UNIVERSITY OF CALIFORNIA SAN DIEGO

Cognitive Consequences of Physical Assembly

A dissertation submitted in partial satisfaction of the
requirements for the degree Doctor of Philosophy

in

Cognitive Science

by

William P. McCarthy

Committee in charge:

Professor Judith E. Fan, Co-Chair
Professor David J. Kirsh, Co-Chair
Professor Timothy F. Brady
Professor Anastasia Kiyonaga
Professor Haijun Xia

2024

The Dissertation of William P. McCarthy is approved, and it is acceptable in quality and form for publication on microfilm and electronically.

University of California San Diego

2024

DEDICATION

For Joanna Beazley.

EPIGRAPH

"What I cannot create, I do not understand"
Richard Feynman (1988)

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

Over the course of my PhD I have had the absurdly good fortune to work exclusively with scientists who are both brilliant and kind. None exhibit these virtues more than Judy Fan, who mentors her students with unparalleled thoughtfulness and generosity, and by being the exemplar of a dedicated and assiduous scientist. It is hard to quantify the impact Judy has had on me as a researcher and as a person (although my impulse to define metrics to measure this does say something to that effect). For this, and for the heroic amount of effort she puts into everything, I extend my most heartfelt thanks.

I would like express a deep gratitude to David Kirsh, for his generous and patient guidance, and for sharpening my thinking with incisive questions and admirable insistence on understanding. David has had a profound impact on the way that I think about cognition, and I count myself extremely lucky to have learned from a truly original thinker. I would also like to thank my committee– Tim Brady, Anastasia Kiyonaga, Haijun Xia– as well as Ed Vul and Marcelo Mattar, for invaluable feedback that nudged these projects in more meaningful directions.

Thank you to my incredible collaborators, especially Lio Wong, Gabe Grand, Yoni Friedman, Jacob Andreas, and Yewen Pu, for making science more fun. Special thanks to Robert Hawkins, for his generous mentorship on all things language.

I am immensely proud to have caught the first wave of the Cognitive Tools Lab: Holly Huey, Haoliang Wang, Felix Binder, Sebastian Holt, Cameron Holdaway, Erik Brockbank, Lauren Oey, Hannah Lloyd, Justin Yang, Zoe Tait, Xuanchen Lu, Arnav Verma, Jack Terwilliger, and Jacob Schenberg, thank you for setting such a high standard and for your support. Thanks also to my colleagues in Cognitive Science, particularly James Michaelov, Stephan Kaufhold, and Michael Allen.

I doubt I would have made it this far if it were not my fellow surfers, musicians, brewery crawlers and burrito connoisseurs that made San Diego feel like home. I am particularly grateful to Philip Belzeski, who fits all of these categories, and is as good a

friend and flatmate as anyone can hope for. Thanks also to my friends back in England, for making it feel like I never left when I visit home, and for essential emotional support.

An enormous thank you to my partner, Katja Lazar, for her sagacious advice, insatiable humor, and for helping me rediscover the joy of making things.

Finally, mostly, I would like to thank my family. I would not be here if it were not for you.

Chapter 1, in full, is a reprint of material as it appears in McCarthy, W. P., Kirsh, D., & Fan, J. E. (2023). Consistency and variation in reasoning about physical assembly. *Cognitive Science*, *47*(12), e13397. An earlier version of this project was published as McCarthy, W., Kirsh, D., & Fan, J. (2020). Learning to build physical structures better over time. *Proceedings of the 42nd Annual Meeting of the Cognitive Science Society*. The dissertation author was the primary investigator and author of this material.

Chapter 2, in full, is currently being prepared for submission for publication of the material. Earlier versions of this project were published as McCarthy, W. P., Anderson, S. P., & Fan, J. E. (2024). How does assembling an object affect memory for it? [In press]. *Proceedings of the Annual Meeting of the Cognitive Science Society*, *46*, and as McCarthy, W. P., & Fan, J. E. (2023). Exploring the impact of a constructive encoding task on visual recognition memory. *Journal of Vision*, *23*(9), 5977–5977. The dissertation author was the primary investigator and author of this material.

Chapter 3, in full, is in currently in review for publication of the material. An earlier version of this project was published as *McCarthy, W. P., *Hawkins, R. D., Wang, H., Holdaway, C., & Fan, J. E. (2021). Learning to communicate about shared procedural abstractions. *Proceedings of the Annual Meeting of the Cognitive Science Society*, *43*. The dissertation author was the primary investigator and author of this material.

VITA

2016	Bachelor of Computer Science and Philosophy, University of Oxford
2017	Master of Computer Science and Philosophy, University of Oxford
2024	Doctor of Philosophy, Cognitive Science, University of California San Diego

PUBLICATIONS

McCarthy, W. P., Anderson, S. P., & Fan, J. E. (n.d.). How does assembling an object affect memory for it? [In preparation].

*McCarthy, W. P., *Hawkins, R. D., Wang, H., & Fan, J. E. (n.d.). Learning to communicate about shared procedural abstractions [In review].

McCarthy, W. P., Anderson, S. P., & Fan, J. E. (2024). How does assembling an object affect memory for it? [In press]. *Proceedings of the Annual Meeting of the Cognitive Science Society*, *46*.

McCarthy, W. P., & Fan, J. E. (2023). Exploring the impact of a constructive encoding task on visual recognition memory. *Journal of Vision*, *23*(9), 5977–5977.

McCarthy, W. P., Kirsh, D., & Fan, J. E. (2023). Consistency and variation in reasoning about physical assembly. *Cognitive Science*, *47*(12), e13397.

*Wong, C., *McCarthy, W. P., *Grand, G., Friedman, Y., Tenenbaum, J. B., Andreas, J., Hawkins, R. D., & Fan, J. E. (2022). Identifying concept libraries from language about object structure. *Proceedings of the Annual Meeting of the Cognitive Science Society*, *44*.

*McCarthy, W. P., *Hawkins, R. D., Wang, H., Holdaway, C., & Fan, J. E. (2021). Learning to communicate about shared procedural abstractions. *Proceedings of the Annual Meeting of the Cognitive Science Society*, *43*.

McCarthy, W. P., Mattar, M. G., Kirsh, D., & Fan, J. E. (2021). Connecting perceptual and procedural abstractions in physical construction. *Proceedings of the Annual Meeting of the Cognitive Science Society*, *43*.

McCarthy, W., Kirsh, D., & Fan, J. (2020). Learning to build physical structures better over time. *Proc. 42nd Annu. Meet. Cogn. Sci. Soc.*

ABSTRACT OF THE DISSERTATION

Cognitive Consequences of Physical Assembly

by

William P. McCarthy

Doctor of Philosophy in Cognitive Science

University of California San Diego, 2024

Professor Judith E. Fan, Co-Chair
Professor David J. Kirsh, Co-Chair

The modern world is densely populated by physical structures that were designed and made by people, from transient arrangements like stacks of books and sandwiches, to enduring constructions like bridges and skyscrapers. *Physical assembly*– the construction of a new object from existing parts– accounts for some of the most complex acts of human cognition. What are the core cognitive processes that underlie this ability? The study of physical assembly presents unique opportunities and challenges because it relies on interactions between multiple cognitive processes, including perception, working memory, planning, and action selection. This dissertation introduces new experimental methods to

investigate these processes, by characterizing the impact of physical assembly experience on our ability to build. Moreover, it explores far-reaching consequences of assembly experience on cognition, including the ability to remember objects, as well as the language we use to communicate about them. Over three chapters, I investigate three consequences of assembly experience. In Chapter 1, I investigate how practice assembling objects changes the *procedures* we use to construct them, finding that people learn to build more quickly and accurately, and use increasingly consistent procedures to do so. In Chapter 2, I explore how assembling objects impacts our *memory* of those objects, finding that how well we remember an object depends crucially on the way we encode it during assembly. In Chapter 3, I go beyond the consequences of assembly for individuals to ask how shared assembly experience impacts how collaborators *communicate* about objects, and find that people coordinate on linguistic conventions for referring to increasingly abstract procedures over time. These results comprise a set of ways that people– individually and collectively– leverage prior assembly experience to improve their ability to build, elucidating one of the most pervasive and complex human behaviors. By clarifying the impact of a creative, generative behavior on our representations of the things we make, these findings have implications for understanding how we relate to the world of constructed objects we inhabit, and the development of technologies that help people create more effectively.

Introduction

0.1 Physical construction

The modern world is densely populated by physical structures that were designed and made by people, from transient arrangements like stacks of books and sandwiches, to enduring constructions like bridges and skyscrapers. *Physical construction* encompasses a range of behaviors, from simple actions to complex construction efforts coordinated across large groups of people, and can be a critical act of survival or an act of creative expression (Korn, 2015). Physical construction is also the primary means by which humans reorganize their environments. We live in worlds of constructed objects, so much so that it is rare to be in an environment that does not contain many objects that were built by people. Characterizing the cognitive mechanisms that underlie physical construction is not only crucial for understanding how we perform this pervasive activity, but also how we relate to the world around us.

As well as being an important activity in its own right, physical construction presents a unique set of opportunities for scientists studying the mind. Even the simplest acts of creation draw on a wide range of cognitive mechanisms, including planning, perception, working memory, and motor control, raising the possibility of studying the interactions between these mechanisms in the context of a more complex behavior. The fact that people can acquire construction skills over relatively short timescales (compared to say, language) is also beneficial– studying how construction ability improves in response to specific kinds of experience can help isolate the contribution of specific cognitive mechanisms. Moreover, physical construction– by definition– generates durable outputs, which can be used as objective indicators of ability, as well as rich readouts of underlying cognitive processes (Bainbridge et al., 2019; Fan et al., 2023). These features make physical construction an ideal test bed for studying how cognitive processes interact to give rise to complex behaviors.

0.2 The challenges of studying physical construction

The study of physical construction has been held back by a number of methodological challenges. The computational complexity of construction tasks is particularly problematic. Even in simple, discretized domains, the number of valid "moves" from any state can be vast (Cortesa et al., 2017). For example, consider deciding your next move when a building a lego structure– there are tens of possible blocks that can be placed in several orientations in hundreds of possible locations. Branching factors as large as these lead to a combinatorial explosion in the state space, making them challenging even for state of the art algorithms (Bapst et al., 2019), let alone for interpretable algorithms that might be used to model human cognition (van Opheusden & Ma, 2019). It is quite telling that, while physical construction was well represented in early AI research (Winograd, 1971; Winston, 1970), the attention it received dwindled until recently (Bapst et al., 2019), relative to more formal domains like games.

The relationships between the myriad cognitive processes involved in construction make it a potentially fruitful domain to study but also make it a thorny one. Building something not only requires planning a sequence of actions, but also reasoning about the physical properties of the intermediate states: Will it be able to stand unsupported? How firmly will you need to push to attach the next piece? Can it handle that force without breaking? Keeping track of what you have built so far requires perception, and memory– two processes that are known to interact with each other (Lee & Rudebeck, 2010; Silva et al., 2006) and both change with expertise (Chase & Simon, 1973). Disentangling the individual contributions of cognitive capacities is typically achieved using a series of controlled experiments, however the feasibility of running such experiments in the context of actual physical construction is questionable. Experimental control is difficult to achieve in a domain where participants are constantly interacting with and changing their environments (Kirsh, 1995). Consider trying to control for perceptual exposure. People will

see different things depending on how they approach the problem and what actions they take. Measuring the effects of their actions on their perceptual input without intervening on their behavior is also impractical. For these reasons, cognitive psychology typically studies processes like perception and memory outside the context of complex, interactive behaviors, which does provides valuable clues about how they operate in isolation, but not whether these findings generalize to more complex domains.

On the other hand, prior investigations into physical construction have typically relied on observational studies (Cortesa et al., 2017; Wolfgang et al., 2001), which have provided key insights into the cognition of people situated in construction environments (Zheng & Tversky, 2024), but also required significant recording and coding efforts to quantify, limiting the precision of the inferences we can draw from these data, and the rates at which these experiments can scale. In sum, the study of construction tasks present a trade-off between ecological validity, on the one hand, and controllability, measurement, and quantifiability, on the other.

0.3 An approach to studying physical construction

What would be needed to study the cognitive mechanisms underlying physical construction in context, while retaining the ability to test hypotheses about these mechanisms using controlled experiments? We want study construction behaviors complex enough to evoke a range of cognitive phenomena. On the other hand, to be able to isolate specific mechanisms, as well as provide precise computational theories of behavior, we need a way of restricting the complexity of the behaviors we intend to study. A promising direction is to take a middle ground and employ construction tasks of a medium-level complexity– complex enough to evoke a wide range of cognitive phenomena, but tractable for algorithmic theories (Bapst et al., 2019; van Opheusden & Ma, 2019). In doing so, we would intentionally lay aside some aspects of real-world construction in order to reduce

the size of the state space.

Physical assembly– the construction of a specific object from a pre-defined set of parts– is a promising domain for doing so. Compared to free construction, the limited set of available parts substantially reduces the number of actions available in any step, massively reducing the state space (Shelton et al., 2022). Consider, for example, the number of ways someone can construct specific IKEA desk, compared to constructing any desk from planks of wood. The goal of creating a specific object may rule out mechanisms crucial for more open-ended construction, but in doing so allows for precise quantification of accuracy. Assembly is also particularly amenable to experimental control; task difficulty can be systematically varied by changing the target object and sets of provided parts and the compositionality of assembled parts also allows for hierarchical relationships between parts, objects, and sets of objects, making it an ideal stimulus domain for measuring learning and generalization.

Compared to 'yes-no' or categorical responses typical in cognitive psychology experiments, and even to the outputs of generative tasks, construction behavior is immensely complex. Quantifying variance in *the way people build* therefore poses a challenge. The behavior of two people precisely following the same set of IKEA instructions might involve the same parts being attached in the same order, but *appear* very different to an observer. The way each person implements an action through motor commands might be very different, depending on factors like body size, strength, and physical ability. Judging two actions as equivalent requires considering them at some particular level of abstraction, for example treating all actions where part A is connected to part B as the same. By choosing a level of abstraction we make the explicit decision to abstract away from, say, the motor planning involved in physical construction, simplifying the task in a way that makes observing relationships between other cognitive mechanisms tractable.

Even after simplifying construction sequences into sequences of discrete actions, measurement and quantification of those actions is still a methodological challenge. *Virtual*

environments provide a simple solution to this problem, by making it trivial to record and measure anything occurring on screen, including the entire sequences of actions performed on every part, the locations of all parts, and timing between actions. Virtual environments also make it easy to compare performance of humans and AI systems, and enable online deployment, making it feasible to collect the large datasets needed to measure variance in behavior. The particular choice of virtual environment and user interface– 2D or 3D, whether physics is implemented, how objects are interacted with– determines which cognitive processes get abstracted out. Most however, will rule out complex motor control and various kinds of situated cognition (Hutchins, 1995), which despite playing an important role in physical construction, are a source of confounds in controlled experiments that can are better to eliminate when asking targeted questions. Furthermore, experimenters retain control over what is visually present on screen at any time, making it easy to control for things like visual exposure at a specified level of abstraction (e.g. at the level of which parts were visible at any time).

0.4 What can we learn using this approach?

If we had the tools to study physical construction in this way, what questions would be most pressing to ask?

As well as making physical construction challenging to study, the computational complexity of physical construction also makes it cognitively interesting: how are people able to plan far into the future when there are so many different actions we could take at any point in the construction process? Focusing on tasks of mid-level complexity allows us to consider interpretable algorithmic theories (van Opheusden & Ma, 2019), including those explored for other domains involving multi-step decision making (Huys et al., 2015; Solway & Botvinick, 2012, 2015). Even in these simpler domains, researchers have identified that people adopt strategies to make efficient use of their limited cognitive resources (Callaway

et al., 2018; Lieder & Griffiths, 2020), and, more generally, the way you represent a task has a crucial implications for the way you plan (Ho et al., 2022). A common finding in these areas is that *expertise* in a domain changes your representation of the task (Botvinick, 2008; Simon & Chase, 1988; Tomov et al., 2020) enabling people to plan further ahead (van Opheusden et al., 2023). How does expertise affect how you represent a construction task and plan how to build something? Generalizing from other domains is risky, because the way we represent a construction task is highly intertwined with our representations of the objects involved. Expertise is also known to change the structure of perceptual representations (Chase & Simon, 1973; Gobet & Simon, 1998; Sheridan & Reingold, 2017), providing multiple routes for experience to shape our behavior. The most valuable step towards understanding how experience shapes assembly behavior is to gather data in this domain.

If assembly experience really does impact our representations of objects it could have far-reaching implications for how we relate to the world of constructed objects we inhabit. As even passive visual exposure is enough to the way people represent objects and their parts (Austerweil & Griffiths, 2013; Orbán et al., 2008), it will be critical to disentangle the contributions of visual exposure and other cognitive processes involved in construction. Recent lines of work have explored how aspects of other kinds of *generative behavior*– drawing, writing, etc.,– can impact perceptual abilities. Handwriting experience, for example, improves children's ability to recognize letters (Li & James, 2016), and drawing objects can lead people to discriminate more accurately discriminate between them (Fan et al., 2020). The mechanisms behind these improvements vary. Handwriting experience is effective both because it leads to the generation of variable visual output (Li & James, 2016), and because it links visual processing with motor experience (James, 2017; Zemlock et al., 2018). While suggesting that other generative tasks may result in perceptual changes, these findings also suggest that the mechanisms underlying these changes could be domain specific, highlighting the importance of studying representational

changes in context. Studying physical assembly directly allows us to do so, while also asking more targeted questions about how generative behaviors impact perceptual abilities. Making things has been shown to not only impact how we *perceive* things, but also how we *remember* them (James, 2017; Wammes et al., 2016). Does physical construction impact memory, and if so, how? Building something is not just a behavior but also a highly immersive way of interacting with the world. Multiple lines of prior work have shown that active engagement impacts our ability to remember the things we interact with (Bonwell & Eison, 1991; Chi, 2009; Craik & Lockhart, 1972; Markant et al., 2016), suggesting that building something may have substantial impact on our memory for objects we create. Disentangling the contributions of active engagement and perceptual exposure on our representations of objects will be crucial for understanding how assembly experience impacts objects.

The most impressive feats of human construction are made not by individuals, but by people working together. Characterizing the full range of human construction will therefore require an understanding of how we are able to coordinate in these tasks. Perhaps our most powerful tool for coordinating behavior is natural language, which we use to define shared goals (Clark, 1996), identify objects relevant to achieving our goals (Clark & Wilkes-Gibbs, 1986), and even convey a means by which to achieve them through instructions. Furthermore, the language of people who work together also becomes more efficient over time, both on large timescales through acquisition of technical language (Goodwin, 2015), and on shorter timescales through formation of conventions (Clark & Wilkes-Gibbs, 1986; Garrod & Doherty, 1994). In the context of construction tasks, people use language, and more implicit communication like gesture, to coordinate on shared actions (Zheng & Tversky, 2024). However, the ways in which collaborators refer to more extended sequences of actions have been underexplored, as has the role of experience in shaping this process.

0.5 Contribution of the dissertation

In this dissertation I present a methodological approach for studying the cognitive mechanisms that underlie our ability to build. I present a novel task paradigm for studying physical assembly– concretely, a simulated 2D physics environment in which people can construct block towers in a web browser. This tool enables the collection of precise behavioral data in the context of controlled experiments involving physical construction tasks. In the following three chapters, I leverage variants of this tool probe cognitive consequences of physical assembly.

In Chapter 1, I introduce the physical construction environment, along with a suite of metrics to measure fine-grained changes in assembly behavior that go beyond the state of the final reconstruction. I ask how practice assembling the same object multiple times changes the way people build it. I find that people are able to learn to build more accurately and quickly across repeated attempts, and that these improvements reflected group-level convergence on a tiny fraction of possible viable procedures. This chapter demonstrates that the way people build things depends on their assembly experience, and that even a small amount of practice is enough to shift their behavior. In doing so, it also validates the use of our tool as a way of influencing and measuring assembly behavior in the context of a controlled experiment. This work was originally published in and is re-printed here with minor edits.

In Chapter 2, I leverage the physical assembly task domain to ask how downstream representations of objects are affected by experience making things. At a high level, this chapter tests the hypothesis that active engagement with visual objects leads to better memory of those objects, compared to more passive viewing. The use of a generative task domain provides a unique perspective on this question; as well as asking whether or not participants *recognize* the objects they have built, I can get more precise readouts of the *contents* of their memory by asking them to build towers from memory. Generalizing

findings from word and concept memory (Bonwell & Eison, 1991; Chi, 2009; Craik & Lockhart, 1972; Markant et al., 2016), we would predict that assembling an object leads to stronger memories than passive viewing. In initial experiments I find precisely the opposite; building a tower appears to lead to weaker recognition and recall. I explore why over a series of further experiments, and find that mnemonic benefits only accrue when creators form holistic visual representations of the thing they have built. The work in this chapter is in preparation for submission and is reprinted here with minor edits.

In Chapter 3, I explore how people are able to collaborate on complex physical assembly tasks, by asking how people are able to coordinate on mental representations of things they are making together. Extending my tools for studying physical assembly to a collaborative setting, I ask people to work together to solve related physical assembly tasks. By limiting communication to a one-way linguistic messages, I encourage participants to form compact representations of target scenes, and expose a human-interpretable channel for observation of these representations. I find that the language participants use to refer to objects they know how to build becomes more efficient over time, and present a computational model that simultaneously captures the internal acquisition of procedural abstractions, as well as coordination of language used to refer to these abstractions. The work in this chapter has been submitted for publication and is currently in review and is reprinted here with minor edits.

In isolation, these three chapters further our understanding of cognitive mechanisms that underlie our ability to build simple objects. Together, they paint a picture of how we bootstrap our ability to build things by leveraging prior experience, providing a crucial piece of the story of how the most impressive feats of human construction are possible.

References

Austerweil, J. L., & Griffiths, T. L. (2013). A nonparametric Bayesian framework for constructing flexible feature representations. *Psychological Review*, *120*(4), 817.

Bainbridge, W. A., Hall, E. H., & Baker, C. I. (2019). Drawings of real-world scenes during free recall reveal detailed object and spatial information in memory. *Nature communications*, *10*(1), 5.

Bapst, V., Sanchez-Gonzalez, A., Doersch, C., Stachenfeld, K., Kohli, P., Battaglia, P., & Hamrick, J. (2019). Structured agents for physical construction. *International conference on machine learning*, 464–474.

Bonwell, C. C., & Eison, J. A. (1991). *Active learning: Creating excitement in the classroom. 1991 ashe-eric higher education reports.* ERIC.

Botvinick, M. M. (2008). Hierarchical models of behavior and prefrontal function. *Trends in cognitive sciences*, *12*(5), 201–208.

Callaway, F., Lieder, F., Das, P., Gul, S., Krueger, P. M., & Griffiths, T. (2018). A resource-rational analysis of human planning. *CogSci*.

Chase, W. G., & Simon, H. A. (1973). Perception in chess. *Cognitive psychology*, *4*(1), 55–81.

Chi, M. T. (2009). Active-constructive-interactive: A conceptual framework for differentiating learning activities. *Topics in cognitive science*, *1*(1), 73–105.

Clark, H. H. (1996). *Using language*. Cambridge University Press.

Clark, H. H., & Wilkes-Gibbs, D. (1986). Referring as a collaborative process. *Cognition*, *22*(1), 1–39.

Cortesa, C. S., Jones, J. D., Hager, G. D., Khudanpur, S., Shelton, A. L., & Landau, B. (2017). Characterizing spatial construction processes: Toward computational tools to understand cognition. *CogSci*.

Craik, F. I., & Lockhart, R. S. (1972). Levels of processing: A framework for memory research. *Journal of verbal learning and verbal behavior*, *11*(6), 671–684.

Fan, J. E., Bainbridge, W. A., Chamberlain, R., & Wammes, J. D. (2023). Drawing as a versatile cognitive tool. *Nature Reviews Psychology*, *2*(9), 556–568.

Fan, J. E., Wammes, J. D., Gunn, J. B., Yamins, D. L., Norman, K. A., & Turk-Browne, N. B. (2020). Relating visual production and recognition of objects in human visual cortex. *Journal of Neuroscience*, *40*(8), 1710–1721.

Garrod, S., & Doherty, G. (1994). Conversation, co-ordination and convention: An empirical investigation of how groups establish linguistic conventions. *Cognition*, *53*(3), 181–215.

Gobet, F., & Simon, H. A. (1998). Expert chess memory: Revisiting the chunking hypothesis. *Memory*, *6*(3), 225–255.

Goodwin, C. (2015). Professional vision. In *Aufmerksamkeit: Geschichte-theorie-empirie* (pp. 387–425). Springer.

Ho, M. K., Abel, D., Correa, C. G., Littman, M. L., Cohen, J. D., & Griffiths, T. L. (2022). People construct simplified mental representations to plan. *Nature*, *606*(7912), 129–136.

Hutchins, E. (1995). *Cognition in the wild*. MIT press.

Huys, Q. J., Lally, N., Faulkner, P., Eshel, N., Seifritz, E., Gershman, S. J., Dayan, P., & Roiser, J. P. (2015). Interplay of approximate planning strategies. *Proceedings of the National Academy of Sciences of the United States of America*, *112*(10), 3098–3103. https://doi.org/10.1073/pnas.1414219112

James, K. H. (2017). The importance of handwriting experience on the development of the literate brain. *Current Directions in Psychological Science*, *26*(6), 502–508.

Kirsh, D. (1995). The intelligent use of space. *Artificial intelligence*, *73*(1-2), 31–68.

Korn, P. (2015). *Why we make things and why it matters: The education of a craftsman*. Random House.

Lee, A. C., & Rudebeck, S. R. (2010). Investigating the interaction between spatial perception and working memory in the human medial temporal lobe. *Journal of cognitive neuroscience*, *22*(12), 2823–2835.

Li, J. X., & James, K. H. (2016). Handwriting generates variable visual output to facilitate symbol learning. *Journal of Experimental Psychology: General*, *145*(3), 298.

Lieder, F., & Griffiths, T. L. (2020). Resource-rational analysis: Understanding human cognition as the optimal use of limited computational resources. *Behavioral and brain sciences*, *43*, e1.

Markant, D. B., Ruggeri, A., Gureckis, T. M., & Xu, F. (2016). Enhanced memory as a common effect of active learning. *Mind, Brain, and Education*, *10*(3), 142–152.

Orbán, G., Fiser, J., Aslin, R. N., & Lengyel, M. (2008). Bayesian learning of visual chunks by human observers. *Proceedings of the National Academy of Sciences of the United States of America*, *105*(7), 2745–2750. https://doi.org/10.1073/pnas.0708424105

Shelton, A. L., Davis, E. E., Cortesa, C. S., Jones, J. D., Hager, G. D., Khudanpur, S., & Landau, B. (2022). Characterizing the details of spatial construction: Cognitive constraints and variability. *Cognitive Science*, *46*(1), e13081.

Sheridan, H., & Reingold, E. M. (2017). Chess players' eye movements reveal rapid recognition of complex visual patterns: Evidence from a chess-related visual search task. *Journal of vision*, *17*(3), 4–4.

Silva, M. M., Groeger, J. A., & Bradshaw, M. F. (2006). Attention–memory interactions in scene perception. *Spatial Vision*, *19*(1).

Simon, H., & Chase, W. (1988). Skill in chess. In *Computer chess compendium* (pp. 175–188). Springer.

Solway, A., & Botvinick, M. M. (2012). Goal-directed decision making as probabilistic inference: A computational framework and potential neural correlates. *Psychological Review*, *119*(1), 120.

Solway, A., & Botvinick, M. M. (2015). Evidence integration in model-based tree search. *Proceedings of the National Academy of Sciences*, *112*(37), 11708–11713.

Tomov, M. S., Yagati, S., Kumar, A., Yang, W., & Gershman, S. J. (2020). Discovery of hierarchical representations for efficient planning. *PLoS computational biology*, *16*(4), e1007594.

van Opheusden, B., Kuperwajs, I., Galbiati, G., Bnaya, Z., Li, Y., & Ma, W. J. (2023). Expertise increases planning depth in human gameplay. *Nature*, *618*(7967), 1000–1005.

van Opheusden, B., & Ma, W. J. (2019). Tasks for aligning human and machine planning. *Current Opinion in Behavioral Sciences*, *29*, 127–133.

Wammes, J. D., Meade, M. E., & Fernandes, M. A. (2016). The drawing effect: Evidence for reliable and robust memory benefits in free recall. *Quarterly Journal of Experimental Psychology*, *69*(9), 1752–1776.

Winograd, T. (1971). *Procedures as a representation for data in a computer program for understanding natural language* (tech. rep.). Massachusetts Institute of Technology.

Winston, P. H. (1970). *Learning Structural Descriptions From Examples* (tech. rep.). USA, Massachusetts Institute of Technology.

Wolfgang, C. H., Stannard, L. L., & Jones, I. (2001). Block play performance among preschoolers as a predictor of later school achievement in mathematics. *Journal of Research in Childhood Education*, *15*(2), 173–180.

Zemlock, D., Vinci-Booher, S., & James, K. H. (2018). Visual–motor symbol production facilitates letter recognition in young children. *Reading and Writing*, *31*(6), 1255–1271.

Zheng, C., & Tversky, B. (2024). Putting it together, together. *Cognitive Science*, *48*(2), e13405.

Chapter 1

Consistency and variation in reasoning about physical assembly

Abstract

The ability to reason about how things were made is a pervasive aspect of how humans make sense of physical objects. Such reasoning is especially useful for a range of everyday tasks, from assembling a piece of furniture to making a sandwich and knitting a sweater. What enables people to reason in this way even about novel objects, and how do people draw upon prior experience with an object to continually refine their understanding of how to create it? To explore these questions, we developed a virtual task environment to investigate how people come up with step-by-step procedures for recreating block towers whose composition was not readily apparent, and analyzed how the procedures they used to build them changed across repeated attempts. Specifically, participants (N=105) viewed 2D silhouettes of 8 unique block towers in a virtual environment simulating rigid-body physics, and aimed to reconstruct each one in less than 60 seconds. We found that people built each tower more accurately and quickly across repeated attempts, and that this improvement reflected both group-level convergence upon a tiny fraction of all possible viable procedures, as well as error-dependent updating across successive attempts by the same individual. Taken together, our study presents a scalable approach to measuring consistency and variation in how people infer solutions to physical assembly problems.

Keywords: planning; spatial reasoning; intuitive physics; construction; action

Humans have populated much of the world with physical artifacts of their own design, from sand castles to skyscrapers. Taken together, these structures exemplify the human capacity to interact with the physical world in creative, yet goal-directed ways. This creative capacity also manifests in many everyday tasks, from assembling a piece of furniture to making a sandwich and knitting a sweater. In these scenarios, people rely upon their ability to not only judge the static properties of objects (e.g., their size, shape, weight), but also to infer the process by which objects are made (e.g., the parts they consist of and how to arrange them). What cognitive mechanisms enable people to engage in such reasoning about complex objects, and how do people draw upon prior experience with an object to continually refine their understanding of how to create it?

Perhaps the most basic requirement is a general-purpose and intuitive understanding of how material objects interact in the physical world, a suite of abilities known as intuitive physics (McCloskey, 1983). That is, even without performing formal calculations, people can make reasonably accurate predictions about how objects will behave in a variety of settings (Kubricht et al., 2017; Smith et al., 2018). A prominent proposal argues that generating these predictions relies on mental simulation, perhaps reflecting a noisy approximation to real-world physical dynamics (Battaglia et al., 2013; Hamrick et al., 2015; Hegarty, 2004; Sanborn et al., 2013; Schwartz & Black, 1999; Smith & Vul, 2013). Recent work has explored the role that simulation plays when people plan single interventions on physical scenes — for example, joining two blocks together to stabilize a block tower (Hamrick et al., 2018) or causing an object to move into a target zone (Allen et al., 2020; Dasgupta et al., 2018).

However, the role of physically grounded mental simulation has yet to be fully explored in the context of the multi-step action sequences required to assemble a complex object (Kirsh, 1995; Kurth-Nelson et al., 2023; Schwartenbeck et al., 2021). This gap in knowledge at least in part reflects the methodological challenges posed by measuring behaviors as open-ended as physical assembly while maintaining a sufficient degree of

experimental control (Cortesa et al., 2017, 2018; Wolfgang et al., 2001).

Recent advances in the study of multi-step planning and decision making in other settings suggest promising ways forward (Daw et al., 2011; Huys et al., 2015; Solway & Botvinick, 2012, 2015). To the degree that these "grid-world" environments used often in this work sacrifice physical realism, they do so in favor of empirical and formal tractability (van Opheusden & Ma, 2019; van Opheusden et al., 2017, 2023). Nevertheless, as the state space grows, the computational cost of conducting thorough mental simulations over the full set of possibilities becomes prohibitive (Callaway et al., 2018; Hamrick et al., 2015; Huys et al., 2015; Solway & Botvinick, 2012, 2015). Prior work has found evidence that humans use a variety of strategies to reduce the cost of planning, such as pruning the search space (i.e., circumventing expensive but irrelevant action sequences (Huys et al., 2012)) and learning procedural abstractions to generate hierarchically organized plans (Botvinick & Weinstein, 2014; Dezfouli & Balleine, 2013; Éltető & Dayan, 2023; Huys et al., 2015; Xia & Collins, 2020). However, it remains unknown which, if any, of these strategies are ones that humans use when attempting to solve physical assembly problems, in which transitions between states are governed by physical constraints (e.g., stability, friction) rather than arbitrary rules (Daw et al., 2011). A valuable step towards bridging this gap would be the development of experimental methods for exploring human assembly behavior in task environments with a greater degree of physical realism than those commonly used to probe multi-step decision making.

An additional benefit of developing such methods would be the opportunity to investigate the impact of experience, building on a long tradition of work investigating changes in problem solving behavior accompanying the acquisition of expertise (Campitelli & Gobet, 2004; Chase & Simon, 1973; Sheridan & Reingold, 2017; Van Harreveld et al., 2007). For example, experience might be linked to changes in both how people encode state information and how they search over the space of possible solutions. Classic and contemporary work using board games suggests that experts display both a pronounced

ability to plan further ahead in games than novices and to mentally represent the configuration of game pieces in visual memory with higher fidelity (Chase & Simon, 1973; Gobet & Simon, 1998; Sheridan & Reingold, 2017; van Opheusden et al., 2023). Moreover, prior work using video games that impose substantial demands on rapid spatial reasoning (e.g., Tetris) has found that experience might also improve the fluency with which participants explore alternative states and determine the value of potential actions (Maglio & Kirsh, 1996). In principle, these experience-dependent changes might also apply to the domain of physical assembly, which would suggest that the underlying learning mechanisms generalize beyond the problem contexts in which they were initially proposed. On the other hand, it might be that there are important differences between reasoning domains: for example, problem-solving experience might have a stronger impact on how state information is encoded in less physically realistic game environments, such as board games, but a more modest impact in physical settings, where the mechanisms for encoding physical state are more stable across the lifespan (Baillargeon, 1995; Spelke & Kinzler, 2007).

Here we introduce a task paradigm for investigating how people reason about physical assembly in a virtual environment that is simple enough to provide a high degree of experimental control and formal tractability, but expressive enough to engage multi-step planning and understanding of core physical concepts (e.g., stability, mass, and friction). We report our findings from an exploratory study in which participants aimed to construct a series of 2D block towers from a set of rectangular blocks of varying sizes. We restricted the set of possible actions to placements of a fixed set of parts, enabling straightforward comparison of building procedures across participants. We further investigated how practice reconstructing a tower impacts the procedures participants subsequently used to build that tower across repeated attempts. Our approach takes inspiration from recent studies in which participants were asked to build copies of actual LEGO structures from LEGO bricks (Cortesa et al., 2018; Shelton et al., 2022). Findings from this line of work suggest

that people converged upon shared strategies for building these LEGO structures layer by layer, consistent with a bias towards shared layer-wise subgoals that also corresponded to physical subunits of the structures themselves (Shelton et al., 2022).

As in this prior work, we go beyond simple measures of assembly performance to characterize the action-by-action procedures people used to build each structure. However, our methodological approach differs in three key ways: First, because the current study aims to investigate the role of experience in assembly behavior, here we ask participants to build the same structures multiple times, allowing us to ask how practice influences the procedures that people use. Second, in order to put greater pressure on participants' ability to reason about how an object could be made, in a context where there is a large number of possible solutions, we presented participants with *silhouettes* of the block towers they sought to build. Third, in order to support high-throughput measurement of these open-ended behaviors, we developed a virtual assembly environment embedded in a web application to enable the concurrent participation of many individuals.

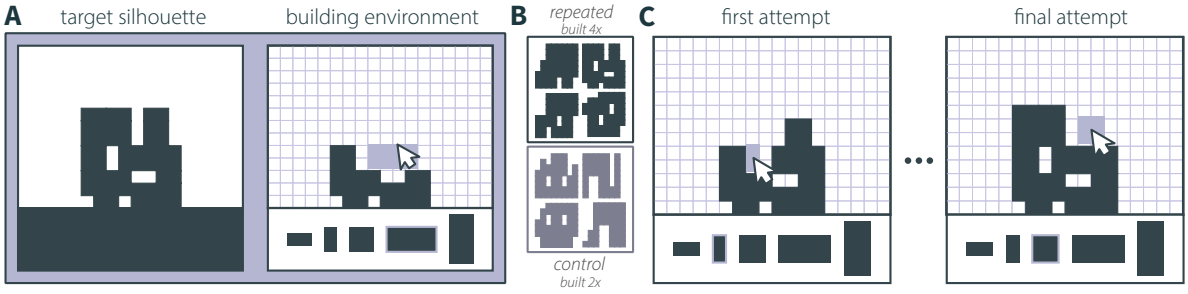

Figure 1.1. (A) Schematic of task display. The left window contained a target silhouette, and the right contained a building environment with gridlines. (B) For each participant the 8 silhouettes were randomly assigned to conditions, 4 in repeated and 4 in control. (C) Repeated towers were attempted 4 times, interleaved among other towers. Control towers were attempted twice, once at the beginning and once at the end of each session.

1.1 Method

The goal of our experiment was to investigate how people's strategies for solving physical reasoning tasks shift as they gain experience. To achieve this goal, we developed a web-based environment in which people could construct various block towers under simulated rigid-body physics. To provide participants with a specific goal, we considered the space of *physical assembly* tasks– namely, those in which people must create an exact replica of a target structure given the set of components used to construct it. However, such straightforward assembly tasks permit only a small range of solutions and can be solved using a simple strategy of copying block for block (Cortesa et al., 2017). To explore how strategies change with experience, we needed a task that permitted a large range of solutions. Therefore, rather than display target towers as a configuration of blocks that could be copied, we showed participants *silhouettes* of target towers and asked them to create *any configuration of blocks that matched the silhouette* shown. This required participants to infer which blocks to use, where to place them, and in what order. On each trial, participants aimed to reconstruct a target tower in less than 60 seconds using a fixed inventory of rectangular blocks. Over the course of an experimental session, participants built each tower either two or four times, allowing us to assess whether additional practice reconstructing a specific tower led to greater improvement than general practice with the task.

1.1.1 Participants

Based on data from pilot studies we estimated that between 100-150 participants would be sufficient to obtain reasonably precise estimates of our measures of consistency and variability. In the end, we successfully recruited 107 U.S.-based participants from Amazon Mechanical Turk. After accounting for technical issues during data acquisition (i.e., missing data), data from 105 participants were retained (49 female, mean age =

36.8 years). Participants provided informed consent in accordance with the institution's Institutional Review Board.

1.1.2 Stimuli

To identify a set of block towers that were non-trivial to reconstruct, we randomly sampled a large number of stable configurations of 8-16 blocks, then manually selected 8 of these that could be reconstructed in many different ways (Fig. 1.1B). We started with an inventory of five types of rectangular blocks that varied in their dimensions (i.e., `1x2`, `2x1`, `2x2`, `2x4`, `4x2`). To generate configurations of blocks, we partially filled an 8x8 rectilinear grid, bottom to top, by sampling random blocks in random x-locations, then randomly selected several blocks to be removed. We simulated the construction of each tower in a physics engine (Pybox2d), rejecting any tower that was unstable at any point during the construction process. To select towers that required planning ahead, we manually identified 8 configurations that included holes and/or overhanging blocks, and verified that these towers could be reconstructed in many different ways (59-7128 minimum unique solutions, $mean = 2418$).

1.1.3 Design

To more thoroughly characterize the effects of practice on physical construction ability, we wanted to be able to distinguish improvements resulting from general task experience from improvements resulting from practice reconstructing a specific tower. For each participant, we therefore randomly split the 8 block towers into 2 sets containing 4 towers each: a *control* set and a *repeated* set (Fig. 1.1 B). Participants reconstructed towers over four consecutive rounds. In the *first* (1st) and *final* (4th) rounds, participants reconstructed all 8 towers in a randomized order. In the middle two rounds (2nd and 3rd), participants reconstructed only the 4 *repeated* towers, also in a randomized order. Thus there were a total of 24 trials in each session: 8 *first* attempts, 2 rounds of 4 *repeated*

attempts, and 8 *final* attempts. In subsequent comparisons between the first and final attempts on each tower, we combine data from both the repeated sets (built 4 times) and control sets (built 2 times). In analyses of fine-grained changes in behavior across successive attempts on the same tower, we restrict our analysis to the repeated sets.

1.1.4 Task Procedure

On each trial, participants were presented with two adjacent display windows: On the left, a target block tower was presented as a silhouette centered on the floor in a 18x13 rectilinear grid environment (Fig. 1.1A); on the right, they were provided with an empty building environment and the inventory of blocks that was used to generate the towers.

Participants' goal was to build a tower that matched the shape of the target silhouette in less than 60 seconds using any combination of the blocks provided. To select a specific block type, they clicked on its image in the block inventory. Then, by hovering the mouse cursor over the building environment, a translucent block would appear, showing where the block would be placed when they clicked again. Blocks could be placed on any level surface in the building environment (i.e., either the floor or on top of another block). To minimize the intrusion of low-level motor noise in block placement, the location of each block 'snapped' to a visible grid.

After the placement of each block, participants' towers became subject to gravity, simulated using `Matter.js`. Thus, if their tower was not sufficiently stable, single blocks or even the entire tower could fall over. After 60 seconds had elapsed or if any block fell, the trial immediately ended and participants moved onto the next tower. We truncated trials on which any block fell for two main reasons: first, to ensure that all recorded block placements could in principle form part of a forward plan to build the target silhouette, rather than reflect online corrections for error; and second, to strongly incentivize the production of stable towers. Participants were rewarded for both accuracy and speed: the more accurate their reconstructions, the larger the monetary bonus they received. If

participants perfectly reconstructed the target silhouette, they could earn an additional bonus for speed.

1.1.5 Statistical Analysis Procedure

Our primary statistical approach involved fitting linear mixed-effects models mirroring, as close as possible, the structure of the experimental design. This included fixed effects for round and condition, as well as their interaction, and random intercepts for participant and tower. We then compared this full model to a series of nested models that had some of the predictors removed, typically starting with the interaction term, then the effect of condition. To select a model we calculated the Akaike Information Criterion (AIC) for each model, selecting the most complex model for which AIC substantially dropped relative to the subsequent simpler model. Full parameter estimates for selected models are reported in Supplemental Materials. For statistics outside of the model, we report confidence intervals generated using bootstrap resampling over 1000 iterations. In each bootstrap iteration we resampled participants with replacement from the entire sample, including all data from each participant every time they were sampled.

1.2 Results

1.2.1 Change in reconstruction accuracy across attempts

We first needed a measure of reconstruction accuracy that tracked how well the towers participants built matched the silhouette they were attempting to reconstruct. Reconstructions are accurate insofar as they coincide with the same region as the target silhouette, while not extending beyond it. We therefore selected a metric that takes into account both *recall* (i.e., the proportion of the target silhouette that coincided with the participants' reconstruction) and *precision* (i.e., the proportion of participants' reconstruction that coincided with the target silhouette). As stable towers existed in a

gridworld, we could compute precision and recall directly by comparing the bitmaps of squares occupied by the target silhouette and reconstruction. The F_1 score takes the harmonic mean of these values to provide a measure that lies in the range $[0, 1]$ and reflects the degree to which the participants' reconstruction coincided with the target silhouette:

$$F_1 = \frac{2}{(recall^{-1} + precision^{-1})}$$

In their first attempts, participants' reconstructions were moderately accurate, suggesting that they were engaged with the task but not at ceiling performance (control: $F_1 = 0.790$, 95% CI: $[0.776, 0.803]$; repeated: $F_1 = 0.800$, 95% CI: $[0.786, 0.814]$). To evaluate changes in reconstruction accuracy over time, we fit a linear mixed-effects model predicting F_1 score from attempt (first, final) and condition (repeated, control) as fixed effects, including random intercepts for participant and tower (Supplemental Table 1.1). We found a main effect of attempt ($b = 0.0759$, $t = 6.99$, $p < 0.001$), showing that participants' reconstruction accuracy reliably improved between their first and final attempts (Fig. 1.2A). We found no reliable effect of condition ($b = 0.00803$, $t = 0.737$, $p = 0.461$), and no evidence of an interaction between attempt and condition ($b = 0.0182$, $t = 1.19$, $p = 0.235$), suggesting that these improvements were at least in part explained by general effects of task practice.

In particular, participants may have learned how to more consistently place blocks that are fully contained within the silhouettes, resulting in fewer 'off-by-one' errors. To explore this possibility, we visualized the spatial distribution of block placements by constructing a heatmap of block placements, averaged across participants (Fig. 1.3). This heatmap suggested that participants did place a greater proportion of blocks outside of target locations in their first attempts than in their final attempts. To evaluate this possibility, we defined the spatial error for a given tower on a given attempt as the root-mean-squared cityblock distance between each location in the heatmap and the edge

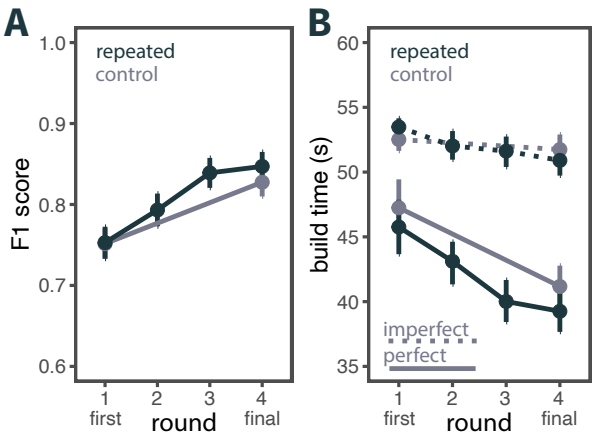

Figure 1.2. (A) Reconstruction accuracy across all four rounds, with first and final attempts of each tower labeled. (B) Build time across attempts, separated by perfect ($F_1 = 1$) and imperfect reconstructions. Error bars represent 95% CI.

of the target silhouette (zero if within the silhouette), weighted by the value at each location in the heatmap. We then computed the mean change in spatial error between their first and final attempts, which revealed that participants generally made fewer and less extreme errors in their final attempts than in their first attempts ($m = -0.625$, 95% CI: $[-1.08, -0.209]$, $p = 0.012$).

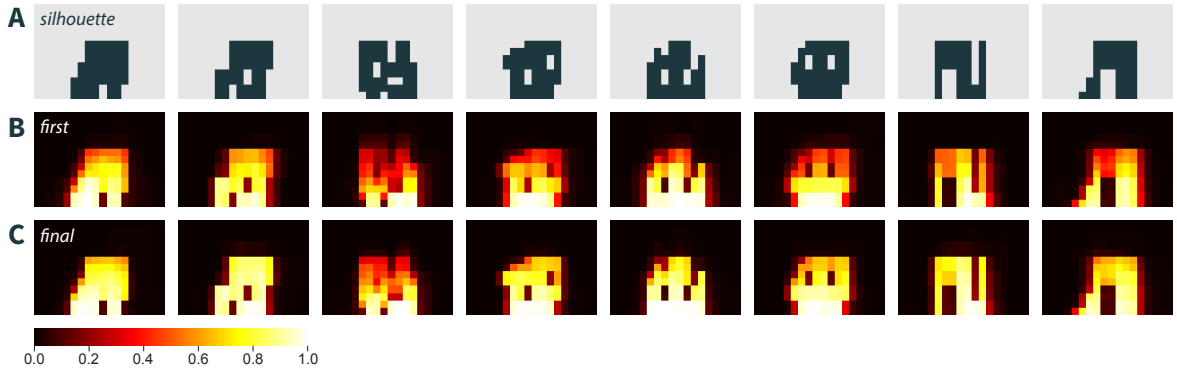

Figure 1.3. (A) 8 target silhouettes used in the experiment. (B,C) Heatmap representations of the spatial distribution of block placements for each tower, for first and final attempts. The intensity of each cell reflects the proportion of participants who placed a block in that location.

1.2.2 Change in reconstruction fluency across attempts

In addition to placing blocks more precisely, participants may have also produced more accurate reconstructions by improving in their ability to place more blocks within the time available on each trial. To evaluate this possibility, we modeled the change in the number of blocks used between the first and final attempts using a linear mixed-effects model otherwise identical in structure to that previously used to analyze accuracy, however we excluded trials which were truncated due to blocks falling (Supplemental Tables 1.2-1.3). This analysis revealed a strong main effect of attempt ($b = 1.19$, $t = 7.41$, $p < 0.001$), showing that participants were able to consistently use more blocks in their final attempt. There was no evidence of an effect of condition ($b = 0.0425$, $t = 0.264$, $p = 0.792$) nor of an interaction between attempt number and condition ($b = 0.167$, $t = 0.735$, $p = 0.463$).

There are at least two potential explanations for how participants were able to place more blocks in their final attempt: *first*, their fluency with the construction task interface may have improved, allowing them to select and place more blocks per unit of time; *second*, they may have been able to recall previously used procedures for building a given tower, and thus required less preparation time to devise an action plan prior to placing their first block. We estimated task fluency by computing the mean time between successive block placements within a single trial. We estimated preparation time by computing the time between trial onset and the placement of the first block. We found that task fluency increased ($b = -1.34$, $t = -13.548$, $p < 0.001$; Supplemental Table 1.4) and preparation time decreased ($b = -2.24$, $t = -8.64$, $p < 0.001$; Supplemental Table 1.5) between first and final attempts, suggesting that participants' improved accuracy may reflect changes in both.

To measure how quickly participants completed their reconstructions, we measured the amount of time elapsed between the start of each trial and the final block placement on that trial, again omitting trials which were truncated due to falling blocks. In their

first attempts, participants used nearly all of the time allotted (control: $51.8s$, 95% CI: $[51.1, 52.7]$; repeated: $52.2s$, 95% CI: $[51.6, 52.8]$), and appeared to use less time to build each tower across attempts (Fig. 1.2B). To evaluate changes in build time between the first and final attempt, we fit a linear mixed-effects model including attempt (first, final) and condition (repeated, control) as fixed effects, including random intercepts for participant and tower (Supplemental Table 1.6). This analysis revealed a main effect of attempt ($b = -1.92$, $t = -4.25$, $p < 0.001$) but not of condition ($b = -0.704$, $t = -1.80$, $p = 0.0725$). In exploratory analyses, we discovered that 22.4% of all trials contained perfect reconstructions (i.e., $F_1 = 1$) of the target silhouette. When we included an additional binary variable in our regression model indicating whether a trial contained a perfect reconstruction, we discovered that these 'perfect' reconstructions took reliably less time than imperfect reconstructions ($b = -3.81$, $t = -4.47$, $p < 0.001$). Moreover, a reliable interaction between attempt number and this binary variable revealed that decreases in build time from first to final attempts were greater for perfect reconstructions ($b = -5.04$, $t = -5.10$, $p < 0.001$). Together, these findings suggest that the greatest increases in speed occurred once participants had discovered a way of producing a perfect reconstruction.

1.2.3 Change in reconstruction procedures across attempts

Having established that participants build more accurately and quickly across successive attempts, we then investigated the changes to participants' construction procedures that underlie this improved performance. An increase in speed and decrease in preparation time are consistent with the possibility that participants reused previous procedures to successfully reconstruct each tower; however, these holistic measures only indirectly bear on this question. We therefore derived two measures of similarity between the actions performed across different building attempts (Fig. 1.4A).

Each *action* consists of an individual block placement, represented by a 4-vector

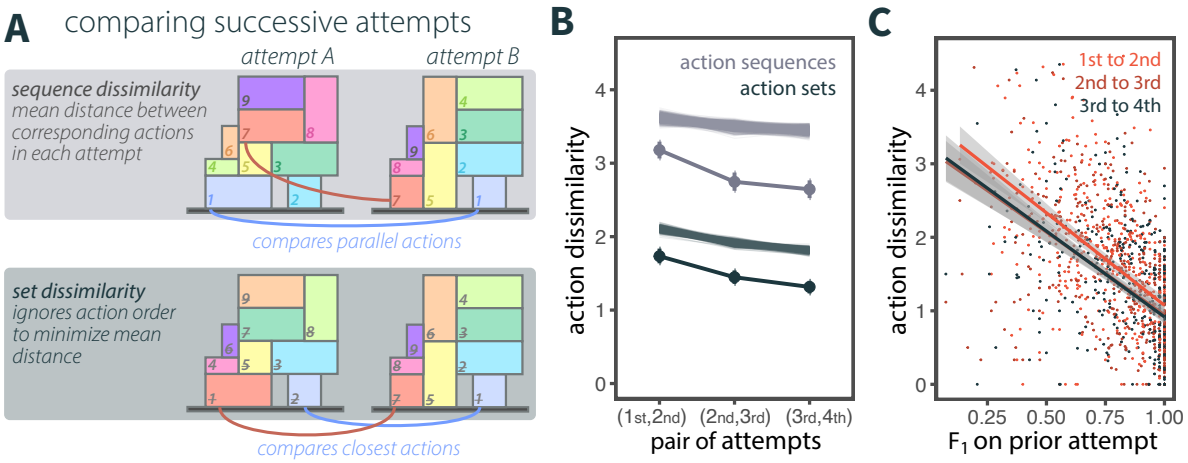

Figure 1.4. (A) Example comparison between building procedures on successive attempts. Numbers on blocks indicate the order in which they were placed. Sequence dissimilarity (top) compares building procedures on an action-by-action basis (i.e. block n to block n). Actions involving different sized blocks placed further apart are judged as more distant. Set dissimilarity (bottom) minimizes the mean distance between actions by ignoring order and pairing similar actions together. (B) Magnitude of change in sequences of actions (gray) and sets of actions (dark green) across successive build attempts. Shaded area represents baseline distributions. (C) Magnitude of change in sets of actions as a function of accuracy (F_1) on previous attempt, for each pair of successive attempts of a given tower.

$[x, y, w, h]$, where $0 \leq x \leq 15$, $0 \leq y \leq 13$ represents the coordinates of the bottom-left corner of the current block and where $(w, h) \in \{(1, 2), (2, 1), (2, 2), (2, 4), (4, 2)\}$ represent its width and height, respectively. Each procedure consists of the full *sequence* of such actions performed on a given reconstruction attempt. We define the "sequence dissimilarity" between any *pair* of action sequences as the mean Euclidean distance between corresponding pairs of $[x, y, w, h]$ action vectors (Fig. 1.4A, top). When two sequences are of different lengths, we evaluate this metric over the first k actions in both, where k represents the length of the shorter sequence. This 'sequence' measure compares the dissimilarity of procedures on an action-by-action basis, and hence assumes that when 'similar' plans are executed, actions are performed in exactly the same order. However, we might also consider procedures to be 'similar' when they involve similar shaped blocks placed in similar locations, regardless of the order in which blocks are placed. To obtain a measure of similarity between procedures that is robust to differences in the order in which actions are performed, we also derived a measure of dissimilarity between the *sets* of actions performed, using the Kuhn-Munkres algorithm to identify the one-to-one mapping between actions from each attempt that minimizes the mean Euclidean distance between them (Fig. 1.4A, bottom). This "set dissimilarity" measure has the advantage of being sensitive to correspondences between similar actions performed in different attempts, even when they were performed in a different order.

We first sought to determine whether participants reused aspects of their own prior attempts when reconstructing towers. We calculated the sequence and set dissimilarities between participants' consecutive attempts at each tower (Fig. 1.4B, solid). To estimate the expected dissimilarity between attempts, we created a baseline distribution of dissimilarity values between participants' 2nd, 3rd, and 4th attempts at a each tower with 'prior attempts' (i.e. 1st, 2nd, and 3rd) from a different, randomly sampled participant. We repeated the process 1000 times, permuting participants separately for each tower (Fig. 1.4B, shaded). We found that participants' procedures were more similar to their own prior attempts

than to other participants' ($p < 0.001$ for each pair of consecutive repetitions, for both sequence and set dissimilarity), suggesting that participants did reuse aspects of their own prior solutions to reconstruct each tower.

To assess whether participants used increasingly similar procedures across consecutive attempts, we fit both sequence and set action dissimilarities with a linear mixed-effects model including fixed effects for attempt pair, the accuracy of the previous attempt, and the dissimilarity type (sequence or set), as well as random intercepts for tower and participant (Supplemental Table 1.7). We found that attempt pair was negatively related to dissimilarity for both dissimilarity measures ($b = -0.186$, $t = -7.40$, $p < 0.001$; Fig. 1.4B), suggesting that participants became increasingly consistent in the procedures they used to reconstruct each tower across repeated attempts. In other words, the actions in participants' later attempts (i.e. attempts 3 and 4) were more similar to each other than the actions in earlier attempts (i.e. 1 and 2). As this result holds for set as well as sequence dissimilarity, it suggests a genuine increase in the consistency between the actions taken by participants, regardless of the specific order in which they performed.

A potential explanation for this convergence in procedures is that, as participants uncover increasingly successful procedures for recreating a tower, they may be less likely to dramatically change their strategy in later attempts. To the extent that accuracy on prior attempts is related to how much participants alter their procedure in subsequent attempts, we would predict that more successful procedures are more likely to be reused than unsuccessful ones. Consistent with this prediction, we found a strong negative relationship between accuracy on the most recent attempt and how much they changed their procedure ($b = -0.6426$, $t = -4.054$, $p < 0.001$; Fig. 1.4C), such that participants updated their procedure to a greater extent when their previous attempt was less successful. Taken together, these results suggest that people can make efficient use of prior experience to update their procedures accordingly.

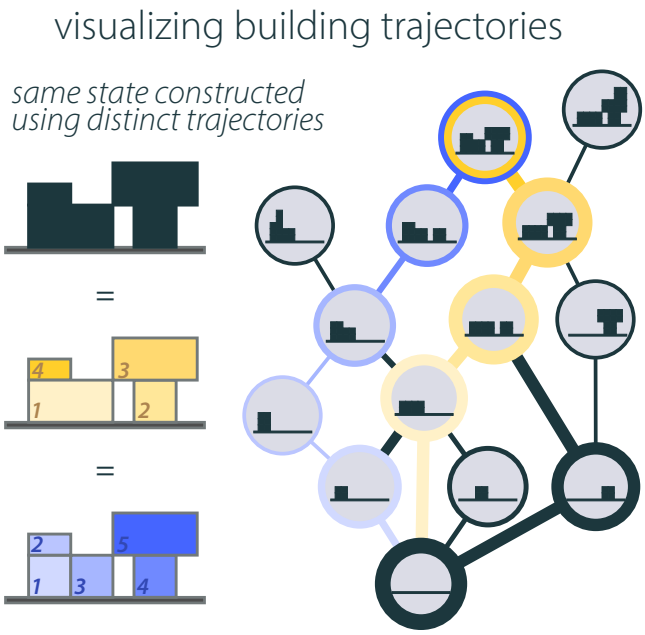

Figure 1.5. To visualize the set of trajectories taken by participants, we constructed graphs of the intermediate states visited during a reconstruction attempt. Larger nodes indicate a greater number of participants constructing that intermediate state, and thicker edges indicate a greater number of participants who transitioned between two world states with a single block placement. Intermediate states are defined by their outline shape and are independent of the underlying blocks used to create them. Two distinct trajectories leading to the same state are highlighted in blue and yellow.

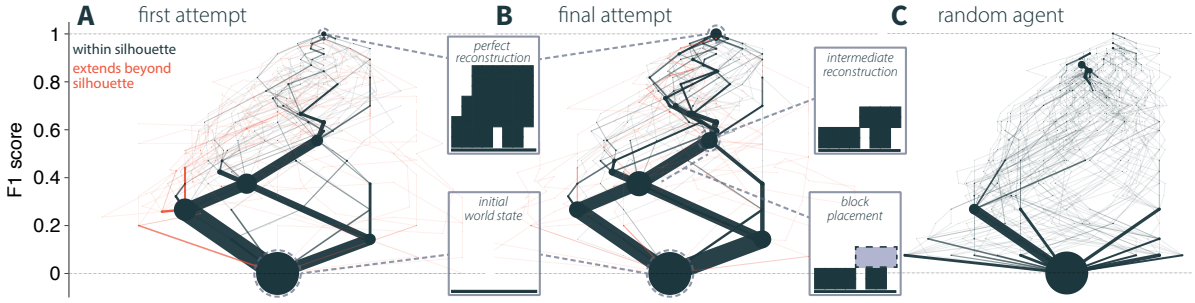

Figure 1.6. Distribution of state trajectories for first attempts (A), final attempts (B) , and an artificial agent (C) employing a random action-selection policy to reconstruct an example tower. Each trajectory consists of a sequence of states (nodes) connected by actions (edges), beginning from the initial world state ($F_1 = 0$) and directed upwards toward complete reconstructions ($F_1 = 1$). Node size represents the number of times a state was visited. Edge thickness represents the number of times a state-state transition was traversed.

1.2.4 Consistency and variability in procedures across individuals

Our results so far show that participants employ increasingly accurate and internally consistent procedures for reconstructing each tower, raising a natural question concerning the degree to which procedures used by different participants coincide with one another. While the analyses above suggest some variation in the actions that participants performed, they do not reveal whether participants were biased towards a small set of solutions for each tower, or whether they instead discovered a wide variety of completely different solutions. We therefore visualized the distribution of procedures used by all participants by constructing a map of *trajectories* over intermediate *states* visited between the start and end of their reconstruction (Fig. 1.5), where each state is defined by the shape of the reconstruction up to that point. Under this definition, reconstructions that are composed of different blocks but share the same shape are treated as occupying the same state, but are reached by taking distinct trajectories.

Even in their first attempts, many participants appeared to traverse the same states when reconstructing each target silhouette (Fig. 1.6), hinting at broad consistency in the procedures people use to perform this task. Additional simulations suggested that at most 2.2% of the total number of possible solutions to each tower were represented in our dataset (i.e., 435 unique trajectories across all towers out of 19,677 discovered via random sampling). To estimate how strongly participants were biased towards the same of subsequences, we computed the Gini index (G) over the number of traversals of each sequence of states across all participants:

$$G = \sum_{i=1}^{n} \sum_{j=1}^{n} |x_i - x_j| * (2 \sum_{i=1}^{n} \sum_{j=1}^{n} x_j)^{-1}$$

where n is the number of states and x_i and x_j represents the number of times states i and j were visited, respectively. G can be thought of as the average difference in

the number of times each subsequence was traversed, normalized by the total number of sequences of that length (summed twice to account for differences in both directions) to lie in the range $[0, 1]$. It largest when there are a small number of frequently traversed subsequences, and smallest when all subsequences were traversed an equal number of times.

To estimate how strongly human procedures concentrate on the same sequences of states at different timescales, we next extracted n-gram representations for all state trajectories, each defined by n successive states, for $1 \leq n \leq 10$, then calculated G_n for each of these n-gram frequency distributions (Fig. 1.7 A). To provide a baseline, we also constructed a random-policy agent that samples blocks and viable locations (i.e., within silhouette, maintaining stability) with equal probability. We used this random-policy agent to generate a null distribution of 1000 Gini values, each computed from 105 random-policy agents identified by unique random seeds. When comparing the mean observed G for human trajectories to this null distribution, we found that human state trajectories were reliably more concentrated on fewer n-grams than the random-policy agents, across n-grams of all lengths, for both first attempts (Z-score = 21.6) and final ones (mean Z-score = 42.7; Fig. 1.7B). These results show that a policy of selecting random viable actions is insufficient to explain patterns of human action selection in this task.

Insofar as participants are biased to discover similar solutions over time, we may expect the Gini index to grow between the first and final attempts. To evaluate this possibility, we fit human Gini values with a linear mixed-effects model including attempt number, linear and quadratic terms for n, as well as random intercepts for target towers and participants (Supplemental Table 1.8). This analysis revealed a positive effect of attempt number ($b = 0.112$, $t = 6.02$, $p < 0.001$), suggesting that participants converged on a smaller set of procedures across attempts, and this convergence applied to n-grams over action sequences of all lengths (Fig. 1.7B).

While participants' convergence towards a smaller number of state sequences might

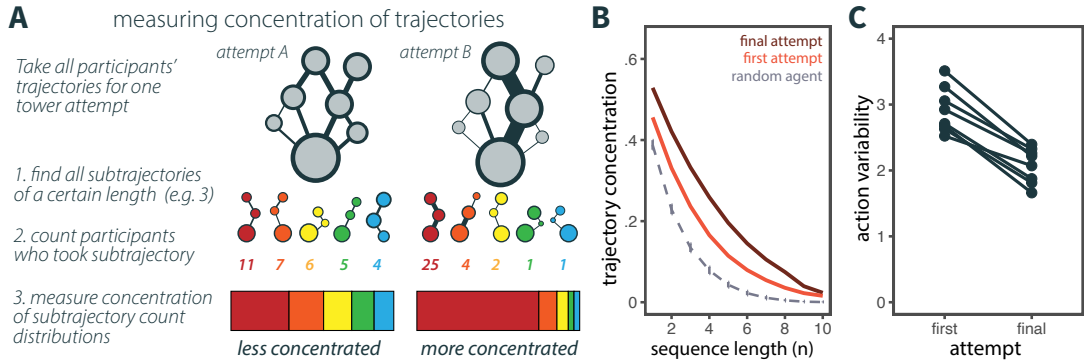

Figure 1.7. (A) To estimate the degree of bias towards certain trajectories we extracted all subsequences of states of a certain length and measured how concentrated construction behavior was on a small number of sequences. (B) Gini index for n-grams of action sequences in first and final attempts, compared to those of a random-policy agent. Higher Gini index reflects a smaller number of frequently appearing action sequences. (C) Variability between sets of actions performed by different participants in first and final attempts. Each line segment represents a different tower.

point to greater consistency in the parts of the towers participants were choosing to reconstruct, it might instead be a necessary consequence of participants building more accurate and hence more consistent reconstructions. To distinguish these possibilities, we repeated the previous analysis but only on trials where participants perfectly reconstructed the target tower. We found that Gini values still increased from first to final attempt ($b = 0.175$, $t = 5.68$, $p < 0.001$; Supplemental Table 1.9), confirming that convergence in trajectories was not simply a consequence of more accurate reconstructions, but also reflected more consistent ways of reconstructing each tower.

Although such convergence is one signature of using similar procedures, the above analysis is insensitive to cases where two participants reconstruct a silhouette by placing the same blocks in the same locations, yet place these blocks in a different order. To address this limitation, we examined the distribution of dissimilarities between the *sets of actions* performed by different participants, and found that the variance of this distribution was smaller on final attempts than in first attempts, for all target towers ($t(7) = 10.603$, $p < 0.001$; Fig. 1.7C). Taken together, these results indicate that despite the relatively

high state-space complexity of this task, people share systematic biases toward similar solutions even in their first attempts, and tend to update their strategies across repeated attempts in similar ways, converging on a more similar set of solutions over time.

1.3 Discussion

In this paper, we investigated how people reason about physical assembly problems and update their approach to solving them over time. Specifically, we developed a web-based environment where participants aimed to reconstruct a set of 2D block towers, and measured how accurately and quickly they could do so across successive attempts at building each tower. We found that participants achieved strong performance even on their first attempts and improved substantially with additional practice. Moreover, our findings suggest that low-level changes in motor fluency were insufficient to fully explain this improvement. Instead, improvement was driven by genuine changes in the decisions made by participants about how to build each tower, with participants updating their procedures to a greater degree when their prior attempt had been less successful. In addition, although there were many possible ways of reconstructing each tower, we found that the procedures participants used to initially construct these towers were strikingly consistent across individuals. Moreover, participants converged on increasingly similar procedures across attempts, suggesting shared biases toward similar approaches to solving these assembly problems.

What accounts for the consistency in participants' assembly behavior, especially given that for some towers there were as many as several thousand valid ways to reconstruct them? One possibility is that shared mechanisms for physical understanding lead to similar mental simulations in planning (Proffitt & Gilden, 1989; Smith & Vul, 2013; Spelke & Kinzler, 2007). Alternatively, the consistency we see in people's initial strategies might have been driven more by participants' use of simple rules and heuristics (e.g., to

build layer by layer; (Shelton et al., 2022)). While our random agent baseline simulates the minimum level of consistency expected under the physical constraints of the task, alternative algorithms could be used to evaluate specific hypotheses concerning the source of homogeneity in participants' solutions. For example, one possibility is that people build "greedily," initially prioritizing larger blocks that cover more of the silhouette, but gradually updating the value of these initial actions in light of whether their reconstruction was ultimately successful (Barto et al., 1995).

Another possibility is that the consistency we observed reflects a tendency for participants to decompose these towers into visual parts in similar ways, and that these parts form the basis for how they then build these towers. Supposing visual organization does serve to structure construction behavior, what characterizes the parts that people favor? Identifying the parts that people use to parse visual objects has long been a central target for classical theories of perceptual organization, which have emphasized spatial and shape-based cues to parthood (Hoffman & Richards, 1984; Palmer, 1977; Schyns & Murphy, 1994; Tversky & Hemenway, 1984; Wertheimer, 1923). Building on this tradition, a related notion is that the parts people use to parse a complex visual object are those that are easy to identify and remember (e.g., according to Gestalt or other principles), and can be used to form more compressed representations of other, similar objects (Biederman, 1987). In other words, people confronting an assembly problem may invoke a mental library containing these useful part concepts to imagine a compact motor program that could be executed to generate the target object from those parts (Ellis et al., 2020; Lake et al., 2015; Tian et al., 2020; *Wong et al., 2022). On this view, the value placed on parts that appear in different objects suggests a route by which prior experience with specific objects guides the kinds of representational primitives that emerge. Future studies could test these ideas by manipulating the prevalence of different parts in the set of objects people are asked to build, and measuring the impact of exposure to these parts on the assembly procedures they converge upon.

A major focus of the current study was on how practice building an object affects a person's approach to building it later. To what degree does such building experience not only affect how people build it later, but also its underlying mental representation, such that they perceive or remember it differently? This question has been explored in prior work investigating other visual production modalities, such as drawing (Fan et al., 2018; Wammes et al., 2016) and handwriting (James, 2017; James, 2010). For example, in one recent study, participants who repeatedly produced drawings of similar objects (e.g., beds and chairs) were better able to discriminate them in a subsequent categorization task, relative to control objects that were not repeatedly drawn (Fan et al., 2018). Moreover, this drawing practice was accompanied by changes in patterns of connectivity between visual and parietal cortex, suggesting a potential neural substrate for experience to intervene upon as people improve their ability to transform the contents of a perceptual representation into representational actions (Fan et al., 2020). A promising direction for future work is to test the degree to which practice plays a similar role in the context of physical assembly, thus providing a measure of how strongly these production-driven learning consequences generalize beyond the domain of drawing and handwriting (Schwartenbeck et al., 2021). Insofar as they do, such findings would lend support to the notion that, at least in some contexts, how people internally represent an object is characterized by a fundamental duality — its correspondence to a static entity with certain perceptual properties, but also to a generative process that gives rise to it (Fan et al., 2018; Fernandes et al., 2018; James, 2017; Lake et al., 2015). Regardless, the results of such studies will be invaluable for advancing our mechanistic understanding of how active and constructive behaviors relate to learning more generally (Chi & Wylie, 2014).

One limitation of our study as it pertains to real-world physical assembly is the focus on building 2D block towers in a virtual environment. While our virtual building environment retained some key aspects of building objects in the physical world, including the relevance of gravity and friction for reasoning about physical stability, there were

many other aspects that were not retained in this environment, such as depth information and the biomechanical details governing how a person would actually need to grip a 3D object in order to maneuver it into place. Future work exploring physical assembly could overcome these drawbacks by using recently developed 3D virtual environments to investigate more realistic forms of interaction (Gan et al., 2020, 2021) and could further connect with research in robotics exploring how data from sight and touch might be integrated in order to plan complex actions in the real world (Erdogan et al., 2014; Fazeli et al., 2019; Mason, 2018). The generality and scope of our findings might also be extended by using a larger and more diverse set of towers, which would support investigation of the relationship between various properties of these towers (e.g., size, presence of 'holes') and how difficult they are to build. Moreover, in order to test the specific hypotheses raised earlier concerning the use of hierarchical representations during physical assembly it will be advantageous to use more complex objects in future studies that more clearly support hierarchical decomposition (*McCarthy et al., 2021; *Wong et al., 2022). Another limitation of the current study is the focus on accurate reconstruction of existing physical structures, rather than reasoning about how to build new ones that satisfy more abstract design criteria, such as the need to provide "shelter" for another object (Bapst et al., 2019). Expanding the suite of physical assembly tasks to include these more open-ended design challenges may provide more direct insight into how humans deploy their general-purpose understanding of how the physical world works to create new things.

Finally, our paper introduces and validates an approach for investigating how people learn how to solve physical assembly problems, providing a window into how physical reasoning and planning interact to achieve specific behavioral goals. Such platforms are especially valuable for advancing mechanistic theories of cognition because they support large-scale measurement of complex human behaviors and the evaluation of candidate cognitive models within the same environment. We hope that our findings will inspire further development of mechanistic models that display these and other richly complex

behaviors, and direct comparison of these models' behavior to that of humans. In the long run, strong alignment between empirical studies of human and model behavior may lead to both more robust artificial intelligence and a deeper understanding of human cognition.

1.4 Supplementary Material

1.4.1 Model Parameter Estimates

Accuracy

$$F1Score \sim \ attempt * condition + (1|participant) + (1|target)$$

Supplemental Table 1.1 Parameter estimates for linear mixed effects model used to predict F1 score from attempt (first and final) and condition.

Predictor	Estimate	Std. Error	df	t value	$Pr(>\|t\|)$
Intercept	$7.48e-01$	$2.37e-02$	12.7	31.5	$1.87e-13^{***}$
attemptFinal	$7.59e-02$	$1.09e-02$	1560	7.00	$3.95e-12^{***}$
conditionRepeated	$8.03e-03$	$1.09e-02$	1570	0.737	0.461
attemptFinal:conditionRepeated	$1.82e-02$	$1.54e-02$	1560	1.19	0.235

Number of blocks

$$numBlocks \sim \ attempt * condition + (1|participant) + (1|target)$$

Supplemental Table 1.2 Parameter estimates for linear mixed effects model used to predict number of blocks from attempt (first and final) and condition.

Predictor	Estimate	Std. Error	df	t value	$Pr(>\|t\|)$
Intercept	8.26	0.355	11.1	23.3	$8.83e-11^{***}$
attemptFinal	1.19	0.160	$1.57e3$	7.41	$2.11e-13^{***}$
conditionRepeated	0.0420	0.161	$1.57e3$	0.264	0.792
attemptFinal:conditionRepeated	0.167	0.227	$1.57e3$	0.735	0.463

Supplemental Table 1.3 Parameter estimates for linear mixed effects model used to predict number of blocks from attempt (first and final) and condition, excluding trials in which the trial ended early due to a block falling.

Predictor	Estimate	Std. Error	df	t value	$Pr(>\|t\|)$
Intercept	8.83	0.333	12.8	26.5	$1.46e-12^{***}$
attemptFinal	1.02	0.150	1320	6.83	$1.31e-11^{***}$
conditionRepeated	-0.0160	0.153	1320	-0.105	0.917
attemptFinal:conditionRepeated	0.207	0.212	1320	0.978	0.328

Mean time between block placements

$$timeBetweenBlocks \sim \ attempt + condition + (1|participant) + (1|target)$$

Supplemental Table 1.4 Parameter estimates for linear mixed effects model used to predict the mean time (seconds) between block placements from attempt (first and final) and condition.

Predictor	Estimate	Std. Error	df	t value	$Pr(>\|t\|)$
Intercept	6.40	0.327	11.6	19.6	$3.18e-10^{***}$
attemptFinal	-1.34	0.0990	1570	-13.5	$<2e-16^{***}$
conditionRepeated	-0.161	0.0998	1570	-1.62	0.106

Preparation time

$$preparationTimeSeconds \sim \ attempt * condition + (1|participant) + (1|target)$$

Supplemental Table 1.5 Parameter estimates for linear mixed effects model used to predict preparation time from attempt (first and final) and condition.

Predictor	Estimate	Std. Error	df	t value	$Pr(>\|t\|)$
Intercept	9.190	0.392	14.7	23.420	$4.84e-13^{***}$
attemptFinal	-2.24	0.260	1570	-8.639	$<2e-16^{***}$
conditionRepeated	0.0955	0.261	1570	0.366	0.714
attemptFinal:conditionRepeated	-0.618	0.367	1570	-1.684	0.0924

Build time

$$buildTimeSeconds \sim attempt * perfectScore + condition + (1|participant) + (1|target)$$

Supplemental Table 1.6 Parameter estimates for linear mixed effects model used to predict total build time (in seconds) from attempt (first and final), condition, and variable indicating whether the reconstruction was perfect.

Predictor	Estimate	Std. Error	df	t value	$Pr(>\|t\|)$
Intercept	53.0	0.853	14.0	62.1	$<2e-16^{***}$
attemptFinal	-1.92	0.451	1330	-4.25	$2.25e-05^{***}$
perfectScore	-3.81	0.854	1420	-4.47	$8.63e-06^{***}$
conditionRepeated	-0.704	0.392	1330	-1.80	0.0725
attemptFinal:perfectScore	-5.04	0.988	1350	-5.10	$3.97e-07^{***}$

Action dissimilarity between consecutive attempts

$$dissimilarity \sim phasePair + measureType * previousF1 + (1|participant) + (1|target)$$

Supplemental Table 1.7 Parameter estimates for linear mixed effects model used to predict action dissimilarity from attempt pair, dissimilarity measure, and F1 score of the previous attempt.

Predictor	Estimate	Std. Error	df	t value	$Pr(>\|t\|)$
Intercept	3.73	0.146	263.1	25.6	$< 2e-16^{***}$
phasePair	-0.186	0.0251	2420	-7.40	$1.91e-13^{***}$
measureTypeSet	-0.482	0.163	2390	-2.96	0.00315
previousF1	-0.643	0.159	2420	-4.05	$5.20e-05^{***}$
measureTypeSet:previousF1	-1.10	0.199	2390	-5.54	$3.37e-08^{***}$

Gini coefficients

$$giniCoefficient \sim \ nlevel * attempt + poly(nlevel,2)$$

Supplemental Table 1.8 Parameter estimates for linear model used to predict difference in Gini coefficients from attempt (first and final) and length of action sequence considered (linear and quadratic) (all trials). df = 155.

Predictor	Estimate	Std. Error	t value	$Pr(>\|t\|)$
Intercept	0.401	0.0132	30.3	$< 2e-16^{***}$
sequenceLength	-0.0454	0.00213	-21.3	$< 2e-16^{***}$
attemptFinal	0.112	0.0187	6.02	$1.24e-08^{***}$
poly(sequenceLength, 2)	0.557	0.0547	10.2	$< 2e-16^{***}$
nlevel:attemptFinal	-0.00926	0.00301	-3.08	0.00247

Supplemental Table 1.9 Parameter estimates for linear model used to predict difference in Gini coefficients from attempt (first and final) and length of action sequence considered (linear and quadratic) (perfect reconstructions only). df = 156.

Predictor	Estimate	Std. Error	t value	$Pr(>\|t\|)$
Intercept	0.411	0.0218	18.9	$< 2e-16^{***}$
sequenceLength	-0.0411	0.00351	-11.7	$< 2e-16^{***}$
attemptFinal	0.175	0.0308	5.68	$6.32e-08^{***}$
sequenceLength:attemptFinal	-0.0125	0.00497	-2.52	0.0129

1.5 Acknowledgments

I thank my co-author Judith Fan for her expert guidance and unparalleled patience during the preparation of this manuscript. I also thank David Kirsh and members of the Cognitive Tools Lab at University of California San Diego for their thoughtful feedback on earlier versions of the manuscript.

Chapter 1, in full, is a reprint of material as it appears in McCarthy, W. P., Kirsh, D., & Fan, J. E. (2023). Consistency and variation in reasoning about physical assembly. *Cognitive Science*, *47*(12), e13397. An earlier version of this project was published as McCarthy, W., Kirsh, D., & Fan, J. (2020). Learning to build physical structures better over time. *Proceedings of the 42nd Annual Meeting of the Cognitive Science Society*. The dissertation author was the primary investigator and author of this material.

References

Allen, K. R., Smith, K. A., & Tenenbaum, J. B. (2020). Rapid trial-and-error learning with simulation supports flexible tool use and physical reasoning. *Proceedings of the National Academy of Sciences*, *117*(47), 29302–29310.

Baillargeon, R. (1995). Physical reasoning in infancy. *The cognitive neurosciences*, 181–204.

Bapst, V., Sanchez-Gonzalez, A., Doersch, C., Stachenfeld, K., Kohli, P., Battaglia, P., & Hamrick, J. (2019). Structured agents for physical construction. *International conference on machine learning*, 464–474.

Barto, A. G., Bradtke, S. J., & Singh, S. P. (1995). Learning to act using real-time dynamic programming. *Artificial intelligence*, *72*(1-2), 81–138.

Battaglia, P. W., Hamrick, J. B., & Tenenbaum, J. B. (2013). Simulation as an engine of physical scene understanding. *Proceedings of the National Academy of Sciences*, *110*(45), 18327–18332.

Biederman, I. (1987). Recognition-by-components: A theory of human image understanding. *Psychological review*, *94*(2), 115.

Botvinick, M., & Weinstein, A. (2014). Model-based hierarchical reinforcement learning and human action control. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *369*(1655), 20130480.

Callaway, F., Lieder, F., Das, P., Gul, S., Krueger, P. M., & Griffiths, T. (2018). A resource-rational analysis of human planning. *CogSci*.

Campitelli, G., & Gobet, F. (2004). Adaptive expert decision making: Skilled chess players search more and deeper. *ICGA Journal*, *27*(4), 209–216.

Chase, W. G., & Simon, H. A. (1973). Perception in chess. *Cognitive psychology*, *4*(1), 55–81.

Chi, M. T., & Wylie, R. (2014). The icap framework: Linking cognitive engagement to active learning outcomes. *Educational psychologist*, *49*(4), 219–243.

Cortesa, C. S., Jones, J. D., Hager, G. D., Khudanpur, S., Landau, B., & Shelton, A. L. (2018). Constraints and development in children's block construction. *CogSci*.

Cortesa, C. S., Jones, J. D., Hager, G. D., Khudanpur, S., Shelton, A. L., & Landau, B. (2017). Characterizing spatial construction processes: Toward computational tools to understand cognition. *CogSci*.

Dasgupta, I., Smith, K. A., Schulz, E., Tenenbaum, J. B., & Gershman, S. J. (2018). Learning to act by integrating mental simulations and physical experiments. *BioRxiv*, 321497.

Daw, N. D., Gershman, S. J., Seymour, B., Dayan, P., & Dolan, R. J. (2011). Model-based influences on humans' choices and striatal prediction errors. *Neuron*, *69*(6), 1204–1215.

Dezfouli, A., & Balleine, B. W. (2013). Actions, Action Sequences and Habits: Evidence That Goal-Directed and Habitual Action Control Are Hierarchically Organized. *PLoS Computational Biology*, *9*(12). https://doi.org/10.1371/journal.pcbi.1003364

Ellis, K., Wong, C., Nye, M., Sable-Meyer, M., Cary, L., Morales, L., Hewitt, L., Solar-Lezama, A., & Tenenbaum, J. B. (2020). Dreamcoder: Growing generalizable, interpretable knowledge with wake-sleep bayesian program learning. *arXiv preprint arXiv:2006.08381*.

Éltető, N., & Dayan, P. (2023). Habits of mind: Reusing action sequences for efficient planning. *arXiv preprint arXiv:2306.05298*.

Erdogan, G., Yildrim, I., & Jacobs, R. A. (2014). Transfer of object shape knowledge across visual and haptic modalities. *Proceedings of the Annual Meeting of the Cognitive Science Society*, *36*(36).

Fan, J. E., Wammes, J. D., Gunn, J. B., Yamins, D. L., Norman, K. A., & Turk-Browne, N. B. (2020). Relating visual production and recognition of objects in human visual cortex. *Journal of Neuroscience*, *40*(8), 1710–1721.

Fan, J. E., Yamins, D. L., & Turk-Browne, N. B. (2018). Common object representations for visual production and recognition. *Cognitive science*, *42*(8), 2670–2698.

Fazeli, N., Oller, M., Wu, J., Wu, Z., Tenenbaum, J. B., & Rodriguez, A. (2019). See, feel, act: Hierarchical learning for complex manipulation skills with multisensory fusion. *Science Robotics*, *4*(26), eaav3123.

Fernandes, M. A., Wammes, J. D., & Meade, M. E. (2018). The surprisingly powerful influence of drawing on memory. *Current Directions in Psychological Science*, *27*(5), 302–308.

Gan, C., Schwartz, J., Alter, S., Schrimpf, M., Traer, J., De Freitas, J., Kubilius, J., Bhandwaldar, A., Haber, N., Sano, M., et al. (2020). Threedworld: A platform for interactive multi-modal physical simulation. *arXiv preprint arXiv:2007.04954*.

Gan, C., Zhou, S., Schwartz, J., Alter, S., Bhandwaldar, A., Gutfreund, D., Yamins, D. L., DiCarlo, J. J., McDermott, J., Torralba, A., et al. (2021). The threedworld transport challenge: A visually guided task-and-motion planning benchmark for physically realistic embodied ai. *arXiv preprint arXiv:2103.14025*.

Gobet, F., & Simon, H. A. (1998). Expert chess memory: Revisiting the chunking hypothesis. *Memory*, *6*(3), 225–255.

Hamrick, J. B., Allen, K. R., Bapst, V., Zhu, T., McKee, K. R., Tenenbaum, J. B., & Battaglia, P. W. (2018). Relational inductive bias for physical construction in humans and machines. *arXiv preprint arXiv:1806.01203*.

Hamrick, J. B., Smith, K. A., Griffiths, T. L., & Vul, E. (2015). Think again? the amount of mental simulation tracks uncertainty in the outcome. *CogSci*.

Hegarty, M. (2004). Mechanical reasoning by mental simulation. *Trends in Cognitive Sciences*, *8*(6), 280–285.

Hoffman, D. D., & Richards, W. A. (1984). Parts of recognition. *Cognition*, *18*(1-3), 65–96.

Huys, Q. J., Eshel, N., O'Nions, E., Sheridan, L., Dayan, P., & Roiser, J. P. (2012). Bonsai trees in your head: How the pavlovian system sculpts goal-directed choices by pruning decision trees. *PLoS Computational Biology*, *8*(3).

Huys, Q. J., Lally, N., Faulkner, P., Eshel, N., Seifritz, E., Gershman, S. J., Dayan, P., & Roiser, J. P. (2015). Interplay of approximate planning strategies. *Proceedings of the National Academy of Sciences of the United States of America*, *112*(10), 3098–3103. https://doi.org/10.1073/pnas.1414219112

James, K. H. (2017). The importance of handwriting experience on the development of the literate brain. *Current Directions in Psychological Science*, *26*(6), 502–508.

James, K. H. (2010). Sensori-motor experience leads to changes in visual processing in the developing brain. *Developmental science*, *13*(2), 279–288.

Kirsh, D. (1995). The intelligent use of space. *Artificial intelligence*, *73*(1-2), 31–68.

Kubricht, J. R., Holyoak, K. J., & Lu, H. (2017). Intuitive physics: Current research and controversies. *Trends in cognitive sciences*, *21*(10), 749–759.

Kurth-Nelson, Z., Behrens, T., Wayne, G., Miller, K., Luettgau, L., Dolan, R., Liu, Y., & Schwartenbeck, P. (2023). Replay and compositional computation. *Neuron*, *111*(4), 454–469.

Lake, B. M., Salakhutdinov, R., & Tenenbaum, J. B. (2015). Human-level concept learning through probabilistic program induction. *Science*, *350*(6266), 1332–1338.

Maglio, P. P., & Kirsh, D. (1996). Epistemic action increases with skill. *Proceedings of the Eighteenth Annual Conference of the Cognitive Science Society*, 391–396.

Mason, M. T. (2018). Toward robotic manipulation. *Annual Review of Control, Robotics, and Autonomous Systems*, *1*, 1–28.

McCarthy, W., Kirsh, D., & Fan, J. (2020). Learning to build physical structures better over time. *Proceedings of the 42nd Annual Meeting of the Cognitive Science Society*.

*McCarthy, W. P., *Hawkins, R. D., Wang, H., Holdaway, C., & Fan, J. E. (2021). Learning to communicate about shared procedural abstractions. *Proceedings of the Annual Meeting of the Cognitive Science Society*, *43*.

McCarthy, W. P., Kirsh, D., & Fan, J. E. (2023). Consistency and variation in reasoning about physical assembly. *Cognitive Science*, *47*(12), e13397.

McCloskey, M. (1983). Intuitive physics. *Scientific American*, *248*(4), 122–131.

Palmer, S. E. (1977). Hierarchical structure in perceptual representation. *Cognitive Psychology*, *9*(4), 441–474.

Proffitt, D. R., & Gilden, D. L. (1989). Understanding natural dynamics. *Journal of Experimental Psychology: Human Perception and Performance*, *15*(2), 384.

Sanborn, A. N., Mansinghka, V. K., & Griffiths, T. L. (2013). Reconciling intuitive physics and newtonian mechanics for colliding objects. *Psychological Review*, *120*(2), 411.

Schwartenbeck, P., Baram, A., Liu, Y., Mark, S., Muller, T., Dolan, R., Botvinick, M., Kurth-Nelson, Z., & Behrens, T. (2021). Generative replay for compositional visual understanding in the prefrontal-hippocampal circuit. *bioRxiv*.

Schwartz, D. L., & Black, T. (1999). Inferences through imagined actions: Knowing by simulated doing. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *25*(1), 116.

Schyns, P. G., & Murphy, G. L. (1994). The ontogeny of part representation in object concepts. *The psychology of learning and motivation*, *31*, 305–349.

Shelton, A. L., Davis, E. E., Cortesa, C. S., Jones, J. D., Hager, G. D., Khudanpur, S., & Landau, B. (2022). Characterizing the details of spatial construction: Cognitive constraints and variability. *Cognitive Science*, *46*(1), e13081.

Sheridan, H., & Reingold, E. M. (2017). Chess players' eye movements reveal rapid recognition of complex visual patterns: Evidence from a chess-related visual search task. *Journal of vision*, *17*(3), 4–4.

Smith, K. A., Battaglia, P. W., & Vul, E. (2018). Different physical intuitions exist between tasks, not domains. *Computational Brain & Behavior*, *1*(2), 101–118.

Smith, K. A., & Vul, E. (2013). Sources of uncertainty in intuitive physics. *Topics in Cognitive Science*, *5*(1), 185–199.

Solway, A., & Botvinick, M. M. (2012). Goal-directed decision making as probabilistic inference: A computational framework and potential neural correlates. *Psychological Review*, *119*(1), 120.

Solway, A., & Botvinick, M. M. (2015). Evidence integration in model-based tree search. *Proceedings of the National Academy of Sciences*, *112*(37), 11708–11713.

Spelke, E. S., & Kinzler, K. D. (2007). Core knowledge. *Developmental science*, *10*(1), 89–96.

Tian, L., Ellis, K., Kryven, M., & Tenenbaum, J. (2020). Learning abstract structure for drawing by efficient motor program induction. *Advances in Neural Information Processing Systems*, *33*, 2686–2697.

Tversky, B., & Hemenway, K. (1984). Objects, parts, and categories. *Journal of Experimental Psychology: General*, *113*(2), 169.

Van Harreveld, F., Wagenmakers, E.-J., & Van Der Maas, H. L. (2007). The effects of time pressure on chess skill: An investigation into fast and slow processes underlying expert performance. *Psychological research*, *71*, 591–597.

van Opheusden, B., Galbiati, G., Bnaya, Z., Li, Y., & Ma, W. J. (2017). A computational model for decision tree search. *CogSci*.

van Opheusden, B., Kuperwajs, I., Galbiati, G., Bnaya, Z., Li, Y., & Ma, W. J. (2023). Expertise increases planning depth in human gameplay. *Nature*, *618*(7967), 1000–1005.

van Opheusden, B., & Ma, W. J. (2019). Tasks for aligning human and machine planning. *Current Opinion in Behavioral Sciences*, *29*, 127–133.

Wammes, J. D., Meade, M. E., & Fernandes, M. A. (2016). The drawing effect: Evidence for reliable and robust memory benefits in free recall. *Quarterly Journal of Experimental Psychology*, *69*(9), 1752–1776.

Wertheimer, M. (1923). Laws of organization in perceptual forms. *Psycologische Forschung*, *4*.

Wolfgang, C. H., Stannard, L. L., & Jones, I. (2001). Block play performance among preschoolers as a predictor of later school achievement in mathematics. *Journal of Research in Childhood Education*, *15*(2), 173–180.

*Wong, C., *McCarthy, W. P., *Grand, G., Friedman, Y., Tenenbaum, J. B., Andreas, J., Hawkins, R. D., & Fan, J. E. (2022). Identifying concept libraries from language about object structure. *Proceedings of the Annual Meeting of the Cognitive Science Society*, *44*.

Xia, L., & Collins, A. G. E. (2020). Temporal and state abstractions for efficient learning, transfer and composition in humans. *bioRxiv*.

Chapter 2

How does assembling an object affect memory for it?

INTERIM SUMMARY

In Chapter 1, I presented a set of methodological tools for studying physical construction, including a virtual environment for running physical assembly experiments in a web browser, and a series of metrics for measuring construction performance and behavior. These tools allowed me to quantify changes in the *procedures* people used built things as they gained experience assembling those objects. I found that people's ability to assemble objects improved with experience, and that these improvements reflected convergence on a small set of strategies. Another way that assembly practice might influence our ability to build is by impacting the way we *represent* the objects we intend to build. In (W. P. McCarthy et al., 2021) I explored the relationship between the procedures people learn for building an object, and their *perceptual decompositions* of those objects. Having found no reliable relationships between the procedures people learn and their perceptual representations, I turn next to a cognitive mechanism that we have more reason to suspect of being influenced by interactive behaviors: *memory*. In the next chapter I explore how experience building an object impacts our memory of it, through a series of experiments designed to disentangle the contributions of active engagement from those of visual exposure.

Abstract

What impacts what we remember about objects we have just encountered? Influential theories of learning suggest that more active engagement leads to stronger memories than passive observation. However, it is not clear which aspects of interaction lead to stronger memories, nor what kinds of memories are supported by active engagement. Here we conduct several experiments to investigate the impact of assembling an object on subsequent recognition and recall performance. We found that reconstructing a block tower by copying it part-by-part could *impair* subsequent memory for that tower, compared to passively viewing that tower. By contrast, when participants initially encoded each tower by building it from working memory, their subsequent recall was enhanced relative to when they held the tower in working memory without building it. Together our results suggest a complex relationship between the nature of our interactions with objects and our subsequent memories of them.

Keywords: memory; working memory; construction; active learning; encoding specificity; procedural memory

2.1 Introduction

To interact with the world in complex ways, we need to remember things about the objects we have interacted with. Sometimes, all we need to remember about an object is whether or not we have seen it before (Brady et al., 2008; Standing, 1973). Other time, we need to remember specific details about our prior interactions. What determines the kinds of information we remember about objects we encounter, and what about our interactions with objects determines how well we remember them?

A substantial body of prior work had found that more *active* forms of encoding, in contrast to more passive observation, lead to stronger memories (Bonwell & Eison, 1991; Chi, 2009; Craik & Lockhart, 1972; Markant et al., 2016). These findings suggest that people will remember more about objects they actively manipulate, compared to those they just see. Indeed, actively rotating 3D objects does lead to better recognition of those objects compared to passively viewing the same sequence of images (Harman et al., 1999). Some forms of interaction may be particularly beneficial to memory. Many memory researchers have identified strong mnemonic benefits of *generation*: people are more likely to remember words (Bertsch et al., 2007; Slamecka & Graf, 1978) and numbers (Crutcher & Healy, 1989) when they have generated them as answers to questions, compared to when those same answers are given to them. These findings suggests that visual memory might also benefit from generative processes, such as altering an object's appearance, or even constructing an object from scratch. Moreover, production of visual objects (i.e. drawing) has been shown to support memory for depicted words and concepts (Fernandes et al., 2018; Wammes et al., 2016), however, whether constructing a visual object strengthens memory of the object itself is less clear.

The experience of constructing an object is a complex physical and cognitive act that could impact memory in various ways, from providing more visual exposure, to "deeper" or embodied processing through multiple sensory channels (Craik & Lockhart, 1972), to

practice "retrieving" objects from memory (Roediger III & Karpicke, 2006; Rowland, 2014; Schuetze et al., 2019). A unique but perhaps critical aspect of construction is the sequence of transformative actions performed. The procedural learning (W. McCarthy et al., 2020; Ryle & Tanney, 2009) that occurs during this process may be intimately related to how we visually represent objects (Lake et al., 2015; Yildirim et al., 2020). On the other hand, our memory of how an object looks might be entirely independent of our memory of how to build an object, which we may only observe in decoding contexts that leverage that information.

In general, the way in which we probe different kinds of memory may have a critical effect on the results we observe. The standard measure of visual recognition memory– asking whether or not someone has seen a stimulus before– may reveal whether someone has stored some aspect of a stimulus in memory, but not which aspects of the stimulus were used to make those judgements (Brady et al., 2008). Theories of verbal and concept memory distinguish between *recognition* (or "familiarity") and *recall* (Yonelinas, 2002), tests of which are able to provide richer readouts of memory. This had led some researchers to explore *visual production* (i.e. drawing) to provide more detailed insight into the *contents* of visual memory (Bainbridge et al., 2019). These generative readouts may be especially sensitive to memories formed during construction, by providing a decoding context that is consistent with how the objects are encoded (Godden & Baddeley, 1975; Tulving & Thomson, 1973).

2.2 General Methods

In this paper, we present a series of 4 experiments designed to assess the impact of generative visual encoding tasks on subsequent memory of objects. We use a task domain with objects that can themselves be constructed– 2D block towers– allowing us to compare the impact of generative experience on recognition as well as recall. All experiments

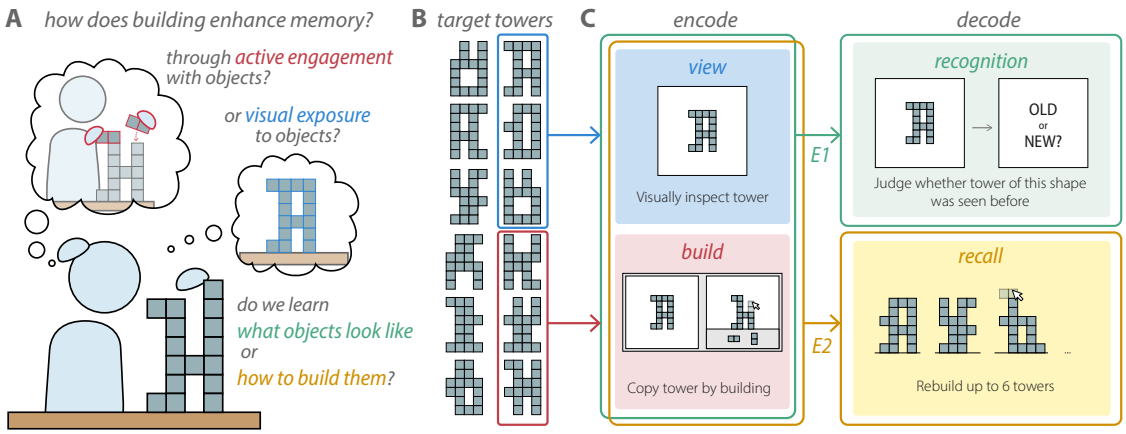

Figure 2.1. Building might impact memory simply by being more "active", but might also require existing memories to strengthen or elaborate. It could impact our ability to recognize the things we build, or our memories of how to build them (A). Target block towers can be built from 8 blocks (B). 3 towers were assigned to each encoding task (C left). In the *View* task, participants inspected the tower for 15 seconds. In the *Build* task, participants rebuilt the tower. We tested recognition (Experiment 1) by asking participants if they had seen each tower before (top-right); we tested recall (Experiment 2) by asking participants to rebuild each tower from memory (bottom-right).

reported consisted of an **encoding phase** and **decoding phase**. In each **encoding phase**, each participant viewed 6 block towers that were randomly split between two encoding conditions, *View* and *Build*. Encoding tasks for each condition varied slightly across experiments. In each **decoding phase**, memory of these towers was tested with an assessment of recognition or recall.

Stimuli

To design a set of visually homogeneous stimuli that could be generated with distinct sequences of actions, we generated a set of 2D block towers (Fig 2.1B). Each tower was constructed out of 8 dominoes, 4 horizontal and 4 vertical (i.e. *2x1* and *1x2* blocks), and fit within a *4x6* grid.

Participants

Participants (18+ years, from USA and UK) were recruited online using the Prolific platform and were paid approximately $16 per hour for their time (20-30 minutes). For

E1 and E2, we recruited participants until 50 participants completed each study without meeting any of our pre-defined exclusion criteria. For E3 and E4, we recruited participants until 50 participants in each group completed the study.

2.3 Experiment 1: Impact of building objects on visual recognition

We manually selected a subset of 12 block towers to be shown to all participants (Fig 2.1B). For each participant, the 12 towers were randomly divided into sets of 6 *target towers* and 6 *foils*. The 6 target towers were randomly split between two conditions– *Build* and *View*– and were all presented in the same color.

Encoding

Participants were informed that their memory for the shape of each tower would be tested later in the experiment. All 6 target towers were presented in a psuedorandom order. *View* towers were displayed on screen for 15000ms, and participants were instructed to "study the shape of the tower" for the entire time it is on screen (Fig 2.1C, upper-left). *Build* towers were presented alongside a building interface: a gridworld environment where blocks could be picked up and placed on any supporting surface by clicking with the mouse. Participants were instructed to "copy the tower" by building it in the environment. Blocks could not be moved once placed, however, the building environment could be reset at any time, and undo/ redo was available. When the participants had perfectly reconstructed the target tower, they automatically proceeded to the next trial (Fig 2.1C, lower-left).

Decoding

Visual recognition memory was measured with an **old-new** task (Fig 2.1C, upper-right). Participants were presented with the target towers one-by-one, randomly interleaved with foils, and asked to indicate whether they had seen the presented tower in the previous phase by keypress.

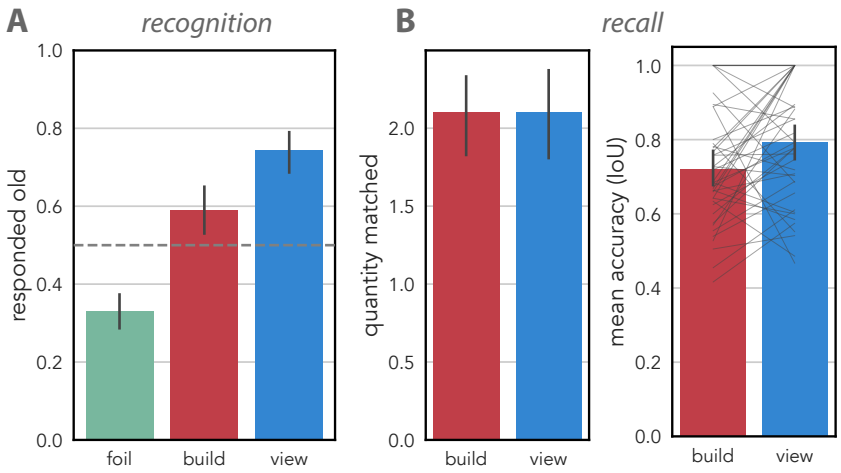

Figure 2.2. Participants correctly responded 'old' to *View* stimuli more often than to *Build* (A). Participants recalled roughly the same amount of *Build* and *View* towers, and those they did recall were of roughly the same accuracy. Error bars in all plots represent 95% CI.

2.3.1 Results

We excluded 8 participants for incomplete data. To determine whether participants had any ability to discriminate between old and new stimuli, we created bootstrapped distributions of the number of times participants responded "old," separately for target towers and foils (Fig 2.2A). Distributions and confidence intervals were created by resampling over 1000 iterations; in each bootstrap iteration we sampled participants with replacement and included all data from a participant every time they were sampled. We found that participants responded "old" more often to target towers ($0.667,\ 95\%\ CI: [0.62,\ 0.708]$) than to foils ($0.33,\ 95\%\ CI: [0.283,\ 0.377]$) ($p = 0$), confirming that they could, in general, discriminate between towers they had seen and those they had not.

We also found that participants were more likely to respond "old" to *View* towers ($0.743,\ 95\%\ CI: [0.683,\ 0.793]$) than to *Build* towers ($0.59,\ 95\%\ CI: [0.527,\ 0.653]$) ($p = 0$). This was particularly surprising given that participants took on average 61.1s ($95\%\ CI: [60.8,\ 61.3]$) to complete each *Build* trial, far longer than the 15s exposure in

the *View* trials. Primarily, however, it conflicts with the prediction that the more active task, building, would lead to stronger memories than the viewing task, which required no overt activity at all.

2.4 Experiment 2: Impact of building objects on visual recall

We had several hypotheses about why building a tower might lead to worse memories, however we first sought to establish whether this phenomena was isolated to visual recognition, or extended to other forms of memory. Recall– the ability to bring an item to mind without a related cue– provides an opportunity for participants to share contents of memory, even if it does not reach threshold for visual recognition. Our task domain provides a natural way of testing visual recall: asking participants to build block towers from memory. Furthermore, this decoding context is highly consistent with the context of encoding (i.e. building towers) (Godden & Baddeley, 1975; Tulving & Thomson, 1973), which may give participants the best chance of leveraging kinds of representations learned during building.

Encoding

The encoding phase was identical to that of Experiment 1, except that participants performed 2 repetitions of each encoding trial. We increased the number of repetitions as we found in piloting that many participants struggled to recall any towers after a single encoding trial, consistent with prior findings that visual recall demands a stronger memory signal than recognition (Yonelinas, 2002).

Decoding

Participants were presented with a building environment almost identical to the one available to them in the *Build* encoding task, minus the target tower. Participants were asked to reconstruct as many towers as they could remember from the previous part of

the study, in any order (Fig 2.1C, lower-right). The experiment ended when a participant submitted 6 towers, or pressed a button indicating that they could not remember any more towers.

2.4.1 Results

We excluded 11 participants for incomplete data. After removing duplicate submissions of towers, participants submitted an average of 4.2 towers ($95\%\ CI : [3.7,\ 4.64]$). On average, 1.46 ($95\%\ CI : [1.06, 1.84]$) of these towers were perfect reconstructions of a target tower, suggesting that accurately recalling towers of this complexity was a difficult task. Fewer *Build* towers ($0.56, 95\%\ CI : [0.34, 0.78]$) were perfectly recalled than *View* towers ($0.9, 95\%\ CI : [0.62, 1.22]$) ($p = 0.020$), providing initial evidence that building did not benefit recall memory.

To measure accuracy of the imperfect reconstructions we calculated the "Intersection Over Union" (IoU): the area of overlap between target and reconstruction, divided by the total area covered by both, allowing for horizontal translation. Imperfect reconstructions present a challenge for analysis: how should we identify which target towers participants were attempting to reconstruct? We made an assumption– that each unique tower built in the recall phase corresponded to a genuinely recalled target tower. To map these recalled towers to their intended targets, we calculated the IoU between every reconstruction and target, then found the mapping that maximizes the mean score. We found no reliable difference between the number of towers paired to targets from the *Build* ($2.1, 95\%\ CI : [1.82,\ 2.34]$) and *View* ($2.1, 95\%\ CI : [1.8,\ 2.38]$) conditions ($p = 0.440$) (Fig 2.2B). However, we did find that participants who recalled towers from both conditions generally built more accurate reconstructions of *View* condition towers ($p = 0.0208$, Cohen's $d = 0.433$), revealed by a paired t-test between reconstruction means in each condition (Fig 2.2C).

In sum, these results point to a moderate recall advantage for towers in the *View*

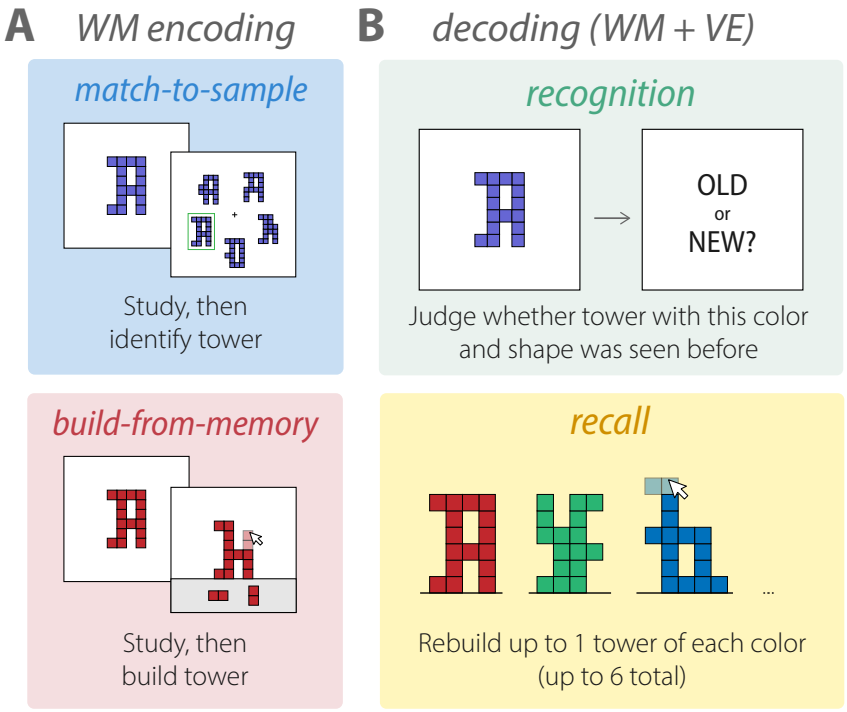

Figure 2.3. Experiments 3 and 4 compared the effect of our original encoding tasks with two new encoding tasks (A). In both tasks, participants studied a tower until it disappeared. They were then asked to either identify the tower from a group (top) or to rebuild the tower (bottom). Different colors were used for each tower, allowing us to measure recognition and recall of specific towers (B).

condition, compared to *Build*, which is also at odds with the prediction that more active engagement leads to stronger memories. This also happened despite the highly similar encoding and decoding contexts in the *Build* condition, suggesting that if generative encoding can actually benefit visual memory, something about the generative experience our participants are engaging in is failing to induce this effect, or is interfering with memory in some way.

2.5 Experiment 3: Impact of building from working memory on visual recognition

Why did participants not remember the towers they built better than the ones they viewed? Much of the prior work demonstrating mnemonic benefits of "generation" investigates processes of reconstructing or generating an example or word from memory or from an internal thought process. Retrieval from an internal representation may serve to reinforce prior representations through retrieval (Fan & Turk-Browne, 2013; Roediger III & Karpicke, 2006; Rowland, 2014; Schuetze et al., 2019), or link these representations to novel experiences. Our building task, in contrast, asks people to copy an object that already exists in the world, meaning it could in principle be completed without any holistic representation of the object. If, for example, participants reconstructed towers by iteratively determining which one block should be placed next, they may have never associated their actions with a representation of what any particular tower looked like. Moreover, if participants learned that they only needed to attend to individual blocks, they may have stopped attending to the entire the tower.

In Experiments 3 and 4, we aimed to test whether a pre-existing visual memory is a prerequisite for a mnemonic advantage of building. We introduced two new encoding tasks that each required participants to hold a representation of an entire tower in working memory before performing some an adapted *Build* or *View* task. Similarly to Experiments 1 and 2, Experiment 3 tests visual recognition and Experiment 4 tests recall.

Stimuli

For each of the target towers used in Experiments 1 and 2, we derived a set of 5 *distractors* by performing the following transformations: horizontal flip, vertical flip, 180 degree rotation, lower half swapped with upper half, and left half swapped with right half. We sampled one of these distractors to act as the *foil* in the old-new decoding task. The

remaining 5 became distractors in the *match-to-sample* encoding task, described below. We randomly sampled sets of 6 target towers until all of the target towers and derived distractors were distinct, and presented this set to all participants in Experiments 3 and 4. Each target tower and its corresponding distractors were assigned one of six colors. As with Experiments 1 and 2, target towers were randomly split between *Build* and *View* conditions for each participant.

Encoding

Participants in the **Visual Exposure** group performed the same *Build* and *View* tasks from Experiments 1 and 2. Participants in the **Working Memory** group performed *modified Build* and *View* tasks that required participants to visually encode each tower before responding. Prior to each Working Memory task, the target tower was displayed on screen for 8000ms and participants were prompted to "study" the shape of the tower. Then, for towers in the *View* condition, participants performed a **match-to-sample task**: they were presented with a centered fixation cross, followed by a circular array of 5 towers– the 4 sampled distractor towers plus the target tower. Participants were instructed to select the tower they had just studied by clicking on it, after which they received feedback. For towers in the *Build* condition, participants performed a **build-from-memory task**: they were presented with an empty building environment, with blocks in the same color as the tower they had just viewed, and prompted to build the target tower from memory. They could submit a tower once they had placed 8 blocks. They received feedback after submission (correct or incorrect), and the target tower was revealed in an adjacent window to allow comparison with their reconstruction.

Decoding

Experiment 3 used the same old-new task from Experiment 1, except that participants saw two trials of each color: one *target* and the randomly sampled *foil* generated from that target.

2.5.1 Results

We excluded 11 participants for failing to complete all trials, leaving 50 in each group. We first analyze performance in the Working Memory encoding phase. In the *match-to-sample* task, participants correctly selected the target tower from the 5 distractors on 91.5% of trials (95% CI : [86.3, 95.8]), suggesting that they successfully encoded the target towers in working memory. In the *build-from-memory* task, participants perfectly reconstructed the target tower on 73.3% of trials (95% CI : [0.688, 0.774]), consistent with this being a more difficult task.

As with Experiment 1, the Visual Exposure group responded "old" to target towers (0.807, 95% CI : [0.76, 0.853]) more often than to foils (0.29, 95% CI : [0.243, 0.34]) ($p = 0$). However, while *View* towers (0.833, 95% CI : [0.753, 0.9]) were remembered marginally more often than *Build* (0.78, 95% CI : [0.713, 0.847]) ($p = 0.173$), we did not see a reliable difference between responses (Fig 2.4A left). Convergence between conditions may have been driven by ceiling effects, as the introduction of colors and increased number of repetitions did appear to result in stronger recognition performance overall (75.8% correct, 95% CI : [71.8, 79.7]), relative to Experiment 1.

This explanation is supported by the fact that in the Working Memory condition, where participants' responses were marginally more accurate again (80.1% correct, 95% CI : [76.3, 83.8]), the difference in responses between *Build* (0.88, 95% CI : [0.827, 0.927]) and *View* (0.873, 95% CI : [0.827, 0.92]) was even less distinct ($p = 0.565$) (Fig 2.4A right). In sum, we find no evidence that building from working memory reliably led to better or worse recognition.

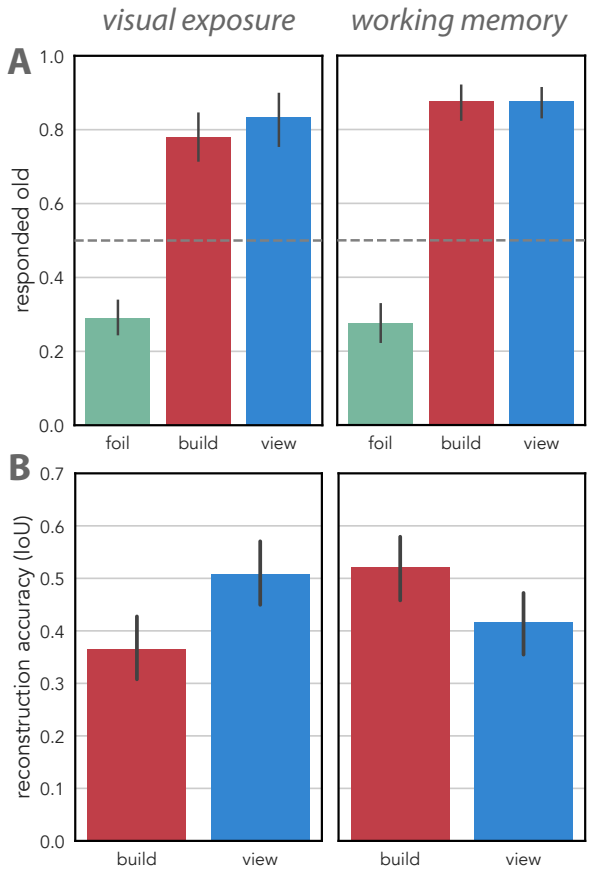

Figure 2.4. Recognition performance was similar for *Build* and *View* (A left), regardless of whether the tower was encoded in working memory (A right). As in Experiment 1, participants *recalled* towers they viewed more accurately than towers they built (B left), unless those towers were first encoded in working memory (B right).

2.6 Experiment 4: Impact of building from working memory on visual recall

Finally, we asked whether building from memory impacts visual recall. We introduced block towers of different colors to provide a way of inferring which towers participants were attempting to recall, as well as provide an additional channel by which participants could discriminate between towers.

Encoding

The encoding phase was identical to the encoding phase in Experiment 3.

Decoding

As in Experiment 2, participants were presented with an empty building environment, and asked to recall as many towers as they could remember from the encoding phase. This time, however, participants first had to select the color of the tower they wanted to build. Once they had placed 8 blocks of their chosen color, then pressed a button to submit their tower and remove that color as an option. The experiment ended when a participant submitted towers of all 6 colors, or pressed a button indicating that they could not remember any more towers.

2.6.1 Results

We excluded 6 participants for failing to complete all trials, and 1 failing to start the decoding task within 10 minutes of finishing the encoding task, leaving 50 participants in each group.

Similarly to Experiment 3, the Working Memory group correctly selected the target tower on 86.7% of *match-to-sample* trials (95% CI : [81.3, 91.7]), and perfectly reconstructed the target tower on 73.9% of *build-from-memory* trials (95% CI : [0.7, 0.778]).

Participants submitted towers on 3.78 towers on average (95% CI : [3.44, 4.11]). The colors of recalled towers provided a mechanism for us to match recalled towers with their intended target stimuli. To compare how different encoding tasks affected recall memory, we fit a mixed-effects logistic regression predicting whether or not a participant submitted a perfect reconstruction of the target tower. We included fixed effects for encoding group (*Visual Exposure* vs. *Working Memory*), encoding context (*Build* vs. *View*), and their interaction; plus random intercepts for participant and tower. We found no evidence that the *Working Memory* tasks reliable led to a better or worse ability to perfectly recall towers ($b = -0.595$, $z = -1.35$, $p = 0.177$). While we did see evidence for a main effect of encoding context, where *Build* towers were recalled less frequently than *View* ($b = -0.879$, $z = -2.52$, $p = 0.0117$), this effect was small compared to a reliable

crossover interaction between encoding task and context ($b = 1.62$, $z = 3.33$, $p < 0.001$): *Build* towers were recalled *more* often than *View* towers when encoded in the *Working Memory* tasks. That is, we see evidence for stronger memories of built towers than viewed towers when building follows prior encoding of the tower.

To verify this finding, we fit a model of the same structure, predicting the *accuracy* of each reconstruction for every target tower, treating towers that were not reconstructed as $IoU = 0$. Again, we found no reliable effect of encoding condition ($b = -0.09261$, $t = -1.58$, $p = 0.116$), a small negative main effect of the *Build* condition ($b = -0.143$, $t = -3.00$, $p = 0.00346$), and a crossover interaction ($b = 0.247$, $t = 3.67$, $p < 0.001$) suggesting that *Build* towers were recalled more accurately than *View* towers in the Working Memory condition (Fig 2.4 B) (and less in the Visual Exposure condition). Together, these results suggest that building a tower from working memory facilitates visual recall, relative to simply viewing a tower.

2.7 Discussion

We asked how generating block towers impacts our subsequent memory of them. We initially compared memory for block towers that participants copied with block towers that they simply viewed on screen, and found that the towers people copied were recognized less frequently and recalled less accurately. We suspected that building block towers while they were still on screen prevented participants from forming holistic representations of them, and that these might be critical for generation to facilitate memory. Consistent with this interpretation, we found that when participants built towers from working memory, they did remember them better later on. Moreover, this relative memory boost was only apparent in visual recall, not visual recognition, suggesting that generative experience impacted some but not all aspects of memory for the object.

Our work has implications for the applicability of active and generative learning to

visual memory (Crutcher & Healy, 1989; Markant et al., 2016; Slamecka & Graf, 1978). It suggests that more active engagement does not necessarily translate to better memory of a visual stimulus– that the kind of engagement matters. Our finding that building from memory supports recall but not recognition, as well as hinting at distinct processes underlying these two forms of memory (Yonelinas, 2002), suggests that active engagement differentially affects different kinds of memory. Why is recall prioritized in this way? A possible reason is suggested by theories of situated cognition (Roth & Jornet, 2013), that have long stressed that internal representations do not always present the most efficient solution to a cognitive problem: why remember what something looks like when you can easily check by looking? Actions are not perceivable in this way, making it more worthwhile to dedicate cognitive resources to remembering them.

Another key question raised by our study is how building from working memory leads to stronger memories. One possibility is that building from memory requires a large volume of queries of working memory, consolidating any pre-existing representations in longer-term memory through retrieval practice (Roediger III & Karpicke, 2006; Rowland, 2014; Schuetze et al., 2019). Alternatively, generative experience may result in a distinct *kind* of action-based representation, akin to procedural knowledge or "knowledge how" (Anderson, 2013; Ryle & Tanney, 2009). Such representations may elaborate on existing perceptual representations, facilitating processing at a deeper level (Bradshaw & Anderson, 1982; Craik & Lockhart, 1972), or simply constitute a distinct memory trace that can be accessed in future generative contexts. Our results do provide one reason to be skeptical of additional memory formats– a seemingly limited capacity to recall objects. Participants in the Working Memory group did not, in general, recall more towers than the Visual Exposure group, suggesting that the build from memory task served to prioritize memory for certain towers above others, more so than it did to boost memory strength overall.

Our study also raise the question of how goals at encoding time affect memory. We chose not tell participants which Working Memory task they would perform until the

stimulus they were encoding had disappeared. However, goals guide visual attention and attention is crucial for determining what gets encoded in memory (Chun & Turk-Browne, 2007). A straightforward way to test whether goals at encoding time impacted memory would be to tell people in advance what task they will perform, potentially cueing different ways of seeing (Goodwin, 2015) and leading to measurable memory effects downstream.

Finally, the hierarchical structure of our stimuli raises the possiblity of relating fine-grained differences in encoding behavior to downstream memory. One well documented strategy for remembering something is to break it down into memorable "chunks" (Chase & Simon, 1973; Miller, 1956; Orbán et al., 2008), a process that may have occurred implicitly as participants built towers. By analyzing the kinds of errors participants made, we may be able to identify subtowers that they did remember, even when they failed to remember the entire tower. Doing so may help to shed light on the structure of the representations used to support visual recognition and recall (Yonelinas, 2002), and tease apart the impact of generative experience on these representations.

2.8 Acknowledgments

I thank my co-authors Judith Fan and Sean Anderson for their invaluable contributions to this manuscript. I also thank David Kirsh, Marcelo Mattar and members of the Cognitive Tools Lab for insightful discussion at various points in the lifecycle of this project.

Chapter 2, in full, is currently being prepared for submission for publication of the material. Earlier versions of this project were published as McCarthy, W. P., Anderson, S. P., & Fan, J. E. (2024). How does assembling an object affect memory for it? [In press]. *Proceedings of the Annual Meeting of the Cognitive Science Society*, *46*, and as McCarthy, W. P., & Fan, J. E. (2023). Exploring the impact of a constructive encoding task on visual recognition memory. *Journal of Vision*, *23*(9), 5977–5977. The dissertation author was the primary investigator and author of this material.

References

Anderson, J. R. (2013). *The architecture of cognition*. Psychology Press.

Bainbridge, W. A., Hall, E. H., & Baker, C. I. (2019). Drawings of real-world scenes during free recall reveal detailed object and spatial information in memory. *Nature communications*, *10*(1), 5.

Bertsch, S., Pesta, B. J., Wiscott, R., & McDaniel, M. A. (2007). The generation effect: A meta-analytic review. *Memory & cognition*, *35*, 201–210.

Bonwell, C. C., & Eison, J. A. (1991). *Active learning: Creating excitement in the classroom. 1991 ashe-eric higher education reports.* ERIC.

Bradshaw, G. L., & Anderson, J. R. (1982). Elaborative encoding as an explanation of levels of processing. *Journal of Verbal Learning and Verbal Behavior*, *21*(2), 165–174.

Brady, T. F., Konkle, T., Alvarez, G. A., & Oliva, A. (2008). Visual long-term memory has a massive storage capacity for object details. *Proceedings of the National Academy of Sciences*, *105*(38), 14325–14329.

Chase, W. G., & Simon, H. A. (1973). Perception in chess. *Cognitive psychology*, *4*(1), 55–81.

Chi, M. T. (2009). Active-constructive-interactive: A conceptual framework for differentiating learning activities. *Topics in cognitive science*, *1*(1), 73–105.

Chun, M. M., & Turk-Browne, N. B. (2007). Interactions between attention and memory. *Current opinion in neurobiology*, *17*(2), 177–184.

Craik, F. I., & Lockhart, R. S. (1972). Levels of processing: A framework for memory research. *Journal of verbal learning and verbal behavior*, *11*(6), 671–684.

Crutcher, R. J., & Healy, A. F. (1989). Cognitive operations and the generation effect. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *15*(4), 669.

Fan, J. E., & Turk-Browne, N. B. (2013). Internal attention to features in visual short-term memory guides object learning. *Cognition*, *129*(2), 292–308.

Fernandes, M. A., Wammes, J. D., & Meade, M. E. (2018). The surprisingly powerful influence of drawing on memory. *Current Directions in Psychological Science*, *27*(5), 302–308.

Godden, D. R., & Baddeley, A. D. (1975). Context-dependent memory in two natural environments: On land and underwater. *British Journal of psychology*, *66*(3), 325–331.

Goodwin, C. (2015). Professional vision. In *Aufmerksamkeit: Geschichte-theorie-empirie* (pp. 387–425). Springer.

Harman, K. L., Humphrey, G. K., & Goodale, M. A. (1999). Active manual control of object views facilitates visual recognition. *Current Biology*, *9*(22), 1315–1318.

Lake, B. M., Salakhutdinov, R., & Tenenbaum, J. B. (2015). Human-level concept learning through probabilistic program induction. *Science*, *350*(6266), 1332–1338.

Markant, D. B., Ruggeri, A., Gureckis, T. M., & Xu, F. (2016). Enhanced memory as a common effect of active learning. *Mind, Brain, and Education*, *10*(3), 142–152.

McCarthy, W., Kirsh, D., & Fan, J. (2020). Learning to build physical structures better over time. *Proc. 42nd Annu. Meet. Cogn. Sci. Soc.*

McCarthy, W. P., Anderson, S. P., & Fan, J. E. (2024). How does assembling an object affect memory for it? [In press]. *Proceedings of the Annual Meeting of the Cognitive Science Society*, *46*.

McCarthy, W. P., & Fan, J. E. (2023). Exploring the impact of a constructive encoding task on visual recognition memory. *Journal of Vision*, *23*(9), 5977–5977.

McCarthy, W. P., Mattar, M. G., Kirsh, D., & Fan, J. E. (2021). Connecting perceptual and procedural abstractions in physical construction. *Proceedings of the Annual Meeting of the Cognitive Science Society*, *43*.

Miller, G. A. (1956). The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological review*, *63*(2), 81.

Orbán, G., Fiser, J., Aslin, R. N., & Lengyel, M. (2008). Bayesian learning of visual chunks by human observers. *Proceedings of the National Academy of Sciences of the United States of America*, *105*(7), 2745–2750. https://doi.org/10.1073/pnas.0708424105

Roediger III, H. L., & Karpicke, J. D. (2006). Test-enhanced learning: Taking memory tests improves long-term retention. *Psychological science*, *17*(3), 249–255.

Roth, W.-M., & Jornet, A. (2013). Situated cognition. *Wiley Interdisciplinary Reviews: Cognitive Science*, *4*(5), 463–478.

Rowland, C. A. (2014). The effect of testing versus restudy on retention: A meta-analytic review of the testing effect. *Psychological bulletin*, *140*(6), 1432.

Ryle, G., & Tanney, J. (2009). *The concept of mind*. Routledge.

Schuetze, B. A., Eglington, L. G., & Kang, S. H. (2019). Retrieval practice benefits memory precision. *Memory*, *27*(8), 1091–1098.

Slamecka, N. J., & Graf, P. (1978). The generation effect: Delineation of a phenomenon. *Journal of experimental Psychology: Human learning and Memory*, *4*(6), 592.

Standing, L. (1973). Learning 10000 pictures. *Quarterly Journal of Experimental Psychology*, *25*(2), 207–222.

Tulving, E., & Thomson, D. M. (1973). Encoding specificity and retrieval processes in episodic memory. *Psychological review*, *80*(5), 352.

Wammes, J. D., Meade, M. E., & Fernandes, M. A. (2016). The drawing effect: Evidence for reliable and robust memory benefits in free recall. *Quarterly Journal of Experimental Psychology*, *69*(9), 1752–1776.

Yildirim, I., Belledonne, M., Freiwald, W., & Tenenbaum, J. (2020). Efficient inverse graphics in biological face processing. *Science Advances*, *6*(10), eaax5979.

Yonelinas, A. P. (2002). The nature of recollection and familiarity: A review of 30 years of research. *Journal of memory and language*, *46*(3), 441–517.

Chapter 3

Learning to communicate about shared procedural abstractions

INTERIM SUMMARY

In the preceding chapters, I explored the consequences of assembling objects for the procedures we use to build those objects, and for our memory of those objects. The results so far suggest that assembly experience helps individuals build better, by helping us build things more accurately and quickly. They also suggest that our ability to recall how something is made can be improved with practice. Together, these results suggest that our mental representations of how something is made can shift in response to experience. The dynamic nature of these representations pose a challenge for real-world construction, which is frequently carried out by groups of collaborators, rather than individuals. How are people able to communicate effectively about how something is made, if their internal representation of that process is constantly shifting? In this chapter, I explore how collaborators in assembly tasks use natural language to coordinate behavior, even amid changing representations of construction procedures.

Abstract

Successful collaboration requires people to coordinate their behavior in pursuit of a shared goal. In addition to breaking down a task into the same components, collaborators often need to *communicate* about those components in the same way. We investigate the joint coordination of *ad hoc* task abstractions and communicative conventions in a collaborative physical assembly paradigm. One participant (the architect) saw the blueprints for scenes containing block towers, and sent assembly instructions to the other participant (the builder) using a natural-language chat interface. Participants converged on increasingly effective instructions across repeated attempts, using more abstract referring expressions capturing each scene's hierarchical structure. To explain these findings, we present a computational model that integrates recent probabilistic accounts of *ad hoc* convention formation with a neurosymbolic account of procedural chunking. Our results shed light on the fundamental mechanisms that enable intelligent agents to communicate and collaborate so flexibly.

Keywords: language, collaboration, social cognition, program induction, learning

3.1 Introduction

Many real-world tasks are too complex for any one individual to accomplish alone. Instead, multiple people must coordinate their behavior and work as a team (Eccles & Tenenbaum, 2004; Grosz & Kraus, 1996; Salas & Fiore, 2004; Stone et al., 2010; Tannenbaum & Salas, 2020). To work together effectively, team members need to think about what they are doing in the same way, sharing similar mental representations of the relevant procedures at the appropriate level of abstraction for their joint goals (DeChurch & Mesmer-Magnus, 2010; Mathieu et al., 2000; Stout et al., 1999; Waller et al., 2004). For example, consider a group of cooks working together at a restaurant (Fine, 2008; Strouse et al., 2021; R. E. Wang et al., 2021). When a new cook is training in the kitchen, they may need to follow step-by-step instructions at the level of individual ingredients, like *melt 30g butter in the small pan, then stir in 30g of flour.* As they gain more experience, however, they may just *make a roux*, efficiently executing the entire procedure as a single routine. When all cooks are using the same unified *roux* abstraction, it is easier to plan and execute complementary actions without clashing. Similar benefits of shared abstractions are found in other domains, from doubles tennis (Blickensderfer et al., 2010) to nursing care (Apker et al., 2006) and operating rooms (Bogdanovic et al., 2015; Klein et al., 2006; Sexton et al., 2006)

In many cases, however, the relevant abstractions are not available to agents in advance, and achieving the collective benefit of shared abstraction requires *ad hoc* coordination between interacting agents as they individually learn the procedures required for the task at hand (Cooke, 2015; Entin & Serfaty, 1999; S. I. Wang et al., 2017). The ability to communicate using natural language is a powerful tool for solving this coordination problem, allowing people to verbally negotiate roles and instructions (Clark, 1996; Suhr et al., 2019; Tellex et al., 2020). Yet for a communication protocol to be effective in novel task settings, these protocols must *also* be able to update over the course of a

group interaction to refer to newly relevant concepts. For example, if the cooks were asked to make a "vegan roux," they might have some uncertainty over what recipe, exactly, is being referred to. What is expected to replace the butter? The process of forming *common ground* or *pacts* to resolve this uncertainty has been central to psycholinguistics (Clark, 1996; Hawkins, Frank, & Goodman, 2020) and natural language processing (Hawkins, Kwon, et al., 2020; Takmaz et al., 2020; Udagawa & Aizawa, 2021).

In this paper, we address a fundamental question concerning the learning mechanisms that enable teams to meet these challenges: *how are people able to simultaneously coordinate on a shared set of concepts as well as the language for talking about them?* We approach this question by building a computational cognitive model capable of capturing both conceptual and communicative dimensions of successful coordination at once. To model *conceptual* coordination, we draw upon recent developed neurosymbolic models emerging from the classical program synthesis literature (Barsalou, 1999; Dehaene et al., 2022; Goodman et al., 2015; Gulwani et al., 2017). These models formalize concepts as structured, executable *programs*. Through a process known as *library learning*, agents are able to supplement an initial set of primitive concepts with more complex abstractions, or "chunks," as they learn more about a task (Ellis et al., 2021; Kumar et al., 2022; Wong et al., 2021). To model *linguistic* coordination, we draw upon a recently proposed of model of linguistic convention formation as probabilistic social inference over those underlying abstractions (Hawkins et al., 2021, 2023). These two model components are complementary to one another: Library learning provides a mechanism for how individuals acquire new concepts by combining existing ones, but cannot explain how individuals bind words to these new concepts, nor how the same concepts would come to be shared between collaborators. On the other hand, linguistic convention formation provides a computational mechanism whereby teams can coordinate on ways of talking about *existing* concepts, but cannot explain how new concepts arise. When combined, however, these two mechanisms generate specific and testable predictions concerning how mental representations change when a

team encounters a new task.

We evaluate these predictions in a physical assembly domain (Bapst et al., 2019; Bramley & Xu, 2023; McCarthy et al., 2020; Walsman et al., 2022). Participants encounter visual scenes populated by a recurring set of block towers. These scenes are hierarchically organized and can thus be validly represented at multiple levels of abstraction. For instance, a scene might be represented holistically or it might be represented as an assemblage of simpler structural units. As participants are presented with multiple such scenes, the library learning component predicts that certain "chunks" should be preferred, grouping primitive elements (i.e., individual blocks) into more complex units (i.e., configurations of multiple blocks; Aslin et al., 1998; Austerweil & Griffiths, 2013; Christiansen & Chater, 2016). However, these newly formed abstractions are only useful for collaboration if they can be shared with other people—for example, by using language. And using language to communicate about these abstractions requires overcoming the inherent risk of miscommunication that accompanies the use of new terms. As such, the model we propose requires both bootstrapping new abstractions (driven by the functional pressure to efficiently represent structure in the world) as well as coordinating on new links between these abstractions and tokens of language (driven by the functional pressure to be understood). In sum, our paper contributes a novel empirical paradigm, computational model, and set of evaluation metrics that expose core principles of successful teamwork and can be used to guide applications such as the ongoing development of artificial agents that collaborate as flexibly as people do.

3.2 Method

3.2.1 Participants

73 dyads ($N = 146$ human participants) were recruited from Amazon Mechanical Turk and automatically paired up to perform a collaborative assembly task. We excluded

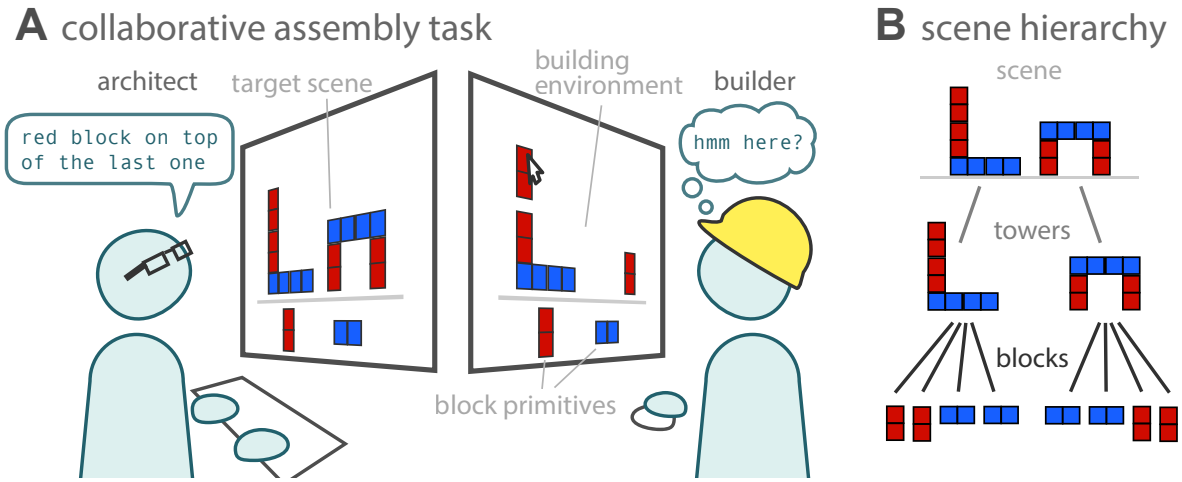

Figure 3.1. Collaborative assembly task. (A) The Architect was shown a target scene and provided assembly instructions to the Builder, who aimed to reconstruct it. (B) Each scene was composed of two towers, which were each composed of four domino-shaped blocks.

24 dyads who failed to meet preregistered criteria (i.e., failing to achieve at least 75% reconstruction accuracy on at least 75% of the trials, self-reporting confusion about task instructions, self-reporting non-fluency in English). Each session lasted approximately 30-50 minutes, and participants were provided with a minimum compensation of $5.00 for task completion, plus a performance bonus of up to $3.00 (see Supplemental Methods for further details). All participants provided informed consent in accordance with IRB.

3.2.2 Procedure

Each participant was assigned a fixed role of *Architect* or *Builder* and proceeded with their partner through a series of twelve trials. On each trial, the Architect was shown a target scene containing block towers (Figure 3.1A). The Builder could not see the target scene, but was shown an empty grid where they could click to place individual domino-like blocks. The Architect was asked to send step-by-step assembly instructions through a free response text box, which the Builder could use to reconstruct the target scene as accurately as possible. The Architect and Builder took as many turns as they needed to reconstruct each scene. On the Architect's turn, they sent a single message containing a

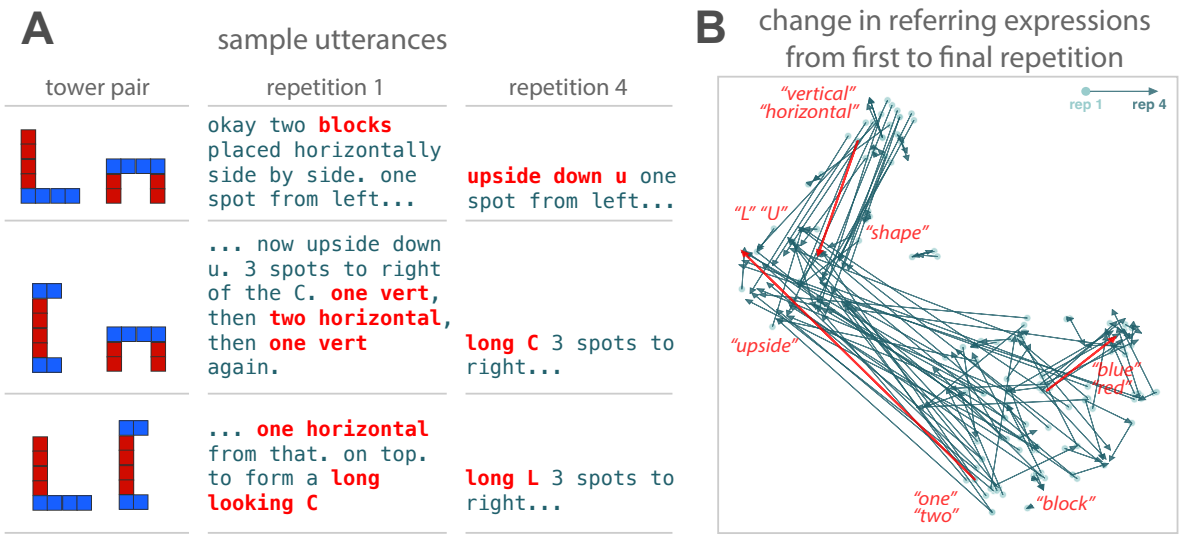

Figure 3.2. A: Example messages from earlier and later repetitions of tower pairs, showing the emergence of tower-level expressions (`upside down U`, `long C`). B: t-SNE visualization of referring expressions mentioned in each repetition. Each line represents the change in the embedding space from a participant's first to final repetition of the same tower pair. Red arrows are highlighted examples with annotated endpoints.

maximum of 100 characters; on the Builder's turn, they placed as many blocks as they wish (including zero) before pressing the "done" button and passing control back to the Architect. Blocks could be placed anywhere so long as they were supported from beneath, and could not be moved once placed. The Architect could see the placement of each block in real time but the communication channel was otherwise unidirectional: the Builder was unable to send messages back to the Architect. During each turn, there was a timer counting down from 30 seconds to encourage the Architect and Builder to work quickly, but there was no time-out or penalty for exceeding this time limit. Once all eight blocks have been placed, both participants received feedback about the mismatch between the target scene and reconstruction before advancing to the next trial.

3.2.3 Stimuli and Design

Each scene was composed hierarchically from two block towers that appeared side by side. Each tower, in turn, was composed hierarchically from four domino-shaped

blocks– two vertical and two horizontal (Figure 3.1B). There were three unique towers. To evaluate changes in collaboration behavior over time, we employed a *repeated* design where each tower appeared multiple times. All three possible pairs of these towers appeared, in randomized sequence, in each of four *repetition blocks* for a total of twelve trials. All towers appeared in both the left and right positions an equal number of times, such that there was no association between any given tower and its location in the scene.

3.2.4 Reconstruction accuracy improves across repetitions

Although each interaction only spanned twelve trials, we hypothesized that human dyads would be able to leverage this small amount of experience to rapidly develop shared task representations, resulting in increasingly successful and efficient collaboration over time. Before turning to our fine-grained predictions about the language used by the Architect (Figure 3.2), we first needed to verify that human dyads were able to work together in the assembly task at all. We used reconstruction accuracy as a measure of overall performance, quantifying the mismatch between the reconstructed tower and the target silhouette. Specifically, we computed the F_1 score, a standard metric of overlap that accounts for both parts of the target silhouette that the reconstruction failed to cover (a *recall* term) as well as parts of the reconstructed tower that lay outside the silhouette (a *precision* term). F_1 is normalized by the total size of the shapes, with 0 representing no overlap at all, and 1 representing perfect overlap. We found that even initial reconstructions were highly accurate with mean $F_1 = 0.88$ (95% CI $= [0.85, 0.90]$), which roughly corresponds to having just one block out of place, while the final reconstructions were near ceiling at $F_1 = 0.98$ (95% CI $= [0.96, 0.99]$). We estimated this increase using a linear mixed-effects model that predicted the F_1 score on each trial, with a fixed effect of repetition number and random intercepts and slopes for each dyad (see Supplementary Methods for more details about the model specification). We found that dyads improved significantly across repetitions ($\beta = 0.92$, $t(54.84) = 6.22$, $p < 0.001$; Figure 3.3A).

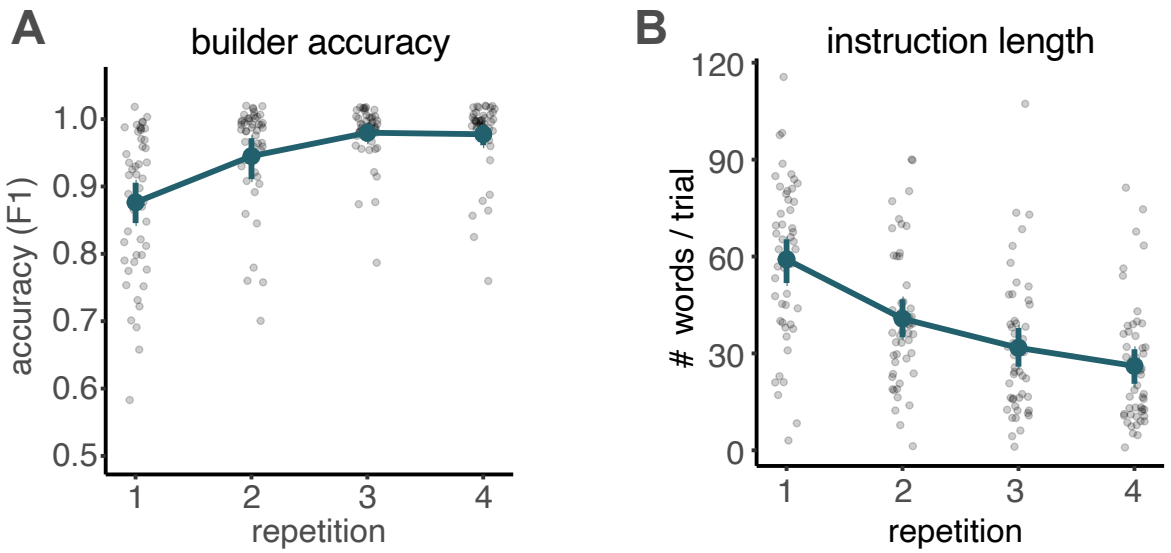

Figure 3.3. (A) Mean reconstruction accuracy improved across repetitions and (B) mean instruction length required on each trial decreased across repetitions as dyads became more effective at collaborating. Points represent values for individual dyads ($N = 49$). Error bars represent 95% CIs using dyads as the unit for bootstrap resampling.

3.2.5 Communicative efficiency improves across repetitions

Having established that the Builder was able to successfully reconstruct the tower from the Architect's descriptions, we could then examine the most basic signature of increasing abstraction in language. Given that the same towers recurred throughout the interaction, we hypothesized that Architects would exploit these regularities to provide more concise instructions over time, conveying the same information in fewer words. To test this hypothesis, we analyzed both changes in the total number of words produced by the Architect within each trial as well as the total number of separate messages sent (where each message may contain more or less words). We estimated this effect using a mixed-effects model containing a fixed effect of repetition, as well as maximal random effects for both items and participants (see Supplemental Methods). Consistent with our hypothesis, we found that Architects sent messages containing significantly fewer words

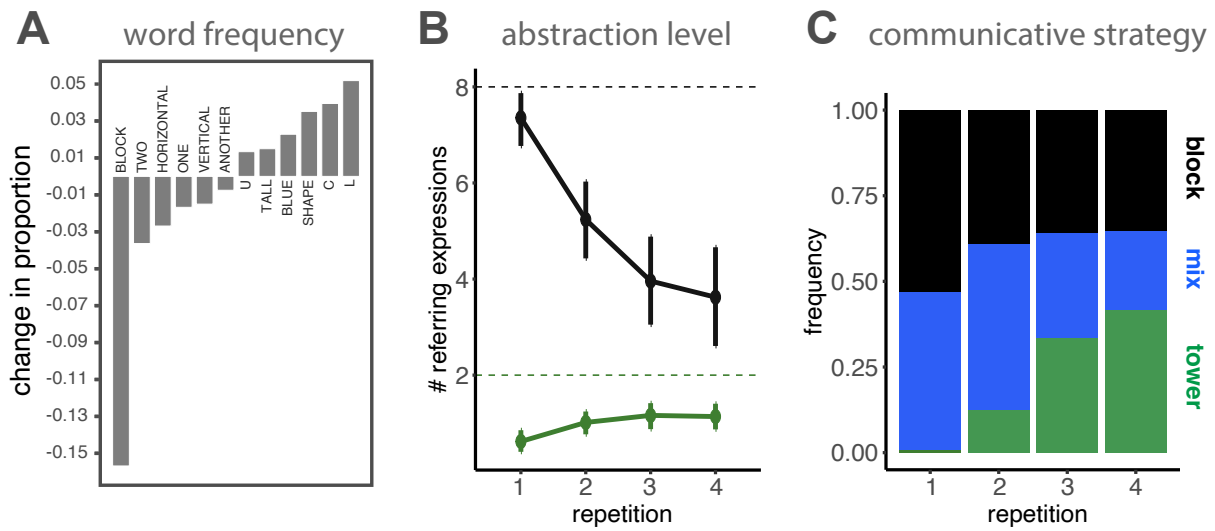

Figure 3.4. (A) Words with largest positive and negative changes in frequency between first and final repetitions. (B) Change in number of block-level and tower-level references across repetitions. Darker and lighter dashed lines represent maximum possible number of blocks and maximum number of towers, respectively. error bars represent bootstrapped 95% CIs. (C) The proportion of referring expressions in each trial that exclusively refer to blocks, towers, or scenes.

over time ($\beta = -8.53$, $t(36.9) = -9.58$, $p < 0.001$; Figure 3.3B), which were themselves contained within fewer discrete messages within each trial ($\beta = -18.1$, $t(24) = -7.11$, $p < 0.001$).

Results

3.2.6 Level of referential abstraction increases across repetitions

What allowed dyads to perform better while also using fewer words? We hypothesized that the increase in communicative effectiveness is due to the ability of each dyad to gradually shift to giving and receiving instructions at a higher level of abstraction — in particular, one corresponding to entire towers rather than individual blocks. We first conducted a qualitative analysis to explore this possibility. Specifically, we tokenized all of the Architect's messages into individual words and examined, across our entire dataset, which words changed the most in frequency from the beginning to the end of

the experiment (Figure 3.4A). We observed that the frequency of low-level nouns like "block" and block-level modifiers like "horizontal" or "red" decreased the most, while that of high-level nouns like "L" or "C" and adjectives like "tall" increased the most.

We followed up this initial exploration with a more systematic analysis of the content in each message. We recruited a group of four annotators who were unaware of the study design and hypotheses to tag each referring expression with the number of references to block-level vs. tower-level entities they contained. Annotations were highly consistent between raters (intraclass correlation, $ICC = 0.83$; 95% CI = $[0.82, 0.84]$, see Supplementary Materials for further analysis). We constructed a mixed-effects model that included fixed effects of repetition (integer: 1 to 4), expression type (categorical: tower vs. block), their interaction, as well as maximal random effects for each dyad. We found a significant interaction ($b = 0.53$, $t(47.5) = 4.8$, $p < 0.001$; Figure 3.4B), providing further evidence that block-level referring expressions became reliably less common while tower-level ones became more prevalent. The mean number of block-level references strictly decreased by half, from approximately 7.3 at the beginning to 3.6 at the end, while the mean number of tower-level messages nearly doubled, from around 0.6 at the beginning to 1.1 at the end. Because many messages contained a mixture of both levels, we further annotated whether each trial contained only block-level (e.g. "horizontal blue block," "vertical red block"), only tower-level (e.g. "C shape," "L shape"), or a mixture of both levels of expressions. We observed that the shift across repetitions is primarily driven by an increase in the proportion of tower-level references and a decrease in the proportion of both mixed and block-level references (Figure 3.4C).

3.2.7 Both conceptual and linguistic coordination are required in a model

Our experiment provided strong evidence that dyads shift to higher levels of conceptual abstraction as their assembly performance improved. However, this finding

raises several questions: Why did tower-level expressions gradually displace block-level ones, rather than the other way around? And why did participants change the way they communicated about the scenes at all, given that initial reconstruction accuracy was already so high? In this section, we argue that a dual-coordination model provides a more satisfying answer to these questions than simpler existing models. The dual-coordination model explains an Architect's use of referring expressions and the Builder's understanding of them in terms of two basic ingredients: (1) the procedural abstractions available to each agent at a given time and (2) a communicative trade-off between informativity and message length (a.k.a. verbosity) given common ground with their partner. We integrated this pair of ingredients in a computational model that integrates a state-of-the-art library learning algorithm (Ellis et al., 2021) with a recently proposed probabilistic model of communication under uncertainty (Hawkins et al., 2023). We conducted several ablation studies and found that both mechanisms are required to explain the patterns observed in our empirical data (see Supplementary Material for a full specification of the model and details of our simulations).

We ran simulated Architect-Builder pairs with different combinations of these components through the same trial sequences used in our human behavioral experiment. For each trial, we first sampled a scene representation and corresponding sequence of instructions from the Architect agent and then sampled a set of corresponding actions from the Builder agent's distribution conditioned on this utterance. The Architect maintains some representation of the target scene using their own library of "chunks," but maintains uncertainty in their beliefs over whether their partner shares that representation and whether they share the linguistic conventions used to refer to it. Given their uncertainty, they generate an instruction balancing between its informativeness (the expected probability that the Builder will build the intended tower) and its length (with the relative weighting determined by a fixed parameter $\beta \in [0, 1]$). After each trial, each agent updated (1) their concept library, (2) their beliefs about their partner's lexicon, or (3) both, the latter

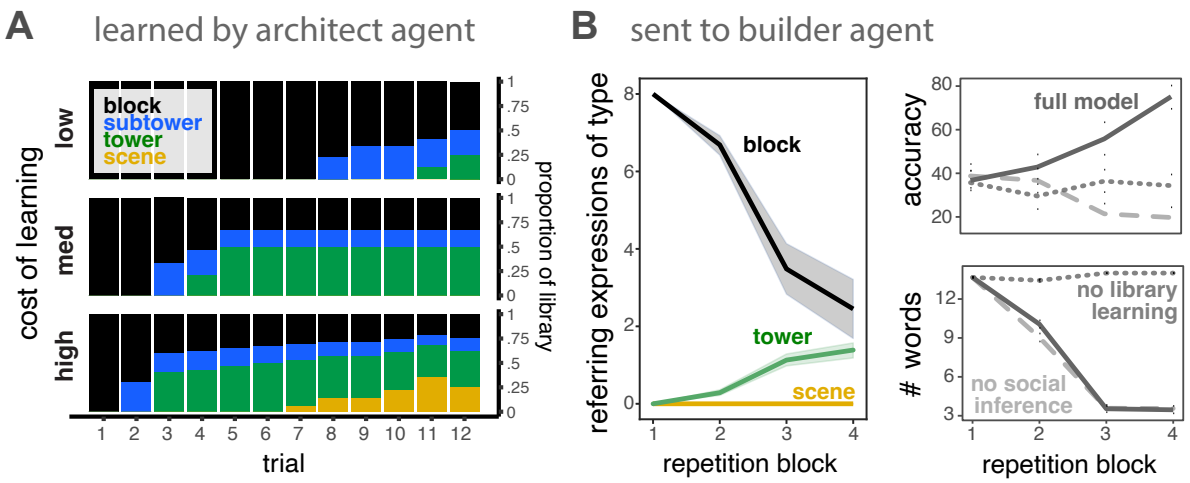

Figure 3.5. (A) The composition of an agent's internal concept library changes as new procedural abstractions are added over time. Bars represent the proportion of library components at different levels, shown for different size penalties (low: $w = 1.5$, medium: $w = 3.2$, high: $w = 9.6$.)(B) Results of simulated interactions. The lefthand plot shows the shift in referring expressions at different levels of abstraction (cf. 3.4). The righthand pair of plots shows lesioned variants, where only the full model displays both improved accuracy (top) and efficiency (bottom). Curves shown for parameters set to $\beta = 0.3$, $\alpha = 5$, $\epsilon = 0.075$ (see Supplemental Material for more information).

constituting our *full* model the previous two representing "lesioned" variants.

The model with both components fully intact displayed two behavioral signatures that were qualitatively consistent with those displayed by human participants. First, repeated exposure to target towers across trials increased the likelihood that chunked subroutines at the tower-level would be independently discovered by each agent, as a consequence of library learning (Fig. 3.5A). Second, as the Architect received feedback about the success of their instructions, thereby reducing their uncertainty that more abstract referring expressions would be interpreted correctly, these instructions no longer seemed as risky and became preferred for their shorter length. Thus our model provides an explanation for why Architects may increasingly prefer shorter, abstract messages over longer, concrete messages: new "chunks" come online and uncertainty is reduced about whether the Builder will understand instructions that refer to them.

By contrast, the lesioned models did not reproduce this pattern of behavior. Without

the library learning component, *efficiency* cannot not improve; there are no more abstract chunks for conventions to bind to, only the same primitives that were available at the beginning. Without the convention formation component, *accuracy* cannot improve; new conceptual abstractions are introduced into each agent's library, but there is no way to update beliefs about which words correspond to it. Finally, within the full model, we found that the tradeoff between informativeness and verbosity (controlled by β) was crucial. If sensitivity to verbosity is too high (i.e. $\beta > 0.5$), Architects would always used the most concise programs available to them – by the third repetition nearly all block-level instructions were replaced by descriptions at higher levels of abstraction, even though these descriptions were more likely to result in Builder errors (Fig. 3.5B). Meanwhile, without any sensitivity to verbosity at all ($\beta = 0$), Architects stuck to a safer strategy, continuing to use longer but less ambiguous descriptions composed solely of block-level instructions. We found that only at intermediate values of β could the model reproduce key aspects of Architect behavior observed in human participants.

Discussion

Successful teamwork relies on team members to both coordinate how they think about and how they talk about what they are doing. This paper explored a "dual coordination" theory of teamwork using a naturalistic collaborative assembly task that required teams of participants to ground their communication in shared abstractions of the scenes they were building. Over the course of an extended interaction, we found that dyads communicated with increasing efficiency by shifting to *ad hoc* labels at higher levels of conceptual abstraction. Our computational model explains this trend by invoking learning mechanisms at two levels: a perceptual "chunking" mechanism (based on Bayesian program learning) and a social inference mechanism (based on a probabilistic model of communication).

Beyond the core implications for coordination in teams, our work also contributes to two influential lines of work on a more basic question: where do abstractions come from in the first place? On one hand, "chunking" strategies are used to reduce cognitive costs across many domains including planning (Ho et al., 2019, 2022), perceptual organization (Palmer, 1977), memory encoding (Ding et al., 2017), concept learning (Tversky & Hemenway, 1984; *Wong et al., 2022), and motor learning (Chaffin & Imreh, 2002). Given their ubiquity, these chunking processes may form a set of implicitly shared inductive biases between partners in collaborative tasks – each individual may justifiably assume others are chunking in a similar way. The functional demands of social coordination emphasized in our work may thus play an important role in the pressures shaping abstraction processes in individual minds more broadly (Gilead et al., 2020). On the other hand, the co-existence of multiple layers of abstraction in the lexicon (e.g. *poodle*, *dog*, *animal*, *thing*) has raised longstanding questions about why some abstractions get lexicalized in language while others do not (Brochhagen et al., 2023; Leising et al., 2014; Rosch et al., 1976; Snefjella & Kuperman, 2015). A dual coordination account provides another perspective: *ad hoc* linguistic conventions co-evolve with *ad hoc* concepts based on the needs of particular communicative contexts. Thus, as our language and concepts adapt to our environment, new conventions dynamically bind to newly relevant concepts.

While our model captures the key signatures of adaptive collaboration behavior that were the focus of our experiment, it is limited in several ways that would be valuable to address in future work. First, our agents did not explicitly reason about the contents of their partner's conceptual library, only the contents of their lexicon. That is, our agents engaged in social inference at the level of language, but were "egocentric" at the level of concepts. While an "egocentric" strategy is functionally equivalent to social reasoning in our task, as both agents experienced the exact same sequence of towers as input, more general settings will require social reasoning about conceptual mismatches, as when an expert must give instructions to a novice. Second, we constructed our task and model to

build in an asymmetry between the Architect and Builder roles, but in practice, these roles often self-organize dynamically, requiring a third layer of coordination above shared concepts and language (Goldstone et al., 2023; Mauro et al., 2009). Third, while we focused on qualitative effects, there is substantial quantitative variation in the natural language strategies chosen by different Architects, posing important challenges for future work using more realistic lexical priors (e.g. given by neural language models). For example, some Architects anaphorically refer to actions from previous trials (e.g. "the same C again") or use disjunctively combine multiple valid expressions (e.g. "like a big C or a tower with the right side cut out"). For all of these reasons, we emphasize that while our model provided one precise instantiation of the dual coordination hypothesis, other instantiations could make similar predictions. In the long run, comparing different models may shed light on the inductive biases that enable such rapid coordination upon shared abstractions during social interaction between intelligent, autonomous agents.

3.3 Acknowledgments

I thank my co-authors Judith Fan, Robert Hawkins, Haoliang Wang, and Cameron Holdaway for their invaluable contributions to this and earlier versions of the manuscript. Special thanks to Robert Hawkins, for his generous mentorship throughout this project.

Chapter 3, in full, is in currently in review for publication of the material. An earlier version of this project was published as *McCarthy, W. P., *Hawkins, R. D., Wang, H., Holdaway, C., & Fan, J. E. (2021). Learning to communicate about shared procedural abstractions. *Proceedings of the Annual Meeting of the Cognitive Science Society, 43*. The dissertation author was the primary investigator and author of this material.

S1 Supplementary Material

S1.1 Sampling procedure and incentive structure across data collection

We pre-registered a sample of 50 dyads (100 participants), and collected data across several batches to achieve this number. Small changes were implemented between batches as we realized that some sessions were taking longer than expected and the payment was not proportionate. After the first 8 dyads, we increased the maximum size of the bonus from \$1.80 (up to 15 cents per trial) to \$2.40 (up to 20 cents per trial). After 4 additional dyads, we further increased the maximum bonus to \$3.00 (up to 25 cents per trial) and also added a solo practice trial as an early off-ramp to catch and remove unreliable participants (or bots) prior to being paired up. These changes did not substantively change the task procedure and were mostly logistical in nature (paying a fairer hourly rate, reducing frustration for reliable participants who happened to get paired with a bot at the outset rather than during post-processing). Hence, we used the full pre-registered sample, collapsing across these changes.

S1.2 Mixed effects model specification for reconstruction accuracy

Formally, the F_1 overlap metric used as our dependent variable is the harmonic mean of the *precision* and *recall*, normalized by the total size of the towers. We analyzed builder reconstruction accuracy using the following model:

```
lmer(f1score ~ poly(rep,2)
             + (1 + poly(rep,2) | gameid)
```

where `f1score` $\in[0,1]$. We make two observations about this model. First, we included a quadratic predictor of repetition number to account for non-linearities in the trend. The model with the non-linear component fit significantly better and allowed for a more

interpretable linear component. Second, we included the full random effect structure for `gameid`, but did not include random effects for items (the three tower scenes), as the variance captured by item effects was too small for the model to be fit without singularities. There may be a potential concern that about this use of a model with normal error terms for a bounded dependent variable (as $F_1 \in [0, 1]$). However, similar effects were observed for a stricter binary metric of performance that simply coded whether or not the builder's reconstruction perfectly matched the target silhouette or not (i.e. no "partial credit"). For this stricter metric, the assumptions of logistic regression are justified.

S1.3 Mixed effects model specification for instruction length

We analyzed the architect's verbosity using the following mixed-effects model, which was the maximal model that converged:

```
lmer(words ~ poly(rep, 2)
             + (1 + poly(rep, 2) | gameid)
             + (1 + poly(rep, 1) | sceneid))
```

Note that here we again included an orthogonal quadratic term to account for the non-linearity in the word count trend across repetitions, and here we were able to support linear slopes for scenes as well as dyads. As in the previous section, there is a potential objection about the use of a linear error function when the number of words is a count variable (i.e. integer valued and always greater than one). Technically, the appropriate model should use a Poisson linking function for count data:

```
glmer(words ~ rep
              + (1 + rep | gameid) +
              + (0 + rep | sceneid)),
      family = 'poisson')
```

which yields a similarly strong effect $\beta = -9.2$, $z = -8.7$, $p < 0.001$.

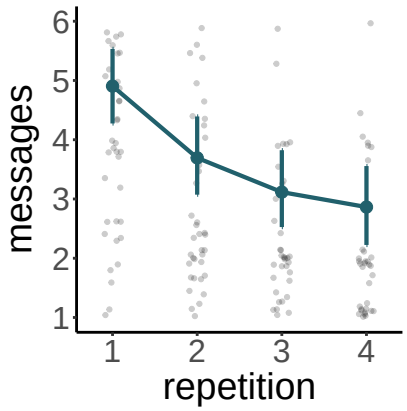

Figure S1. Mean number of discrete messages sent per trial decreases across repetitions. Points represent means for individual dyads, error bars represent bootstrapped 95% CIs.

S1.4 Mixed effects model specification for number of messages

Finally, our model of the number of discrete messages sent was specified as follows:

```
lmer(n_messages ~ poly(rep,2)
               + (1 + poly(rep,2) | gameid)
               + (1 + poly(rep,2) | sceneid))
```

We reported coefficients from the standard linear model in the main text for interpretability and familiarity, but as the number of messages on each trial are count data, we checked our findings with the more appropriate Poisson linking function:

```
glmer(n_messages ~ rep +
               (1 + rep | gameid) +
               (1 + rep | sceneid),
      family = 'poisson')
```

with estimated effect $\beta = -7.2$, $z = -8.4$, $p < 0.001$ (see Figure S1).

S1.5 Annotation of referring expressions

To identify the referential content of participants' messages we asked four naive raters to count the number of a) block-level and b) tower-level expressions in each messages. We decided not to ask raters to identify sub-tower or scene level expressions, as we did not identify any such expressions in the data ourselves, and wanted to keep the rating task straightforward. Raters annotated messages sent by all participants, dyad by dyad, in the order they were sent. Raters were instructed not to count references to locations (e.g. "next to the previous block").

To measure reliability between raters we calculated the intraclass coefficient (ICC), using the pingouin python package. As all observations were rated by the same, fixed set of raters, we report ICC3, which assumes a two-way mixed-effects model and aims to measure the actual agreement between the set of raters involved, rather than estimating agreement in a larger population. In the main text we report this measure for tower and block level expressions together. Agreement for block-level expressions $ICC = 0.815$, 95% CI:$[0.8, 0.83]$, was lower than for tower-level expressions $ICC = 0.935$, 95% CI:$[0.93, 0.94]$, although variance in block-level expressions is expected to be higher due to the larger number of blocks per scene than towers.

S1.6 Referential conventions diverge across dyads

The overall increase in tokens resembling entire towers ("C" and "L" shapes) in the final repetition suggests that different dyads may have arrived at similar tower-level abstractions and similar ways of talking about them. To what extent did different dyads converge on the same set of labels for each tower, rather than settle on distinct, but internally consistent solutions to the coordination problem? To explore this question, we estimated how dissimilar the language used by different dyads was within each repetition, by computing the Jensen-Shannon divergence (JSD) between their word

frequency distributions, aggregating language from all trials in a repetition block. We found that the mean pairwise JSD increased significantly between the first and final repetitions ($d = 0.080$, 95% CI:$[0.041, 0.118]$, $p = 0.004$), consistent with divergence between dyads.

To visualize these distances, we first ran Principal Component Analysis (PCA) on the collection of binary vectors of length 247 (representing whether or not each distinct referring expression was present in each) to reduce their sparsity. We then extracted the top 21 principle components and fed them into a t-SNE embedding (Fig. 3.2, right). Two participants were removed for empty messages or highly-idiosyncratic messages, which compressed other data points onto a small region of space. This analysis revealed widespread convergence on messages that involve tower-level expressions in the final repetition (e.g. "L" and "U"), as well as a collection of more idiosyncratic strategies that were more consistent throughout the experiment (e.g. consistently using "blue" and "red" to refer to blocks). These findings suggest that even in this relatively simple task domain, human dyads manage to discover a diverse array of solutions for mapping tokens of natural language to components of each scene.

Modeling details: Library learning component

In the following sections, we specify our computational model in full detail. We begin by specifying how each agent's procedural knowledge is represented and modified over the course of learning in the task. Following (Ellis et al., 2021), we assume that each agent maintains a library $\mathcal{L}$ of conceptual primitives that can be combined to generate simple block structures in a domain-specific language (DSL). We assume the library is initialized with the following primitives: `h` (place a horizontal block), `v` (place a vertical block), `l` (move hand to the left), `r` (move hand to the right) and digits `1`$\sim$`9`. This DSL is small but fully expressive: any possible tower can be written by combining together these basic commands.

In the Bayesian program learning framework, the DSL is updated over time by

expanding the library with new primitives. As an agent progresses through multiple trials of tower scenes $\{T_n\}_1^N$ expressed as programs, they may extract common subroutines that would allow them to re-represent the data more efficiently. Formally, the model proposes a set of candidate sub-routine fragments f after each trial and updates a posterior distribution over possible ways of extending the library (including $f = \emptyset$, which would maintain the current library):

$$P(\mathcal{L} \cup \{f\} | \{T_n\}_1^N) \propto \underbrace{P(\mathcal{L} \cup \{f\})}_{\text{description-length prior}} \times \underbrace{\prod_{n=1}^{N} P(T_n | \mathcal{L} \cup \{f\})}_{\text{likelihood}} \tag{3.1}$$

This posterior distribution weighs two competing criteria for a good library: the likelihood and the prior. The *likelihood* in 3.1 captures the ability of an extended library efficiently to explain previous towers:

$$P(T_n | \mathcal{L} \cup \{f\}) = \mathsf{exp}(-\mathsf{MDL}(T_n \mid \mathcal{L} \cup \{f\}))$$

where MDL is a function evaluating the *Minimum Description Length*. Intuitively, the Minimum Description Length is the most compact version of T_n that can possibly be written in the updated library $\mathcal{L} \cup \{f\}$. This term is therefore maximized by sets of fragments $\{f\}$ that allow the existing data to be expressed most efficiently. The *prior* over libraries, meanwhile, instantiates an Occam's razor preferring smaller libraries, all else being equal,

$$P(\mathcal{L} \cup \{f\}) = \mathsf{exp}(-w \cdot \mathsf{size}(\mathcal{L} \cup \{f\}))$$

where $\mathsf{size}(\mathcal{L} \cup \{f\})$ represents the number of primitives in the updated library. The strength of this preference is controlled by a parameter w. We explore several values of w in our simulations (Figure 3.5A). Intuitively, when $w = 0$, there is no penalty for having a larger library, so the library that best explains the observations would simply be the

exhaustive set of scenes T_n observations themselves. As $w \to \infty$, any expansion of the library is considered too costly, preventing library learning entirely. The expressiveness and simplicity objectives balance out in the posterior distribution (Eq. 3.1) such that the fragments f with the highest posterior probability are those that provide maximal compression of input tower programs while minimizing expansion of the library. We make the simplifying assumption that both participants update their libraries at the same rate (using the same w). While we believe this is a reasonable assumption for our task, it is likely glossing over real individual differences. Collaborators in the real world are likely to discover useful abstractions at different rates, due to differences in prior knowledge or from approaching the task from different perspectives.

In practice, we selected the single highest posterior-probability set of fragments at each point in the task, conditioning on the previous trials (Fig. 3.5A). The resulting DSL was supplied to both the Architect and Builder agent model as the set of primitives they are able to represent. In other words, we assume that the Builder and Architect learn abstractions at the same rate throughout the experiment. We further assume that when the Architect agent is presented with a scene, they are able to synthesize a set of 1 to 4 possible candidate programs for representing that scene in their current DSL. For example, the Architect agent may simultaneously recognize that a scene may be constructed by placing eight primitive blocks, `(h (l 1) v v (r 2) ...)`, or by combining two higher-level primitives `(chunk1 (r 2) chunk2)` and must choose which of these to convey to the Builder.

S1.7 Probabilistic model of communication as social reasoning

Now we are ready to embed the library learning module in the previous section inside a model of *communicative grounding* where each agent's DSL serves as a basis for grounding structured linguistic meanings. We assume the Architect is a cooperative speaker who aims to produce utterances that will allow the Builder agent to accurately

re-produce the target tower. For simplicity, we also assume the Architect generates natural language instructions *sequentially*, aiming to produce an utterance for each step t_i of a full procedural sequence T written in their current DSL (in principle, this sequence could be planned jointly). Following recent probabilistic models of communication as social reasoning (e.g. Goodman & Frank, 2016), the speaker chooses an utterance according to a utility function that trades off informativity against verbosity.

$$
\begin{aligned}
P_{S_1}(u_i|t_i, \mathcal{L}) &\propto \exp\{-\alpha \cdot U(u_i; t_i, \mathcal{L})\} \\
U(u; t_i, \mathcal{L}) &= (1-\beta) \cdot \ln P_{L_0}(t_i|u_i, \mathcal{L}) - \beta \cdot \text{cost}(u_i) \\
P_{L_0}(t_i|u_i, \mathcal{L}) &\propto \mathcal{L}(u_i, t_i)
\end{aligned}
\tag{3.2}
$$

where $\alpha \in [0, \infty]$ is the soft-max temperature, $\beta \in [0, 1]$ controls the relative sensitivity to verbosity, $\ln P_{L_0}(t_i|u_i)$ is a measure of information gain to a literal builder, and $\mathcal{L}(u_i, t_i)$ is the literal meaning function that a literal Builder agent is expected to use, evaluating to 1 when u_i is true of the primitive t_i in the agent's lexicon $\mathcal{L}$ and 0 otherwise. When β is high, note that the length of the required description dominates the Architect agent's decision-making; when it is low, the Architect's decisions are dominated by informativity to the Builder.

The key effect we aimed to explain with this model is the Architect's increasing preference for more abstract descriptions without sacrificing Builder accuracy (i.e. Figure 3.4 in the main text). Eq. 3.2 gives the Architect's preferences for conveying each instruction of a fixed program T under a fixed lexicon $\mathcal{L}$. However, once we plug in our neurosymbolic model from the previous section, an Architect on later trials in fact has multiple ways of representing the raw scene T^* available to them, drawing upon different primitives in their library. We must extend our model to explicitly model the Architect's joint decision over which of these *realizations* T^k of the raw scene T^* they should attempt to transmit,

alongside what utterance they should use to transmit it:

$$P_{S_1}(u, T^k | T^*, \mathcal{L}) \propto \exp\{-\alpha \cdot U(u, T^k; \mathcal{L}) \tag{3.3}$$

$$U(u, T^k; \mathcal{L}) = \sum_i \left[(1-\beta) \cdot \ln P_{L_0}(t_i^k | u_i, \mathcal{L}) - \beta \cdot \text{cost}(u, T^k)\right] \tag{3.4}$$

where we are now accounting for the full sequence of instructions $T^k = \{t_1^k, \ldots, t_M^k\}$ and so taking the sum of the utility over all steps of the sequence.

Finally, to account for the last condition of our hypothesis, that the Architect is sensitive to the risks of introducing novel descriptions, we must specify their expectations about how to refer to a new chunk given their partner's lexicon: $\mathcal{L}(u, \texttt{chunk})$. Following recent models of lexical coordination (Hawkins et al., 2023), we assume that the Architect's lexicon $\mathcal{L}$ is *dynamic* rather than static or fixed. Each agent maintains uncertainty over the possible lexical mappings their partner may be using between words and primitives in their DSL $P(\mathcal{L})$ and marginalizes over this distribution when evaluating their utility:

$$P_{s_1}(u, T^k | T^*) \propto \exp\{-\alpha \cdot U(u, T^k)\} \tag{3.5}$$

$$U(u, T^k) = \sum_i \left[(1-\beta) \cdot \sum_j P(\mathcal{L}_j | D) \cdot \ln P_{L_0}(t_i^k | u_i, \mathcal{L}_j) - \beta \cdot \text{cost}(u, T^k)\right] \tag{3.6}$$

where D represents the shared history of feedback from the Builder agent's previous actions (e.g. their placement of blocks in response to different instructions) and $P(\mathcal{L}|D)$ represents the agent's updated posterior beliefs over the lexicons given these observations (see Hawkins et al., 2023, for additional details). We assume the lexical bindings for the starting primitives of the DSL are fixed and deterministic, e.g. $P(\{\texttt{h}: \text{“place a horizontal block”}\}) = 1$. But for learned abstractions (`chunk1`, `chunk2`) coming online through the library learning

process, we assume uncertainty over a space of other utterances ("phraseA", "phraseB") that can be emitted:

$$P(\{\texttt{chunk1}: \text{“phraseA”}\}) = 0.50$$

$$P(\{\texttt{chunk1}: \text{“phraseB”}\}) = 0.25$$

$$P(\{\texttt{chunk1}: \text{“phraseC”}\}) = 0.25$$

and similarly for the other chunks. Note that, following the simulations in Hawkins et al. (2023), we used a biased lexical prior to capture the idea that participants are using natural language and hence not beginning with a completely blank slate: there is ambiguity over exactly which tower shape the utterance "build a C" might correspond to (there are many shapes that look like a C), but not all shapes are equally likely. Future work extending this model from this artificial proof-of-concept domain to full natural language could instead use an elicited empirical prior or neural prior over all possible shapes. Indeed, while our architecture posits a clean separation between the discovery of conceptual abstractions and their subsequent communication, this is likely a two-way street. People may leverage their partner's language to discover new abstractions, a possibility suggested by language-guided library-learning (Wong et al., 2021).

S1.8 Analyzing the library learning component in model simulations

We examined the trajectory of procedural abstractions that were acquired by the model over the 49 trial sequences presented to participants, while varying the penalty on library size, w (Figure 3.5A). We manually categorized the resulting fragments based on their level of abstraction at the *sub-tower* level (e.g. a routine producing a configuration of 2-3 blocks that co-occur within multiple towers), the *tower* level (e.g. a routine generating four block placements that exactly reproduce one of the tower stimuli), or the *scene* level

(e.g. a routine generating eight block placements in the exact configuration that appeared on a trial). First, we found that the statistical structure of the trial sequence did indeed allow our library learning algorithm to acquire full *tower-level* primitives across a wide range of w, although higher (e.g. $w = 9.6$) significantly delayed learning. Surprisingly, the discovery of tower-level fragments was always preceded by sub-tower fragments. For example, the pair of blocks forming the lower left of the 'L' and 'C' towers was frequently added, and many more such fragments were added at lower values of w. There are several possible reason why these sub-tower abstractions were rare in our behavioral data, and additional work is required to determine whether Architects failed to represent them as perceptual configurations, or whether they simply suppressed the production of referring expressions for such structures.

S1.9 Analyzing reconstruction accuracy in model simulations

We measured the extend to which the programs inferred by the Builder agent matched the intended programs of the Builder. As our primary goal was to model the emergence of compositional abstractions in response to varying demands for efficiency and accuracy, we decided to initialise the Architect's and Builder's lexicons with deterministic mappings from DSL primitives to natural language expressions. This meant that before abstractions were learned and used by the Architect, the Builder would always perfectly interpret the Architect's language, and execute the program they had intended. When new abstractions are introduced, the Builder must guess the referent of the new expression, in this case by sampling from a uniform distribution over possible referring expressions. We thus expect, and see, a steep drop in accuracy following the inclusion of the first abstraction. In contrast, human participants often chose words such as 'C' or 'L-shape', for which likely have sharp priors about the possible programs they represent. After the initial drop in accuracy, the proportion of Builder actions that matched the intended program of the Architect steadily increased.

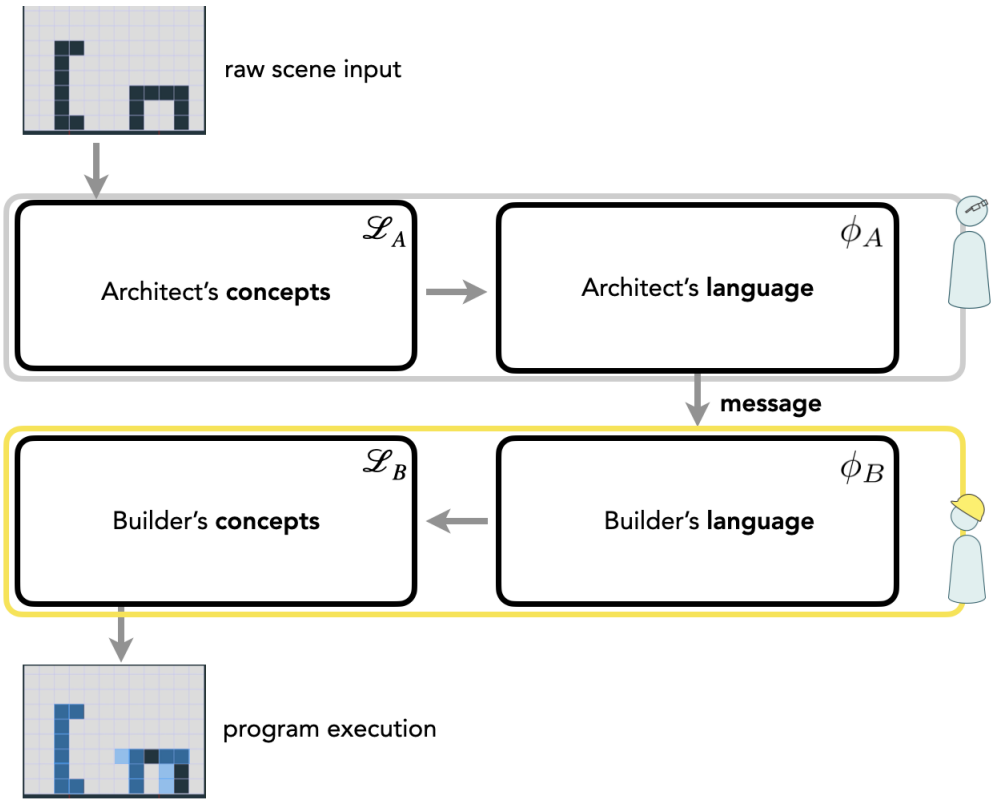

Figure S2. Schematic of model pipeline on an example trial. The Architect agent receives as input a scene presented as a grid, which is encoded into a program representation (using their current library of concepts $\mathcal{L}_A$) and then into a linguistic representation (using their current lexicon parameters ϕ_A.) The grey box indicates that these processes take place within the Architect, such that the Builder has no access to them. Then the Architect produces a message from the linguistic representation, which is interpreted by the Builder (using their current lexicon parameters ϕ_B) and translated back into a program (using their current library of concepts $\mathcal{L}_B$) which can be executed to produce a series of block placements back in the grid environment. The yellow box indicates processes taking place within the Builder, such that the Architect has no access to them.

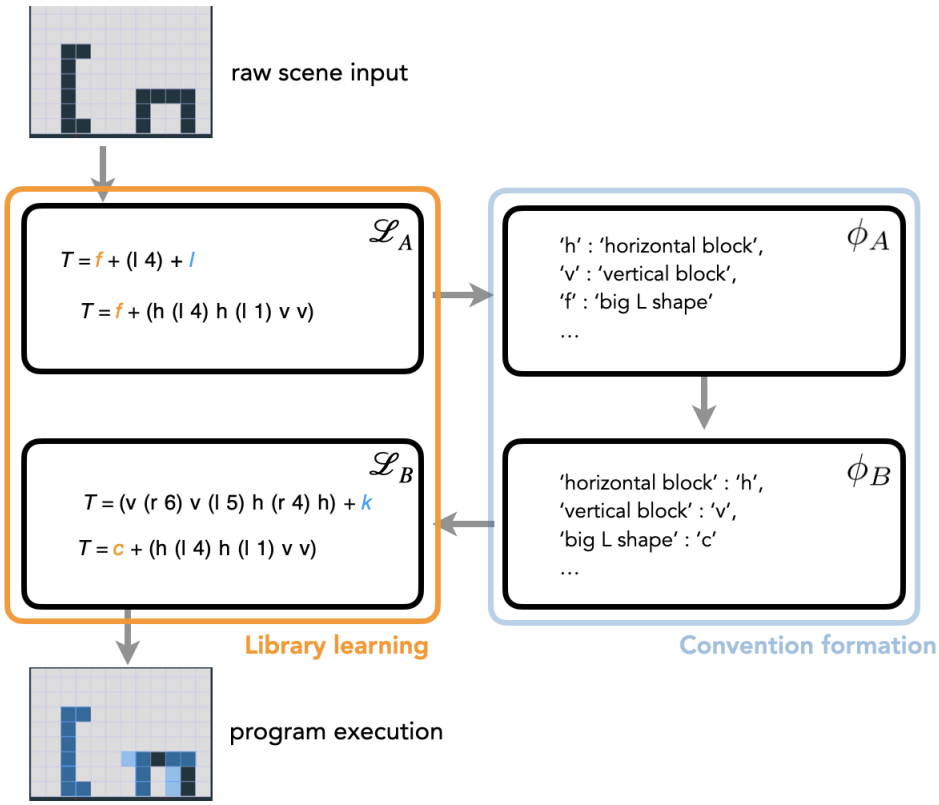

Figure S3. Critically, the library and lexicon shown in Figure S2 are not static but are updated over time. The orange box on the left highlights library learning updating $\mathcal{L}_A$ and $\mathcal{L}_B$ in both the Architect and Builder. Statistically reliable chunks are pulled out as new concepts in a library of primitives (e.g. `f` and `l`). The light blue box on the right depicts convention formation coordinating the lexicons ϕ_A and ϕ_B as the Architect and Builder update their respective beliefs in light of their shared history of successful or unsuccessful trials.

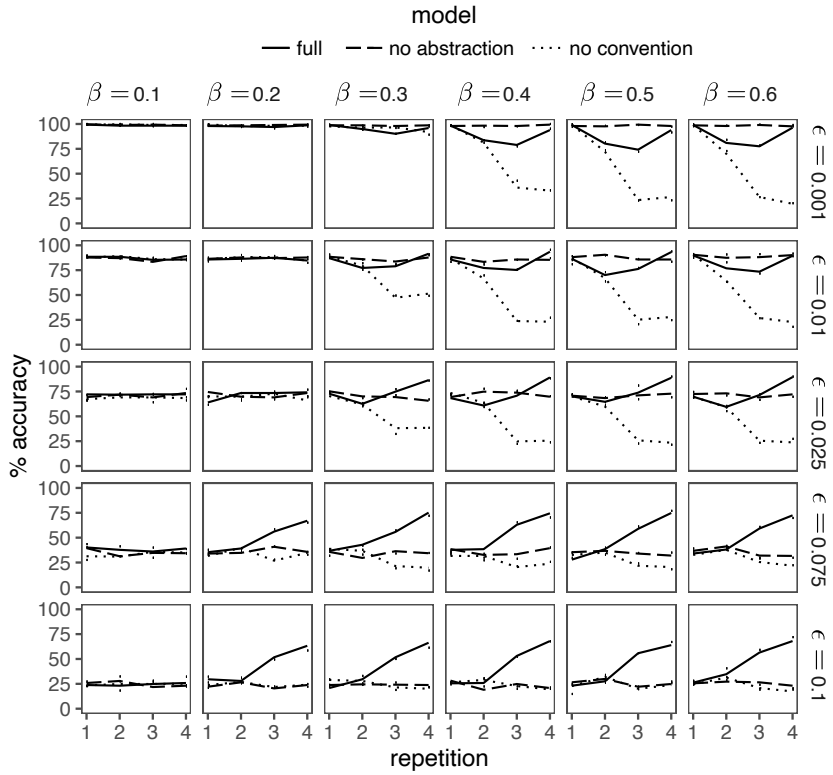

Figure S4. Accuracy of simulated agents across four repetitions shown for a range of β and ϵ parameter values. We ran two iterations for each of the 49 empirical trial sequences, so each curve is estimated from approximately 100 simulated games. The optimality parameter is set to $\alpha = 5$ for all simulations as it does not affect qualitative behavior.

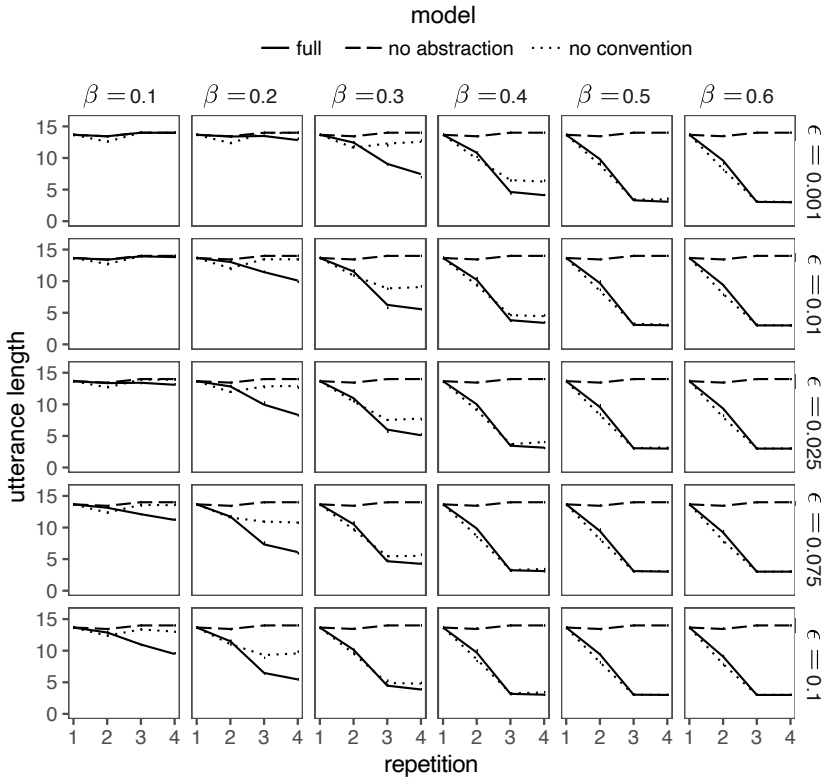

Figure S5. Instruction lengths produced by simulated Architects across four repetitions shown for a range of β and ϵ parameter values. We ran two iterations for each of the 49 empirical trial sequences, so each curve is estimated from approximately 100 simulated games. The optimality parameter is set to $\alpha = 5$ for all simulations.

References

Apker, J., Propp, K. M., Ford, W. S. Z., & Hofmeister, N. (2006). Collaboration, credibility, compassion, and coordination: Professional nurse communication skill sets in health care team interactions. *Journal of Professional Nursing*, *22*(3), 180–189.

Aslin, R. N., Saffran, J. R., & Newport, E. L. (1998). Computation of conditional probability statistics by 8-month-old infants. *Psychological Science*, *9*(4), 321–324.

Austerweil, J. L., & Griffiths, T. L. (2013). A nonparametric Bayesian framework for constructing flexible feature representations. *Psychological Review*, *120*(4), 817.

Bapst, V., Sanchez-Gonzalez, A., Doersch, C., Stachenfeld, K., Kohli, P., Battaglia, P., & Hamrick, J. (2019). Structured agents for physical construction. *International conference on machine learning*, 464–474.

Barsalou, L. W. (1999). Perceptual symbol systems. *Behavioral and Brain Sciences*, *22*(4), 577–660.

Blickensderfer, E. L., Reynolds, R., Salas, E., & Cannon-Bowers, J. A. (2010). Shared expectations and implicit coordination in tennis doubles teams. *Journal of Applied Sport Psychology*, *22*(4), 486–499.

Bogdanovic, J., Perry, J., Guggenheim, M., & Manser, T. (2015). Adaptive coordination in surgical teams: An interview study. *BMC Health Services Research*, *15*, 1–12.

Bramley, N. R., & Xu, F. (2023). Active inductive inference in children and adults: A constructivist perspective. *Cognition*, *238*, 105471.

Brochhagen, T., Boleda, G., Gualdoni, E., & Xu, Y. (2023). From language development to language evolution: A unified view of human lexical creativity. *Science*, *381*(6656), 431–436.

Chaffin, R., & Imreh, G. (2002). Practicing perfection: Piano performance as expert memory. *Psychological Science*, *13*(4), 342–349.

Christiansen, M. H., & Chater, N. (2016). The now-or-never bottleneck: A fundamental constraint on language. *Behavioral and Brain Sciences*, *39*, e62.

Clark, H. H. (1996). *Using language*. Cambridge University Press.

Cooke, N. J. (2015). Team cognition as interaction. *Current Directions in Psychological Science*, *24*(6), 415–419.

DeChurch, L. A., & Mesmer-Magnus, J. R. (2010). The cognitive underpinnings of effective teamwork: A meta-analysis. *Journal of Applied Psychology*, *95*(1), 32.

Dehaene, S., Al Roumi, F., Lakretz, Y., Planton, S., & Sablé-Meyer, M. (2022). Symbols and mental programs: A hypothesis about human singularity. *Trends in Cognitive Sciences*, *26*, 751–766.

Ding, X., Gao, Z., & Shen, M. (2017). Two equals one: Two human actions during social interaction are grouped as one unit in working memory. *Psychological Science*, *28*(9), 1311–1320.

Eccles, D. W., & Tenenbaum, G. (2004). Why an expert team is more than a team of experts: A social-cognitive conceptualization of team coordination and communication in sport. *Journal of Sport and Exercise Psychology*, *26*(4), 542–560.

Ellis, K., Wong, C., Nye, M., Sablé-Meyer, M., Morales, L., Hewitt, L., Cary, L., Solar-Lezama, A., & Tenenbaum, J. B. (2021). Dreamcoder: Bootstrapping inductive program synthesis with wake-sleep library learning. *Proceedings of the 42nd ACM SIGPLAN International Conference on Programming Language Design and Implementation*, 835–850.

Entin, E. E., & Serfaty, D. (1999). Adaptive team coordination. *Human Factors*, *41*(2), 312–325.

Fine, G. A. (2008). *Kitchens: The culture of restaurant work*. University of California Press.

Gilead, M., Trope, Y., & Liberman, N. (2020). Above and beyond the concrete: The diverse representational substrates of the predictive brain. *Behavioral and Brain Sciences*, *43*, e121.

Goldstone, R. L., Andrade-Lotero, E. J., Hawkins, R. D., & Roberts, M. E. (2023). The emergence of specialized roles within groups. *Topics in Cognitive Science*.

Goodman, N. D., & Frank, M. C. (2016). Pragmatic language interpretation as probabilistic inference. *Trends in Cognitive Sciences*, *20*(11), 818–829.

Goodman, N. D., Tenenbaum, J. B., & Gerstenberg, T. (2015). Concepts in a probabilistic language of thought. *The conceptual mind: New directions in the study of the concepts*, 623–654.

Grosz, B., & Kraus, S. (1996). Collaborative plans for complex group action. *Artificial Intelligence*, *86*(2), 269–357.

Gulwani, S., Polozov, O., Singh, R., et al. (2017). Program synthesis. *Foundations and Trends in Programming Languages*, *4*(1-2), 1–119.

Hawkins, R. D., Frank, M. C., & Goodman, N. D. (2020). Characterizing the dynamics of learning in repeated reference games. *Cognitive Science*, *44*(6), e12845.

Hawkins, R. D., Franke, M., Frank, M. C., Goldberg, A. E., Smith, K., Griffiths, T. L., & Goodman, N. D. (2023). From partners to populations: A hierarchical Bayesian account of coordination and convention. *Psychological Review*.

Hawkins, R. D., Goldberg, A. E., & Griffiths, T. L. (2021). Respect the code: Speakers expect novel conventions to generalize within but not across social group boundaries. *Proceedings of the 43rd Annual Conference of the Cognitive Science Society*, 2232–2238.

Hawkins, R. D., Kwon, M., Sadigh, D., & Goodman, N. (2020). Continual adaptation for efficient machine communication. *Proceedings of the 24th Conference on Computational Natural Language Learning*, 408–419.

Ho, M. K., Abel, D., Correa, C. G., Littman, M. L., Cohen, J. D., & Griffiths, T. L. (2022). People construct simplified mental representations to plan. *Nature*, *606*(7912), 129–136.

Ho, M. K., Abel, D., Griffiths, T. L., & Littman, M. L. (2019). The value of abstraction. *Current Opinion in Behavioral Sciences*, *29*, 111–116.

Klein, K. J., Ziegert, J. C., Knight, A. P., & Xiao, Y. (2006). Dynamic delegation: Shared, hierarchical, and deindividualized leadership in extreme action teams. *Administrative Science Quarterly*, *51*(4), 590–621.

Kumar, S., Correa, C. G., Dasgupta, I., Marjieh, R., Hu, M. Y., Hawkins, R. D., Cohen, J. D., Narasimhan, K., Griffiths, T., et al. (2022). Using natural language and program abstractions to instill human inductive biases in machines. *Advances in Neural Information Processing Systems*, *35*, 167–180.

Leising, D., Scharloth, J., Lohse, O., & Wood, D. (2014). What types of terms do people use when describing an individual's personality? *Psychological Science*, *25*(9), 1787–1794.

Mathieu, J. E., Heffner, T. S., Goodwin, G. F., Salas, E., & Cannon-Bowers, J. A. (2000). The influence of shared mental models on team process and performance. *Journal of Applied Psychology*, *85*(2), 273.

Mauro, R., Pierro, A., Mannetti, L., Tory Higgins, E., & Kruglanski, A. W. (2009). The perfect mix: Regulatory complementarity and the speed-accuracy balance in group performance. *Psychological Science*, *20*(6), 681–685.

McCarthy, W., Kirsh, D., & Fan, J. (2020). Learning to build physical structures better over time. *Proceedings of the 42nd Annual Meeting of the Cognitive Science Society*.

*McCarthy, W. P., *Hawkins, R. D., Wang, H., Holdaway, C., & Fan, J. E. (2021). Learning to communicate about shared procedural abstractions. *Proceedings of the Annual Meeting of the Cognitive Science Society*, *43*.

Palmer, S. E. (1977). Hierarchical structure in perceptual representation. *Cognitive Psychology*, *9*(4), 441–474.

Rosch, E., Mervis, C. B., Gray, W. D., Johnson, D. M., & Boyes-Braem, P. (1976). Basic objects in natural categories. *Cognitive Psychology*, *8*(3), 382–439.

Salas, E. E., & Fiore, S. M. (2004). *Team cognition: Understanding the factors that drive process and performance.* American Psychological Association.

Sexton, J. B., Makary, M. A., Tersigni, A. R., Pryor, D., Hendrich, A., Thomas, E. J., Holzmueller, C. G., Knight, A. P., Wu, Y., & Pronovost, P. J. (2006). Teamwork in the operating room: Frontline perspectives among hospitals and operating room personnel. *The Journal of the American Society of Anesthesiologists*, *105*(5), 877–884.

Snefjella, B., & Kuperman, V. (2015). Concreteness and psychological distance in natural language use. *Psychological Science*, *26*(9), 1449–1460.

Stone, P., Kaminka, G. A., Kraus, S., Rosenschein, J. S., et al. (2010). Ad hoc autonomous agent teams: Collaboration without pre-coordination. *Proceedings of the AAAI Conference on Artificial Intelligence*, 1504–1509.

Stout, R. J., Cannon-Bowers, J. A., Salas, E., & Milanovich, D. M. (1999). Planning, shared mental models, and coordinated performance: An empirical link is established. *Human Factors*, *41*(1), 61–71.

Strouse, D., McKee, K., Botvinick, M., Hughes, E., & Everett, R. (2021). Collaborating with humans without human data. *Advances in Neural Information Processing Systems*, *34*, 14502–14515.

Suhr, A., Yan, C., Schluger, J., Yu, S., Khader, H., Mouallem, M., Zhang, I., & Artzi, Y. (2019). Executing instructions in situated collaborative interactions. *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*, 2119–2130.

Takmaz, E., Giulianelli, M., Pezzelle, S., Sinclair, A., & Fernández, R. (2020). Refer, reuse, reduce: Generating subsequent references in visual and conversational contexts. *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, 4350–4368.

Tannenbaum, S., & Salas, E. (2020). *Teams that work: The seven drivers of team effectiveness*. Oxford University Press.

Tellex, S., Gopalan, N., Kress-Gazit, H., & Matuszek, C. (2020). Robots that use language. *Annual Review of Control, Robotics, and Autonomous Systems*, *3*, 25–55.

Tversky, B., & Hemenway, K. (1984). Objects, parts, and categories. *Journal of Experimental Psychology: General*, *113*(2), 169.

Udagawa, T., & Aizawa, A. (2021). Maintaining common ground in dynamic environments. *Transactions of the Association for Computational Linguistics*, *9*, 995–1011.

Waller, M. J., Gupta, N., & Giambatista, R. C. (2004). Effects of adaptive behaviors and shared mental models on control crew performance. *Management Science*, *50*(11), 1534–1544.

Walsman, A., Zhang, M., Kotar, K., Desingh, K., Farhadi, A., & Fox, D. (2022). Break and make: Interactive structural understanding using lego bricks. *European Conference on Computer Vision*, 90–107.

Wang, R. E., Wu, S. A., Evans, J. A., Tenenbaum, J. B., Parkes, D. C., & Kleiman-Weiner, M. (2021). Too many cooks: Coordinating multi-agent collaboration through inverse planning. *Topics in Cognitive Science*, *13*, 414–432.

Wang, S. I., Ginn, S., Liang, P., & Manning, C. D. (2017). Naturalizing a programming language via interactive learning. *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 929–938.

Wong, C., Ellis, K. M., Tenenbaum, J., & Andreas, J. (2021). Leveraging language to learn program abstractions and search heuristics. *International Conference on Machine Learning*, 11193–11204.

*Wong, C., *McCarthy, W. P., *Grand, G., Friedman, Y., Tenenbaum, J. B., Andreas, J., Hawkins, R. D., & Fan, J. E. (2022). Identifying concept libraries from language about object structure. *Proceedings of the Annual Meeting of the Cognitive Science Society*, *44*.

Chapter 4

Discussion

In this dissertation I presented methods for studying the cognitive processes that underlie physical assembly, and applied them to discover several cognitive consequences of assembling objects.

In Chapter 1, I introduced a novel task environment for studying physical assembly, specifically a web-based environment where participants could recreate 2D block towers. This tool enabled a detailed investigation of how individuals develop procedural knowledge and enhance task performance through repeated attempts. I developed metrics sensitive to variance in both the outcomes and the evolving processes of assembly, allowing me to track changes in both. My findings indicated that participants not only improved in constructing the target structures more accurately and quickly but also showed a significant convergence toward a limited set of effective strategies. These results demonstrate that the procedures people use build something can change when they practice, and suggest that the cognitive processes underlying this ability shift also. This work establishes a robust paradigm for impacting assembly behavior over a handful of experimental trials, validating the efficacy of our methodology for future studies investigating cognitive changes associated with physical assembly tasks.

In Chapter 2, I leveraged the tools introduced in Chapter 1 to examine how building objects impacts our memory of them. Using a related block tower assembly domain, I compared participants' memory for tower they had built with ones they had viewed more passively. Surprisingly, despite requiring more active engagement with the towers, building them did not necessarily lead to better memories than viewing them. However, additional experiments revealed that building could enhance recall of a tower, contingent on the participants' ability to form a detailed, holistic representation of the tower during the building process. As well as furthering our understanding of how active engagement impacts our memory, this chapter provides another means by which people learn to build more effectively– by remembering the process of building something. Together, Chapters 1 and 2 provide evidence that the way we mentally represent how something is made can

shift as we gain more assembly experience.

In Chapter 3 I asked how people are able to communicate so effectively in construction tasks, even when their mental representations of construction processes are likely changing. By extending the tower assembly domain to a collaborative setting, I explored how procedural knowledge and communication strategies co-evolve during an interaction. Behavioral results indicated that the language people used to describe objects and procedures became more efficient as they gained shared experience, and that the shift was largely due to the introduction of words referring to higher-level *procedural abstractions*. Formalizing the learning of these abstractions as the acquisition of program abstractions in a mental concept library, I introduced a novel computational model to simultaneously track shifting representations and the formation of ad-hoc linguistic conventions for referring to them. The model was able to capture the shift to a lexicon more abstract expressions, laying the groundwork for theories of how shared experience leads to more efficient collaboration during assembly. At a high level, these results show that collaborators are able to leverage shared assembly experience to infer more efficient ways of talking about what they are building.

The results presented in this dissertation comprise a set of that people get better at building when they practice: by learning to build the same objects in better ways, by recalling procedures for building them, and by inferring more efficient ways of communicating about the objects they are building. Studying the mechanisms of learning, memory, and communication in the context of this generative task enabled the discovery of relationships between these processes that would not have been present studying each mechanism in isolation. In doing so, we intentionally laid aside other mechanisms that might be relevant, like planning and physical simulation, as well as situated and embodied strategies that we know are prevalent in assembly and other generative behaviors (Kirsh, 1995). While some aspects of construction behavior may always be out of reach in simulated environments, a large number of affordances could be incorporated, in principle, particularly if environments

were extended to 3D objects, virtual reality, and more realistic physics and haptics. Studies in these more complex environments could reveal interactions between other cognitive processes.

Our ability to learn procedures for creating something was at the core of all three chapters, and I presented evidence for relationships between this ability and memory, communication, and construction ability as a whole. These findings have implications for our understanding of the format, or formats, of our mental representations of objects. In Chapter 2, I found that the effect of assembling objects on memory differed depending on the particular readouts of memory we used, suggesting multiple kinds of representation that are differentially sensitive to experience. Machine learning distinguishes between generative and discriminative representations, a dichotomy that may track the distinction between recognition and recall, and perhaps be reflected in the ventral and dorsal pathways (Chao & Martin, 2000). On the other hand, our ability to reason about objects in various ways– what they look like, how to use them, as well as how they are made– suggests a role for general-purpose object representations, perhaps akin to symbolic representations in a language of thought (Fodor et al., 1975; Lake et al., 2017). Such representations would need to be capable of capturing detailed information about objects' structure and their parts. Several researchers have proposed that *generative programs* may fill this role, due to their flexibility and expressive power (Lake et al., 2015, 2017; Yildirim et al., 2020). Likewise, in Chapter 3, I leverage such representations because they support abstraction (Ellis et al., 2020). Insofar as solving an assembly problem invokes a mental library of part concepts to generate the target object (Tian et al., 2020), our results suggest that these libraries are not fixed. Furthermore, while I used language to probe the libraries of part concepts people learn, language may in fact play a causal role in the the kinds of representations people learn (Wong et al., 2021), suggesting a further role for language as a tool for investigating the dynamic structure of object representations (*Wong et al., 2022).

The experimental methods developed for studying how people collaborate in assembly tasks could easily be extended to explore collaboration in other generative tasks. Investigating other collaborative domains, such as product design, architecture, and computer aided design, might reveal similar linguistic trends towards higher-level abstractions, or domain-specific strategies for communicating creative intent. To truly understand the range of communicative strategies in these highly visual domains will require going beyond language; as well as speaking and writing, designers use *drawing* to communicate their designs (Lawson, 2006; Williams & Cowdroy, 2002). Incorporating additional communicative modalities into these experiments could allow researchers to investigate how professional creators make efficient use of these modalities to convey design intent. Understanding the nuances of how people communicate in different domains could fuel the development of creative tools– by suggesting novel user interfaces supporting common goals, and via the development of AI agents that can understand and execute naturalistic instructions.

References

Chao, L. L., & Martin, A. (2000). Representation of manipulable man-made objects in the dorsal stream. *Neuroimage*, *12*(4), 478–484.

Ellis, K., Wong, C., Nye, M., Sable-Meyer, M., Cary, L., Morales, L., Hewitt, L., Solar-Lezama, A., & Tenenbaum, J. B. (2020). Dreamcoder: Growing generalizable, interpretable knowledge with wake-sleep bayesian program learning. *arXiv preprint arXiv:2006.08381*.

Fodor, J. A., et al. (1975). *The language of thought* (Vol. 5). Harvard university press Cambridge, MA.

Kirsh, D. (1995). The intelligent use of space. *Artificial intelligence*, *73*(1-2), 31–68.

Lake, B. M., Salakhutdinov, R., & Tenenbaum, J. B. (2015). Human-level concept learning through probabilistic program induction. *Science*, *350*(6266), 1332–1338.

Lake, B. M., Ullman, T. D., Tenenbaum, J. B., & Gershman, S. J. (2017). Building machines that learn and think like people. *Behavioral and brain sciences*, *40*, e253.

Lawson, B. (2006). *How designers think*. Routledge.

Tian, L., Ellis, K., Kryven, M., & Tenenbaum, J. (2020). Learning abstract structure for drawing by efficient motor program induction. *Advances in Neural Information Processing Systems*, *33*, 2686–2697.

Williams, A., & Cowdroy, R. (2002). How designers communicate ideas to each other in design meetings. *DS 30: Proceedings of DESIGN 2002, the 7th International Design Conference, Dubrovnik*.

Wong, C., Ellis, K. M., Tenenbaum, J., & Andreas, J. (2021). Leveraging language to learn program abstractions and search heuristics. *International Conference on Machine Learning*, 11193–11204.

*Wong, C., *McCarthy, W. P., *Grand, G., Friedman, Y., Tenenbaum, J. B., Andreas, J., Hawkins, R. D., & Fan, J. E. (2022). Identifying concept libraries from language

about object structure. *Proceedings of the Annual Meeting of the Cognitive Science Society*, *44*.

Yildirim, I., Belledonne, M., Freiwald, W., & Tenenbaum, J. (2020). Efficient inverse graphics in biological face processing. *Science Advances*, *6*(10), eaax5979.