

competing hypotheses that have been expressed, a priori, as similarity matrices calculated from a separate set of data. In this case, such a procedure may be reasonable. For example, Nilli et al. (2014) consider the case where each of many competing computational models produces a dissimilarity matrix, as a “matrix model” about the structure underlying some observed data. Considering latent factor analysis and RSA side-by-side may illustrate the importance of this procedure. On the one hand, the latent factor approach starts with a representation of structure in terms of raw data, and then uses parameter estimation to derive an implied similarity matrix and goodness-of-fit. Its flexibility depends on the number of free parameters in the model, which needs to be taken into account when performing model selection. On the other hand, the RSA approach takes an implied similarity matrix as a “matrix model”, and then estimates goodness-of-fit through a measure of association. While some overall flexibility of these “matrix models” comes from the measure of association chosen, requiring that they be similarity matrices effectively constrains them to have comparable flexibility in fitting the data. In terms of association measures, Pearson’s Correlation cares about linear correspondence, as in a latent factor model with a single scaling parameter, while Kendall’s Tau only cares about relative ordering, as in a latent factor model with many order constraints put on parameters. From this perspective, RSA is useful in that it uses measures of association to provide a definition of the data structure implied by “matrix models”, which results in an estimate goodness of fit. Such an approach is valuable because it allows for the assessment of many competing hypotheses with relative simplicity, and little computational demand, by requiring hypotheses to have similar levels of complexity.

On the other hand, the success of RSA may have encouraged the overuse of “model matrices” to represent hypotheses. For instance, Example 1 brought into question the within- vs

between-group contrast, which is equivalent to procedures using a “model matrix” with ones and zeros corresponding to within- and between-group correlations, respectively. Importantly, such uses of RSA overlook the potential issues of interpretation raised by explicit data models, such as the RFM. Moreover, while models like the RFM may appear to be a constrained approach to RSA, in that it fits a covariance matrix using potentially few parameters, it should be noted that--in some ways--the RFM offers more flexibility in fitting a set of covariance matrices. For example, suppose we know data is being generated by an RFM with two-latent factors that are correlated at .7, and factor loadings constrained to be equal within each group. If using RSA with Kendall’s Tau as the measure of association, what “model matrix” best represents this RFM? The answer varies depending on the values of the loadings for each group, as that will influence the relative order of the model’s implied within- and between-group covariances. For example, three sets of values for the loadings of such an RFM result in distinct orderings of the within- and between-group correlations implied by the model (Table 4-1). An alternative set of hypotheses to those currently incorporated into RSA might usefully have this kind of complexity.

Table 4-1. Two-factor RFM allows flexible ordering of correlations

Parameters	Order of Model-Implied Correlations
$\rho = .7, \lambda_A = .3, \lambda_B = .5$	between(A,B) < within(A) < within(B)
$\rho = .7, \lambda_A = .5, \lambda_B = .3$	between(A,B) < within(B) < within(A)
$\rho = .7, \lambda_A = .5, \lambda_B = .2$	within(B) < between(A,B) < within(A)

Note. The relative orderings of model-implied correlation values in a Two-factor RFM, as a function of its parameters. Two factors, A and B, may have different loadings, but their correlation is fixed at $\rho = .7$. Model-implied correlations: $\text{between}(A,B) = \rho * \lambda_A * \lambda_B$, $\text{within}(A) = \lambda_A^2$, $\text{within}(B) = \lambda_B^2$. A “model matrix” must define an ordering for the three blocks of model-implied correlations.

Limitations and Future Directions

While the account in this paper lays out important considerations for covariance analysis, it should be noted that we did not consider the role of the temporal dynamics expected in inter-subject correlation analysis (comparisons of temporal trajectories), or spatial covariance between voxels in pattern similarity analysis (comparisons of spatial patterns). However, latent factor analysis has been adapted to account for temporal or spatial covariance in the form of dynamic factor analysis and spatial factor analysis. Below, we briefly summarize resources for these two approaches, followed by considerations of exploratory factor analysis models, implementations of factor analysis in neuroscience, and alternative approaches.

Dynamic factor analysis. This method represents factors as underlying responses along a temporal dimension, as is emphasized in inter-subject correlation analysis, and is often applied in the economics literature. One application that relates to the RFM introduced here is sometimes referred to as a dynamic factor model with block structure. Here, block structure refers to restrictions on the factor loadings, as the RFM has done by only allowing a factor to have non-zero loadings on a single group of trials / items (for review, see Bai & Wang, 2016, 2015; Moench, Ng, & Potter, 2013). In addition, dynamic factor models may incorporate time-varying parameters, meaning that the parameters of the model, such as factor loadings or variance terms, may change across time (for an EEG example, see Nakajima & West, 2016). Moreover, for research such as inter-subject correlation, external measures of how similar subjects are to one another may be incorporated within a bayesian framework, by modeling the structure of the factor loadings (spatial dynamic factor analysis; see Lopes, Gamerman, & Salazar, 2011; Lopes, Salazar, & Gamerman, 2008).

In addition, dynamic factor models are also discussed in detail in psychology, where longitudinal data may be collected across time--for example, during a series of brief experimental tasks, or over a series of years. In addition, this area is relevant to work in neuroscience, where subjects might supply, in addition to neural data, behavioral data over the course of an experimental session (see Ram, Brose, & Molenaar, 2013). Finally, it should be noted that these models share a close relationship to state space models, and have been used to model neuronal spike responses (Pillow & Scott, 2012).

Spatial factor analysis. This method represents factors as underlying responses along a spatial dimension, as is emphasized in pattern similarity analysis, and is often applied in epidemiology. This is in contrast to the spatial *dynamic* factor analysis mentioned above, where the factors are latent timecourses, and spatial dependence arises through the factor loadings (Lopes et al., 2008, p. 760). Several researchers consider cases analogous to using the RFM for pattern similarity analysis, where factor loadings model the relationship between items or trials. The key difference is that these models also account for spatial covariance between voxels, through conditional autoregressive models on the factors, the residuals, or both (Hogan & Tchernis, 2004; Mezzetti, 2012). Moreover, solutions exist for data with large numbers of items or spatial locations (i.e. voxels), which may become computationally burdensome (see Ren & Banerjee, 2013).

Exploratory factor analysis. Another important limitation of this paper is that the number of factors are considered known. In some cases, researchers may want to use an empirical approach to estimate the number of factors that does not require grouping items beforehand. This process, sometimes called exploratory factor analysis, has a long history, and is an integral part of many papers on dynamic and spatial factor analysis (). Moreover, principal components

analysis may be viewed as a special case of this type of exploratory factor analysis. Finally, certain restrictions may be placed on the covariance between latent variables, as well as factor loadings, in order to ensure that model estimation has a unique solution (see [let me think about this]).

Factor Models in Neuroscience. Many insightful adaptations of latent factor analysis, or similar models, have arisen in neuroscience. For example Klami, Virtanen, Leppäaho, & Kaski (2015) present an innovative bayesian extension of early psychometric methods for grouping of data along three-dimensions (or multiple groupings) via factor analysis (e.g. Tucker, 1958), which they apply to inter-subject correlation data. Their procedure, Group Factor Analysis, was demonstrated on functional connectivity data, where a voxel-voxel covariance matrix may be grouped into ROIs, with items being the voxels within each ROI. In this case, for a single subject, the procedure estimates factors specific to each ROI, as well as factors relating ROIs to one another. However, this Group Factor Analysis may also be applied in inter-subject correlation research, where (within an ROI) factors explain variation within and across groups of subjects. Their paper provides a review of comparable machine learning approaches, as well as a description of cases in which the model reduces to (bayesian) principle components analysis, canonical components analysis, or exploratory factor analysis.

Another interesting approach is taken in the Shared Response Model that was proposed for hyper-alignment based on inter-subject correlation (Chen et al. 2015). Within a subject, this model treats voxels as items that load on latent timecourses¹⁹. While factor loadings may differ,

¹⁹ We have discussed inter-subject correlation under the assumption that ROI selection has been used to select a single timecourse for each subject, but the logic for averaging voxels to create one timecourse per subject is similar to that addressed on averaging subject timecourses (section 2)

the latent timecourses are taken to be the shared across subjects. Thus, the factor loadings serve to align voxels across subjects, based on the shared factors. Such an approach is valuable because it provides a flexible alternative to simply averaging voxels within an anatomical ROI, which can be seen as approximating an (assumed) latent ROI timecourse, by assigning voxels equal factor loadings (Marsh et al., 2013). However, unlike the Group Factor Model, the Shared Response Model uses the same factors to explain covariance within and between all subjects. Such an approach would be interesting to compare to Group Factor Analysis, when each subject is represented as group of voxel timecourses, and factors may explain variation within only a single subject, or between sets of subjects.

Alternative Approaches. While this paper has focused on factor analysis, it should be emphasized that it is not the only approach, and there are many thoughtful considerations of how to measure and test pattern similarity in the literature. For example, Allefeld & Haynes (2014) consider pattern similarity analysis from the perspective of multivariate analysis of variance (MANOVA), which accounts for the spatial covariance between voxels (for similarities between MANOVA and latent factor models, see Cole, Maxwell, Arvey, & Salas, 1993). While this paper considered three common practices in pattern similarity analysis through the perspective of latent factor analysis, it is clear that in all cases, they would benefit from the use of any explicit modeling approach.

Conclusion

Covariance analyses are an important tool for using neuroimaging to understanding the processing and representation of information in the brain. While simple approaches, such as contrasting blocks of correlations, averaging across items, and correlating across splits of the data, have an intuitive appeal, care needs to be taken in considering their inferential value and

assumptions. While a null hypothesis significance test may yield low p-values, the interpretation of this test depends on a clear understanding of what null hypothesis is being employed. Latent factor models allow explicit consideration of simple approaches and the hypotheses they may test. The RFM provides a general rationale for these approaches, encouraging separate consideration of two underlying properties of interest, consistency and pattern similarity. Future work is needed to derive robust procedures that maintain the simple, intuitive appeal of previous approaches, while preserving interpretation that is consistent with simple simulations and models of the data.

Abrams, D. A., Ryali, S., Chen, T., Chordia, P., Khouzam, A., Levitin, D. J., & Menon, V. (2013).

Inter-subject synchronization of brain responses during natural music listening.

European Journal of Neuroscience, 37(9), 1458–1469. <http://doi.org/10.1111/ejn.12173>

Aho, K., Derryberry, D., & Peterson, T. (2014). Model selection for ecologists: the worldviews of

AIC and BIC. *Ecology*, 95(3), 631–636.

Allefeld, C., & Haynes, J.-D. (2014). Searchlight-based multi-voxel pattern analysis of fMRI by cross-validated MANOVA. *NeuroImage*, 89, 345–357.

<http://doi.org/10.1016/j.neuroimage.2013.11.043>

Aly, M., & Turk-Browne, N. B. (2015). Attention Stabilizes Representations in the Human

Hippocampus. *Cerebral Cortex*. <http://doi.org/10.1093/cercor/bhv041>

Ames, D. L., Honey, C. J., Chow, M. A., Todorov, A., & Hasson, U. (2015). Contextual alignment of cognitive and neural dynamics. *Journal of Cognitive Neuroscience*, 27(4), 655–664.

http://doi.org/10.1162/jocn_a_00728

Bach, F. R., & Jordan, M. I. (2005). A probabilistic interpretation of canonical correlation analysis. Retrieved from <http://www.stat.berkeley.edu/~jordan/688.pdf>

Bai, J., & Wang, P. (2015). Identification and Bayesian Estimation of Dynamic Factor Models.

Journal of Business & Economic Statistics, 33(2), 221–240.

<http://doi.org/10.1080/07350015.2014.941467>

Bai, J., & Wang, P. (2016). Econometric Analysis of Large Factor Models. *Annual Review of*

Economics, 8(1), null. <http://doi.org/10.1146/annurev-economics-080315-015356>

Barber, D. (2012). *Bayesian reasoning and machine learning*. Cambridge University Press.

- Ben-Yakov, A., Honey, C. J., Lerner, Y., & Hasson, U. (2012). Loss of reliable temporal structure in event-related averaging of naturalistic stimuli. *NeuroImage*, 63(1), 501–506.
<http://doi.org/10.1016/j.neuroimage.2012.07.008>
- Bien, J., & Tibshirani, R. J. (2011). Sparse estimation of a covariance matrix. *Biometrika*, 98(4), 807–820. <http://doi.org/10.1093/biomet/asr054>
- Boettiger, C. (2015). An introduction to Docker for reproducible research. *ACM SIGOPS Operating Systems Review*, 49(1), 71–79.
- Boker, S., Neale, M., Maes, H., Wilde, M., Spiegel, M., Brick, T., ... others. (2011). OpenMx: an open source extended structural equation modeling framework. *Psychometrika*, 76(2), 306–317.
- Bollen, K. (1989). *Structural equations with latent variables*. Wiley.
- Bollen, K., & Lennox, R. (1991). Conventional wisdom on measurement: A structural equation perspective. *Psychological Bulletin*, 110(2), 305.
- Borsboom, D. (2006). The attack of the psychometricians. *Psychometrika*, 71(3), 425–440.
<http://doi.org/10.1007/s11336-006-1447-6>
- Brants, M., Baeck, A., Wagemans, J., & Op de Beeck, H. P. (2011). Multiple scales of organization for object selectivity in ventral visual cortex. *NeuroImage*, 56(3), 1372–1381. <http://doi.org/10.1016/j.neuroimage.2011.02.079>
- Breiman, L. (2001). Statistical modeling: The two cultures (with comments and a rejoinder by the author). *Statistical Science*, 16(3), 199–231.

- Bullmore, E., Horwitz, B., Honey, G., Brammer, M., Williams, S., & Sharma, T. (2000). How Good Is Good Enough in Path Analysis of fMRI Data? *NeuroImage*, 11(4), 289–301.
<http://doi.org/10.1006/nimg.2000.0544>
- Chen, G., Shin, Y.-W., Taylor, P. A., Glen, D. R., Reynolds, R. C., Israel, R. B., & Cox, R. W. (2016). Untangling the relatedness among correlations, part I: Nonparametric approaches to inter-subject correlation analysis at the group level. *NeuroImage*.
<http://doi.org/10.1016/j.neuroimage.2016.05.023>
- Chen, J., Honey, C. J., Simony, E., Arcaro, M. J., Norman, K. A., & Hasson, U. (2015). Accessing Real-Life Episodic Information from Minutes versus Hours Earlier Modulates Hippocampal and High-Order Cortical Dynamics. *Cerebral Cortex*, bhv155.
<http://doi.org/10.1093/cercor/bhv155>
- Cole, D. A., Maxwell, S. E., Arvey, R., & Salas, E. (1993). Multivariate group comparisons of variable systems: MANOVA and structural equation modeling. *Psychological Bulletin*, 114(1), 174–184. <http://doi.org/10.1037/0033-2909.114.1.174>
- Congdon, P. (2014). *Applied bayesian modelling* (Vol. 595). John Wiley & Sons.
- Cronbach, L. J. (1942). Studies of acquiescence as a factor in the true-false test. *Journal of Educational Psychology*, 33(6), 401–415. <http://doi.org/10.1037/h0054677>
- Diedrichsen, J., Ridgway, G. R., Friston, K. J., & Wiestler, T. (2011). Comparing the similarity and spatial structure of neural representations: A pattern-component model. *NeuroImage*, 55(4), 1665–1678. <http://doi.org/10.1016/j.neuroimage.2011.01.044>

- Diedrichsen, J., Wiestler, T., & Ejaz, N. (2013). A multivariate method to determine the dimensionality of neural representation from population activity. *NeuroImage*, 76, 225–235. <http://doi.org/10.1016/j.neuroimage.2013.02.062>
- Dietz, E. J. (1983). Permutation Tests for Association between Two Distance Matrices. *Systematic Zoology*, 32(1), 21–26.
- Eid, M. (2000). A multitrait-multimethod model with minimal assumptions. *Psychometrika*, 65(2), 241–261.
- Ezzyat, Y., & Davachi, L. (2014). Similarity Breeds Proximity: Pattern Similarity within and across Contexts Is Related to Later Mnemonic Judgments of Temporal Proximity. *Neuron*, 81(5), 1179–1189. <http://doi.org/10.1016/j.neuron.2014.01.042>
- Favila, S. E., Chanales, A. J. H., & Kuhl, B. A. (2016). Experience-dependent hippocampal pattern differentiation prevents interference during subsequent learning. *Nat Commun*, 7. Retrieved from <http://dx.doi.org/10.1038/ncomms11066>
- Furr, R. M., & Bacharach, V. R. (2013). *Psychometrics: an introduction*. Sage.
- Garrido, L., Vaziri-Pashkam, M., Nakayama, K., & Wilmer, J. (2013). The consequences of subtracting the mean pattern in fMRI multivariate correlation analyses. *Frontiers in Neuroscience*, 7. <http://doi.org/10.3389/fnins.2013.00174>
- Geiser, C., Bishop, J., & Lockhart, G. (2015). Collapsing factors in multitrait-multimethod models: examining consequences of a mismatch between measurement design and model. *Frontiers in Psychology*, 6. <http://doi.org/10.3389/fpsyg.2015.00946>

- Graham, J. M. (2006). Congeneric and (Essentially) Tau-Equivalent Estimates of Score Reliability: What They Are and How to Use Them. *Educational and Psychological Measurement*, 66(6), 930–944. <http://doi.org/10.1177/0013164406288165>
- Graves, T. A., Beier, P., & Royle, J. A. (2013). Current approaches using genetic distances produce poor estimates of landscape resistance to interindividual dispersal. *Molecular Ecology*, 22(15), 3888–3903. <http://doi.org/10.1111/mec.12348>
- Green, S. B., & Yang, Y. (2009). Reliability of summed item scores using structural equation modeling: An alternative to coefficient alpha. *Psychometrika*, 74(1), 155–167.
- Greicius, M. D., Krasnow, B., Reiss, A. L., & Menon, V. (2003). Functional connectivity in the resting brain: a network analysis of the default mode hypothesis. *Proceedings of the National Academy of Sciences*, 100(1), 253–258.
- Guillot, G., & Rousset, F. (2013). Dismantling the Mantel tests. *Methods in Ecology and Evolution*, 4(4), 336–344.
- Hasson, U. (2004). Intersubject synchronization of cortical activity during natural vision. *Science*, 303(5664), 1634–1640. <http://doi.org/10.1126/science.1089506>
- Hasson, U., Avidan, G., Gelbard, H., Vallines, I., Harel, M., Minshew, N., & Behrmann, M. (2009). Shared and idiosyncratic cortical activation patterns in autism revealed under continuous real-life viewing conditions. *Autism Research*, 2(4), 220–231. <http://doi.org/10.1002/aur.89>
- Hasson, U., Malach, R., & Heeger, D. J. (2010). Reliability of cortical activity during natural stimulation. *Trends in Cognitive Sciences*, 14(1), 40–48. <http://doi.org/10.1016/j.tics.2009.10.011>

- Hasson, U., Yang, E., Vallines, I., Heeger, D. J., & Rubin, N. (2008). A hierarchy of temporal receptive windows in human cortex. *The Journal of Neuroscience*, 28(10), 2539–2550.
- Haxby, J. V. (2001). Distributed and Overlapping Representations of Faces and Objects in Ventral Temporal Cortex. *Science*, 293(5539), 2425–2430.
<http://doi.org/10.1126/science.1063736>
- Haxby, J. V. (2012). Multivariate pattern analysis of fMRI: The early beginnings. *NeuroImage*, 62(2), 852–855. <http://doi.org/10.1016/j.neuroimage.2012.03.016>
- Herbec, A., Kauppi, J.-P., Jola, C., Tohka, J., & Pollick, F. E. (2015). Differences in fMRI intersubject correlation while viewing unedited and edited videos of dance performance. *Cortex*, 71, 341–348.
- Hogan, J. W., & Tchernis, R. (2004). Bayesian Factor Analysis for Spatially Correlated Data, With Application to Summarizing Area-Level Material Deprivation From Census Data. *Journal of the American Statistical Association*, 99(466), 314–324.
<http://doi.org/10.1198/016214504000000296>
- Honey, C. J., Thompson, C. R., Lerner, Y., & Hasson, U. (2012a). Not Lost in Translation: Neural Responses Shared Across Languages. *Journal of Neuroscience*, 32(44), 15277–15283.
<http://doi.org/10.1523/JNEUROSCI.1800-12.2012>
- Honey, C. J., Thompson, C. R., Lerner, Y., & Hasson, U. (2012b). Not Lost in Translation: Neural Responses Shared Across Languages. *Journal of Neuroscience*, 32(44), 15277–15283.
<http://doi.org/10.1523/JNEUROSCI.1800-12.2012>

- Huang, J. Z., Liu, N., Pourahmadi, M., & Liu, L. (2006). Covariance matrix selection and estimation via penalised normal likelihood. *Biometrika*, 93(1), 85–98.
<http://doi.org/10.1093/biomet/93.1.85>
- Iordan, M. C., Greene, M. R., Beck, D. M., & Fei-Fei, L. (2015). Basic Level Category Structure Emerges Gradually across Human Ventral Visual Cortex. *Journal of Cognitive Neuroscience*, 27(7), 1427–1446. http://doi.org/10.1162/jocn_a_00790
- Jääskeläinen, I. P., Koskentalo, K., Balk, M. H., Autti, T., Kauramäki, J., Pomren, C., & Sams, M. (2008). Inter-Subject Synchronization of Prefrontal Cortex Hemodynamic Activity During Natural Viewing. *The Open Neuroimaging Journal*, 2, 14–19.
<http://doi.org/10.2174/18744440000802010014>
- Karlsson Wirebring, L., Wiklund-Hornqvist, C., Eriksson, J., Andersson, M., Jonsson, B., & Nyberg, L. (2015). Lesser Neural Pattern Similarity across Repeated Tests Is Associated with Better Long-Term Memory Retention. *Journal of Neuroscience*, 35(26), 9595–9602.
<http://doi.org/10.1523/JNEUROSCI.3550-14.2015>
- Kauppi. (2010). Inter-subject correlation of brain hemodynamic responses during watching a movie: localization in space and frequency. *Frontiers in Neuroinformatics*.
<http://doi.org/10.3389/fninf.2010.00005>
- Kauppi, J.-P., Pajula, J., & Tohka, J. (2014). A versatile software package for inter-subject correlation based analyses of fMRI. *Frontiers in Neuroinformatics*, 8.
<http://doi.org/10.3389/fninf.2014.00002>

- Kievit, R. A., Frankenhuys, W. E., Waldorp, L. J., & Borsboom, D. (2013). Simpson's paradox in psychological science: a practical guide. *Frontiers in Psychology, 4*.
<http://doi.org/10.3389/fpsyg.2013.00513>
- Klami, A., Virtanen, S., Leppäaho, E., & Kaski, S. (2015). Group factor analysis. *IEEE Transactions on Neural Networks and Learning Systems, 26*(9), 2136–2147.
- Knafl, G. J., & Grey, M. (2007). Factor analysis model evaluation through likelihood cross-validation. *Statistical Methods in Medical Research, 16*(2), 77–102.
- Kriegeskorte, N. (2008). Representational similarity analysis – connecting the branches of systems neuroscience. *Frontiers in Systems Neuroscience*.
<http://doi.org/10.3389/neuro.06.004.2008>
- Kriegeskorte, N., Mur, M., Ruff, D. A., Kiani, R., Bodurka, J., Esteky, H., ... Bandettini, P. A. (2008). Matching Categorical Object Representations in Inferior Temporal Cortex of Man and Monkey. *Neuron, 60*(6), 1126–1141. <http://doi.org/10.1016/j.neuron.2008.10.043>
- Krijnen, W. P. (1996). Algorithms for unweighted least-squares factor analysis. *Computational Statistics & Data Analysis, 21*(2), 133–147. [http://doi.org/10.1016/0167-9473\(95\)00011-9](http://doi.org/10.1016/0167-9473(95)00011-9)
- Kruschke, J. K. (2013). Bayesian estimation supersedes the t test. *Journal of Experimental Psychology: General, 142*(2), 573–603. <http://doi.org/10.1037/a0029146>
- Lahnakoski, J. M., Glerean, E., Jääskeläinen, I. P., Hyönä, J., Hari, R., Sams, M., & Nummenmaa, L. (2014). Synchronous brain activity across individuals underlies shared psychological perspectives. *NeuroImage, 100*, 316–324.
<http://doi.org/10.1016/j.neuroimage.2014.06.022>

LaRocque, K. F., Smith, M. E., Carr, V. A., Witthoft, N., Grill-Spector, K., & Wagner, A. D. (2013).

Global Similarity and Pattern Separation in the Human Medial Temporal Lobe Predict Subsequent Memory. *Journal of Neuroscience*, 33(13), 5466–5474.

<http://doi.org/10.1523/JNEUROSCI.4293-12.2013>

Lee, S.-Y. (2007). *Structural equation modeling: A Bayesian approach* (Vol. 711). John Wiley & Sons.

Lee, S.-Y., & Song, X.-Y. (2004). Evaluation of the Bayesian and Maximum Likelihood Approaches in Analyzing Structural Equation Models with Small Sample Sizes. *Multivariate Behavioral Research*, 39(4), 653–686. http://doi.org/10.1207/s15327906mbr3904_4

Legendre, P., Fortin, M.-J., & Borcard, D. (2015). Should the Mantel test be used in spatial analysis? *Methods in Ecology and Evolution*, 6(11), 1239–1247.

Le Rest, K., Pinaud, D., Monestiez, P., Chadoeuf, J., & Bretagnolle, V. (2014). Spatial leave-one-out cross-validation for variable selection in the presence of spatial autocorrelation: Spatial leave-one-out cross-validation. *Global Ecology and Biogeography*, 23(7), 811–820. <http://doi.org/10.1111/geb.12161>

Lerner, Y., Honey, C. J., Katkov, M., & Hasson, U. (2014). Temporal scaling of neural responses to compressed and dilated natural speech. *Journal of Neurophysiology*, 111(12), 2433–2444. <http://doi.org/10.1152/jn.00497.2013>

Lerner, Y., Honey, C. J., Silbert, L. J., & Hasson, U. (2011). Topographic mapping of a hierarchy of temporal receptive windows using a narrated story. *The Journal of Neuroscience*, 31(8), 2906–2915.

- Little, T. D., Rhemtulla, M., Gibson, K., & Schoemann, A. M. (2013). Why the items versus parcels controversy needn't be one. *Psychological Methods, 18*(3), 285–300.
<http://doi.org/10.1037/a0033266>
- Looser, C. E., Guntupalli, J. S., & Wheatley, T. (2013). Multivoxel patterns in face-sensitive temporal regions reveal an encoding schema based on detecting life in a face. *Social Cognitive and Affective Neuroscience, 8*(7), 799–805.
<http://doi.org/10.1093/scan/nss078>
- Lopes, H. F., Gamerman, D., & Salazar, E. (2011). Generalized spatial dynamic factor models. *Computational Statistics & Data Analysis, 55*(3), 1319–1330.
<http://doi.org/10.1016/j.csda.2010.09.020>
- Lopes, H. F., Salazar, E., & Gamerman, D. (2008). Spatial dynamic factor analysis. *Bayesian Analysis, 3*(4), 759–792. <http://doi.org/10.1214/08-BA329>
- Lord, F. M., Novick, M. R., & Birnbaum, A. (1968). Statistical theories of mental test scores.
- Maas, C. J. M., Lensvelt-Mulders, G. J. L. M., & Hox, J. J. (2009). A Multilevel Multitrait-Multimethod Analysis. *Methodology, 5*(3), 72–77. <http://doi.org/10.1027/1614-2241.5.3.72>
- Mantel, N. (1967). The detection of disease clustering and a generalized regression approach. *Cancer Research, 27*(2 Part 1), 209–220.
- Marsh, H. W., Lüdtke, O., Nagengast, B., Morin, A. J. S., & Von Davier, M. (2013). Why item parcels are (almost) never appropriate: Two wrongs do not make a right—Camouflaging misspecification with item parcels in CFA models. *Psychological Methods, 18*(3), 257–284. <http://doi.org/10.1037/a0032773>

- Maydeu-Olivares, A., & Coffman, D. L. (2006). Random intercept item factor analysis. *Psychological Methods*, 11(4), 344–362. <http://doi.org/10.1037/1082-989X.11.4.344>
- Mellenbergh, G. J. (1994). Generalized linear item response theory. *Psychological Bulletin*, 115(2), 300.
- Mezzetti, M. (2012). Bayesian factor analysis for spatially correlated data: application to cancer incidence data in Scotland. *Statistical Methods & Applications*, 21(1), 49–74. <http://doi.org/10.1007/s10260-011-0177-9>
- Moench, E., Ng, S., & Potter, S. (2013). Dynamic Hierarchical Factor Models. *Review of Economics and Statistics*, 95(5), 1811–1817. http://doi.org/10.1162/REST_a_00359
- Monin, B., & Oppenheimer, D. M. (2005). Correlated averages vs. averaged correlations: Demonstrating the warm glow heuristic beyond aggregation. *Social Cognition*, 23(3), 257.
- Mumford, J. A., & Poldrack, R. A. (2007). Modeling group fMRI data. *Social Cognitive and Affective Neuroscience*, 2(3), 251–257. <http://doi.org/10.1093/scan/nsm019>
- Murphy, K., Birn, R. M., Handwerker, D. A., Jones, T. B., & Bandettini, P. A. (2009). The impact of global signal regression on resting state correlations: Are anti-correlated networks introduced? *NeuroImage*, 44(3), 893–905. <http://doi.org/10.1016/j.neuroimage.2008.09.036>
- Muthén, B., & Asparouhov, T. (2012). Bayesian structural equation modeling: A more flexible representation of substantive theory. *Psychological Methods*, 17(3), 313–335. <http://doi.org/10.1037/a0026802>

- Muthén, L. K., & Muthén, B. O. (2015). Mplus User's Guide. Seventh Edition. Retrieved from <https://www.statmodel.com>
- Nakajima, J., & West, M. (2016). Dynamics and sparsity in latent threshold factor models: A study in multivariate EEG signal processing. *arXiv Preprint arXiv:1606.08292*. Retrieved from <http://arxiv.org/abs/1606.08292>
- Nili, H., Wingfield, C., Walther, A., Su, L., Marslen-Wilson, W., & Kriegeskorte, N. (2014). A Toolbox for Representational Similarity Analysis. *PLoS Computational Biology*, 10(4), e1003553. <http://doi.org/10.1371/journal.pcbi.1003553>
- Novick, M. R. (1966). The axioms and principal results of classical test theory. *Journal of Mathematical Psychology*, 3(1), 1 – 18. [http://doi.org/http://dx.doi.org/10.1016/0022-2496\(66\)90002-2](http://doi.org/http://dx.doi.org/10.1016/0022-2496(66)90002-2)
- Nummenmaa, L., Saarimäki, H., Glerean, E., Gotsopoulos, A., Jääskeläinen, I. P., Hari, R., & Sams, M. (2014). Emotional speech synchronizes brains across listeners and engages large-scale dynamic brain networks. *NeuroImage*, 102, 498–509. <http://doi.org/10.1016/j.neuroimage.2014.07.063>
- Nummenmaa, L., Smirnov, D., Lahnakoski, J. M., Glerean, E., Jaaskelainen, I. P., Sams, M., & Hari, R. (2014). Mental Action Simulation Synchronizes Action-Observation Circuits across Individuals. *Journal of Neuroscience*, 34(3), 748–757. <http://doi.org/10.1523/JNEUROSCI.0352-13.2014>
- Op de Beeck, H. P., Brants, M., Baeck, A., & Wagemans, J. (2010). Distributed subordinate specificity for bodies, faces, and buildings in human ventral visual cortex. *NeuroImage*, 49(4), 3414–3425. <http://doi.org/10.1016/j.neuroimage.2009.11.022>

- Pearl, J. (2014). Comment: Understanding Simpson's Paradox. *The American Statistician*, 68(1), 8–13. <http://doi.org/10.1080/00031305.2014.876829>
- Pillow, J. W., & Scott, J. G. (2012). Fully Bayesian inference for neural models with negative-binomial spiking. In *NIPS* (pp. 1907–1915). Retrieved from http://pillowlab.princeton.edu/pubs/Pillow_Scott_NIPS2012.pdf
- Poldrack, R. A., Mumford, J. A., & Nichols, T. E. (2011). *Handbook of functional MRI data analysis*. New York: Cambridge University Press.
- Pollet, T. V., Stulp, G., Henzi, S. P., & Barrett, L. (2015). Taking the aggravation out of data aggregation: A conceptual guide to dealing with statistical issues related to the pooling of individual-level observational data: Data Aggregation and the Ecological Fallacy. *American Journal of Primatology*, 77(7), 727–740. <http://doi.org/10.1002/ajp.22405>
- Ram, N., Brose, A., & Molenaar, P. C. (2013). Dynamic factor analysis: Modeling person-specific process. *The Oxford Handbook of Quantitative Methods*, 2, 441–457.
- Regev, M., Honey, C. J., Simony, E., & Hasson, U. (2013). Selective and Invariant Neural Responses to Spoken and Written Narratives. *Journal of Neuroscience*, 33(40), 15978–15988. <http://doi.org/10.1523/JNEUROSCI.1580-13.2013>
- Ren, Q., & Banerjee, S. (2013). Hierarchical Factor Models for Large Spatially Misaligned Data: A Low-Rank Predictive Process Approach: Hierarchical Factor Models for Large Spatially Misaligned Data. *Biometrics*, 69(1), 19–30. <http://doi.org/10.1111/j.1541-0420.2012.01832.x>
- Rice, G. E., Watson, D. M., Hartley, T., & Andrews, T. J. (2014). Low-Level Image Properties of Visual Objects Predict Patterns of Neural Response across Category-Selective Regions of

- the Ventral Visual Pathway. *Journal of Neuroscience*, 34(26), 8837–8844.
<http://doi.org/10.1523/JNEUROSCI.5265-13.2014>
- Rosseel, Y. (2012). lavaan: An R Package for Structural Equation Modeling. *Journal of Statistical Software*, 48(2), 1–36.
- Ross, S. M. (2014). *Introduction to probability and statistics for engineers and scientists*. Academic Press.
- Rubin, D. B., & Thayer, D. T. (1982). EM algorithms for ML factor analysis. *Psychometrika*, 47(1), 69–76. <http://doi.org/10.1007/BF02293851>
- Ruxton, G. D. (2006). The unequal variance t-test is an underused alternative to Student's t-test and the Mann-Whitney U test. *Behavioral Ecology*, 17(4), 688–690.
<http://doi.org/10.1093/beheco/ark016>
- Saad, Z. S., Gotts, S. J., Murphy, K., Chen, G., Jo, H. J., Martin, A., & Cox, R. W. (2012). Trouble at Rest: How Correlation Patterns and Group Differences Become Distorted After Global Signal Regression. *Brain Connectivity*, 2(1), 25–32.
<http://doi.org/10.1089/brain.2012.0080>
- Salmi, J., Roine, U., Glerean, E., Lahnakoski, J., Nieminen-von Wendt, T., Tani, P., ... Sams, M. (2013). The brains of high functioning autistic individuals do not synchronize with those of others. *NeuroImage: Clinical*, 3, 489–497. <http://doi.org/10.1016/j.nicl.2013.10.011>
- Savalei, V., & Falk, C. F. (2014). Recovering Substantive Factor Loadings in the Presence of Acquiescence Bias: A Comparison of Three Approaches. *Multivariate Behavioral Research*, 49(5), 407–424. <http://doi.org/10.1080/00273171.2014.931800>

Schapiro, A. C., Rogers, T. T., Cordova, N. I., Turk-Browne, N. B., & Botvinick, M. M. (2013).

Neural representations of events arise from temporal community structure. *Nature Neuroscience*, 16(4), 486–492. <http://doi.org/10.1038/nn.3331>

Sijtsma, K. (2009a). Correcting Fallacies in Validity, Reliability, and Classification. *International Journal of Testing*, 9(3), 167–194. <http://doi.org/10.1080/15305050903106883>

Sijtsma, K. (2009b). On the use, the misuse, and the very limited usefulness of Cronbach's alpha. *Psychometrika*, 74(1), 107–120.

Simony, E., Honey, C. J., Chen, J., Lositsky, O., Yeshurun, Y., Wiesel, A., & Hasson, U. (2016).

Dynamic reconfiguration of the default mode network during narrative comprehension. *Nat Commun*, 7. Retrieved from <http://dx.doi.org/10.1038/ncomms12141>

Spearman, C. (1910). Correlation calculated from faulty data. *British Journal of Psychology*, 1904-1920, 3(3), 271–295.

Stephens, G. J., Silbert, L. J., & Hasson, U. (2010). Speaker-listener neural coupling underlies successful communication. *Proceedings of the National Academy of Sciences*, 107(32), 14425–14430. <http://doi.org/10.1073/pnas.1008662107>

Tucker, L. R. (1958). An inter-battery method of factor analysis. *Psychometrika*, 23(2), 111–136.

van den Heuvel, M. P., & Hulshoff Pol, H. E. (2010). Exploring the brain network: A review on resting-state fMRI functional connectivity. *European Neuropsychopharmacology*, 20(8), 519–534. <http://doi.org/10.1016/j.euroneuro.2010.03.008>

van de Schoot, R., Kluytmans, A., Tummers, L., Lugtig, P., Hox, J., & Muthén, B. (2013). Facing off with Scylla and Charybdis: a comparison of scalar, partial, and the novel possibility of

approximate measurement invariance. *Frontiers in Psychology*, 4.

<http://doi.org/10.3389/fpsyg.2013.00770>

Van De Schoot, R., Schmidt, P., De Beuckelaer, A., Lek, K., & Zondervan-Zwijnenburg, M. (2015).

Editorial "Measurement Invariance." *Frontiers in Psychology*, 6(1064).

<http://doi.org/10.3389/fpsyg.2015.01064>

Walther, A., Nili, H., Ejaz, N., Alink, A., Kriegeskorte, N., & Diedrichsen, J. (2015). Reliability of

dissimilarity measures for multi-voxel pattern analysis. *NeuroImage*.

<http://doi.org/10.1016/j.neuroimage.2015.12.012>

Wasserstein, R. L., & Lazar, N. A. (2016). The ASA's statement on p-values: context, process, and purpose. *The American Statistician*, 00–00.

<http://doi.org/10.1080/00031305.2016.1154108>

Wickham, H. (2007). Reshaping Data with the reshape Package. *Journal of Statistical Software*, 21(12), 1–20.

Wickham, H., & Francois, R. (n.d.). *dplyr: A Grammar of Data Manipulation*. Retrieved from

<https://github.com/hadley/dplyr>

Wolosin, S. M., Zeithamova, D., & Preston, A. R. (2013). Distributed hippocampal patterns that discriminate reward context are associated with enhanced associative binding. *Journal of Experimental Psychology: General*, 142(4), 1264–1276.

<http://doi.org/10.1037/a0033609>

Ximénez, C. (2009). Recovery of weak factor loadings in confirmatory factor analysis under conditions of model misspecification. *Behavior Research Methods*, 41(4), 1038–1052.

<http://doi.org/10.3758/BRM.41.4.1038>

Xue, G., Dong, Q., Chen, C., Lu, Z., Mumford, J. A., & Poldrack, R. A. (2010). Greater Neural Pattern Similarity Across Repetitions Is Associated with Better Memory. *Science*, 330(6000), 97–101. <http://doi.org/10.1126/science.1193125>