CONTEXTUAL INFLUENCES ON UNDERGRADUATE BIOLOGY STUDENTS' REASONING AND REPRESENTATIONS OF EVOLUTIONARY CONCEPTS

By

Joelyn de Lima

A DISSERTATION

Submitted to
Michigan State University
in partial fulfillment of the requirements
for the degree of

Plant Biology—Doctor of Philosophy

2021

ABSTRACT

CONTEXTUAL INFLUENCES ON UNDERGRADUATE BIOLOGY STUDENTS' REASONING AND REPRESENTATIONS OF EVOLUTIONARY CONCEPTS

By

Joelyn de Lima

Context is the background or the settings of an event or idea. It is only when events or ideas are considered within the context in which they occur that they can be fully understood. In education, the application of knowledge communicated in one context to a different one is a central feature of learning. However, knowledge transfer can be affected by multiple factors including contexts used. Context plays a vital role in both shaping students' learning and in eliciting their knowledge. Therefore, understanding how context can help or hinder learning and how context impacts knowledge assessment is important for improving science learning outcomes.

For my dissertation, I studied contextual influences on the ways students reason and represent their knowledge. My studies explored two types of contexts: surface features of prompts provided to students (e.g., organism used) and the mode of response requested (e.g., written narratives vs constructed models). I analysed the effect of prompt surface features on the content of students' written responses and on the architecture of models they constructed to explain evolution by natural selection. I also analysed the effect of mode on the content and level of scientific plausibility of students' responses. In addition, I explored the association between instruction and prior achievement and susceptibility to contextual influences.

My results indicate that prompt contextual features and mode of response are eliciting differences in the content of students' representations. Contextual susceptibility decreased with instruction and higher prior academic achievement. This could indicate that they are novice learners and have a fragile understanding of either the subject matter (evolution), the alternative representation that was required (constructing models), or of both the subject matter and the representation. Incorporating multiple contexts and modes of assessment has potential to generate a more holistic view of students' understanding and may promote greater transfer by requiring students to think and reason across contexts.

Dedicated to Mridul, Monica, Marcelia, and Maria Ignes – with deepest gratitude

ACKNOWLEDGEMENTS

The journey that culminates in this dissertation started a long time ago, even before I joined graduate school. This journey was a learning experience not only in research and academia, but also in the resilience of an individual in the face of adversity, in the value of collaborations when building knowledge and community, and in the strength of relationships that nourish both survival and growth. It was not a journey I could have possibly walked alone, and it is truly my privilege to honour and thank all those who made it possible.

First of all, to the world's best advisor – Tammy Long. The only reason I even considered joining grad school is you, and I cannot even imagine a PhD without you. Thank you for giving me this opportunity to be your first PhD student (even though the initial offer was for a postdoc! Is that offer still open btw?). The conversations we had, the questions you asked, and the advice you gave, made me a better scientist. The example you set by being so supportive, compassionate, and uplifting has made me a better mentor. And having you in my life has made me a better person. I admire you as a scientist, value you as a mentor and I treasure you as a lifelong friend and family member.

I also owe a huge debt of gratitude to the three strong scientists on my committee, Amelia Gotwals, Melanie Cooper, and Kay Gross. The conversations I had with you and all the feedback you gave me improved not only this product, but also my ability to think

critically about my work. Amelia, you allowed me to soar while keeping me firmly grounded. Melanie, you challenged me to become a better scientist every time we spoke. Kay, your infectious enthusiasm about my research kept me going. I am a better scientist because you all pushed me, but I was able to reach this day because you all were truly understanding and put my well-being before all else.

I have been very blessed to have inspiring mentors my whole life; from my school teachers who still support and encourage me; to my college lecturers who went out of their way to see that I got resources to help me succeed. Ma'am Arina Frank, you saw something in me that I had not yet seen, and you knew I could get here way before I did. You were the first person who gave me a scientific journal and told me to read it. I still remember all those afternoons I spent at your place studying for exams. Ma'am Delia Antao, you were unflinching and unwavering in your belief in me, that belief kept me going when mine faltered. You fought battles to smoothen my path and inspired me to keep fighting. For that, and so much more, I will always be grateful.

During these past five years, I have been fortunate to share time and space with an amazing bunch of people who collectively call ourselves the Long Lab. You all were a constant source of practical help, emotional support, and mental calmness. Thank you for bringing joy and laughter in my life. To the best postdoc, Caleb Trujillo, thank you for always being willing to make time for me, talking about my research and answering all my questions patiently – no matter how stupid they might have seemed to you.

To the best lab managers I have seen, Patrycja Zdziarska, Etiowo Usoro, and Socheatha Chan, thank you for all the logistical support, for making sure that my data collections went without any problems and for dealing with all my anxieties. Trisha – thank you for forcing me to take breaks and colour old patent designs that we then used to decorate Joelyn and Trisha's kingdom (or what the uninitiated call S-340). Twoah, thank you for helping me find an apartment that I liked and in which I felt safe. Chetta, thank you for introducing me to the delights of shopping at Sam's Club and getting takeout at buffets (you can get three meals for the price of one!).

To the amazing undergrads in the Long Lab – your enthusiasm and irresistible joie de vivre kept my spirits up. I got excited and inspired to use Evolution as the context for my research because of Mitch Distin's infectious enthusiasm about evolution. I am grateful to Mitch, and to Devin Babi and Hunter Hicks for all their hard work transcribing student data. If you had not transcribed all those models and narratives for me, trying to decipher student handwriting would have probably driven me insane!

And to my fellow Long Lab grads, you were the best lab mates anyone could have asked for. Seth Hunt, you helped me when I was moving in and had nothing, and when I was moving out and had to get rid of everything – thank you. Steve Bennett, I miss your quiet and calming presence so much, thank you for being so supportive and encouraging. And Beth Gettings, my partner in exploring just how resilient you need to be to get a PhD, thank you for always being willing to listen and be there for me, for

making sure that I did not die of heatstroke, and of course, for always knowing what the date was!

Thank you also to everybody else at MSU who mentored me, supported me, and helped me complete this journey. Diane Ebert-May, your enthusiasm is so infectious, Thank you for all your guidance, encouragement and your unique perspective on priorities and change. Rique Campa, I have always appreciated having you in my camp. Thank you for all your encouragement, and for replying to emails within minutes of getting them (how do you always manage to do that?). Amber Peters, thank you for being generous with your class time and allowing me to collect data in your class. Jon Stoltzfus and Louise Mead, thank you for trusting my scientific abilities enough to fund me off your grants. Collaborating with you has helped me grow as a scientist and I hope to continue our collaborations. And Tracey Barner, thank you for always watching out for me and having my back.

While my PhD is based at an American institution, I also had the good fortune of being a visiting student in two international institutions. I was lucky to be a visiting student at Eawag, Switzerland (and being assigned what I think is the best office I will ever have). Francesco Pomatti and Anita Narwani, thank you for having me as a visiting student in your labs, even though I did nothing with phytoplankton. Gioia Matheson and Arianne Maniglia, thank you for being on top of the paperwork every time. To all the people at Eawag, thank you for making my time there so much fun. And to Simone and Dijana

Fontana, thank you so much for everything, especially during these past few months, I can't wait to come to Zurich to see you all soon.

Thank you also to all the amazing people at the Centre for Ocean Life, DTU, Denmark. You all made me feel so welcome and included. Thank you for all the memories and for the lasting friendships! Thank you specially to Ken Andersen and Thomas Kiørboe for allowing me to be a guest student and opening the centre and your homes to me. Thank you also to Lillian Andersen for all your support and of course for making sure the paperwork was always fine.

It is amazing when professional relationships morph into personal friendships and sources of support during tough times. Thank you to all the awesome DBER-SiTers, I miss seeing your faces every month! To my friends at EvoKE, specially to Xana Sá Pinto, I am so grateful for you. My KBS academic family; especially Kara Haas, Misty Klotz, Sarah Carroll, Julie Doll, Tom Getty, Elena Litchman and Chris Klausmeier; thank you for always being there for me. It was because of my interactions with all of you that I realised that scientists can be wonderful people too! I am also grateful for everyone in my cohort, we started this journey together, and we have had so many adventures along the way. I will treasure those memories.

To my KBS extended family; especially Nicole Kokx, Tina Mattson, Roz and Rich Cooper, Joni and Bob VanDenBos, and Carol Marbach; you are my rock. When I went away in 2013, I did not know when I would see you all again, I am so glad that this PhD

gave me an excuse to spend more time with all of you. Thank you for all your care and concern. I bet the PF Chang's in Lansing will miss us!

Living away from family was challenging, especially during these last few months, but I had an awesome support system. Dee Jordan, you have always stood up for me and been my champion, your strength makes me stronger. Lisa Stelzner, it has always been such a source of comfort to know that you lived close by and that I could reach out at any time. Katie Minnix, thank you for literally making sure that I did not starve! I am grateful for all of you and all that you did for me.

To my old hen's club! Megan Shiroda and Kellie Walters. You just light up my day. I miss our weekly dinners. You both have been there with me through some of my lowest moments and you propped me up. I cannot express just how thankful I am for you.

To my girls back home, especially to Mearl Fernandes, Jocelyn Fernandes, Michelle Mascarenhas, Sibyl Fernandes, Candyce Colaco, Shweta Hegde, Kim Lobo, and Tina Zacharia, thank you. You all have known me longer than most and have always been a part of my struggles and my successes. You will always be a part of my heart.

A very special thank you to my in-laws, Malathi and Tom. Your support and encouragement, and your belief in me has kept me fuelled. I am so looking forward to spending more time and going on more adventures with you.

To the two awesome men I lived with! Ravi Ranjan and Victor Felix de Souza Keller. I loved living with you. You helped to make the apartment feel like home and not just a place to watch time go by. Thank you for tolerating my peculiarities. For better or worse, you are stuck with me for life! I miss you, and I am looking forward to kicking your respective asses at Carcassonne again soon!

Some family we choose for ourselves, and I have been so blessed to be able to have so many amazing people as a part of my family. Colin Kremer and Rachel Prunier, you are such an integral part of my life, I cannot imagine crossing this milestone without you. Anne Royer, you are the one who has dealt with me at my lowest and has always figured out a way to lift me up again. Tim Dickson and Jenny Hopwood, my heart smiles just thinking about you. Thank you for being with me on this journey. I hate that you all are so far away, and I love you all so much. Can you all just move to Switzerland please?

Shawna Rowe and Christopher Warneke, I love you. Thank you for buffering all that life threw at me these past five years. I was able to be resilient in the face of the storm because you propped me up. Christopher, your calm strength bolstered me, and your fierce determination to make a difference inspired me. Shawna, your generosity motivated me to be a better person, and your ability to deal with all that life threw at you gave me much needed perspective into my life. My life is richer for having you two in it. I admire you both so much, even though Christopher you are totally wrong about tea and Shawna, you are totally wrong about basically everything I do not agree with.

This journey would definitely not have been possible without the unceasing and unconditional love and support from my family. Ma and Pa, I am so proud to be your daughter. I am resilient because you taught me about courage and persistence by setting such an amazing example. Thank you for encouraging me to make my own choices, giving me the strength to pursue what I wanted, and being willing to make so many sacrifices so that I could get to where I am today. And Bai, thank you for all your love and support, and especially for all your patience with me.

And to my dearest husband, Mridul. Our relationship has been forged by the pressures of all the challenges we have had to face over the years. I could not have made this journey without you, thank you for being there every step of the way, through all the peaks and troughs of grad school, despite often being several thousands of kilometres and several time zones away. I am so glad that after 10 years of marriage, we no longer have to be in a long-distance marriage. I have enjoyed every adventure we have been on and I look forward to so many more.

I want to take a special moment to honour all those walked before me and whose lives and actions paved my path. I especially want to honour two of the most resilient women I know, my grandmothers Maria Santana Marcilia da Cruz and Maria Ignes Severina Pinto. Without them, their sacrifices, and their sheer grit, I would not be here. Dev borem korum mamai ani gharamai.

This journey has been long, and my memory is short. Therefore, I want to thank everyone who in any way has helped me complete it. Thank you for joining heads, hands, and hearts with me.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

KEY TO ABBREVIATIONS

AAAS	American Association for the Advancement of Science
ACARA	Australian Curriculum Assessment and Reporting Authority
ACORNS	Assessing COntextual Reasoning about Natural Selection
CS	Cognitive Structure
GPA	Grade Point Average
HCA	Human/Cheetah Assessment
IRR	Inter-Rater Reliability
KC	Key Concept
KMK	Sekretariat der Ständigen Konferenz der Kultusminister der Länder in der Bundesrepublik Deutschland
MYA	Million Years Ago
NASEM	National Academies of Sciences, Engineering, and Medicine
NGSS	Next Generation Science Standards
NI	Naïve Idea
NRC	National Research Council
SBF	Structure-Behaviour-Function
STEM	Science, Technology, Engineering and Mathematics
TC	Threshold Concepts

INTRODUCTION

The goal of a good science education is to build a scientifically literate populace and to ensure that the skills and knowledge gained are useful in the real world (Lave, 1988; Lobato, 2006). Pedagogical approaches that have gained traction in recent times (e.g., problem-based learning, place-based education, etc.) leverage the idea that students learn best from relatable and relevant examples, which are comparable to situations that they might encounter outside the classroom (Allen & Tanner, 2003; Gentner et al., 2003; Smith, 2002). However, even though learning is situated - i.e., it takes place at a particular time, place, and in a particular setting - the outcomes of that learning are only determined in a different time, place, and setting (Gilbert, 2006). Therefore, the application of knowledge communicated in one context to a different one is a central feature of learning (Barnett & Ceci, 2002; Opfer & Thompson, 2008). In this dissertation, I present my research about how context influences the ways students reason and represent biological systems.

The inability to transfer concepts learned in one context to another prevents students from applying what they have learned in class to their daily lives (Georghiades, 2000; Lobato, 2006). Transferring concepts and principles across contexts is important because the same principles apply to multiple contexts (Bransford & Schwartz, 1999; Morris et al., 1977; Thorndike & Woodworth, 1901), and relearning common principles in each new context is inefficient. There are multiple factors that can influence knowledge transfer, including overlap (degree of similarity between learning context and

assessment context), degree of abstraction (continuum between hyper-contextualised to devoid of any contextual details), mode of transfer (active vs. passive), and the specific nature of the context (e.g., instruction vs. assessment) (Gentner et al., 2003; Jacobson & Spiro, 1995; Loewenstein et al., 1999; Loewenstein & Gentner, 2001; Nehm & Ha, 2011; Vosniadou, 1989).

Contrary to the functionalist view that generalisation of knowledge (and transfer) best happens by “freeing oneself from experience” (Lave, 1988), the consensus is that knowledge is not created in a vacuum, but is affected by the setting in which it is constructed (J. S. Brown et al., 2007; Hall, 1996; Van Oers, 1998). Context plays a vital role not only in shaping, but also in eliciting this knowledge (Hofer, 2006). Understanding how context both helps and hinders learning and transfer is therefore of paramount importance if we are to improve science literacy (National Academies of Sciences, Engineering, and Medicine [NASEM], 2016).

By its nature, ‘context’ is difficult to define (Goodwin & Duranti, 1992). The common language definition in the Oxford English Dictionary (2020) is, “The whole structure of a connected passage regarded in its bearing upon any of the parts which constitute it; the parts which immediately precede or follow any particular passage or ‘text’ and determine its meaning”. However, technical definitions within and across disciplines are substantially broader. It has been described as a type of framework of features that in unison give meaning to an action (Van Oers, 1998). Gilbert (2006) considered the purpose of context to be that of adding clarity and perspective and defined it as, “the

circumstances that form the setting for an event, statement, or idea, and in terms of which it can be fully understood".

Consequently, researchers and educators use the term 'context' in a wide variety of ways. In science education, 'context' generally refers to features of the task stem (i.e., the prompt or question) which could include 'personal perspectives and concrete examples' (Krell et al., 2012; Son & Goldstone, 2009). But it has also been used to refer to: a societal and cultural setting (NASEM, 2016), a focal event (Goodwin & Duranti, 1992), a specific discipline (Nehring et al., 2012; Topcu, 2013), the order in which questions are asked (Federer et al., 2015), the types of questions asked (Driver et al., 1994; Watkins & Elby, 2013), and even the specific words in the question (S. Brown et al., 2011; Krell et al., 2015; Nehm & Ha, 2011).

In science education it is accepted that context influences teaching and learning of both knowledge and skills, but the nature of that influence is still debated (Gobert et al., 2011; Hofer, 2006; Krell et al., 2012, 2014; Muis et al., 2006; Nehring et al., 2012; Op 't Eynde et al., 2006; Topcu, 2013). Studies have shown that context influences both the way knowledge is integrated into and elicited from mental knowledge structures (J. S. Brown et al., 2007; Hall, 1996; Jones et al., 2000; Williams & Hollan, 1981). Teaching concepts in a highly contextualised manner (personalised to the learner) has been linked to both desirable (Allen & Tanner, 2003; Parker & Lepper, 1992; Wason & Shapiro, 1971) and undesirable outcomes (Detterman & Sternberg, 1993; Lave, 1988; Son & Goldstone, 2009). Additionally, assessment context has been shown to influence

the ideas elicited in student responses (Göransson et al., 2020; Kohn et al., 2018; Nehm & Ha, 2011), the way students approach problems and the skills they use (Bennett et al., 2020; Chi et al., 1981; Çikla & Çakiroğlu, 2006; Keller & Hirsch, 1998; Prevost et al., 2013), as well as their performance (Schurmeier et al., 2010) .

There are still many questions that remain unresolved, such as: Should teaching be done in contexts that are 'typical' of the discipline rather than contextualized to the learner (Driver et al., 1994)? Does context affect all learning and understanding? Or, are there generalities that can be used to teach certain concepts (Greeno, 2009; Guerra-Ramos, 2012; Hofer, 2006; Son & Goldstone, 2009)? Is there a best context in which to teach/assess a particular concept for a particular class? Or, should we teach/assess concepts in a wide variety of contexts to overcome biases (Leach et al., 2000)?

This dissertation aims to contribute to our understanding of the way context influences the ways students both reason about and represent biological systems. And because understanding evolution is fundamental to understanding biology, I have used evolution as the theme to investigate students' reasoning and representations. The first two chapters explore the influence of specific features of a question prompt (i.e., the task stem or the item-feature; Krell et al., 2012, 2015; Nehm & Ha, 2011). While the first chapter investigates how changes in the prompt influence the content of students' narrative responses, the second chapter investigates how the same item features influence the content and architecture of student's' model-based responses. In the third chapter, I explore how context, defined as the mode by which students are asked to

represent their reasoning (i.e., narrative vs. model), influences the content of students' responses. Collectively, my work on in this dissertation aims to further our understanding of contextual influences on students reasoning and representations so that we might improve the efficiency and effectiveness of instruction and assessments and make education more inclusive.

REFERENCES

REFERENCES

Allen, D., & Tanner, K. (2003). Approaches to Cell Biology Teaching: Learning Content in Context--Problem-Based Learning. *Cell Biology Education*, 2(2), 73–81. https://doi.org/10.1187/cbe.03-04-0019

Barnett, S. M., & Ceci, S. J. (2002). When and where do we apply what we learn? A taxonomy for far transfer. *Psychological Bulletin*, 128(4), 612–637. https://doi.org/10.1037//0033-2909.128.4.612

Bennett, S., Gotwals, A. W., & Long, T. M. (2020). Assessing students' approaches to modelling in undergraduate biology. *International Journal of Science Education*, 1–18. https://doi.org/10.1080/09500693.2020.1777343

Bransford, J. D., & Schwartz, D. L. (1999). Rethinking Transfer : A Simple Proposal with Multiple Implications. *Review of Research in Education*, 24, 61–100.

Brown, J. S., Collins, A., & Duguid, P. (2007). Situated Cognition and the Culture of Learning. *Educational Researcher*, 18(1), 32–42.

Brown, S., Lewis, D., Montfort, D., & Borden, R. (2011). The Importance of Context in Students ' Understanding of Normal and Shear Stress in Beams. *118th ASEE Annual Conference & Exposition*, Session W522B.

Chi, M. T., Feltovich, P. J., & Glaser, R. (1981). Categorization and Representation of Physics Problems by Experts and Novices. *Cognitive Science*, 5(2), 121–152.

Çikla, O. A., & Çakiroğlu, E. (2006). Seventh grade students' use of multiple representations in pattern related algebra tasks. *Hacettepe Üniversitesi Eğitim Fakültesi Dergisi*, 31(31), 13–24.

Detterman, D. K., & Sternberg, R. J. (Eds.). (1993). *Transfer on trial: Intelligence, cognition, and instruction*. Ablex Publishing.

Driver, R., Asoko, H., Leach, J., Scott, P., & Mortimer, E. (1994). Constructing Scientific Knowledge in the Classroom. *Educational Researcher*, 23(7), 5–12.

Federer, M. R., Nehm, R. H., Opfer, J. E., & Pearl, D. (2015). Using a constructed-response instrument to explore the effects of item position and item features on the assessment of students' written scientific explanations. *Research in Science Education*, 45(4), 527–553. https://doi.org/10.1007/s11165-014-9435-9

Gentner, D., Loewenstein, J., & Thompson, L. (2003). Learning and transfer: A general role for analogical encoding. *Journal of Educational Psychology*, 95(2), 393–408. https://doi.org/10.1037/0022-0663.95.2.393

Georghiades, P. (2000). Beyond conceptual change learning in science education: Focusing on transfer, durability and metacognition. *Educational Research*, 42(2), 119–139. https://doi.org/10.1080/001318800363773

Gilbert, J. K. (2006). On the nature of “context” in chemical education. *International Journal of Science Education*, 28(9), 957–976. https://doi.org/10.1080/09500690600702470

Gobert, J. D., O’Dwyer, L., Horwitz, P., Buckley, B. C., Levy, S. T., & Wilensky, U. (2011). Examining the relationship between students’ understanding of the nature of models and conceptual learning in biology, physics, and chemistry. *International Journal of Science Education*, 33(5), 653–684. https://doi.org/10.1080/09500691003720671

Goodwin, C., & Duranti, A. (1992). *Rethinking Context: Language as an Interactive Phenomenon*. Cambridge University Press.

Göransson, A., Orraryd, D., Fiedler, D., & Tibell, L. A. E. (2020). Conceptual Characterization of Threshold Concepts in Student Explanations of Evolution by Natural Selection and Effects of Item Context. *CBE Life Sciences Education*, 19(1), ar1. https://doi.org/10.1187/cbe.19-03-0056

Greeno, J. G. (2009). A theory bite on contextualizing, framing, and positioning: A companion to son and goldstone. *Cognition and Instruction*, 27(3), 269–275. https://doi.org/10.1080/07370000903014386

Guerra-Ramos, M. T. (2012). Teachers’ Ideas About the Nature of Science: A Critical Analysis of Research Approaches and Their Contribution to Pedagogical Practice. *Science and Education*, 21(5), 631–655. https://doi.org/10.1007/s11191-011-9395-7

Hall, R. (1996). Representation as Shared Activity : Situated Cognition and Dewey ’ s Cartography of Experience. *The Journal of the Learning Sciences*, 5(3), 209–238.

Hofer, B. K. (2006). Domain specificity of personal epistemology: Resolved questions, persistent issues, new models. *International Journal of Educational Research*, 45(1–2), 85–95. https://doi.org/10.1016/j.ijer.2006.08.006

Jacobson, M. J., & Spiro, R. J. (1995). Hypertext Learning Environments, Cognitive Flexibility, and the Transfer of Complex Knowledge: An Empirical Investigation. *Journal of Educational Computing Research*, 12(4), 301–333. https://doi.org/10.2190/4T1B-HBP0-3F7E-J4PN

Jones, M. G., Carter, G., & Rua, M. J. (2000). Exploring the development of conceptual ecologies: Communities of concepts related to convection and heat. Journal of *Research in Science Teaching*, 37(2), 139–159. https://doi.org/10.1002/(SICI)1098-2736(200002)37:2<139::AID-TEA4>3.0.CO;2-1

Keller, B. A., & Hirsch, C. R. (1998). Student preferences for representations of functions. *International Journal of Mathematical Education in Science and Technology*, 29(1), 1–17. https://doi.org/10.1080/0020739980290101

Kohn, K. P., Underwood, S. M., & Cooper, M. M. (2018). Energy connections and misconnections across chemistry and biology. *CBE Life Sciences Education*, 17(1), 1–17. https://doi.org/10.1187/cbe.17-08-0169

Krell, M., Reinisch, B., & Krüger, D. (2015). Analyzing Students' Understanding of Models and Modeling Referring to the Disciplines Biology, Chemistry, and Physics. *Research in Science Education*, 45(3), 367–393. https://doi.org/10.1007/s11165-014-9427-9

Krell, M., Upmeier zu Belzen, A., & Krüger, D. (2012). Students ' Understanding of the Purpose of Models in Different Biological Contexts. *International Journal of Biology Education*, 2(2), 1–34.

Krell, M., Upmeier zu Belzen, A., & Krüger, D. (2014). Students' Levels of Understanding Models and Modelling in Biology: Global or Aspect-Dependent? *Research in Science Education*, 44(1), 109–132. https://doi.org/10.1007/s11165-013-9365-y

Lave, J. (1988). *Cognition in Practice: Mind, Mathematics and Culture in Everyday Life*. Cambridge University Press.

Leach, J., Millar, R., Ryder, J., & Séré, M. . (2000). Epistemological understanding in science learning: The consistency of presentations across contexts. *In Learning and Instruction* (Vol. 10, Issue 6).

Lobato, J. (2006). Alternative Perspectives on the Transfer of Learning: History, Issues, and Challenges for Future Research. *The Journal of the Learning Sciences*, 15(4), 431–449. https://doi.org/10.1207/s15327809jls1504

Loewenstein, J., & Gentner, D. (2001). Spatial Mapping in Preschoolers: Close Comparisons Facilitate Far Mappings. *Journal of Cognition and Development*, 2(2), 189–219. https://doi.org/10.1207/S15327647JCD0202_4

Loewenstein, J., Thompson, L., & Gentner, D. (1999). Analogical encoding facilitates knowledge transfer in negotiation. *Psychonomic Bulletin and Review*, 6(4), 586–597. https://doi.org/10.3758/BF03212967

Morris, C. D., Bransford, J. D., & Franks, J. J. (1977). Levels of processing versus transfer appropriate processing. *Journal of Verbal Learning and Verbal Behavior*, 16(5), 519–533.

Muis, K. R., Bendixen, L. D., & Haerle, F. C. (2006). Domain-generality and domain-specificity in personal epistemology research: Philosophical and empirical reflections in the development of a theoretical framework. *Educational Psychology Review*, 18(1), 3–54. https://doi.org/10.1007/s10648-006-9003-6

National Academies of Sciences, Engineering, and Medicine [NASEM]. (2016). *Science Literacy: Concepts, Contexts, and Consequences*. The National Academies Press. https://doi.org/10.17226/23595

Nehm, R. H., & Ha, M. (2011). Item feature effects in evolution assessment. *Journal of Research in Science Teaching*, 48(3), 237–256. https://doi.org/10.1002/tea.20400

Nehring, A., Nowak, K. H., Upmeier zu Belzen, A., & Tiemann, R. (2012). Doing Inquiry in Chemistry and Biology. The Context's Influence on the Students' Cognitive Load. *La Chimica Nella Scuola*, XXXIV–3(January), 253–258.

Op 't Eynde, P., De Corte, E., & Verschaffel, L. (2006). Epistemic dimensions of students' mathematics-related belief systems. *International Journal of Educational Research*, 45(1–2), 57–70. https://doi.org/10.1016/j.ijer.2006.08.004

Opfer, J. E., & Thompson, C. A. (2008). The trouble with transfer: Insights from microgenetic changes in the representation of numerical magnitude. *Child Development*, 79(3), 788–804. https://doi.org/10.1111/j.1467-8624.2008.01158.x

Oxford English Dictionary. (2020). Oxford English Dictionary Online. *Oxford English Dictionary*. http://dictionary.oed.com

Parker, L. E., & Lepper, M. R. (1992). Effects of Fantasy Contexts on Children's Learning and Motivation : Making Learning More Fun. Journal of Personality and Social Psychology, 62(4), 625. https://doi.org/10.1037/0022-3514.62.4.625

Prevost, L. B., Knight, J., Smith, M. K., & Urban-Lurain, M. (2013). *Student writing reveals their heterogeneous thinking about the origin of genetic variation in populations.* National Association for Research in Science Teaching.

Schurmeier, K. D., Atwood, C. H., Shepler, C. G., & Lautenschlager, G. J. (2010). Using item response theory to assess changes in student performance based on changes in question wording. *Journal of Chemical Education*, 87(11), 1268–1272. https://doi.org/10.1021/ed100422c

Smith, G. A. (2002). Place-based education: Learning to be where we are. *Phi Delta Kappan*, 83(8), 584–594. https://doi.org/10.1111/j.1475-682X.2001.tb01110.x

Son, J. Y., & Goldstone, R. L. (2009). Contextualization in perspective. *Cognition and Instruction,* 27(1), 51–89. https://doi.org/10.1080/07370000802584539

Thorndike, E. L., & Woodworth, R. S. (1901). The influence of improvement in one mental function upon the efficiency of other functions (I). *Psychology Review*, 8, 247–261.

Topcu, M. S. (2013). Preservice teachers' epistemological beliefs in physics, chemistry, and biology: A mixed study. *International Journal of Science and Mathematics Education*, 11(2), 433–458. https://doi.org/10.1007/s10763-012-9345-0

Van Oers, B. (1998). The Fallacy of Decontextualization. *Mind, Culture, and Activity*, 5(2), 135–142. https://doi.org/10.1207/s15327884mca0502-7

Vosniadou, S. (1989). Analogical reasoning as a mechanism in knowledge acqisition: a developmental perspective. *Similarity and Analogical Reasoning*, Technical, 413–437.

Wason, P. C., & Shapiro, D. (1971). Natural and Contrived Experience in a Reasoning Problem. *Quarterly Journal of Experimental Psychology*, 23(1), 63–71.

Watkins, J., & Elby, A. (2013). Context dependence of students' views about the role of equations in understanding biology. *CBE Life Sciences Education*, 12(2), 274–286. https://doi.org/10.1187/cbe.12-11-0185

Williams, M. D., & Hollan, J. D. (1981). The process of retrieval from very long-term memory. *Cognitive Science*, 5(2), 87–119. https://doi.org/10.1207/s15516709cog0502_1

CHAPTER ONE:

Prompt context influences the content of students' narratives of evolution by natural selection

ABSTRACT

Despite the importance given to understanding evolutionary theory by scientists and educators, students graduate with incomplete and incorrect ideas about evolution. The context in which knowledge about evolution and evolutionary processes is obtained and elicited can contribute to facilitating or hindering the learning process. Human evolution in particular, has been controversial and especially difficult for students to accept and understand. Multiple studies have documented the difficulties that arise with respect to the acceptance of human evolution. This study contributes to the field that seeks to understand how context shapes the way students reason about evolution, including human evolution, and the effect of instruction on their reasoning.

We asked students in a large (n=160) introductory biology course how a biologist would explain the evolution of traits in humans and in cheetahs. Structural traits (heel/leg bones) were contrasted with functional traits (abilities to walk upright/run fast); and these were contrasted within each taxon. EvoGrader (an online assessment tool) scored students' responses for the presence of 6 key concepts (variation, heritability, competition, limited resources, differential survival, and non-adaptive) and three naïve ideas (adapt, need, and use/disuse).

We found that taxon was a significant predictor for the number of key concepts in students' responses; responses to prompts about cheetahs had more key concepts and were more likely to have only key concepts (no naïve ideas) than responses to prompts about humans. Responses to questions about functional traits had more naive ideas that than those about structural traits. At the end of a semester of instruction, the number of responses with key concepts as well as the number of key concepts per response increased, and the number of responses with naive ideas and the number of naive ideas per response decreased.

Our results are consistent with prior research that shows a clear effect of contextual influences (taxa and trait) on student reasoning about evolution. This suggests that students are reasoning differently when thinking about evolutionary processes with respect to humans as compared to other non-human animals. Students' explanations are further influenced by instruction, after which students responded with more key concepts, fewer naive ideas, and a modest decrease in the difference seen with respect to the context of the prompt. This is indicative of instruction having contributed to an increased understanding of the evolutionary process which in turn could have resulted in a decreasing susceptibility to contextual influences and an increased potential to transfer knowledge.

INTRODUCTION

Dobzhansky famously wrote, “Nothing in biology makes sense except in the light of evolution” (Dobzhansky, 1973). As such, evolutionary theory provides the needed context, underpinnings, and coherence to understand biological complexity (Alters & Alters, 2001; Blackwell et al., 2003). Evolution is to biology what plate tectonics is to geology, relativity is to time, and heliocentrism is to astronomy (Deniz & Borgerding, 2018b). However, despite its importance, it is perhaps one of the most controversial and polarizing topics in science (Glaze & Goldston, 2015; Pobiner, 2016). Across many countries and cultures, a significant proportion of people do not accept evolution as the unifying theory that explains the origin and diversity of life (Allmon, 2011; Brenan, 2019; Council of Europe, 2017; Deniz & Borgerding, 2018b; Downie & Barron, 2000; Miller et al., 2006; Nehm & Schonfeld, 2007; Oliveira & Cook, 2018; Smith, 2010a, 2010b; Thagard & Findlay, 2010).

Understanding evolution is a critical component of science literacy and its centrality to the biology curriculum is broadly valued by the scientific community. National and international reports that provide guidance for science teaching at both the K-12 and undergraduate level have stressed inclusion of evolution as a foundational concept (American Association for the Advancement of Science [AAAS], 2011; Deniz & Borgerding, 2018a; NGSS Lead States, 2013; UK Department of Education, 2015). However, a significant number of students graduate from college without an understanding of evolution even after rigorous training in science (Alters & Nelson, 2002; Kalinowski et al., 2010; Pobiner et al., 2018).

Students have great difficulty comprehending and explaining evolution and misconceptions often persist despite explicit instruction (Bishop & Anderson, 1990; Bray Speth et al., 2009; Catley & Novick, 2009; Morabito et al., 2010; Nehm & Reilly, 2007; Nehm & Ridgway, 2011; Sinatra et al., 2008; Smith, 2010a, 2010b). Some studies have suggested that students are also less likely to retain concepts related evolution than non-controversial topics such as photosynthesis (Glaze & Goldston, 2015; Nehm & Schonfeld, 2007; Sinatra et al., 2003).

As with any concept, student's knowledge and understanding of evolutionary theory is not created in a vacuum, but is affected by the setting in which it is constructed (Brown et al., 2007; Hall, 1996; Van Oers, 1998). Context plays a vital role not only in shaping, but also in eliciting and activating this knowledge (Clark, 2006; Hofer, 2006; Jones et al., 2000; Sabella & Redish, 2007). Understanding how context both helps and hinders learning and knowledge transfer is therefore of paramount importance if we are to improve science literacy (National Academies of Sciences, Engineering, and Medicine [NASEM], 2016; Nehm & Ha, 2011).

Although 'context' can mean many things (e.g., disciplinary perspective, social and cultural context in which learning occurs, etc.), in this paper, we restrict the use of 'context' to refer to specific features of a prompt or task stem used in assessment. Prompt context has been shown to be an important feature of assessment design. Several studies have shown that alternative phrasing can lead students to interpret a prompt differently (often incorrectly) or to focus on irrelevant or superficial features

unrelated to the intent of an assessment item (diSessa et al., 2004; Nehm & Ha, 2011; Ozdemir & Clark, 2009; Prevost et al., 2013). Seemingly minor changes in the wording of prompts have been shown to result in major changes in student performance on assessments (Chi et al., 1981; Potari & Spiliotopoulou, 1996; Schurmeier et al., 2010). Such contextual influences affect both novice and expert learners. For example, Gros, Sander, & Thibaut (2019) showed that when presented with certain prompt contexts involving daily-life scenarios, even experts could not solve simple mathematical problems.

Numerous studies have shown that context is an important factor contributing to the difficulty associated with learning and assessing students' conceptions of evolution (Beggrow & Sbeglia, 2019; Evans, 2008; Ha et al., 2006; Kampourakis & Zogza, 2008, 2009; Prevost et al., 2013). Clough & Driver (1986) observed inconsistent reasoning in students' explanations of evolution by natural selection when asked to reason about the origin and prevalence of different colours in caterpillars and in foxes. Nehm and Schonfeld (2008) showed performance differences when students were asked conceptually equivalent questions that differed with respect to scale, organism and evolutionary direction. Nehm & Ha (2011) similarly showed that prompt features such as scale, evolutionary direction, lineage, organism, and trait influenced both the number of key concepts (e.g., variation, heritability) and naïve ideas (e.g., need, use) students used in their explanations of evolution by natural selection. For example, students included more scientific concepts when responding to prompts about trait gains and

more naïve ideas when reasoning about trait loss, despite otherwise equivalent prompt construction.

Assessing evolutionary knowledge and understanding becomes even more problematic when dealing with evolution of humans. Ever since Darwin proposed evolutionary theory, human evolution by natural selection has been controversial. The society he lived in, including his peers (A.R. Wallace included) took objection to the fact that humans were not the exception (Mayr, 1982). Even today, people are more willing to accept natural selection as an explanation for evolution for species other than humans (Miller et al., 2006). Such trends are seen even among college educated adults (Brenan, 2019). Many studies that have explored students' acceptance of human evolution have shown that students reason differently in human vs non-human animal contexts (Atran, 1998; Atran et al., 2001; Nettle, 2010) and that acceptance of evolution increases when the organism in question is farther in evolutionary distance from humans (Evans, 2008; Sinatra et al., 2003).

While we know that evolution acceptance can be influenced when considering human vs non-human organisms, we do not know if that difference extends to the content of students' explanations about evolution by natural selection. Beggrow & Sbeglia (2019) showed that disciplinary context was more important than prompt context when explaining differences in student responses to questions about human and non-human evolution.

In this study, we aim to contribute to a growing understanding about the role of context in influencing students' reasoning about evolution. Here, our use of 'context' is consistent with definitions offered by Krell et al. (2015, 2012) and Nehm & Ha (2011), where 'context' refers to the specific features of a question prompt (i.e., the task stem or the item-feature). In particular, we ask whether students reason differently about evolution by natural selection when responding to prompts that vary in context and whether differences persist following a semester of active, learner-centred instruction.

METHODS

Setting and Participants

This study was conducted at a large, public university in the Midwest with highest research activity (The Carnegie Classification of Institutions of Higher Education, n.d.). Data for these analyses came from student responses in a large introductory biology course for majors (n = 194 students enrolled) that focused on content domains of genetics, evolution, and ecology. The course is second in a 2-course sequence required for life science majors; the first course focused on cell and molecular biology. Of the enrolled students, 160 completed all required tasks and were included in the analysis. The study population was 61% female, 21% first generation college students, and 21% non-White, with an average GPA of 3.2 (4.0 scale). The course is targeted toward sophomores (59% of students in study) but also includes a significant number of juniors (31%) and few freshmen (3%) or seniors (7%).

Assessment Design

We designed four isomorphic prompts based on the ACORNS instrument (Nehm et al., 2012) to assess students' explanations about natural selection in human and non-human animals. The prompts contained the following basal structure: *"(Taxon) has (trait). How would biologists explain how a (taxon) with (trait) evolved from an ancestral (taxon) without (trait)?*" Contextual features of prompts varied in taxon (human vs cheetah) and type of trait (structural vs functional). Humans and Cheetahs are not very distant in evolutionary terms (diverged approximately 96 MYA (Kumar et al., 2017)) and prior studies have examined how students reason when the taxon is 'cheetah' (Bishop & Anderson, 1990; Nehm & Ha, 2011; Nehm & Reilly, 2007). 'Structural traits' in this study refer to morphological traits that affect fitness, specifically 'heel bones' in humans and 'leg bones' in cheetahs. 'Functional traits' are behavioural traits or abilities that similarly affect fitness, such as 'walking upright' in humans and 'running fast' in cheetahs. All the prompts dealt with trait gain rather than trait loss. Prompts were designed as open-ended questions because prior research has shown they provide better insights into student thought processes and subject knowledge (Foddy, 1993).

From the four prompts, we created two forms of the assessment, hereafter 'Human/Cheetah Assessment' (or, HCA). Each form contained two prompts that were the same with respect to trait type (structural or functional trait) but differed in taxon (human or cheetah; Table 1.1). To control for potential influences of order (Federer et al., 2015; Schuman & Presser, 1996), each form of the HCA was further divided into

sub-forms that differed in the order of appearance of each taxon (i.e., half of the copies of each form had cheetah first and half had human first).

Table 1.1. Prompts used in the Human/Cheetah Assessment. *Two forms of an assessment were developed that differed in trait type (structural vs functional). Each form prompted students (n=182, Form 1; n=138, Form 2) to explain evolution by natural selection for both human and non-human animals.*

		Trait	
		Form 1: Structural Trait	**Form 2: Functional Trait**
Taxon	**Human**	Modern humans have ***enlarged heels***. How would biologists explain how a species of humans with ***enlarged heels*** evolved from an ancestral human species without ***enlarged heels***?	Modern humans have ***the ability to walk upright***. How would biologists explain how a species of humans with ***the ability to walk upright*** evolved from an ancestral human species without ***the ability to walk upright***?
	Cheetah	A species of cheetah has ***long leg bones***. How would biologists explain how a species of cheetah with ***long leg* bones** evolved from an ancestral cheetah species without ***long leg bones***?	A species of cheetah has ***the ability to run fast***. How would biologists explain how a species of cheetah with ***the ability to run fast*** evolved from an ancestral cheetah species without ***the ability to run fast***?

Students completed the HCA individually during class time for credit. Formative assessments were administered regularly in class and awarded credit for participation/effort. Each student provided responses to the same form of the HCA at the beginning and end of the semester (i.e., form was held constant). Only students who completed the HCA at both times were included in analyses (n = 160). Table 1.1 provides a summary to show how these were distributed between taxa and trait. The

two taxa (human and cheetah) were not referenced during instruction or assessment at any point in the course.

Coding Responses

Students' narrative responses were coded using the online assessment tool EvoGrader (Moharreri et al., 2014). EvoGrader codes for the presence of six key evolution concepts (KCs; Variation, Heritability, Competition, Limited Resources, Differential Survival, and non-adaptive) and 3 naïve ideas (NIs; Adapt, Need, and Use/Disuse) (Table 1.2). EvoGrader's reliability and validity have been established in previous studies (see Moharreri, Ha, & Nehm, 2014) and demonstrated comparable to that of trained human raters (>0.81 Kappa) despite requiring 99% less time for scoring. It is important to note that EvoGrader evaluates presence/absence of concepts; additional analyses and coding approaches are necessary in order to make inferences about correct applications of concepts.

Table 1.2. Description of the six Key Concepts (KCs) and three Naïve Ideas (NIs). *Modified from Moharreri et al. (2014).*

Concept Type	Concept Name	Concept Description
Key Concepts	Variation	The presence and causes of variation (mutation / recombination /sex)
	Heritability	The heritability of variation (The degree to which a trait is transmitted from parents to offspring)
	Competition	A situation in which two or more individuals struggle to get resources that are not available to everyone
	Limited Resources	Limited resources related to survival/reproduction, such as food and predators, and reproduction (such as pollinators)
	Differential Survival/ Reproduction	The differential reproduction and/or survival of individuals
	Non-Adaptive Idea	Genetic drift and related non-adaptive factors contributing to evolutionary change
Naïve Ideas	Adapt / Acclimation	Adjustment or acclimation to circumstances (which may subsequently be inherited)
	Need / Goal	Goal-directed change; needs as a direct cause of evolutionary change
	Use / Disuse	The use (or lack of use) of traits directly causes their evolutionary increase or decrease

Data Analyses

A total of 640 student responses were included in analyses (n=160 students; 4 responses per student). Each student provided responses pre- and post-instruction for either human or cheetah. Within taxon (human or cheetah) students responded to two prompts regarding evolution of both structural and functional traits.

We used three quantitative approaches to determine the influence of context and instruction on student responses.

1. *Abundance and diversity of KCs and NIs*. In ecology, abundance indices measure the relative frequencies of organisms in a community, while diversity indices (e.g., Shannon, Simpson) measure variation in the types of organisms (e.g., species) across different communities. In our analyses, we considered the sum of all students' responses as analogous to a community, and subsets of them representing discrete populations (e.g., the population of responses to a cheetah prompt pre-instruction). We then explored both the diversity (KCs and NIs) and relative abundances of ideas (number of times each KC or NI appears) in the respective populations and in the community in general.

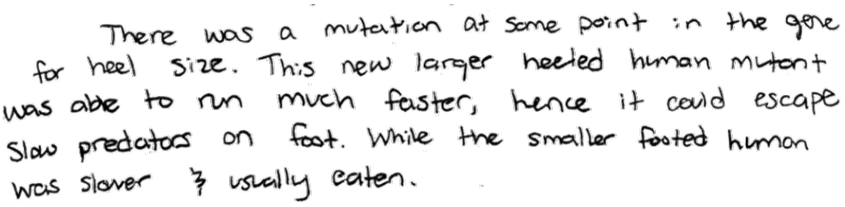
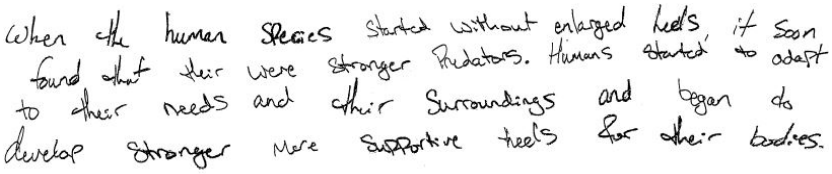
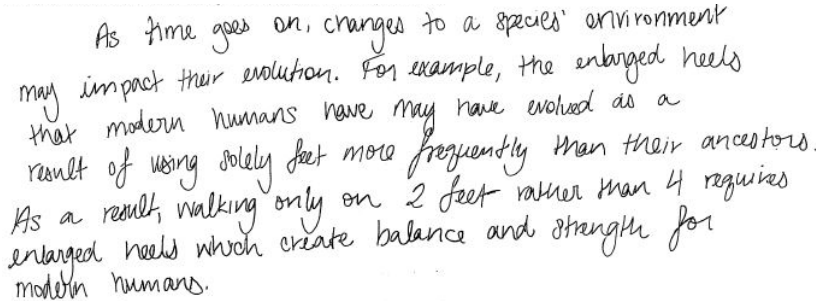
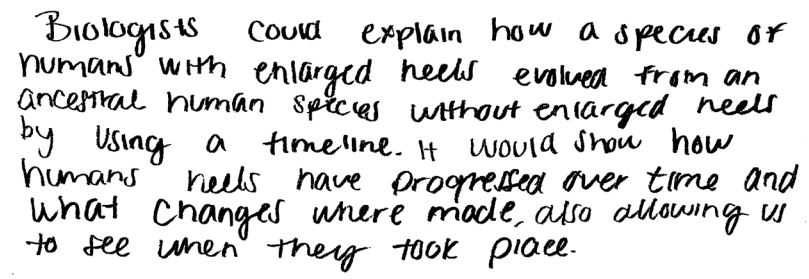
2. *Regression Analyses for total number of KCs and NIs.* We fitted regressions to quantify the effects of taxon, trait type, prompt order, and pre/post instruction on the total number of KCs and NIs. Since the data is not continuous, we used a mixed-effects Poisson distribution. We calculated the 95% confidence intervals

on all parameter estimates based on the model standard errors. The models showed signs of underdispersion, so we refit the models to account for this in two different ways: (i) using a mixed-effects zero-inflated Poisson regression, and (ii) using a mixed-effects Conway Maxwell Poisson regression. In both cases, the estimated coefficient values and p-values were the same, indicating that dispersion was not a major problem. Therefore, we present the results from the mixed-effects Poisson here.

3. *Multiple logistic regression analysis.* Students' responses were sorted into 4 groups based on the EvoGrader output:
    a. KC only: These responses had only key concepts. The maximum number of key concepts measurable by EvoGrader is 6.
    b. Mixed: These responses had both key concepts as well as naïve ideas.
    c. NI only: These responses had only naïve ideas. The maximum number of naïve ideas measurable by EvoGrader is 3.
    d. None: These responses had no key concepts or naïve ideas.

Examples of student responses with accompanying EvoGrader codes and corresponding group assignments are provided in Table 1.3. We performed multiple logistic regressions to understand how our independent variables (e.g., taxon, trait, prompt order, and instruction) contributed to the relative odds of belonging to one of the four groups.

Table 1.3. Examples of student responses belonging to each of the four groups based on their content coded by EvoGrader*. The four groups are: KC only, Mixed (both KCs and NIs present), NI only and None (neither KCs nor NIs present).*

Student Response	Coded by EvoGrader	Group
	<u>3 KCs:</u> Variation, Limited Resources, Differential Reproduction	KC only
	<u>1 KC:</u> Limited Resources. <u>1 NI:</u> Needs, Adapt	Mixed
	<u>2 NIs:</u> Use, Needs	NI Only
		None

Software

All statistical analyses were done using the R statistical environment v 3.6.3 (R Core Team, 2020). We made use of the *dplyr* (Wickham et al., 2020) and *tidyr* (Wickham & Henry, 2020) packages for data processing, *lme4* (Bates et al., 2015) for mixed-effects logistic regressions, *effects* (Fox, 2003) for computing and plotting marginal effects, *DHARMa* (Hartig, 2018) to checking residuals of mixed-effects models for patterns of overdispersion and underdispersion and *glmmTMB* (Brooks et al., 2017) to fit mixed-effects Poisson regression.

RESULTS

Our results showed that students' responses were influenced by both prompt context and instruction. Results of our specific analyses are presented with respect to each of our original research questions.

Prior research has shown that student performance on assessment tasks is affected by the sequence in which the assessment items are presented (Federer et al., 2015; Gray, 2004; Hambleton & Traub, 1974; MacNicol, 1956; Monk & Stallings, 1970), and general recommendations are to take task order into consideration when designing assessments (Schuman & Presser, 1996). However, similar to (Weston et al., 2015), in our study we did not find any such statistically significant effects.

1. How Do Contextual Features Influence the Content of Student Responses to Prompts About Evolution by Natural Selection?

Results of all three analytic approaches indicated that students' responses were significantly influenced by both taxon and trait.

Responses to questions about cheetahs had more KCs and fewer NIs than questions about humans (with the exception of Variation). Figure 1.1 shows the percentage of responses that contained each of the 6 KCs and 3 NIs for each taxon before and after instruction. Limited Resources was the KC most sensitive to the effect of taxon both before and after instruction. Pre-instruction, only 24% (n=38) of responses to the Human prompt mentioned Limited Resources compared to 62% (n=99) of responses to Cheetah. This was virtually unchanged with instruction, with 29% (n=46) and 63% (n=101) of responses to Human and Cheetah, respectively, mentioning Limited Resources. In contrast, Variation increased significantly with instruction, but there was almost no difference due to taxon. Interestingly, post-instruction Variation is the only instance in which we saw a higher frequency of a KC in the Human prompt (~6%, n=10 more) compared with Cheetah.

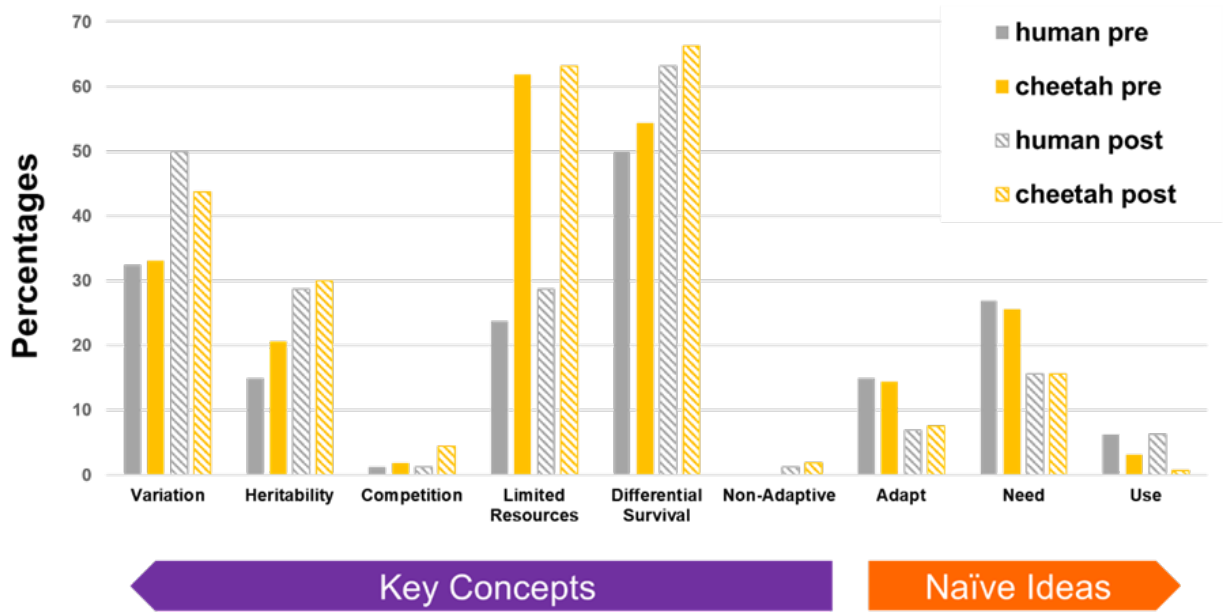


Figure 1.1. Percentage of responses that contain each of the six Key Concepts and three Naïve Ideas*. KCs occurred more frequently in Cheetah responses and responses written at the end of the semester. NIs occurred less frequently at the end of the semester.*

The mixed-effects Poisson regression (Figs. S1.1 and S1.2) and the mixed-effects logistic regression (Figs. S1.3 – S1.8) both show that taxa and trait influence the content of student narratives.

Overall, responses had an average of 1.7 KCs and 0.4 NIs. Number of KCs differed between taxa, with a mean of 1.8 KCs for Cheetah vs 1.4 KCs for Human ($p < 0.001$; Fig. 1.2). Most of the responses did not have any NIs, and the number of NIs differed based on type of trait. Responses had an average of 0.3 NIs when the prompt was about a Functional trait and 0.2 NIs when the prompt was about a Structural trait ($p < 0.05$; Fig. 1.3).

Figure 1.2. Average number of KCs in responses for each of the two taxa, estimated by the fitted model.

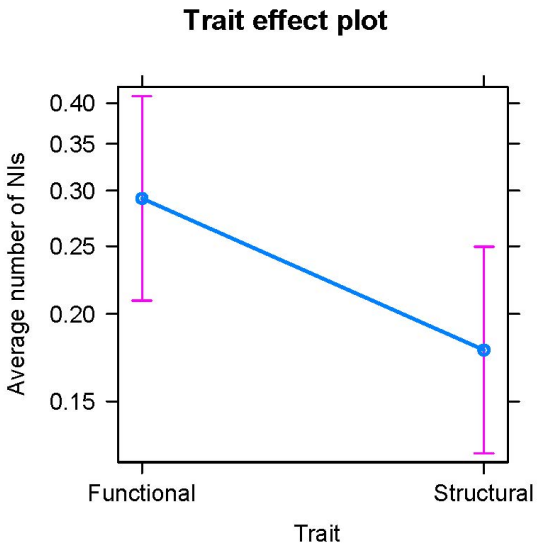


Figure 1.3. Average number of NIs in responses for each of the two traits, estimated by the fitted model.

Table 1.4. Odds ratios of logistic regression analysis for effect of Taxon (using 'Human' as the reference taxon) and Trait (using 'Structural' as the reference trait). *Values with asterisks are statistically significant (*** p < 0.001; ** p < 0.01; * p < 0.05, $^{\lambda}$p < 0.1). Lower and Upper Confidence intervals are provided in the brackets. This table provides the coefficients for 'Taxon' and 'Trait', however the model also included 'Pre/post Instruction' as a predictor.*

	NI only vs None	Mixed vs None	KC only vs None	Mixed vs NI only	KC only vs NI only	KC only vs Mixed
Taxon 'Human'	0.87 [0.17, 3.02]	**0.06*** ** [0.01, 0.22]	**0.19*** ** [0.08, 0.41]	**0.04*** ** [0.00, 0.19]	**0.17*** ** [0.06, 0.40]	1.14 [0.66, 1.99]
Trait 'Structural'	0.89 [0.23, 3.57]	0.27 [0.04, 1.29]	1.21 [0.45, 3.34]	0.31 [0.03, 1.56]	1.06 [0.35, 3.15]	**3.97** ** [1.51, 11.65]

Students' responses were even less likely to have either KCs only or a mixture of KCs and NIs, than no KCs or no ideas (KCs nor NIs) in their responses to the Human prompt (relative to the Cheetah prompt, $p < 0.001$, Table 1.4)

When responding to the Human prompt (relative to the Cheetah prompt) students were:

- 50% less likely to include a mixture of KCs and NIs in their responses, as opposed to only NIs or no ideas at all (Figs. S1.4 and S1.6)
- 6% less likely to include only KCs than only NIs (Fig. S1.7)
- 10% less likely to include only KCs than no ideas (Fig. S1.5)

Students responses were more likely to have only KCs, than a mixture of KCs and NIs when they were responding to prompts about a structural trait (relative to a functional trait, $p < 0.01$, Table 1.4). Students included only KCs ~15% more frequently than they included a mixture of KCs and NIs when writing about structural traits (Fig. S1.8).

2. How Does a Semester of Active, Learner-Centred Instruction Influence the Content of Student Responses to the Same Prompts?

Results of the abundance and diversity of KCs and NIs in students' responses (pre and post instruction) are shown in Figs. 1.4 and 1.5, respectively (n=640 for both). Although 6 KCs are possible, we observed no more than 4 within any response (n=47) and a majority contained at least 2 (n = 398). No KCs were present in 96 student responses. For NIs, the maximum of 3 naïve ideas was present in only a single response and a majority had 0 naïve ideas (n = 466).

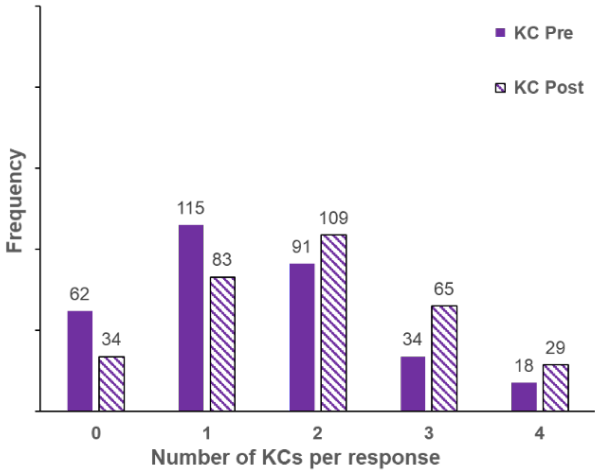


Figure 1.4. Frequency of the total number of KCs pre and post instruction.

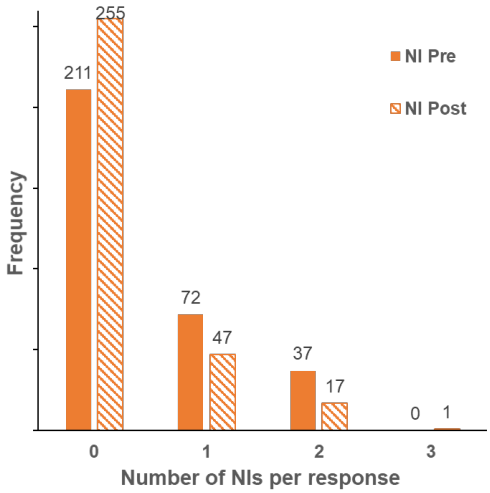


Figure 1.5. Frequency of the total number of NIs pre and post.

Overall, our results show that instruction increases the number of KCs in responses and decreases the number of responses containing no KCs. Similarly, instruction decreases the number of NIs per response and increases the frequency of responses with no NIs.

Differential Survival was the most frequently applied KC; it was present in more than 50% (n=167) of responses pre-instruction and more than 63% (n=207) responses post, irrespective of prompt context. The least used KC was Non-Adaptive, which appeared in only 1.5% (n=5) of the responses post-instruction (Fig. 1.1). Variation was the KC most responsive to instruction, with 33% (n=105) and 47% (n=150) responses including it pre- and post-instruction, respectively. Post-instruction, >93% (n=140) of the responses that mentioned Variation were in the KC Only group; only 6.6% (n=10) of those responses had any NIs at the end of the semester, compared to 14% (n=15) at the beginning of the semester (Fig. 1.1). Taxon-specific differences in KCs decreased moderately with instruction, with the greatest reductions observed for Heritability (4.4%) and Limited Resources (3.7%; Fig. 1.1).

The above trends are further corroborated by regression analyses (Figs. 1.6 and 1.7) that show the results of our mixed effects Poisson regressions for significant fixed effects (pre/post instruction) for KCs and NIs, respectively. Table 1.5 gives the odds ratios of multiple logistic regressions that show the relative odds of belonging to one of the four previously mentioned groups (Table 1.3) based on pre/post instruction (post instruction as the reference value).

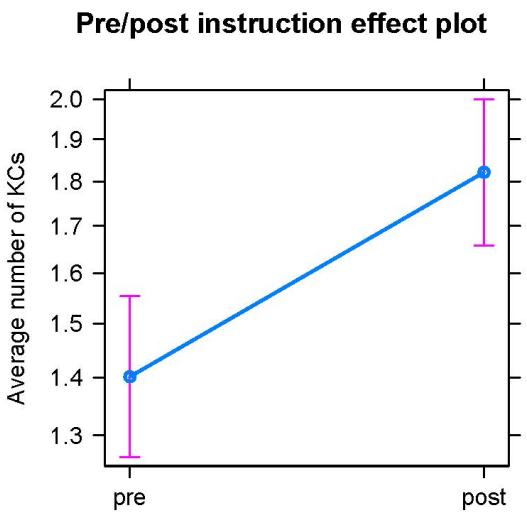


Figure 1.6. Average number of KCs in responses for pre and post instruction, estimated by the fitted model.

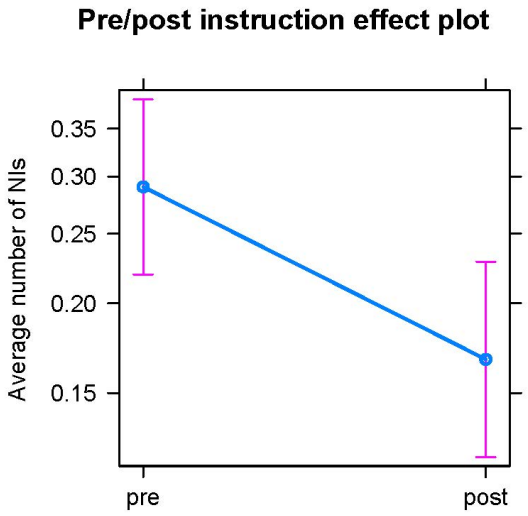


Figure 1.7. Average number of NIs in responses for pre and post instruction, estimated by the fitted model.

Overall, students' responses contained more KCs following instruction, regardless of taxon or trait type. Post instruction had 30% more KCs compared to pre-instruction (1.9 vs 1.4, respectively; $p \leq 0.001$; Fig. 1.6). Additionally, responses had 40% fewer NIs at the end of the semester ($p \leq 0.001$; Fig. 1.7).

Table 1.5. Odds ratios of logistic regression analysis for effect of instruction using 'post instruction' as the reference point. *Values with asterisks are statistically significant (*** p < 0.001; ** p < 0.01; * p < 0.05, $^{\lambda}$p < 0.1). Lower and Upper Confidence intervals are provided in the brackets. This table provides the coefficients for 'Pre/post Instruction', however the model also included 'Taxon' and 'Trait' as predictors.*

	NI only vs None	Mixed vs None	KC only vs None	Mixed vs NI only	KC only vs NI only	KC only vs Mixed
Pre/post instruction 'post'	0.49 [0.14, 1.47]	1.13 [0.36, 3.76]	**2.06*** [1.03, 4.27]	3.35$^{\lambda}$ [0.99, 17.29]	**4.62***** [2.01, 12.47]	**3.06***** [1.75, 5.54]

Students' responses were even more likely to have KCs only, than a mixture of KCs and NIs, or only NIs, or no ideas (KCs nor NIs) in their responses post instruction (relative to pre instruction, *p* ranging from < 0.001 to < 0.05, Table 1.5)

Post instruction (relative to pre instruction) students were:

- 13% more likely to include only KCs in their responses in their responses, as opposed to a mixture of KCs and NIs (Fig. S1.8)
- 3% more likely to include only KCs than only NIs (Fig. S1.7)
- 3% more likely to include only KCs than no ideas (Fig. S1.5)

DISCUSSION

Our results indicate that students' explanations about evolution by natural selection are influenced by both contextual features of prompts (taxon and trait type) and by instruction. In this section, we compare our findings with previous studies and offer

some possible explanations for the patterns we see. Additionally, we will discuss the implications of our findings on instruction and assessment.

Contextual Effects of the Prompt

The isomorphic prompts in our study share a common underlying structure and are intended to assess equivalent knowledge despite minor variations in superficial features unrelated to the construct of interest. Since they have the exact same prompt stem (except for the couple of contextual words) they are designed to go beyond defined standards of equivalency in difficulty and complexity (Kjolsing & Van Den Einde, 2016). Terms such as “explanatory coherence” (Kampourakis & Zogza, 2009), “knowledge coherence” (Nehm & Ha, 2011), and “causal flexibility” (Evans, 2008) refer to one’s ability to produce similar responses to isomorphic prompts and demonstrate an ability to identify a relevant concept despite irrelevant or peripheral details. For example, Weston, Haudek, Prevost, Urban-Lurain, & Merrill (2015), found that changing a species on questions about photosynthesis did not influence students’ responses. They state that students did not consider the species in the prompt to be a relevant detail and therefore, did not consider it when formulating their response. In contrast, many studies, including our own, have shown that students’ explanations about evolution by natural selection are highly susceptible to contextual features of question prompts (Kampourakis & Zogza, 2008; Prevost et al., 2013; Schurmeier et al., 2010). In particular, our findings are consistent with others that have shown taxon to be particularly influential in shaping students’ responses (Beggrow & Sbeglia, 2019; Nehm & Ha, 2011).

In each of our analyses, we found that ‘taxon’ was the most important variable influencing the number and type of KCs in a response and the group to which the response belonged. Responses to prompts about human evolution had fewer KCs and were more likely to have NIs despite instruction. This suggests that students are reasoning differently about humans compared to non-human animals. Beggrow & Sbeglia (2019) found that even students who study humans as a focal organism (e.g., anthropology majors) responded with fewer KCs and more NIs in responses to questions about evolution in humans as compared to non-human animals. Similar results were obtained by Ha et al. (2006) who found that students were less likely to use ‘natural selection after mutation’ as an explanation in response to questions about human evolution as compared to their responses to questions about plant and other animal evolution. It is possible that because students consider humans taxonomically unique (Coley, 2007) and not part of the evolutionary tree (AAAS, 2018; Coley & Tanner, 2015) that they are willing to reason differently about humans in evolution contexts.

Effects of Instruction

Increased use of KCs and decreased sensitivity to prompt contexts can be important indicators of student’s understanding of evolution and acceptable measures of instructional efficacy. Our results show that patterns of KCs and NIs changed between the start and end of the semester. At the beginning of the semester, all but one of the KCs were represented in student responses, albeit fewer per response. With instruction however, we observed an increase in the number of KCs and decrease in the number of

NIs per response, as well as a modest reduction in response differences due to taxon. This indicates that instruction increased the accuracy of students' explanations of evolution but had minimal impact in reducing the influence of context. This finding is consistent with research that has examined a variety of instructional methodologies targeting students understanding of evolution (Bray Speth et al., 2009, 2014; Halldén, 1988; Kampourakis & Zogza, 2009; Nehm & Reilly, 2007; Nettle, 2010; Pobiner et al., 2018).

Of the KCs assessed, Variation was most responsive to instruction. Students' use of Variation increased by 10.6% and 17.5% in cheetah and human contexts, respectively. In the course that was the target of this study, variation was a central theme. Course content was organized around the central questions of: how does biological variation originate at the molecular level? How is molecular-level variation expressed at the organismal level? And what are consequences of organismal variation for evolution of populations and ecosystem function? Our data revealed that students gained an appreciation of variation during the semester (14% more inclusion of Variation on average in the post-semester responses). Our findings are consistent with those of Speth et al. (2014) that observed improvement in students' representations of origin of variation using a similar instructional approach. Additionally, in our study, variation was elicited at a greater extent by the human prompt post-instruction. This could be an artefact of student's general tendency to categorise by species (not recognise individual level variation) when asked about non-human animals as compared to humans (Nettle,

2010), rather than a direct consequence of instruction causing them to appreciate variation differentially between the species.

An appreciation of the causes, consequences, and extent of variation is central to understanding evolution (Emmons & Kelemen, 2015; Gregory, 2009; Halldén, 1988; Shtulman, 2006). Darwin himself recognised the importance of variation (Darwin, 1868, Ch. 20, p. 192) and lamented the lack of understanding of its origin (Darwin, 1859, Ch. 5, p. 167). In our study, few of the responses that included Variation had any naïve ideas (9.8%). This is consistent with the findings of Shtulman & Schulz, (2008) who showed that students who have a better understanding of within-species variation also have an accurate and mechanistic understanding of natural selection.

We observed that although the presence of naïve ideas decreased post instruction, they persisted in students' responses. At the start of the semester 34% of responses had naïve ideas compared to the 20% of responses at the end of the semester. Our results are consistent with many studies that have shown that naïve ideas, a form of intuitive thinking, are remarkable resistant to change and frequently co-exist with correct scientific conceptions that are fundamentally mutually exclusive (Bishop & Anderson, 1990; Bray Speth et al., 2009; Nehm & Reilly, 2007; Nehm & Ridgway, 2011; Shtulman & Valcarcel, 2012; Sinatra et al., 2008; Smith, 2010a, 2010b).

Linking findings with Existing Theory

The literature offers several insights that could account for the difficulties associated with teaching and learning evolution. Here, we discuss three hypotheses that may inform our understanding of the patterns we observed: world-view and intuitive thinking, prior knowledge and experience, and scientific expertise.

Students' world-view and intuitive thinking

A worldview is a set of deeply entrenched beliefs and expectations that form the framework of a person's individuality and define how they see the world around them (Glaze & Goldston, 2015). Worldviews regarding evolution often do not change after instruction (Blackwell et al., 2003; Cavallo & McCall, 2008) and can hinder understanding and acceptance of evolutionary theory (Alters & Nelson, 2002; Evans, 2008; Nehm, 2006). Smith (2010a) proposes an alternative hypothesis: a limited understanding of evolution, including being exposed to inadequate empirical evidence, could be interfering with accepting it as part of one's worldview. Ingram & Nelson (2006) showed that after instruction about evolution students' positive views towards evolution increased and students who showed the greatest gains were those who were initially undecided about evolution. Regardless of the specific mechanism, inconsistencies between students' worldviews and tenets of evolutionary theory (especially with respect to human evolution) could make students more susceptible to contextual influences.

Intuitive ways of thinking about existing and new information can also pose barriers to increasing evolutionary understanding and decreasing contextual susceptibility. Smith

(2010b) describes these predictable ways of thinking as ‘rules of thumb’. These are default approaches that are ingrained into the brain and are used in situations where there is a lack of knowledge. Researchers have documented such expected patterns when students reason about biological entities, processes, and phenomena (Coley & Tanner, 2015; Inagaki & Hatano, 2006). Coley & Tanner (2015), categorised types of biological intuitive thinking into three different types of what they called ‘construals’ namely: teleological thinking, essentialist thinking, and anthropocentric thinking. These patterns of reasoning are powerful and can pose incredible barriers to learning since students do not understand that their reasoning itself is erroneous (Sinatra et al., 2008). These patterns include attributing a purpose to all events and attributing their cause to intentional agency. Such patterns lead students’ to incorporating naïve ideas like need and adapt. The tendency to categorise by type deters students from appreciating variation.

Students’ prior evolutionary knowledge and education.

Students arrive at every course with previous knowledge and prior conceptions about evolution that they have gained through their formal education and lived experiences. This knowledge often includes evolution misconceptions which have been well documented in the literature (e.g., Gregory, 2009; West, El Mouden, & Gardner, 2011). Alters & Nelson (2002), listed several factors inconsistent language usage and contradictory learning can contribute to misconceptions. For example, colloquial terms such as ‘fitness’ and ‘adaptation’ that have distinct meanings in and out of evolution contexts, or seeing humans and dinosaurs coexisting in various media.

By the time students reach the undergraduate classroom, their knowledge about evolution has also been influenced by their formal education, the quantity and quality of which is not consistent. Although evolution is now a part of the required curriculum in many countries, it is not required in some and banned outright in others. Even in countries that require evolution to be taught, the grades at which it is introduced, the perspective from which it is taught, and the focus of evolution education varies widely (Deniz & Borgerding, 2018a). In the United States, 20 states have adopted the Next Generation Science Standards (NGSS), which are generally more comprehensive than other state standards with respect to evolution (Gross et al., 2013). However, in their review of the NGSS. Gross et al. (2013) stated that while these standard were better than many state standards with respect to evolution, they too had some important weaknesses including the way with they dealt with heredity and the links between DNA and evolutionary relationships. In terms of our study, the major drawback we noticed with the NGSS is that they do not even mention human evolution.

Among the states that have not adopted the NGSS, some do not even mention the word ‘evolution’ in their standards and others make superficial references to it (Lerner, 2000; Vazquez, 2017) Additionally, adopting standards for evolution education does not guarantee they are actually being implemented (Glaze & Goldston, 2015) or that they are being implemented consistently. In some cases, students continue to be taught alternative theories in addition to, or at times instead of evolution (Bowman, 2008). Multiple studies have documented troublesome issues with teachers responsible for evolution education, ranging from inadequate preparation for teaching evolution (Smith,

2010b) to de-emphasising or avoiding teaching it (Glaze & Goldston, 2015) to purposefully teaching students that ‘evolution is wrong’ (Boujaoude et al., 2011). Such variability and inconsistency in students’ instruction about evolution make it difficult to make any sort of assumptions about their prior knowledge before entering the undergraduate classroom.

Even at the undergraduate level, evolution is rarely presented as a unifying theme for understanding biology, despite its pervasiveness as an explanatory construct across biological research. Instead, evolution is generally taught as a distinct topic without explicitly making it clear how it plays a role in other biological concepts and processes. This is reflected in the structure of textbooks frequently used in undergraduate biology instruction and in the syllabi derived from them (Nehm et al., 2009). Additionally in the context of this particular study, while ‘Evolution’ is one of the core concepts in Vision and Change (AAAS, 2011), the report does not refer to human evolution either (to be fair – it does not use any other taxa as a reference either).

Such differences in the quantity and quality of students’ prior evolutionary knowledge at both the K-12 and undergraduate levels could explain students’ susceptibility to contextual influences as well as the difficulty associated with changing students’ mental models of evolution that have been shaped by years of exposure and experiences. When it comes to undergraduate education,

Students' scientific expertise

As novice science learners, students may be more sensitive to contextual influences when learning complex concepts, such as evolution. There are major differences between the way experts and novices approach problem solving in any field. Experts have deeper conceptual understanding of their subject matter which enables them to be flexible in identifying and retrieving bits of relevant knowledge. This leads experts to intuitively see patterns that novices are unable to discern (National Research Council [NRC], 2000). Additionally, experts are more able to identify and focus on the abstract principles that underlie a problem's structure while novices tend to focus on more superficial features (Chi et al., 1981; Hmelo-Silver & Pfeffer, 2004; Nehm & Ridgway, 2011). As novices, it is not surprising that students are influenced by prompt contexts. However, our data suggest that instruction can decrease students' sensitivity to context, perhaps indicating they are making progress in their transition from novice to expert.

Implications for Instruction and Assessment

Multiple studies that have looked at different instructional strategies and contexts and have shown varying levels of improvements in students' understanding of evolution in general and human evolution in particular (Alters & Nelson, 2002; Bray Speth et al., 2009, 2014; Kalinowski et al., 2010; Kampourakis & Zogza, 2009; Pobiner et al., 2018).

Many researchers have called for evolution to be taught using humans as a focal organism. Nettle (2010) showed gains in student understanding of evolution in general after students were taught evolution in the context of humans. Pobiner et al. (2018) and

Pobiner (2016), propose teaching about human evolution as a direct and effective way to decrease barriers to accepting and subsequently understanding evolution. However, Beggrow & Sbeglia (2019), did not find any particular affordances offered by teaching evolution in a human context. Deeper learning can result when the learner identifies with the subject matter and finds it relevant (NRC, 2009). Therefore, since students are highly likely to find themselves and their development interesting (Pobiner et al., 2018), so teaching evolution in the human context could mean that students find it relevant and identify with it. We would like to offer the suggestion that while teaching evolution in a solely human context might not be the optimal solution, including humans as one of the contexts is very important.

At various institutions and at the national level, numerous efforts are underway to renovate and align the biology core curriculum. In particular, there is considerable interest in increasing the prominence of science practices as an explicit objective for student learning at all levels so as to teach science as it is practised and to thus encourage students to think and reason about science similar to the practitioners (AAAS, 2011; Cooper et al., 2015). A lack of scientific accuracy in students' reasoning is often not because of a lack of knowledge of the scientific principles, but due to inadequate activation, recruitment, or transfer of those scientific principles across contexts (Nehm & Ha, 2011). Pedagogical techniques that include various forms and degrees of scientific practises (active learning) in the classroom, have been shown to improve learning gains (Freeman et al., 2014). Our study, as well as that of Speth et al. (2014) showed gains in student performance in understanding variation after a

semester of model-based pedagogy. Perhaps by using scientific practises such as data analysis, modelling and argumentation during active learning based instruction, and including humans as an instructional context, will lead to a deeper, more conceptual understanding of evolution, an increased ability to transfer relevant concepts, and thereby decrease susceptibility to contextual influences.

Finally, our findings have clear implications for assessment. A common strategy used in assessment is to frame parallel or ‘isomorphic’ prompts. Parallel prompts are framed to test concepts in usually using contexts that were not used during instruction and it is assumed that they are adequate in measuring learning outcomes. The assumption is that students will be able to identify the concepts they are being tested on, recruit the relevant knowledge, and transfer it to the new context. However, there are multiple difficulties with making such assumptions.

It has been established that ensuring that prompts are parallel in terms of difficulty is both, important and difficult to accomplish (Hamp-Lyons & Mathias, 1994; Lee & Anderson, 2007; Li, 2018; Sydorenko, 2011). Our findings indicate that while making assumptions of equivalence, we should be considering yet another dimension – the contexts that are chosen. The prompts used in our study were designed not just to test for equivalent content but were truly isomorphic in that they utilised the same prompt stem. The words that varies were also remarkable similar, both were mammals, close in terms of evolutionary terms (Kumar et al., 2017), and both were familiar to the students

(Nehm et al., 2012). However, we still saw differences in student responses based on the couple of words that varied.

Perhaps student understanding and reasoning, including with respect to evolution by natural selection (Nehm & Ha, 2011), can be comprehensively and accurately assessed only when context is taken into consideration. When we use multiple contexts both in instruction and in assessment, we can facilitate the students being able to identify the important concepts and transfer them across contexts.

ACKNOWLEDGEMENTS

I thank Mitch Distin and Tammy Long for help in conceptualising the study and for help with data collection; Etiowo Usoro and Patrycja Zdziarska for logistical and technical support throughout the project; Mridul Thomas for help with data analysis; and Melanie Cooper, Amelia Gotwals, and Katherine Gross for comments on earlier versions of the manuscript. I am grateful to Francesco Pomatti and Anita Narwani at the Department of Aquatic Ecology at Eawag, Switzerland for providing me space and resources while writing this manuscript. I express my deep gratitude to the world's best advisor, Tammy Long, for the mentorship and supervision provided during all stages of this project. Finally, I gratefully acknowledge all the anonymous students whose assessments we used as the data for this study.

Funding

This material is based in part upon research supported by the National Science Foundation under grant numbers DRL 1420492, DRL 0910278, and DBI-0939454. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

APPENDIX

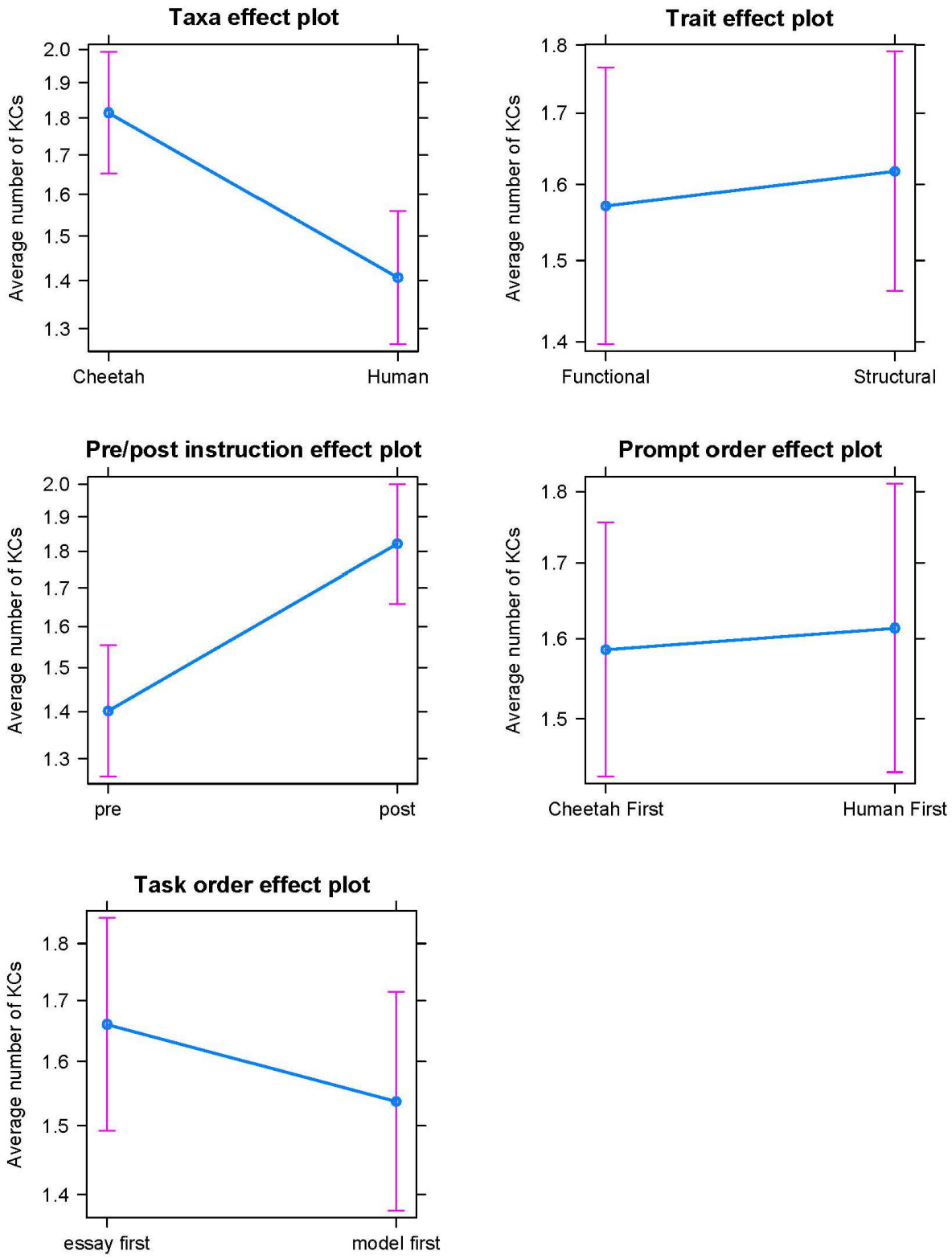


Figure S1.1. Plots showing the effects of taxon, trait type, prompt order, and pre/post instruction on the average number of KCs determined using a mixed-effects Poisson regression.

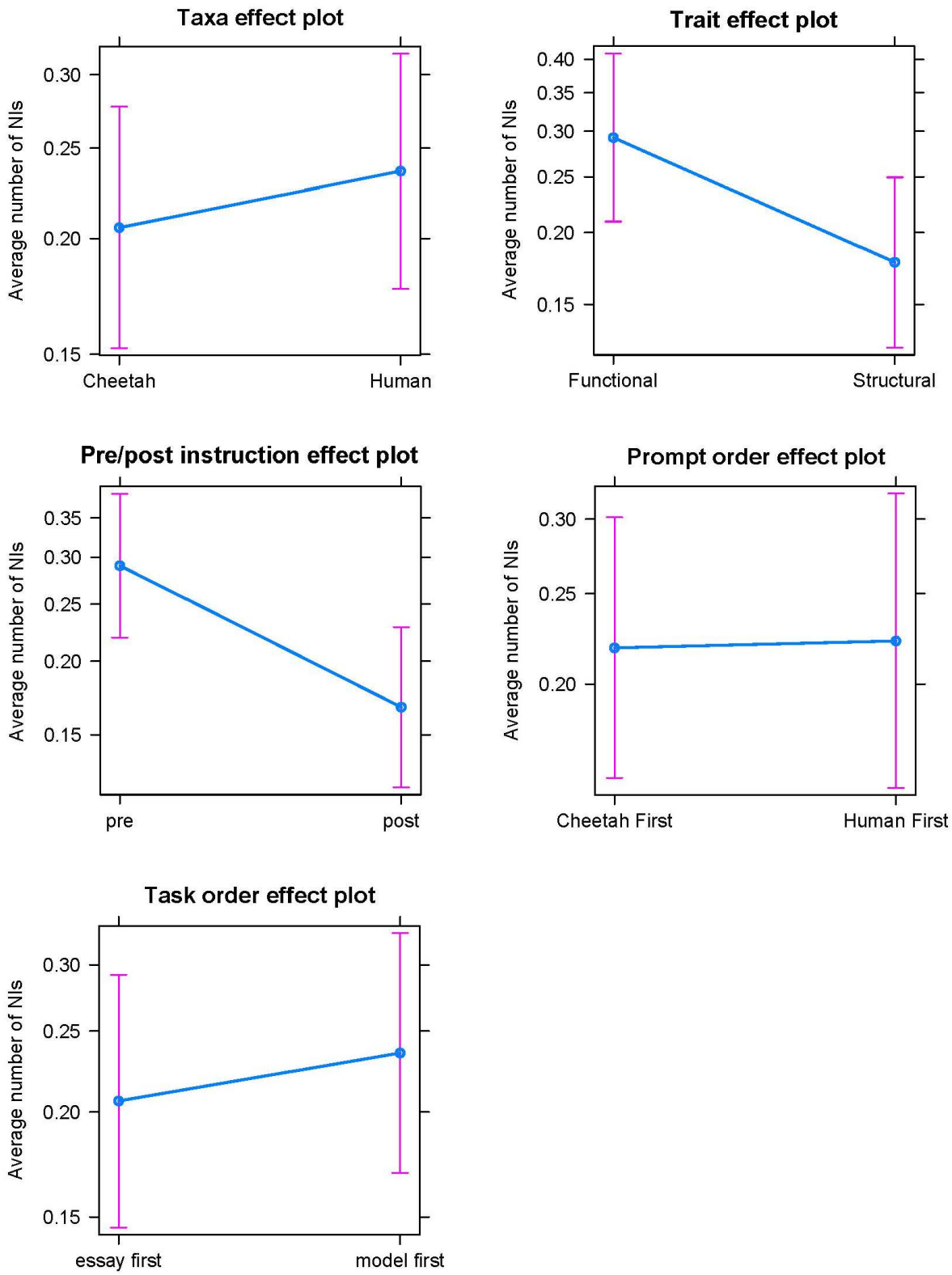


Figure S1.2. Plots showing the effects of taxon, trait type, prompt order, and pre/post instruction on the average number of NIs determined using a mixed-effects Poisson regression.

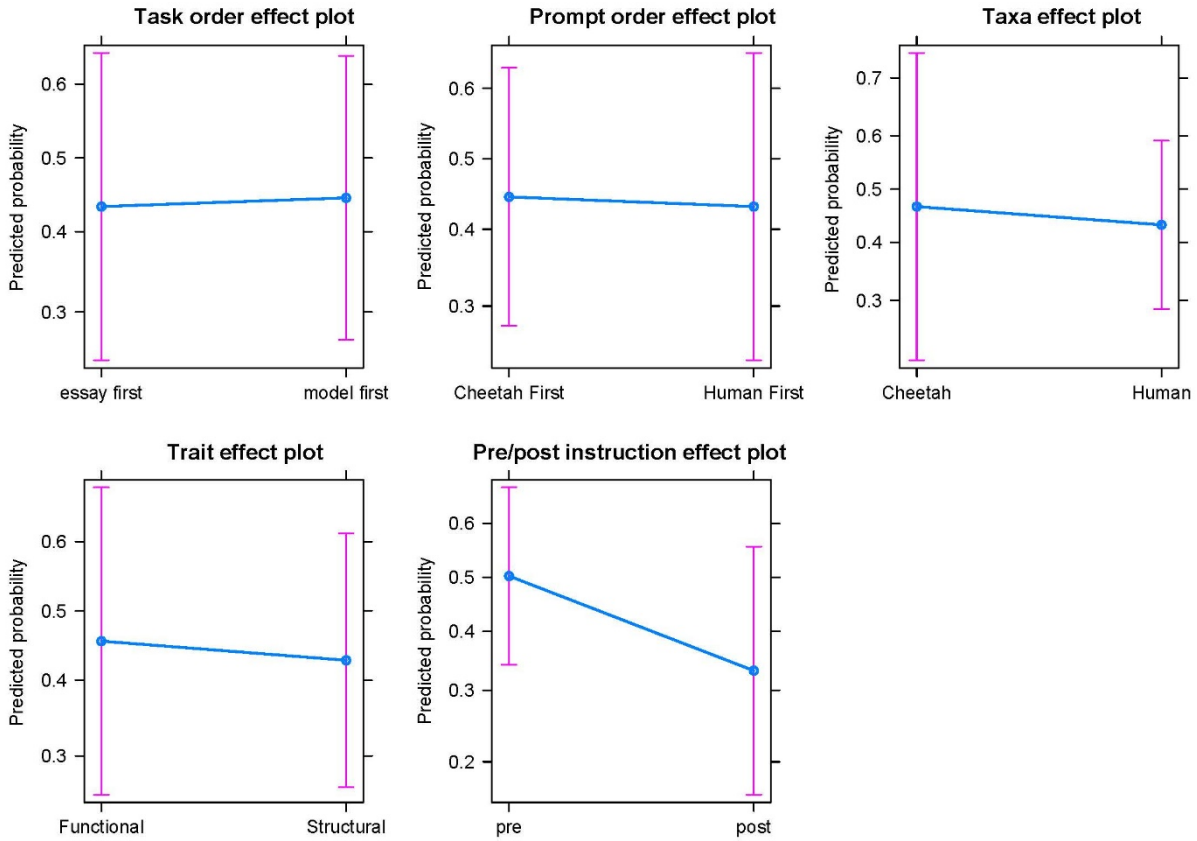


Figure S1.3. Predicted probabilities of being in the NI only group vs None group for each of the predictors in the multiple logistic regression model.

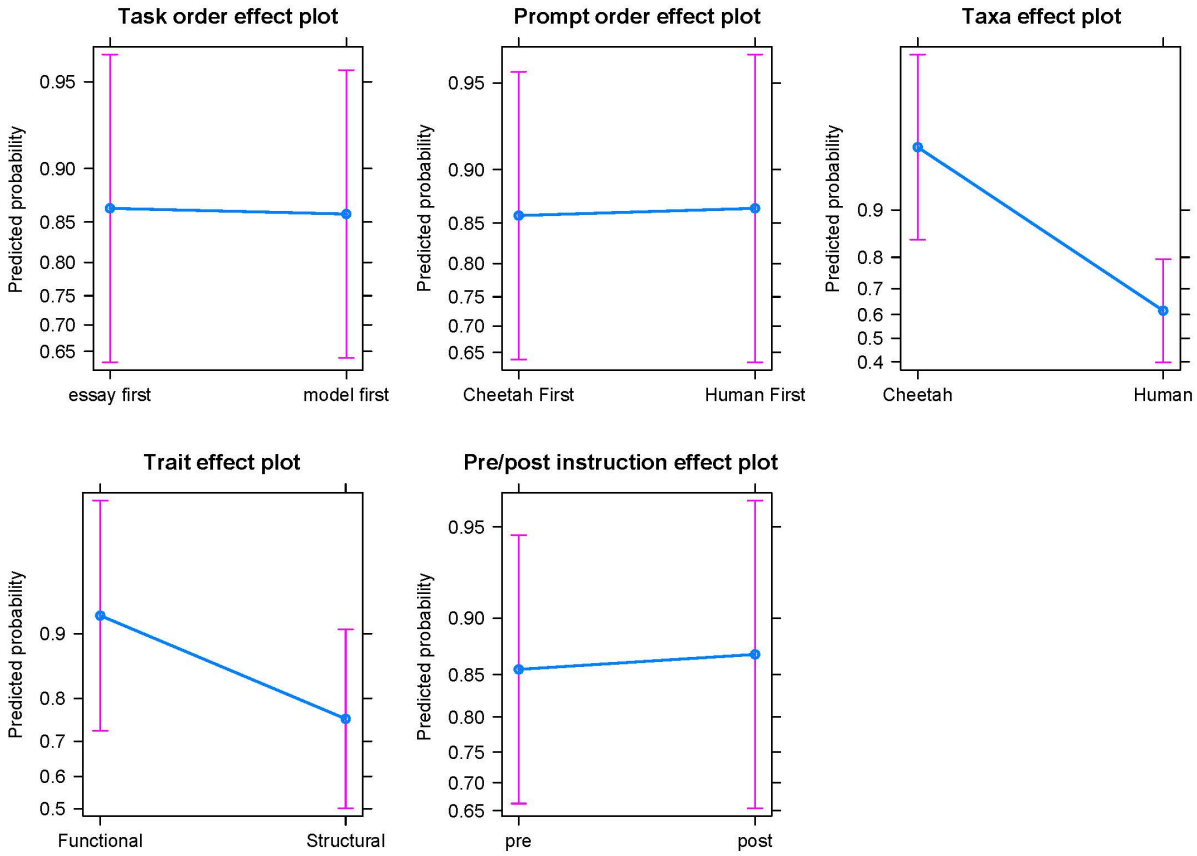


Figure S1.4. Predicted probabilities of being in the Mixed group vs None group for each of the predictors in the multiple logistic regression model.

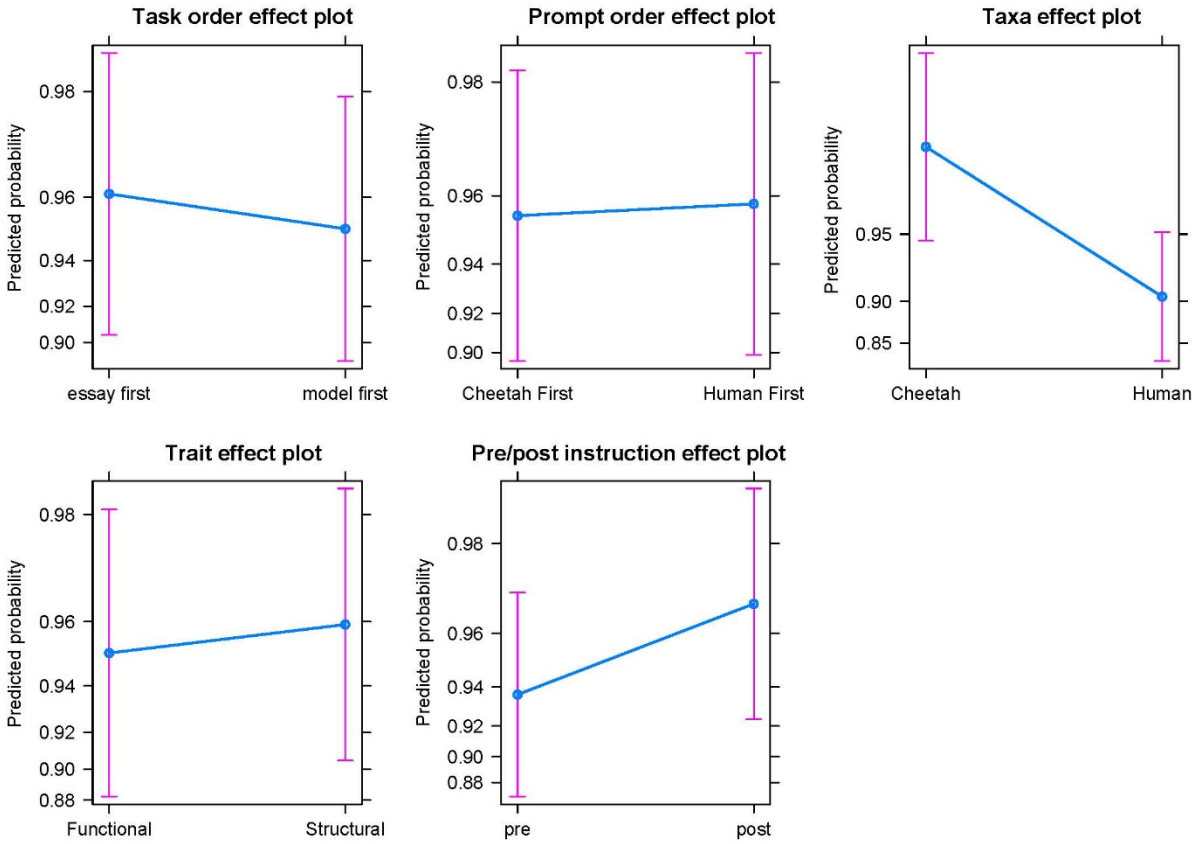


Figure S1.5. Predicted probabilities of being in the KC only group vs None group for each of the predictors in the multiple logistic regression model.

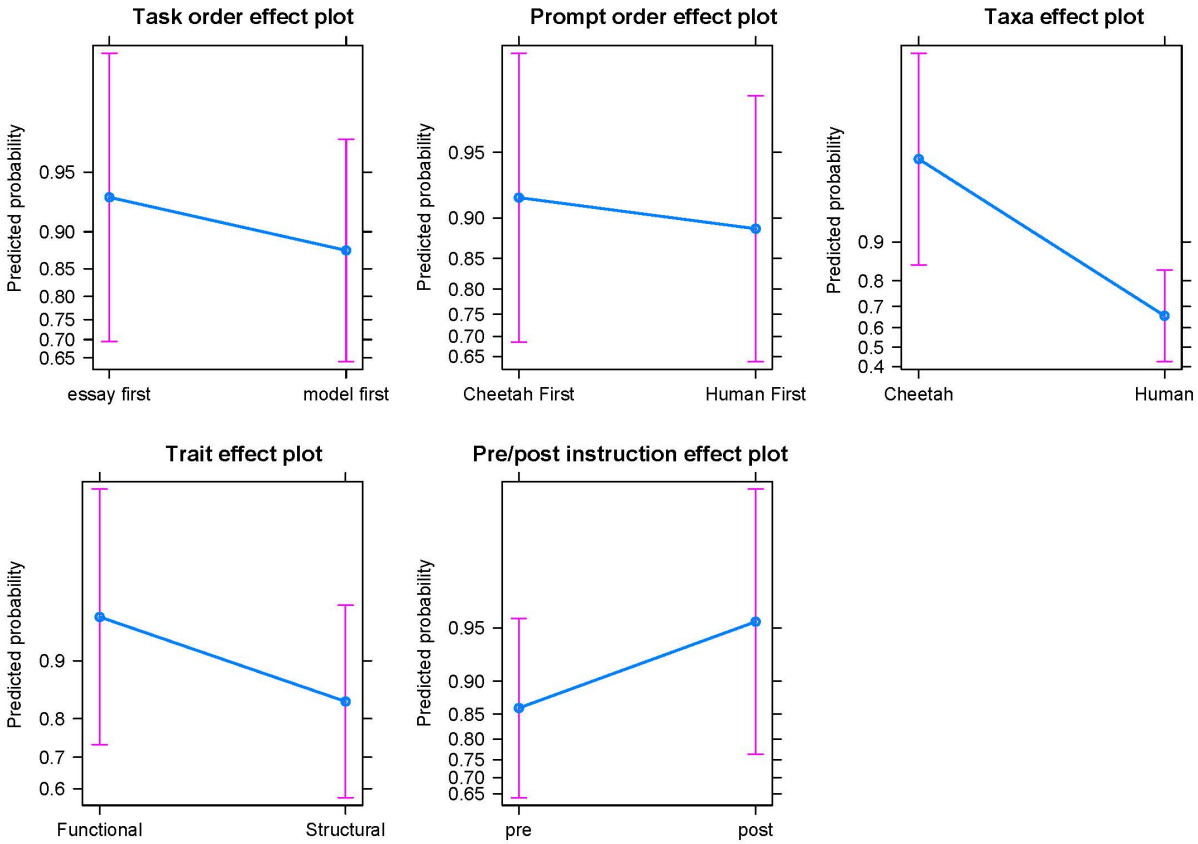


Figure S1.6. Predicted probabilities of being in the Mixed group vs NI only group for each of the predictors in the multiple logistic regression model.

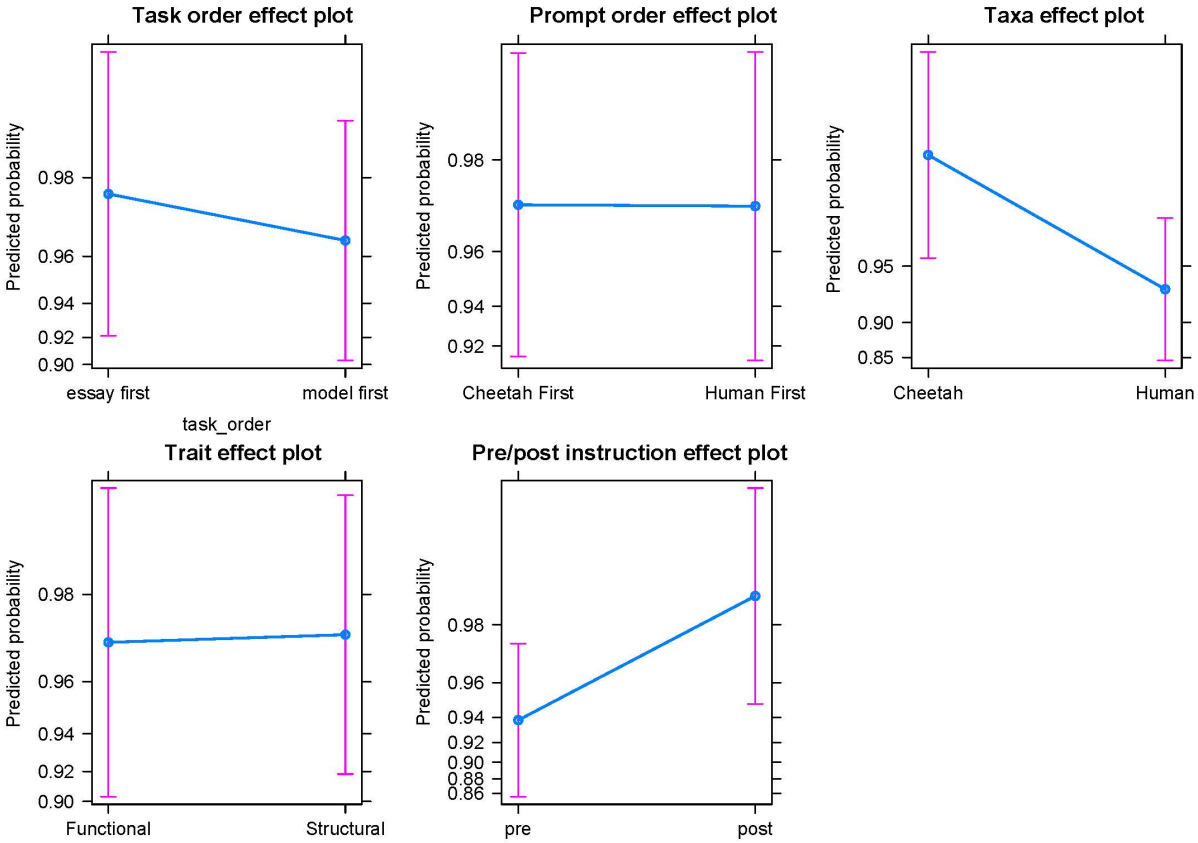


Figure S1.7. Predicted probabilities of being in the KC only group vs NI only group for each of the predictors in the multiple logistic regression model.

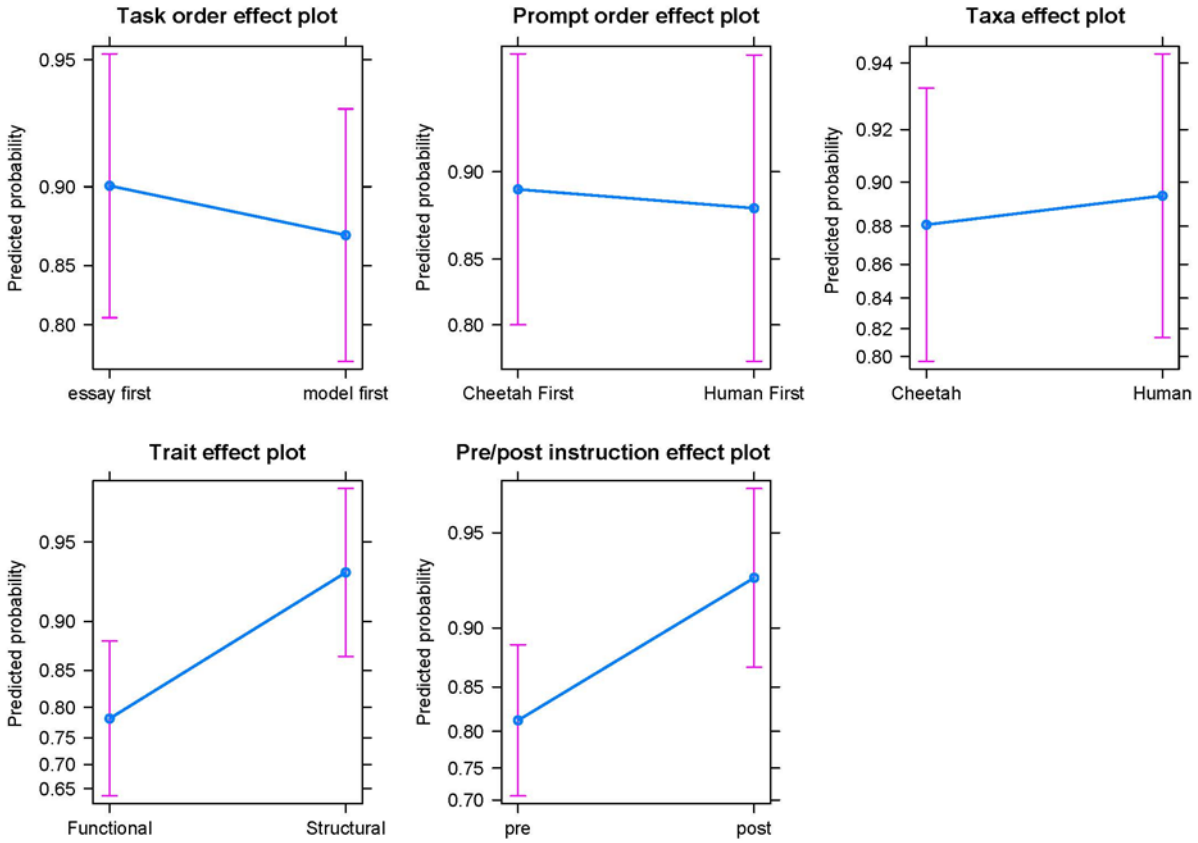


Figure S1.8. Predicted probabilities of being in the KC only group vs Mixed group for each of the predictors in the multiple logistic regression model.

REFERENCES

REFERENCES

Allmon, W. D. (2011). Why Don't People Think Evolution Is True? Implications for Teaching, in and out of the Classroom. *Evolution: Education and Outreach*, 4(4), 648–665. https://doi.org/10.1007/s12052-011-0371-0

Alters, B. J., & Alters, S. (2001). *Defending evolution in the classroom : a guide to the creation/evolution controversy*. Jones & Bartlett Publishers.

Alters, B. J., & Nelson, C. E. (2002). Perspective: Teaching Evolution in Higher Education. *Evolution*, 56(10), 1891–1901.

American Association for the Advancement of Science [AAAS]. (2011). *Vision and Change in Undergraduate Biology Education: a call to action.* http://visionandchange.org

American Association for the Advancement of Science [AAAS]. (2018). *Project 2061: Evolution and Natural Selection. AAAS Science Assessment.* http://assessment.aaas.org/topics/1/EN#/0

Atran, S. (1998). Folk biology and the anthropology of science: cognitive universals and cultural particulars*. Behavioral and Brain Sciences*, 21(4), 547–569.

Atran, S., Medin, D., Lynch, E., Vapnarsky, V., Ucan Ek', E., & Sousa, P. (2001). Folkbiology doesn't Come from Folkpsychology: Evidence from Yukatek Maya in Cross-Cultural Perspective. *Journal of Cognition and Culture*, 1(1), 3–42. https://doi.org/10.1163/156853701300063561

Bates, D., Maechler, M., Bolker, B., & Walker, S. (2015). Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software*, 67(1), 1–48. https://doi.org/10.18637/jss.v067.i01.

Beggrow, E. P., & Sbeglia, G. C. (2019). Do disciplinary contexts impact the learning of evolution? Assessing knowledge and misconceptions in anthropology and biology students. *Evolution: Education and Outreach*, 12(1). https://doi.org/10.1186/s12052-018-0094-6

Bishop, B. A., & Anderson, C. W. (1990). Students conceptions of natural selection and its role in evolution. *Journal of Research in Science Teaching*, 27(5), 415–427.

Blackwell, W. H., Powell, M. J., & Dukes, G. H. (2003). The problem of student

acceptance of evolution. *Journal of Biological Education*, 37(2), 58–67. https://doi.org/10.1080/00219266.2003.9655852

Boujaoude, S., Asghar, A., Wiles, J. R., Jaber, L., & Alters, B. (2011). Biology professors' and teachers' positions regarding biological evolution and evolution education in a Middle Eastern society. *International Journal of Science Educ*ation, 33(7), 979–1000. https://doi.org/10.1080/09500693.2010.489124

Bowman, K. L. (2008). The evolution battles in high-school science classes: who is teaching what ? *Frontiers in Ecology and the Environment*, 6(2), 69–74. https://doi.org/10.1890/070013

Bray Speth, E., Long, T. M., Pennock, R. T., & Ebert-May, D. (2009). Using Avida-ED for Teaching and Learning About Evolution in Undergraduate Introductory Biology Courses. *Evolution: Education and Outreach*, 2(3), 415–428. https://doi.org/10.1007/s12052-009-0154-z

Bray Speth, E., Shaw, N., Momsen, J., Reinagel, A., Le, P., Taqieddin, R., & Long, T. (2014). Introductory biology students' conceptual models and explanations of the origin of variation. *CBE Life Sciences Education*, 13(3), 529–539. https://doi.org/10.1187/cbe.14-02-0020

Brenan, M. (2019). *40% of Americans Believe in Creationism*. Gallup. https://news.gallup.com/poll/261680/americans-believe-creationism.aspx

Brooks, M. E., Kristensen, K., Benthem, K. J. van, Magnusson, A., Berg, C. W., Nielsen, A., Skaug, H. J., Bolker, M. M., & M., B. (2017). glmmTMB Balances Speed and Flexibility Among Packages for Zero-inflated Generalized Linear Mixed Modeling. *The R Journal*, 9(2), 378–400.

Brown, J. S., Collins, A., & Duguid, P. (2007). Situated Cognition and the Culture of Learning. *Educational Researcher*, 18(1), 32–42.

Catley, K. M., & Novick, L. R. (2009). Digging deep: Exploring college students' knowledge of macroevolutionary time. *Journal of Research in Science Teaching*, 46(3), 311–332. https://doi.org/10.1002/tea.20273

Cavallo, A. M. L., & McCall, D. (2008). Seeing May Not Mean Believing : Examining Students ' Understandings. *The American Biology Teacher*, 70(9), 522–531.

Chi, M. T., Feltovich, P. J., & Glaser, R. (1981). Categorization and Representation of Physics Problems by Experts and Novices. *Cognitive Science*, 5(2), 121–152.

Clark, D. B. (2006). Longitudinal conceptual change in students' understanding of thermal equilibrium: An examination of the process of conceptual restructuring. *Cognition and Instruction*, 24(4), 467–563. https://doi.org/10.1207/s1532690xci2404

Clough, E. E., & Driver, R. (1986). A study of consistency in the use of students' conceptual frameworks across different task contexts. *Science Education*, 70(4), 473–496. https://doi.org/10.1002/sce.3730700412

Coley, J. D. (2007). The Human Animal: Developmental Changes in Judgments of Taxonomic and Psychological Similarity Among Humans and Other Animals. *Cognition, Brain, Behavior*, 11(4), 733–756.

Coley, J. D., & Tanner, K. (2015). Relations between intuitive biological thinking and biological misconceptions in biology majors and nonmajors. *CBE Life Sciences Education*, 14(1), 1–19. https://doi.org/10.1187/cbe.14-06-0094

Cooper, M. M., Caballero, M. D., Ebert-May, D., Fata-Hartley, C. L., Jardeleza, S. E., Krajcik, J. S., Laverty, J. T., Matz, R. L., Posey, L. A., & Underwood, S. M. (2015). Challenge faculty to transform STEM learning. *Science*, 350(6258), 281–282. https://doi.org/10.1126/science.aab0933

Council of Europe. (2017). *The dangers of creationism in education*. Parliamentary Assembly Document Number 11375. http://www.assembly.coe.int/nw/xml/XRef/X2H-Xref-ViewHTML.asp?FileID=11751&lang=en

Darwin, C. (1859). On the Origin of the Species. In *J. Murray* (Vol. 5). https://doi.org/10.1016/S0262-4079(09)60380-8

Darwin, C. (1868). *The variation of animals and plants under domestication. Volume II.* John Murray.

Deniz, H., & Borgerding, L. A. (2018a). Evolution Education Around the Globe: Conclusions and Future Directions. In H. Deniz & L. A. Borgerding (Eds.), *Evolution Education Around the Globe* (pp. 449–464). Springer, Cham. https://doi.org/10.1007/978-3-319-90939-4

Deniz, H., & Borgerding, L. A. (2018b). Evolutionary Theory as a Controversial Topic in Science Curriculum Around the Globe. In H. Deniz & L. A. Borgerding (Eds.), *Evolution Education Around the Globe* (pp. 3–11). Springer, Cham. https://doi.org/10.1007/978-3-319-90939-4

diSessa, A. A., Gillespie, N. M., & Esterly, J. B. (2004). Coherence versus fragmentation in the development of the concept of force. *Cognitive Science*, 28(6), 843–900. https://doi.org/10.1016/j.cogsci.2004.05.003

Dobzhansky, T. (1973). Nothing in biology makes sense except in the light of evolution. *The American Biology Teacher*, 35(3), 125–129. https://doi.org/10.2307/4444260

Downie, J. R., & Barron, N. J. (2000). Evolution and religion: Attitudes of Scottish first year biology and medical students to the teaching of evolutionary biology. *Journal of Biological Education*, 34(3), 139–146. https://doi.org/10.1080/00219266.2000.9655704

Emmons, N. A., & Kelemen, D. A. (2015). Young children's acceptance of within-species variation: Implications for essentialism and teaching evolution. *Journal of Experimental Child Psychology*, 139, 148–160. https://doi.org/10.1016/j.jecp.2015.05.011

Evans, E. M. (2008). Conceptual change and evolutionary biology: A developmental analysis. In S. Vosniadou (Ed.), *International Handbook of research on conceptual change* (pp. 263–294).

Federer, M. R., Nehm, R. H., Opfer, J. E., & Pearl, D. (2015). Using a constructed-response instrument to explore the effects of item position and item features on the assessment of students' written scientific explanations. *Research in Science Education*, 45(4), 527–553. https://doi.org/10.1007/s11165-014-9435-9

Foddy, W. (1993). *Constructing questions for interviews and questionnaires: Theory and practice in social research*. Cambridge University Press.

Fox, J. (2003). Effect Displays in R for Generalised Linear Models. *Journal of Statistical Software*, 8(15), 1–27. http://www.jstatsoft.org/v08/i15/

Freeman, S., Eddy, S. L., McDonough, M., Smith, M. K., Okoroafor, N., Jordt, H., & Wenderoth, M. P. (2014). Active learning increases student performance in science, engineering, and mathematics. *Proceedings of the National Academy of Sciences*, 111(23), 8410–8415. https://doi.org/10.1073/pnas.1319030111

Glaze, A. L., & Goldston, M. J. (2015). Education U . S . Science Teaching and Learning of Evolution : A Critical Review of the Literature 2000 – 2014. *Science Education*, 99(3), 500–518. https://doi.org/10.1002/sce.21158

Gray, K. E. (2004). *The Effect of Question Order on Student Responses to Multiple*

Choice Physics Questions (Masters Thesis) [Kansas State University]. https://web.phys.ksu.edu/dissertations/MSThesis-KaraGray.pdf

Gregory, T. R. (2009). Understanding Natural Selection : Essential Concepts and Common Misconceptions. *Evolution Education Around the Globe*, 2(2), 156–175. https://doi.org/10.1007/s12052-009-0128-1

Gros, H., Sander, E., & Thibaut, J. (2019). When masters of abstraction run into a concrete wall : Experts failing arithmetic word problems. *Psychonomic Bulletin & Review*, 1–9.

Gross, P., Buttrey, D., Goodenough, U., Koertge, N., Lerner, L. S., Schwartz, M., & Schwartz, R. (2013). *Final Evaluation of the Next Generation Science Standards*.

Ha, M.-S., Lee, J.-K., & Cha, H.-Y. (2006). A Cross-Sectional Study of Students' Conceptions on Evolution and Characteristics of Concept Formation about It in Terms of the Subjects: Human, Animals and Plants. *Journal of the Korean Association for Science Education*, 26(7), 813–825.

Hall, R. (1996). Representation as Shared Activity : Situated Cognition and Dewey ' s Cartography of Experience. *The Journal of the Learning Sciences*, 5(3), 209–238.

Halldén, O. (1988). The evolution of the species : pupil perspectives and school perspectives. *International Journal of Science Education*, 10(5), 541–552.

Hambleton, R. K., & Traub, R. E. (1974). The Effects of Item Order on Test Performance and Stress. The Journal of Experimental Education, 43(1), 40–46.

Hamp-Lyons, L., & Mathias, S. P. (1994). Examining expert judgments of task difficulty on essay tests. *Journal of Second Language Writing*, 3(1), 49–68. https://doi.org/10.1016/1060-3743(94)90005-1

Hartig, F. (2018). *DHARMa: Residual Diagnostics for Hierarchical (Multi-Level / Mixed) Regression Models*. https://cran.r-project.org/package=DHARMa

Hmelo-Silver, C. E., & Pfeffer, M. G. (2004). Comparing expert and novice understanding of a complex system from the perspective of structures, behaviors, and functions. *Cognitive Science*, 28(1), 127–138. https://doi.org/10.1016/S0364-0213(03)00065-X

Hofer, B. K. (2006). Domain specificity of personal epistemology: Resolved questions, persistent issues, new models. *International Journal of Educational Research*,

45(1–2), 85–95. https://doi.org/10.1016/j.ijer.2006.08.006

Inagaki, K., & Hatano, G. (2006). Young Children's Conception of the Biological World. *Current Directions in Psychological Science*, 15(4), 177–181. https://doi.org/10.1111/J.1467-8721.2006.00431.X

Ingram, E. L., & Nelson, C. E. (2006). Relationship between achievement and students' acceptance of evolution or creation in an upper-level evolution course. *Journal of Research in Science Teaching*, 43(1), 7–24. https://doi.org/10.1002/tea.20093

Jones, M. G., Carter, G., & Rua, M. J. (2000). Exploring the development of conceptual ecologies: Communities of concepts related to convection and heat. *Journal of Research in Science Teaching*, 37(2), 139–159. https://doi.org/10.1002/(SICI)1098-2736(200002)37:2<139::AID-TEA4>3.0.CO;2-1

Kalinowski, S. T., Leonard, M. J., & Andrews, T. M. (2010). Nothing in Evolution Makes Sense Except in the Light of DNA. *CBE—Life Sciences Education*, 9(2), 87–97. https://doi.org/10.1187/cbe.09

Kampourakis, K., & Zogza, V. (2008). Students ' intuitive explanations of the causes of homologies and adaptations. *Science*, 17(1), 27–47. https://doi.org/10.1007/s11191-007-9075-9

Kampourakis, K., & Zogza, V. (2009). Preliminary evolutionary explanations: A Basic Framework for Conceptual Change and Explanatory Coherence in Evolution. *Science and Education*, 18(10), 1313–1340. https://doi.org/10.1007/s11191-008-9171-5

Kjolsing, E., & Van Den Einde, L. (2016). Peer Instruction: Using Isomorphic Questions to Document Learning Gains in a Small Statics Class. *Journal of Professional Issues in Engineering Education and Practice*, 142(4). https://doi.org/10.1061/(ASCE)EI.1943-5541.0000283

Krell, M., Reinisch, B., & Krüger, D. (2015). Analyzing Students' Understanding of Models and Modeling Referring to the Disciplines Biology, Chemistry, and Physics. *Research in Science Education*, 45(3), 367–393. https://doi.org/10.1007/s11165-014-9427-9

Krell, M., Upmeier zu Belzen, A., & Krüger, D. (2012). Students ' Understanding of the Purpose of Models in Different Biological Contexts. *International Journal of Biology Education*, 2(2), 1–34.

Kumar, S., Stecher, G., Suleski, M., & Hedges, S. B. (2017). TimeTree: A Resource for Timelines , Timetrees , and Divergence Times. *Molecular Biology and Evolution*, 34(7), 1812–1819. https://doi.org/10.1093/molbev/msx116

Lee, H. K., & Anderson, C. (2007). Validity and topic generality of a writing performance test. *Language Testing*, 24(3), 307–330. https://doi.org/10.1177/0265532207077200

Lerner, L. S. (2000). Good Science, Bad Science: *Teaching Evolution in the States*. Thomas B. Fordham Foundation.

Li, J. (2018). Establishing Comparability Across Writing Tasks With Picture Prompts of Three Alternate Tests. *Language Assessment Quarterly*, 15(4), 368–386. https://doi.org/10.1080/15434303.2017.1405422

MacNicol, K. (1956). Effects of varying order of item difficulty in an unspeeded verbal test. *Unpublished Manuscript, Educational Testing Service*, Princeton, NJ.

Mayr, E. (1982). *The Growth of Biological Thought*. The Belknap Press of Harvard University Press Cambridge.

Miller, J. D., Scott, E. C., & Okamoto, S. (2006). Public Acceptance of Evolution. *Science*, 313(5788), 765–766. https://doi.org/10.1126/science.1126746

Moharreri, K., Ha, M., & Nehm, R. H. (2014). EvoGrader: an online formative assessment tool for automatically evaluating written evolutionary explanations. *Evolution: Education and Outreach*, 7(15), 1–14. https://doi.org/10.1186/s12052-014-0015-2

Monk, J. J., & Stallings, W. M. (1970). Effects of item order on test scores. *The Journal of Educational Research*, 63(10), 463–465.

Morabito, N. P., Catley, K. M., & Novick, L. R. (2010). Reasoning about evolutionary history: Post-secondary students' knowledge of most recent common ancestry and homoplasy. *Journal of Biological Education*, 44(4), 166–174. https://doi.org/10.1080/00219266.2010.9656217

National Academies of Sciences, Engineering, and Medicine [NASEM]. (2016). *Science Literacy: Concepts, Contexts, and Consequences*. The National Academies Press. https://doi.org/10.17226/23595

National Research Council [NRC]. (2000). How Experts Differ from Novices People. In

How People Learn: Brain, Mind, Experience, and School: Expanded Edition. National Academies Press. https://doi.org/10.17226/9853

National Research Council [NRC]. (2009). *Learning Science in Informal Environments: People, Places, and Pursuits.* The National Academies Press.

Nehm, R. H. (2006). Faith-based Evolution Education? *BioScience*, 56(8), 638–639.

Nehm, R. H., Beggrow, E. P., Opfer, J. E., & Ha, M. (2012). Reasoning About Natural Selection: Diagnosing Contextual Competency Using the ACORNS Instrument. *The American Biology Teacher*, 74(2), 92–98. https://doi.org/10.1525/abt.2012.74.2.6

Nehm, R. H., & Ha, M. (2011). Item feature effects in evolution assessment. Journal of Research in Science Teaching, 48(3), 237–256. https://doi.org/10.1002/tea.20400

Nehm, R. H., Poole, T. M., Lyford, M. E., Hoskins, S. G., Carruth, L., Ewers, B. E., & Colberg, P. J. S. (2009). Does the Segregation of Evolution in Biology Textbooks and Introductory Courses Reinforce Students ' Faulty Mental Models of Biology and Evolution? *Evolution: Education and Outreach*, 2(3), 527–532. https://doi.org/10.1007/s12052-008-0100-5

Nehm, R. H., & Reilly, L. (2007). Biology majors' knowledge and misconceptions of natural selection. *Bioscience*, 57(3), 263–272. https://doi.org/10.1641/b570311

Nehm, R. H., & Ridgway, J. (2011). What Do Experts and Novices "See" in Evolutionary Problems? *Evolution: Education and Outreach*, 4(4), 666–679. https://doi.org/10.1007/s12052-011-0369-7

Nehm, R. H., & Schonfeld, I. S. (2007). Does increasing biology teacher knowledge of evolution and the nature of science lead to greater preference for the teaching of evolution in schools? *Journal of Science Teacher Education*, 18(5), 699–723. https://doi.org/10.1007/s10972-007-9062-7

Nehm, R. H., & Schonfeld, I. S. (2008). Measuring knowledge of natural selection: A comparison of the CINS, an open-response instrument, and an oral interview. *Journal of Research in Science Teaching*, 45(10), 1131–1160. https://doi.org/10.1002/tea.20251

Nettle, D. (2010). Understanding of Evolution May Be Improved by Thinking about People. *Evolutionary Psychology*, 8(2), 205–228.

NGSS Lead States. (2013). *Next Generation Science Standards: For States, By States*.

https://doi.org/10.17226/18290

Oliveira, A. W., & Cook, K. L. (2018). Evolution Education and the Rise of the Creationist Movement in Brazil. In H. Deniz & L. A. Borgerding (Eds*.), Evolution Education Around the Globe* (pp. 119–136). Springer, Cham. https://doi.org/10.1007/978-3-319-90939-4

Ozdemir, G., & Clark, D. (2009). Knowledge structure coherence in Turkish students' understanding of force. *Journal of Research in Science Teaching*, 46(5), 570–596. https://doi.org/10.1002/tea.20290

Pobiner, B. (2016). Accepting, understanding, teaching, and learning (human) evolution: Obstacles and opportunities. *American Journal of Physical Anthropology*, 159, 232–274. https://doi.org/10.1002/ajpa.22910

Pobiner, B., Beardsley, P. M., Bertka, C. M., & Watson, W. A. (2018). Using human case studies to teach evolution in high school A.P. biology classrooms. *Evolution: Education and Outreach*, 11(1). https://doi.org/10.1186/s12052-018-0077-7

Potari, D., & Spiliotopoulou, V. (1996). Children's approaches to the concept of volume. *Science Education*, 80(3), 341–360. https://doi.org/10.1002/(SICI)1098-237X(199606)80:3<341::AID-SCE4>3.0.CO;2-E

Prevost, L. B., Knight, J., Smith, M. K., & Urban-Lurain, M. (2013). Student writing reveals their heterogeneous thinking about the origin of genetic variation in populations. *National Association for Research in Science Teaching*.

R Core Team. (2020). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. https://www.r-project.org/

Sabella, M. S., & Redish, E. F. (2007). Knowledge organization and activation in physics problem solving. *American Journal of Physics*, 75(11), 1017–1029. https://doi.org/10.1119/1.2746359

Schuman, H., & Presser, S. (1996). *Questions and answers in attitude surveys: Experiments on question form, wording, and context*. SAGE Publications, Inc.

Schurmeier, K. D., Atwood, C. H., Shepler, C. G., & Lautenschlager, G. J. (2010). Using item response theory to assess changes in student performance based on changes in question wording. *Journal of Chemical Education*, 87(11), 1268–1272. https://doi.org/10.1021/ed100422c

Shtulman, A. (2006). Qualitative differences between naïve and scientific theories of evolution. *Cognitive Psychology*, 52(2), 170–194. https://doi.org/10.1016/j.cogpsych.2005.10.001

Shtulman, A., & Schulz, L. (2008). The Relation Between Essentialist Beliefs and Evolutionary Reasoning. *Cognitive Science*, 32(8), 1049–1062. https://doi.org/10.1080/03640210801897864

Shtulman, A., & Valcarcel, J. (2012). Scientific knowledge suppresses but does not supplant earlier intuitions. *Cognition*, 124(2), 209–215. https://doi.org/10.1016/j.cognition.2012.04.005

Sinatra, G. M., Brem, S. K., & Evans, E. M. (2008). Changing Minds? Implications of Conceptual Change for Teaching and Learning about Biological Evolution. *Evolution: Education and Outreach*, 1(2), 189–195. https://doi.org/10.1007/s12052-008-0037-8

Sinatra, G. M., Southerland, S. A., McConaughy, F., & Demastes, J. W. (2003). Intentions and beliefs in students' understanding and acceptance of biological evolution. *Journal of Research in Science Teaching*, 40(5), 510–528. https://doi.org/10.1002/tea.10087

Smith, M. U. (2010a). Current status of research in teaching and learning evolution: I. Philosophical/epistemological issues. *Science and Education*, 19(6–8), 523–538. https://doi.org/10.1007/s11191-009-9215-5

Smith, M. U. (2010b). Current status of research in teaching and learning evolution: II. Pedagogical issues. *Science and Education*, 19(6–8), 539–571. https://doi.org/10.1007/s11191-009-9216-4

Sydorenko, T. (2011). Item writer judgments of item difficulty versus actual item difficulty: A case study. *Language Assessment Quarterly*, 8(1), 34–52. https://doi.org/10.1080/15434303.2010.536924

Thagard, P., & Findlay, S. (2010). Getting to Darwin: Obstacles to accepting evolution by natural selection. *Science and Education*, 19(6–8), 625–636. https://doi.org/10.1007/s11191-009-9204-8

The Carnegie Classification of Institutions of Higher Education. (n.d.). Retrieved March 23, 2018, from http://carnegieclassifications.iu.edu/

UK Department of Education. (2015). *National curriculum in England: science*

programmes of study. https://www.gov.uk/government/publications/national-curriculum-in-england-science-programmes-of-study/national-curriculum-in-england-science-programmes-of-study

Van Oers, B. (1998). The Fallacy of Decontextualization. *Mind, Culture & Activity*, 5(2), 135–142. https://doi.org/10.1207/s15327884mca0502

Vazquez, B. (2017). A state-by-state comparison of middle school science standards on evolution in the United States. *Evolution: Education and Outreach*, 10(1). https://doi.org/10.1186/s12052-017-0066-2

West, S. A., El Mouden, C., & Gardner, A. (2011). Sixteen common misconceptions about the evolution of cooperation in humans. *Evolution and Human Behavior*, 32(4), 231–262. https://doi.org/10.1016/j.evolhumbehav.2010.08.001

Weston, M., Haudek, K. C., Prevost, L., Urban-lurain, M., & Merrill, J. (2015). Examining the Impact of Question Surface Features on Students ' Answers to Constructed-Response Questions on Photosynthesis. *CBE—Life Sciences Education*, 14(2), ar19. https://doi.org/10.1187/cbe.14-07-0110

Wickham, H., François, R., Henry, L., & Müller, K. (2020). *dplyr: A Grammar of Data Manipulation*. R package version 0.8.5. https://cran.r-project.org/package=dplyr

Wickham, H., & Henry, L. (2020). *tidyr: Tidy Messy Data*. R package version 1.0.3. https://cran.r-project.org/package=tidyr

CHAPTER TWO:

Item-feature context influences the content and architecture of student-constructed models.

ABSTRACT

Scientific models are specialised external representations that explain or predict a concept, process, or phenomenon. They lend themselves to both authentic instruction and assessment. Student-constructed models are partial representations of their mental models and can give us insights into student thinking and reasoning that are not captured in multiple choice or even narrative responses. Such externalised representations are particularly valuable in gauging students' knowledge and understanding of complex biological phenomena. Additionally, features of model architecture can provide insights into aspects of students' cognitive structures (CSs), such as size and complexity.

In this study, we ask whether item-feature context (i.e., variables in a question prompt) impacts the content and network architecture of students' constructed models of evolution by natural selection. Students in two large (n=384) introductory biology courses were asked to construct models to explain the evolution of traits in two taxa – humans and cheetahs. We coded the model content for the presence/absence of evolutionary ideas and quantified model architecture using network metrics for each model. Model content and architecture were analysed to determine contextual effects.

We also tested for association between prior academic performances and contextual effects.

We found that taxon influenced the content of student-constructed models. Cheetah models were more likely to have key evolutionary concepts and fewer naïve ideas as compared to Human models. Taxon also influenced the architecture of the models - Cheetah models were larger and more complex than the Human models. Prior academic performance (measured by GPA) was a predictor of model content, architecture, and contextual susceptibility.

Our results indicate that contextual features of the prompt are eliciting differences in students' models. This could indicate that students are either using surface cues to access their cognitive structures and build their mental models, or that they have a piecemeal understanding of evolution which results in a non-robust cognitive structure. Decreased susceptibility to context with increasing GPA indicates a progression from novice to expert with respect to both modelling and evolutionary knowledge.

INTRODUCTION

Models in Science

Models are a "strategy for coping with an extraordinarily complex world" (Odenbaugh, 2005). They are simplifications or abstractions that can be used to facilitate reasoning or communicate specific information about an entity (Gilbert, 2004; Seel, 2003; Wilson et al., 2019). An ideal model is an accurate, parsimonious, and coherent representation of

an entity, while at the same time, is also as general and useful as possible (Constantinou et al., 2019; White et al., 2011). Scientific models are specialised external representations of entities (concepts, processes, phenomena, or systems), which help illustrate, explain, or make predictions (Constantinou et al., 2019; Lee et al., 2017; Osbeck & Nersessian, 2006; Schwarz et al., 2009). Such models are routinely used by scientists to generate, evaluate, and communicate science (Gilbert, 2004; Halloun, 2007; Long et al., 2014; Upmeier zu Belzen et al., 2019a). Indeed, the inextricable nature of modelling with the practice of science was captured in Gilbert's (1991) definition of science as a “process of constructing predictive conceptual models”.

Models in Science Education

For some time now, there has been increasing desire and pressure to make the teaching and learning of science more reflective of the way science is practised (Barab, Hay, Barnett, & Keating, 2000; Cooper et al., 2015; Duschl & Gitomer, 1997; Gilbert, 2004; National Research Council [NRC], 1996, p 214; White, 1993). This involves engaging students in scientific skills and practises such as experimental design, quantitative reasoning, and modelling as part of instruction (American Association for the Advancement of Science [AAAS], 2011). Science educators and education researchers agree that models and modelling are not only of great importance in science education but should be considered a required skill in the development of scientific literacy (Achér et al., 2007; Coll et al., 2005; Gilbert et al., 2000; Halloun, 2007; van Driel et al., 2019). Recognising the integral nature of models and modelling in science, and thus the requirement that they become similarly integral to science

pedagogy, models have been included in the standards and required curricula for science at K-12 and university levels in multiple countries (In America: AAAS 2011; NRC, 2012; NGSS Lead States, 2013; in Germany: Sekretariat der Ständigen Konferenz der Kultusminister der Länder in der Bundesrepublik Deutschland [KMK], 2005b, 2005c, 2005a; in the United Kingdom: UK Department of Education, 2015; in Australia: Australian Curriculum Assessment and Reporting Authority [ACARA], 2010)

Several researchers have proposed best practises and described courses that use models and modelling either as a component of, or as a framework for instruction (Achér et al., 2007; Bennett et al., 2020; Bryce et al., 2016; J. J. Clement & Rea-Ramirez., 2008; Constantinou et al., 2019; Hung, 2008; Liu & Hmelo-Silver, 2009; Long et al., 2014; Schwarz et al., 2009). As a tool, models lend themselves to assessment of more than just students' knowledge and understanding of course material (Bray Speth et al., 2014; Hay et al., 2008; Long et al., 2014; Odenbaugh, 2005). For example, models and modelling have been used to assess students' systems thinking (Ben-Zvi Assaraf & Orion, 2005; Bergan-Roller et al., 2018; Hmelo-Silver et al., 2017; Hung, 2008; Tripto et al., 2013), long term retention (Dauer & Long, 2015) and understanding of the 'nature of science' (Boulter & Buckley, 2000; Cheng et al., 2015; Krell et al., 2014; Schwarz, 2002). Additionally, several researchers have shown benefits gained by students when instruction and assessment use a model-based approach (Baze & Gray, 2018; Bierema et al., 2017; Dauer et al., 2013; Louca & Zacharia, 2012; Vattam et al., 2011; Windschitl et al., 2008). A meta-analysis on the effects of active and constant engagement with models in class showed that there were associated gains in

knowledge retention (Nesbit & Adesope, 2006). In all the studies mentioned above, students did not passively view or memorize provided models, but rather actively engaged with the process of modelling. Researchers have argued that having students construct, evaluate, and modify their own models are desirable ways of engaging students in modelling beyond merely reasoning with provided models (Clement, 1989; Gilbert, 2004; Windschitl et al., 2008).

What Can We Learn from Student-Constructed Models?

Student-constructed models are one of the modes by which we can get insight into students' cognitive structures (CSs) (Kinchin et al., 2000; Nesbit & Adesope, 2006). A cognitive structure is a mental framework that serves as a way to store and connect information about concepts (Ausubel, 1963; Ifenthaler, 2011; Ifenthaler et al., 2011; Shavelson, 1974). We add to and modify our CS as we learn and make new connections (Rumelhart & Norman, 1978). As we progress from novice to expert, we increase both the size and the connections in our CSs (Ifenthaler et al., 2011).

When students are asked to construct a model, their working memory accesses the CS in their long term memory using a context that they find relevant and builds a mental model (Dauer et al., 2013; Shell et al., 2010). Student-constructed externalized models are then partial representations of their mental models, which in turn, are products of their CSs (Dauer et al., 2013; Ifenthaler et al., 2011; Seel, 2003). By extension, an external model constructed by a student can be considered a limited representation of their CS (Greca & Moreira, 2000; Hay et al., 2008; Ifenthaler, 2010). As such, student-

constructed external models can serve as a source of insight into student thinking, particularly in relation to how students are connecting concepts and how their thinking changes over time (Dauer et al., 2013; Dauer & Long, 2015; Hmelo-Silver et al., 2007; Ifenthaler et al., 2011). Externalised representations are especially useful when eliciting biological phenomena that can be conceptually, spatially, and temporally complex (Brandstädter et al., 2012; Dauer et al., 2013; Kapteijn, 1990).

Structure-Behaviour-Function (SBF) Models

One type of model that is used in instruction and assessments is the Structure-Behaviour-Function (SBF) Model. It is based on the SBF Theory (Goel & Stroulia, 1996), which originated as way of describing the functions of complex systems as an outcome of the components (structures) of the system and their interactions (behaviours). The SBF framework synthesises three major components of systems: elements, interconnections, and functions (Arnold & Wade, 2015; Meadows, 2008). In SBF models (Fig. 2.1), the *structures* (components, concepts) of a system are put in boxes. Pairs of structures are linked with arrows describing a *behaviour* (relationship) that links them. The model (all the structures and behaviours) is designed with a particular purpose in mind and therefore has a specific *function*.

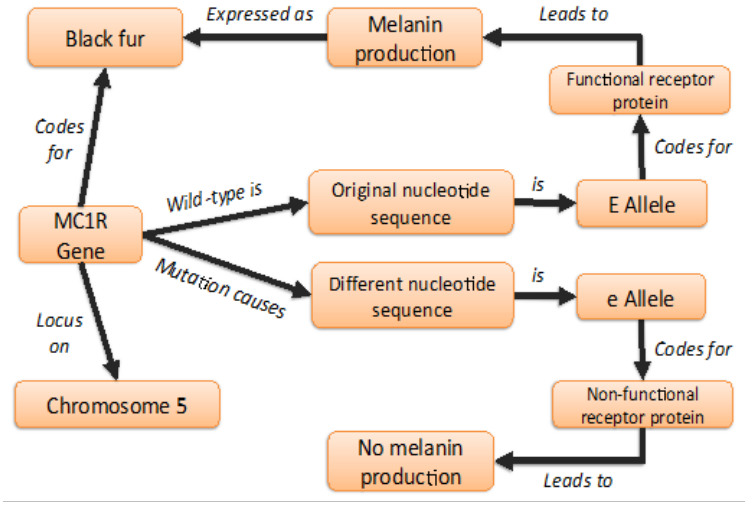


Figure 2.1. SBF model of the melanin system responsible for determining hair colour in mammals constructed by a student. *Structures (boxes) are linked by behaviours (on arrows) that describe relationships. This model was constructed in response to a prompt that required students to build a model that conveyed two functions: (a) cause of genetic variation, and (b) the consequence of genetic variation i.e. phenotypic variation.*

Such models are conducive to representing complex systems, particularly in biological science, and have been used both in instruction and assessment (Dauer et al., 2013; Hmelo-Silver et al., 2007; Lira & Gardner, 2016; Liu & Hmelo-Silver, 2009; Vattam et al., 2011; Wilson et al., 2019). SBF models are similar to concept maps (both consist of components and connections between the components and both are usually constructed as boxes with linking arrows) but have some additional constraints and affordances: they have to have a function (Sommer & Lücken, 2010), they do not have to represent everything the student knows (Jonassen et al., 2005), and do not have to be exclusively hierarchical (Hmelo-Silver & Azevedo, 2006).

SBF models have been used by researchers to gain insights into students CSs. Changes in student-constructed SBF models over the course of a semester have been

used to understand changes to students' CSs as a result of instruction in biology (Dauer et al., 2013). Students' SBF models have been used to investigate change in their understanding of the origin and maintenance of genetic and phenotypic variation (Bray Speth et al., 2014). SBF models have also been used to explore differences between experts and novices (Hmelo-Silver et al., 2007) and specifically to make claims about the differences in their CSs (Hmelo-Silver & Pfeffer, 2004). Ifenthaler et al. (2011) used models similar to SBF to make claims about the size and complexity of students' CS. Additionally, student-constructed SBF models have been used to evaluate long-term conceptual retrieval in students (Dauer & Long, 2015), and to characterize students' deep and surface approaches to modelling (Bennett et al., 2020).

Student-constructed models, such as SBF models, are considered partial representations of their CSs (Dauer et al., 2013; Greca & Moreira, 2000), and therefore represent modes by which we can assess students' mental models (Brandstädter et al., 2012; Evagorou et al., 2009; Ruiz-Primo & Shavelson, 1996; Shute & Zapata-Rivera, 2008). Williams & Hollan (1981) describe the process students use when asked to retrieve knowledge from their CS as progressing from first identifying a context for the desired knowledge, to searching their CS for knowledge related to that context, and then verifying the match between the knowledge possessed with the knowledge requested. Students use cues from question prompts or other instructional contexts as an access point for retrieving information; therefore, contextual cues narrow and focus a student's search within their CS (Dauer & Long, 2015; Reiser et al., 1985; Williams & Hollan, 1981). When students build models that draw from knowledge in their CS, they

explicitly represent and link ideas in an effort to represent the contents of their CS (Nesbit & Adesope, 2006) based on the context they used to access it (Williams & Hollan, 1981). Models may therefore provide insights into both the structure of a student's CS as well as the sensitivity of that structure to different contextual cues.

Contextual Influences and Models

Context plays an important role not just in the construction of knowledge and integration of new information into the CS, but also in the elicitation of that knowledge (Brown et al., 2007; Hall, 1996; Jones et al., 2000; Williams & Hollan, 1981). Previous work on how context affects modelling has explored students' epistemological understandings of modelling using models that were provided to the students (Gobert et al., 2011; Krell et al., 2012, 2015, 2014). Students' explanations about the purpose of models in biology varied across contexts and also varied when presented with contextualised vs decontextualised models (Krell et al., 2012, 2014). Schwarz, (2002) reported that aspects of meta-modelling knowledge also varied with context. Bennett et al. (2020) showed that students' approaches to modelling were not necessarily unique to the student, but dependent on contextual cues from modelling prompts.

In biology instruction, 'context' can be defined in multiple ways. For example, context can be defined by scientific domain (Gobert et al., 2011; Kohn et al., 2018; Krell et al., 2015), type of biological model (e.g., a mathematical curve vs functional model vs a computer simulation) (Krell et al., 2014), biological system (Bennett et al., 2020), or as specific words or examples used in the question prompt (i.e., item-feature) (Krell et al.,

2015; Nehm & Ha, 2011). There have been calls to further understand the effect of context on modelling tasks. For example, van Driel et al. (2019) stated a need to understand the effect of task context with respect to the interaction between task difficulty and task completion. Dauer & Long (2015) called for studies that explore the influence of item-feature context on students' model-based responses and ability to retrieve information from their CS over short- and long-term time frames. However, there have been no controlled experimental studies that explicitly test for the effect of item-feature context on student-*constructed* models.

Contextual effects on the content of student-constructed models

Multiple studies have analysed the effects of item-feature context on students' narrative data (written or through interviews). Many studies have shown that students find it difficult to identify the key principle or content being tested and are easily distracted or influenced by irrelevant contextual information provided in question prompts (diSessa et al., 2004; Nehm & Ha, 2011; Ozdemir & Clark, 2009; Prevost et al., 2013). In STEM, the effect of prompt context on the content of responses has been shown across expertise levels (Gross et al., 2013) and domains, including physics (Chi et al., 1981; diSessa et al., 2004), chemistry (Schurmeier et al., 2010), and biology (Bennett et al., 2020; Göransson et al., 2020; Nehm & Ha, 2011; Prevost et al., 2013). In our previous work, we saw that changes in taxon (humans vs cheetahs) and trait (functional vs structural) used in prompts elicited differences in the number of key evolutionary concepts and naïve ideas in students' narrative responses when responding to questions about evolution by natural selection Studies have shown that students' find it difficult to accept

human evolution (Atran, 1998; Atran et al., 2001; Nettle, 2010) and acceptance of evolution increases with increasing evolutionary distance from humans (Evans, 2008; Sinatra et al., 2003). In our previous study, we found large contextual influences based on the taxon in the prompt even though the evolutionary distance between humans and cheetahs is relatively small (diverged approximately 96 MYA (Kumar et al., 2017)).

Like narrative responses, the content of models can reveal the extent of a student's knowledge about a particular concept (Ben-Zvi Assaraf & Orion, 2010; Brandstädter et al., 2012; Bray Speth et al., 2014; Dauer et al., 2013; Dauer & Long, 2015; Ruiz-Primo & Shavelson, 1996). We can therefore use the differences in the content of students' models constructed in response to different contexts to determine the extent to which context affects retrieval of conceptual information.

Contextual effects on the architecture of student-constructed models

A model is a representation of not only conceptual knowledge, but also the underlying architectural organisation (structure) of that knowledge (Halford, 1993; Ifenthaler et al., 2011). When stripped of their content, SBF models show the network that students conceptualize while constructing a model (Fig. 2.2). Just like a graph, this network has vertices (structures, boxes) and edges (behaviours, arrows). The combination of vertices and edges gives the network a distinctive architecture which approximates the architecture of the student's mental model that they are attempting to externalize.

Researchers have used architectural metrics that are commonly used to analyse and evaluate networks and graphs, like size and complexity, to quantify and compare the architecture of student-constructed models and make claims about the architecture of students' mental model (Ifenthaler, 2008; Ifenthaler et al., 2007; Shute & Zapata-Rivera, 2008). Ifenthaler et al. (2011) quantified the numbers of vertices (structures) and edges (behaviours) in a model and used this metric as an indicator of size of the CSs that were accessed while building the respective models. They estimated development in the size and development of students' CSs by measuring changes in the number of vertices and propositions in their models. They also used model complexity, connectedness, and ruggedness (number of unlinked subgraphs) as proxies for the depth of understanding where higher connectedness and complexity and lower ruggedness indicated greater conceptual depth. Ifenthaler (2011) reported structural differences in student-constructed causal maps in different disciplines – which he called "externalised cognitive structures" – and interpreted this to mean that there were differences in their internal CS. Hmelo-Silver & Pfeffer (2004) used differences in SBF model architecture (number of structures and behaviours) to demonstrate differences in expert/novice understanding about a biological system and showed that while experts had a more functional understanding of the system (meaning, they used the function of the system to describe/list the structures), novices did not tend to progress beyond listing the structures and simple relationships. Dauer et al. (2013) used changes in model complexity (web-like causality index; Plate, 2010) to illustrate changes in students' CS as they added to and fine-tuned their Gene-to-Evolution models over a semester of instruction.

Association of prior academic performance on the content and architecture of student-constructed models.

Work by a number of researchers has shown that has shown that students' prior academic performance has a strong impact on current academic performance (Brookhart, 1997; Cassidy, 2012; Elias & MacDonald, 2007) and is often the strongest or only predictor of future performance (Casillas et al., 2012; Haak et al., 2011; Spinath et al., 2006). Although previous research has similarly shown an influence of prior achievement on the way students engage with modelling tasks, prior achievement does not appear to have the same strength in predicting model-based performance that it does in more traditional assessment contexts. Instead, interactions between prior achievement and model-based performance are more complex and less well established. Bennett et al. (2020) found that although prior performance is associated with the way students approach modelling tasks, it was not a reliable predictor. Dauer et al. (2013) showed that after a semester of modelling-based instruction, the highest relative gains on modelling tasks were seen in the lowest-achieving students. In a follow-up study that assessed long-term knowledge retention in the same student population, Dauer & Long (2015) showed mid-achieving students tended to outperform their peers on model-based tasks that required knowledge retrieval from a course they had taken 2.5 years previously. Model-based instruction and assessment have therefore been advocated as potential strategies for reducing achievement gaps and engaging students who tend to underperform on standard and rote assessments (Bierema et al., 2017; Manthey & Brewe, 2013; Reinagel & Bray Speth, 2016; Verhoeff et al., 2008).

Research Questions:

Prior studies have used the content and the architecture of student-constructed models to make claims about students' cognitive structures and explored the effect of prior achievement on the way students construct models. Other studies have explored the effect of item-feature context on content retrieved from students' CSs by analysing their narrative responses. Particularly, multiple studies have used 'Cheetah' as a taxon when comparing how prompt context influences the content of students' responses (Göransson et al., 2020; Nehm & Ha, 2011; Nehm & Schonfeld, 2008). Other studies, including our own have compared the content of students' narrative responses when asked to reason about evolution by natural selection in Humans vs Cheetahs (Beggrow & Sbeglia, 2019; de Lima, 2020, p 12). This study builds on all those previous strands of research and seeks to further our knowledge about the way students engage with modelling tasks by examining the influence of item-feature context on students' retrieval of conceptual information from their CSs. In particular, we explore how varying the taxon of an organism in a prompt (Human vs Cheetah) influences both the architecture and content of student-constructed models for students across a range of achievement levels.

METHODS

Setting and Participants

This study was conducted at a large, public university in the Midwest with highest research activity (The Carnegie Classification of Institutions of Higher Education, n.d.). Data for these analyses came from student responses (n = 384) in a large introductory

biology course for majors that focused on content domains of genetics, evolution, and ecology. The course is second in a two-course sequence required for life science majors; the first course focused on cell and molecular biology. Data were collected from two sections (n = 190 and 194) taught in different semesters by different instructors. Students receive explicit instruction on how to construct SBF models and regularly construct models on assessments (homework, exams, in-class activities).

Assessment Design

We used prompts from the 'Human/Cheetah Assessment' (or, HCA) described in (de Lima, 2020, p 12). In that study, HCA prompts elicited differences in the content of students' narrative responses. Namely, responses to questions about Cheetah evolution were more scientifically accurate than responses to questions about Human evolution. Each student responded to two prompts that required them to construct models that explained evolution by natural selection. The prompts had the same basal structure but differed in the specific taxon of organism - one prompt was about Cheetahs, the other about Humans (Table 2.1). Students had not seen these particular contexts (Human/Cheetah) on assessments or during instruction.

Table 2.1. Example prompts from the Human/Cheetah Assessment*. Two forms of an assessment item were developed that differed in taxon (Human vs Cheetah).*

Taxon	Human	Cheetah
Prompt	Modern Humans have *(trait)*. How would biologists explain how a species of Humans with *(trait)* evolved from an ancestral Human species without *(trait)*?	A species of Cheetah has *(trait)*. How would biologists explain how a species of Cheetah with *(trait)* evolved from an ancestral Cheetah species without *(trait)*?

In both class sections, the HCA was administered during class hours as part of a routine in-class low-stakes assessment following an instructional module on evolution. At this point in the course, students had constructed multiple SBF models and received feedback. We controlled for potential effects of prompt order by using versions of the HCA that differed in the order in which taxon (Human or Cheetah) was presented to students.

Data Processing

Selecting data

Perhaps owing to the low-stakes nature of the assessment, a large number of students (~54%) were excluded from analyses. A total of 350 model-based responses were included in analyses (2 models from each of 175 students, Table 2.2). Students were included if their responses to both prompts were in the form of models that were codable using SBF criteria. Examples of non-codable responses included narratives/essays, pictures of Humans and Cheetahs, and models that failed to include relationships (i.e., no connecting lines/arrows between structures).

To ensure that we had a representative sample, we compared included and excluded student populations to determine whether they differed in terms of demographics (Gender, Ethnicity, First Generation Learner) and/or prior academic achievement (Number of Credits and cumulative GPA at the start of the semester). Mixed-effects multiple logistic regressions did not detect significant differences between the populations included and excluded from analyses (Table S2.1).

Table 2.2. Demographic characteristics and prior academic achievement of student subgroups: included in study, excluded from study, and total student population*. STEM credits are the number of STEM credits completed at the beginning of the semester. Start GPA is the cumulative GPA of the students (based on a 4-point system) at the beginning of the semester.*

	Students included in the analysis	***Students excluded from the analysis***	***Students in the course***
Gender (% Female)	61%	61%	62%
Ethnicity (% white non-Hispanic)	79%	73%	76%
First Generation Learner (%)	24%	28%	26%
Class Rank (% Sophomore / %Junior)	54% / 33%	54% / 34%	54% / 34%
STEM credits (mean)	43.9	44.8	44.4
Start GPA (mean)	3.3	3.2	3.3

Content of student models

Researchers have used automated semantic mapping to evaluate the content of student-constructed models (Ifenthaler, 2010; Ifenthaler et al., 2011; Luckie et al., 2011). While automation can significantly expedite processing of large numbers of models, it is dependent on the presence or absence of predetermined entities. Specifically, semantic mapping relies on propositional accuracy or comparison with an 'expert model'. In this study, we were interested in capturing the ideas present in a model rather than evaluating its content or comparing it to a reference model

We coded student-constructed models for conceptual content using a rubric that quantifies the presence/absence of concepts related to evolution by natural selection and is based on prior studies that explored the effect of item-feature context on the content of student's narratives (see de Lima, 2020, p 160 for details on rubric development). The rubric assesses Key Concepts (KCs: Variation, Limited Resources and Competition, Differential Survival and Reproduction, and Heritability), Naïve Ideas (NIs: Need, Use, Adapt), and Threshold Concepts (TCs: Probability, Randomness and multiple Levels of Biological Organisation) (Göransson et al., 2020; Moharreri et al., 2014).

Model architecture

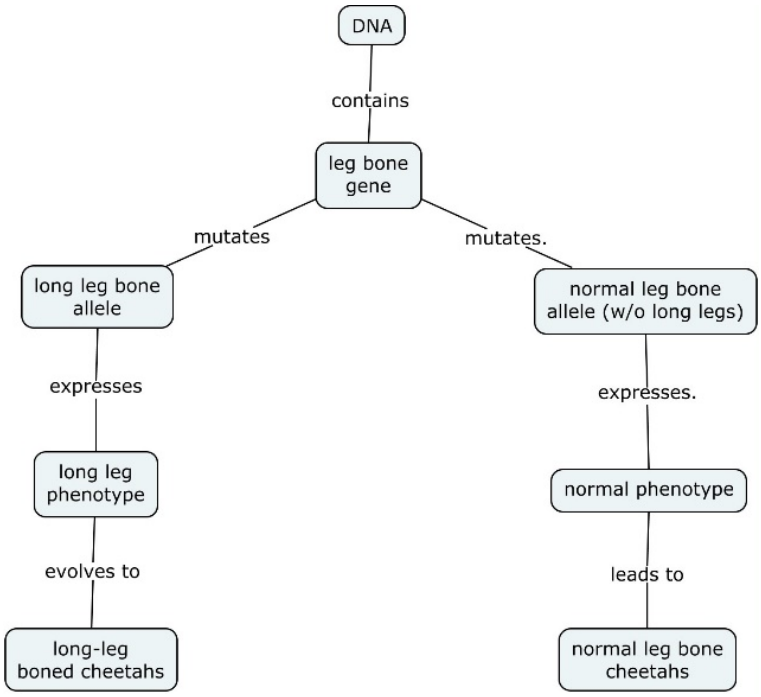


Figure 2.2. Example of a transcribed student-constructed model. *'DNA' is an example of a vertex, 'contains' is a linking phrase that defines an edge, and 'DNA – contains – leg bone gene' is a proposition. See Table 2.3 for details on how such models can be analysed using network metrics and for the specific calculated metrics for this model.*

Student-constructed models were transcribed into a digital format (Fig. 2.2) using CmapTools (https://cmap.ihmc.us/cmaptools/), a freely available software designed for constructing concept maps (Novak & Cañas, 2006). These were then exported into a datasheet as a list of vertices and linking statements. Each vertex (model structure), edge (linking phrase), and proposition (vertex-edge-vertex combination) were uniquely identified.

Data Analyses

We analysed both the content and architecture of students' models to explore the influence of context (taxon) on students' model-based explanations in terms of (1) the specific ideas elicited, and (2) the scope of ideas represented and connections among them.

1. Analysing the content of student-constructed models

a. Effect of prompt context on the content of student-constructed models.

We used mixed-effects multiple logistic regressions and multiple ordinal logistic regressions to evaluate whether prompt context influenced the presence/absence of concepts in student-constructed models. We present the results of these analyses as odds ratios and model-predicted conditional probabilities; these essentially compare the odds and probabilities of each concept occurring in Cheetah models relative to Human models. All models included student ID as a random intercept to account for variation due to individual differences. These models also include a covariate to account for variation in prior performance (see below).

b. Association between prior performance and model content

Among the various measures of academic performance, Grade Point Average (GPA) is one that is frequently used (Freeman et al., 2014; Freudenthaler et al., 2008). In the mixed-effects multiple logistic regression models described above, we included 'GPA tertile' as a covariate to quantify and account for the association between prior performance and the content of student-constructed models. Student's GPA tertile was

determined based on their cumulative GPA at the start of the course. Students were evenly binned into three tertiles: low (GPA< 3.22), medium (GPA between 3.23 and 3.67) and high (GPA > 3.68).

c. Association between prior performance and the consistency in model content across taxa. Students were binned into 3 groups based on the consistency with which they included evolutionary concepts in each of their two models that differed in taxon. Students who included a concept in both models were ‘consistent’, ‘inconsistent’ if they included a concept in one of their models but not the other, and ‘absent’ if a concept of interest did not appear in either model. For each concept, groups are exclusive. We then conducted a descriptive analysis to determine whether differences among consistency groups were associated with prior academic performance (GPA tertile).

2. Analysing the architecture of student-constructed models

a. Effect of prompt context on architecture (size and complexity) of student-constructed models

To explore the effect of prompt context on the architecture of student-constructed models, we analysed the size and complexity of the models by computing four different network metrics that had been used previously in research exploring students’ CSs (Dauer et al., 2013; Ifenthaler et al., 2011; Table 2.3). Number of Vertices and Surface Structure are indicative of the size and development of the CS. Average Degree of Vertices and Web-like Causality Index are indicative of the complexity of the CS. We calculated network metric values using the R package *igraph* (Csardi & Nepusz, 2006).

***Table 2.3. Network metrics used to analyse the architecture of student-constructed models**. Adapted from Ifenthaler et al. (2011) and Dauer et al. (2013). The last column gives the value for that particular metric as calculated for the model shown in Fig. 2.2*

Metric	How is it calculated?	Range	What does it indicate?	Metric value for Fig.2.2
Number of Vertices	Sum of all the Vertices	1 – ∞	Higher values indicate an increase in the size of the CS	8
Surface Structure	Sum of the number of all propositions.	0 – ∞	Higher values indicate higher development of the CS	7
Average Degree of Vertices	Mean number of edges linked to each vertex (incoming and outgoing)	1 – ∞	Higher values Indicate an increase in the complexity of the CS	1.75
Web-like Causality Index (WCI)	Proportion of vertices with more than 1 incoming edge added to the proportion of vertices with more than 1 outgoing edge	0 – 2	Higher values Indicate an increase in the complexity of the CS	0.125

Each student in the analysis created models for both Cheetahs and Humans. We therefore used paired t-tests to evaluate whether the mean value of each network metric differed between models of Cheetahs and Humans.

b. Association between prior performance and model architecture (size and complexity)

We fit mixed-effects linear models to explain variation in the different network metrics. The predictors we used were: prompt taxon, GPA tertile, and the interaction between the two fixed effects. We also used student ID as random intercepts to account for variation among individuals. We quantified the variance explained by the models using marginal R^2 and conditional R^2 values. Marginal R^2 refers to the proportion of variance explained by the fixed effects alone, while the conditional R^2 quantifies the variance explained by the fixed and random effects together (Nakagawa & Schielzeth, 2013).

Software

All analyses were done using the R statistical environment v 3.6.3 (R Core Team, 2020). We used the *dplyr* (Wickham, François, Henry, & Müller, 2020), *tidyr* (Wickham & Henry, 2020), and *stringr* (Wickham, 2019) packages for data processing, *igraph* (Csardi & Nepusz, 2006) for network metric calculations and visualisation, *lme4* (Bates et al., 2015) for mixed effects logistic regressions, *lmerTest* (Kuznetsova et al., 2017) for mixed-effects linear regressions, *ordinal* (Christensen, 2019) for mixed effects ordinal logistic regressions, *MuMIn* (Barton, 2018) for calculation of marginal and conditional R2 values for mixed models, *effects* (Fox, 2003) for calculating and plotting model output, *sjPlot* (Lüdecke, 2020) for generating tables, and *ggplot2* (Wickham, 2016) and *ggmosaic* (Jeppson et al., 2018) for plotting.

RESULTS

We found that student-constructed models were influenced by both prompt context and prior performance. Students' Cheetah models had more Key Concepts (KCs), fewer Naïve Ideas (NIs) and were bigger in size and complexity as compared to their Human models. Results of our specific analyses are presented with respect to each of our original research questions.

1. Analysis of Content of Student-Constructed Models

a. Effect of prompt context on the content of student-constructed models

Context had a significant effect on the presence of KCs and NIs, but not on Threshold Concepts (TCs) in students' models. Student-constructed models in response to the Cheetah prompt had on average 2.4 KCs and 0.04 NIs. However, those in response to the Human prompt had fewer KCs (2.0) and more NIs (0.09) on average. We did not see any difference in the average number of TCs with respect to the prompt taxon – the average was 0.9 TCs per model for both. The frequency with which the concepts were included in student-constructed models differed based on context (Fig. 2.3). Variation was the most frequently used KC and was present in both models for 86.3% (n=151) of the students, but it occurred more frequently in Cheetah models (93%, n=163), than in Human models (87%, n=153). Differential Resources and Competition was the KC that has the next highest frequency of occurrence (present in both the models for 60% (n=105) of students), however it occurred in Cheetah models (n=132) more than in Human models (n=111). Limited Resources and Survival was the most infrequent KC (absent in both the models for 63.4% of the students) and was highly influenced by

prompt context, appearing in nearly twice as many Cheetah models (n=58) as Human models (n=32). Heritability was similarly infrequent overall (occurred in ~50% of the models), but, in contrast, showed little difference based on the prompt context (71 Cheetah models, 62 Human models).

Due to the low frequency of each of the NIs, we present the data for responses that had any NI (Need, Use, Adapt). Most students (91%) did not include any NIs in both their responses, but among those that did, NIs were included more frequently in Human models (n=13) as compared to Cheetah models (n=5). In contrast, TCs were prevalent in student responses, with nearly 70% of responses containing at least one TC. However, unlike NIs, TCs did not appear to be influenced by context.

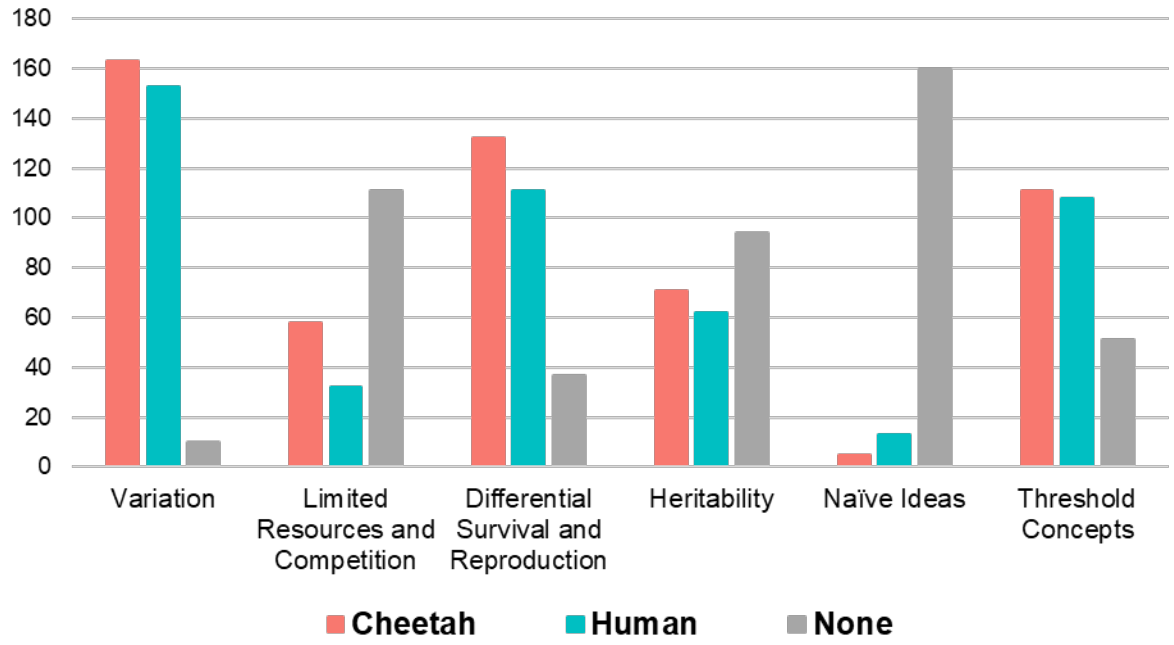


Figure 2.3. The frequency with which concepts were included in student-constructed models. *Bars represent the frequency of responses that contain each of the four Key Concepts (Variation, Limited Resources and Competition, Differential Survival and Reproduction and Heritability), at least one Naive Idea (need, use, adapt), and at least one Threshold Concept (probability, randomness, or at least two Levels of Biological Organisation)*

Results from the mixed-effects logistic regression show that Taxon was a strong predictor of the content in student-constructed models. The odds of students including KCs was lower in their Human models than their Cheetah models (*p* ranges from <0.001 to <0.1 for different KCs). However, the reverse was true for the odds of them including NIs ($p < 0.05$) (i.e., students had more NIs in their Human Models than in their Cheetah models). There was no difference between Cheetah and Human models for TCs (Table 2.4). Although the odds ratios for Variation and Naive Ideas are significant, the results should be interpreted with caution. This is because 90% of the student-constructed models had Variation and 95% did not have any Naive Ideas, posing algorithmic challenges when fitting the statistical models.

Table 2.4. Odds ratios of mixed-effects logistic regression analysis for Taxon using 'Human' as the reference taxon*. Bolded values are statistically significant (*** p < 0.001; ** p < 0.01; * p < 0.05; λ p < 0.1). Lower and Upper Confidence intervals are given in the brackets. This table provides the coefficients for 'Taxon', however the model also included 'Tertile' as a predictor.*

	Key Concepts					
	Variation	**Limited Resources and Competition**	**Differential Survival and Reproduction**	**Heritability**	**Naïve Ideas**	**Threshold Concepts**
Taxon 'Human'	**0.16*** [0.02, 0.61]	**0.19***** [0.07, 0.42]	**0.22**** [0.07, 0.49]	0.51 λ [0.22, 1.1]	**4.82*** [1.31, 25.96]	0.81 [0.38, 1.69]

Students included ideas about Limited Resources and Competition ~4.5 times as often in their Cheetah models than their Human models (Fig. 2.4a). Similarly, they used ideas about Differential Survival and Reproduction about 20% more frequently in their

Cheetah models as in their Human models (Fig. 2.4b). They also included NIs less frequently in their Cheetah models (Fig. 2.4c)

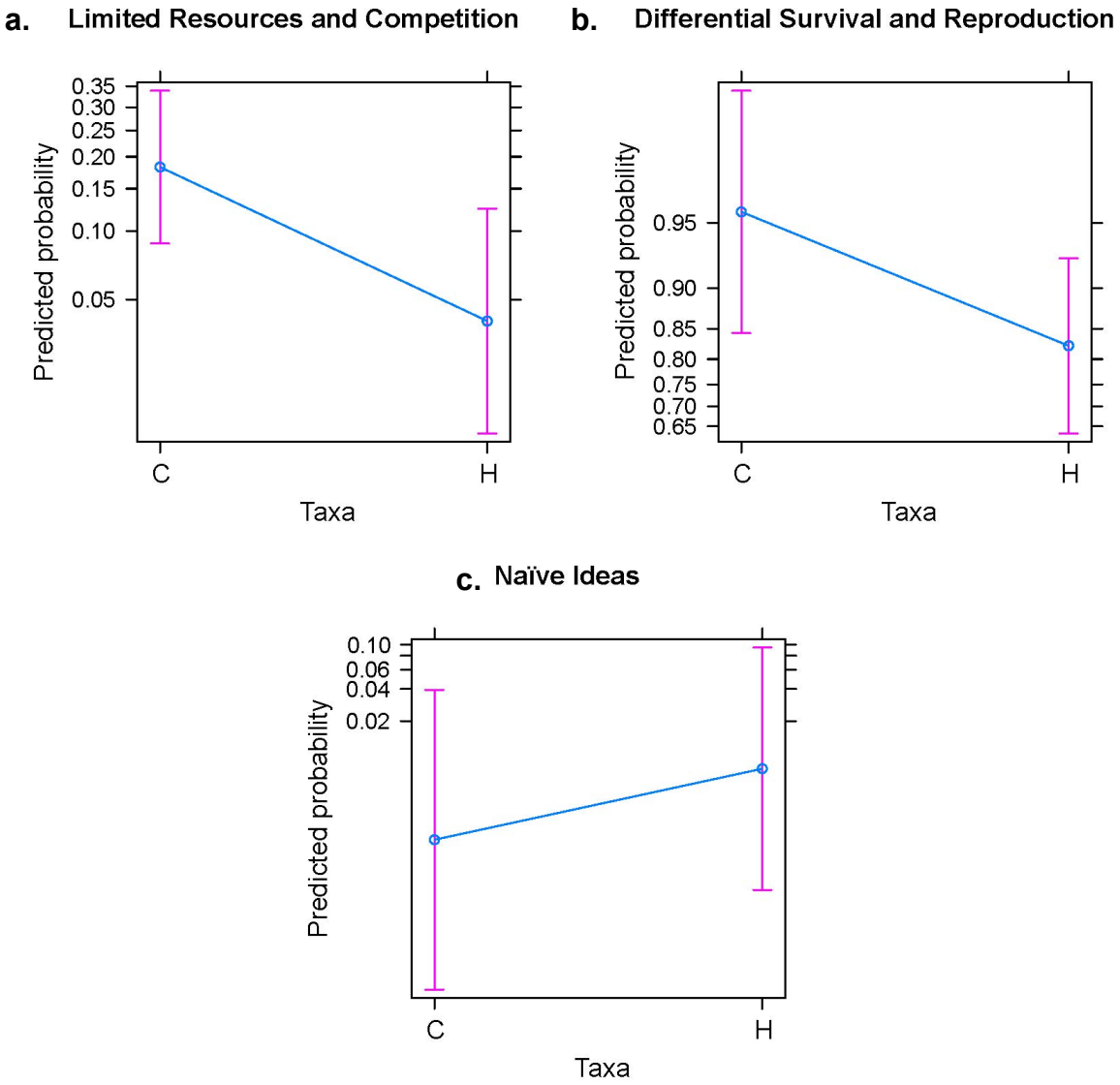


Figure 2.4. Model-inferred marginal probability of concept use in the two taxa Cheetah (C) and Human (H): (a) Limited Resources and Competition, (b) Differential Survival and Reproduction, and (c) Naïve Ideas occurring in student-constructed models. *Note that although the confidence intervals appear very wide, inference based on visualisation of the error bars is not very reliable, especially because of the presence of additional terms in the model (GPA tertile as fixed effect, and student identity as random intercept). Refer to Table 2.5 for additional information relevant to statistical inference.*

b. Association between prior performance and model content.

Results from our mixed effects logistic regression show that prior academic performance (as measured by incoming GPA) was a weak predictor for the presence of two of the KCs (Variation and Differential Survival and Reproduction, $p \leq 0.05$) and for the presence of TCs ($p \leq 0.05$) in student-constructed models. With every increase in tertile (low → medium achieving and medium → high achieving) the odds of students including TCs and ideas related to Variation and Differential Survival and Reproduction increased (Table 2.5).

Table 2.5. Odds ratios of mixed effects logistic regression analysis for effect of prior performance. *The odds shown are for every increase of one tertile. Bolded values are statistically significant (*** p < 0.001; ** p < 0.01; * p < 0.05; $^{\lambda}$ p < 0.1). Lower and Upper Confidence intervals are given in the brackets. This table provides the coefficients for 'Tertile', however the model also included 'Taxon' as a predictor*

	Key Concepts				Naïve Ideas	Threshold Concepts
	Variation	**Limited Resources and Competition**	**Differential Survival and Reproduction**	**Heritability**		
Tertile	**6.8*** [1.61, 120.0]	0.90 [0.39, 1.96]	**3.18***** [1.29, 13.10]	1.53 [0.54, 6.4]	2.02 [0.86, 6.26]	**3.42*** [1.29, 18.23]

The frequency of Differential Survival and Reproduction increased by 13% between the 1st and 3rd tertiles (Fig. 2.5a), and the frequency of TC use increased by 30% (Fig. 2.5b). Because of the same algorithmic challenges discussed earlier, we are cautious in interpreting the results for Variation. The difference in frequency between the 2nd and 3rd tertiles (for both TC and Differential Survival and Reproduction) was narrower than

that between the 1st and 2nd tertiles (Table S2.4). However, this should be interpreted with caution as it may be due to the nature of the model, which is linear on a log-odds scale.

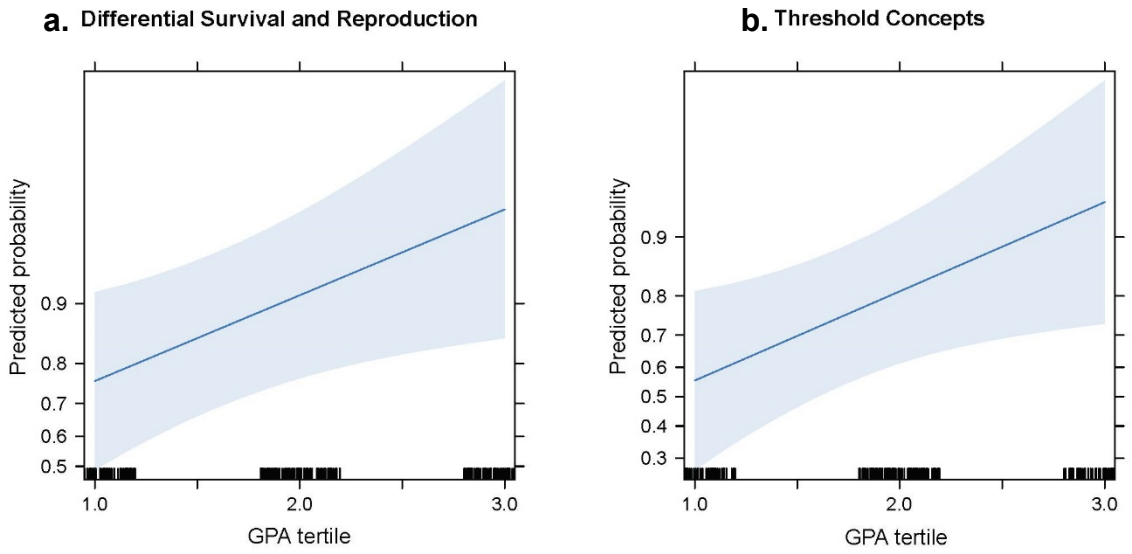


Figure 2.5. Effect of prior performance on the probability of a (a) Key Concept Differential Survival and Reproduction, and (b) Threshold Concepts occurring in student-constructed models*. Prior performance was determined on the basis of incoming GPA and students were binned into three tertiles: (1) Low-achieving, (2) Mid-achieving, and (3) High-achieving. Shaded region represents the confidence bands.*

c. Association between prior performance and consistency in model content across taxa.

Due to the large differences (skew) in the number of individuals in the subpopulations who represented concepts consistently (in both models), inconsistently (in one of the models) or not at all (absent in both models), we examined associations between variables and did not try to make causal statistical inferences.

We found that high-achieving students were more consistent in including KCs and TCs in their responses, regardless of taxon of the prompt (Table 2.6). Students who included

Variation in both their Cheetah and Human models had an average GPA 0.2 points higher than those who included Variation in only one of their models and 0.83 points higher than students who did not include Variation in either model. Similarly, students who included ideas about Differential Survival and Reproduction in both models had an average GPA that was 0.18 points higher than the students who included it in only one model, and 0.21 points higher than those who did not include it in either. Fewer students included ideas about Limited Resources and Competition and Heritability in their responses, and those that included them in both models had average GPAs that were only marginally higher than students who did not include them in either of their models (0.06 points higher for students who included Limited Resources and Competition and 0.14 points higher for students who included Heritability). Students who included TCs in both their models had an average GPA that was 0.24 points higher than those who included TCs in only one model and 0.33 points higher than those who did not include TCs in either model. While our results also indicate that the students who consistently included NIs also have a higher average GPA than their peers who did not have any NIs in their models, this difference is very small (0.03 GPA points), and the number of students who included NIs in all their responses is very low (3 students).

Table 2.6. Comparison of the number and GPA (mean ± Standard Error) of students who expressed an idea consistently (in both Human and Cheetah models), inconsistently (in either Human or Cheetah models) or did not include it in both (absent).

Concept		**Consistently included the concept** (present in both responses)	**Inconsistently used the concept** (present in one of the responses)	**Concept absent from all responses**
Variation	N	151	14	10
	GPA	3.42 (± 0.04)	3.22 (± 0.11)	2.59 (± 0.33)
Limited Resources and Competition	N	26	38	111
	GPA	3.43 (± 0.07)	3.27 (± 0.08)	3.37 (± 0.05)
Differential Survival and Reproduction	N	105	33	37
	GPA	3.43 (± 0.05)	3.25 (± 0.09)	3.22 (± 0.08)
Heritability	N	52	29	94
	GPA	3.41 (± 0.06)	3.56 (± 0.08)	3.27 (± 0.06)
Naïve Ideas	N	3	12	160
	GPA	3.39 (± 0.29)	3.32 (± 0.11)	3.36 (± 0.04)
Threshold Concepts	N	82	42	51
	GPA	3.51 (± 0.04)	3.27 (± 0.08)	3.18 (± 0.09)

2. Analysis of Architecture of Student-Constructed Models

a. Effect of prompt context on architecture (size and complexity) of student-constructed models

The architecture of student-constructed models differed based on prompt context (Cheetah vs Human; Table 2.7). Students' Cheetah models were significantly larger than their Human models (Number of Vertices, $p < 0.01$, Fig. 2.6a; Surface structure, $p < 0.01$, Fig. 2.6b) and were also more complex (WCI, $p < 0.05$, Fig. 2.6d; Average degree of vertices, $p < 0.1$, Fig. 2.6c)

Table 2.7. Paired t-tests comparing the mean value of each network metric in Cheetah and Human models. *Bolded values are statistically significant. CI indicates confidence interval.*

Network Metric	Cheetah models		Human models		Mean differences between Cheetah and Human Models	*CI*	*t*	*df*	*p*
	Mean	***Range***	***Mean***	***Range***					
Number of Vertices	**5.45**	**2-14**	**5.09**	**2-12**	**0.36**	**0.09, 0.62**	**2.66**	**174**	**0.008**
Surface structure	**4.43**	**1-14**	**4.07**	**1-10**	**0.36**	**0.10, 0.36**	**2.70**	**174**	**0.007**
Average degree of vertices	1.58	1-2.27	1.54	1-2	0.04	0, 0.06	1.95	174	0.052
Web-like causality index	**0.10**	**0-1**	**0.07**	**0-0.43**	**0.03**	**0, 0.04**	**2.20**	**174**	**0.029**

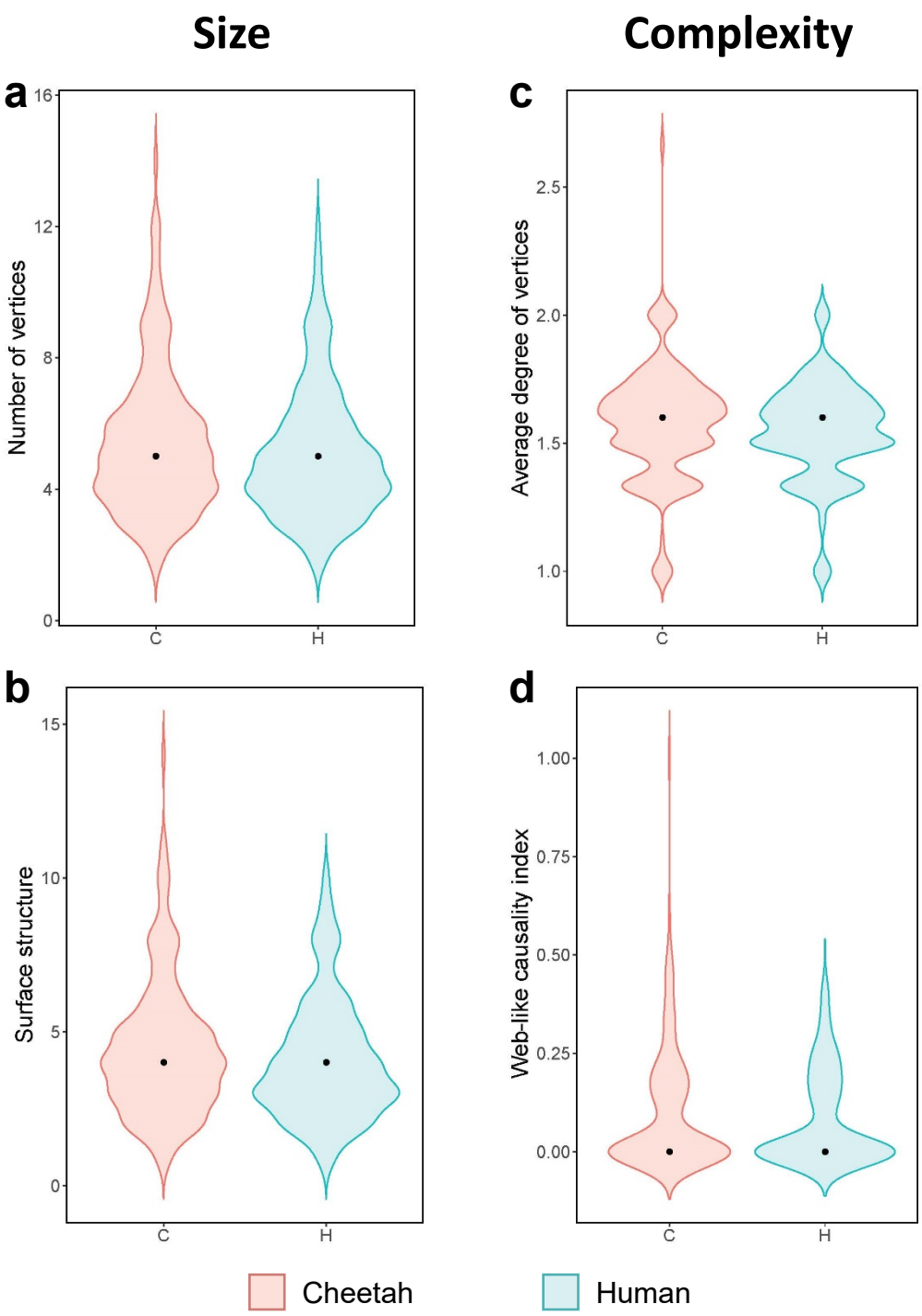


Figure 2.6. Distribution of network metrics for students' models of Cheetah (in red) and Human (in blue). Subplots include: (a) Number of Vertices, (b) Surface structure, (c) Average degree of vertices, and (d) Web-like causality index. *The black dot indicates the mean value of the metric. Number of Vertices and Surface structure are measures of the size. Average degree of vertices and Web-like causality index are measures of complexity.*

b. Association between prior performance and model architecture (size and complexity)

The mixed-effects linear models showed that taxa and tertile both affected the architecture of student-constructed models (Table 2.8). Regression coefficients for the effect of taxa on each network metric show a similar general pattern to that in the paired t-tests.

Table 2.8. Regression coefficients, marginal R^2 and conditional R^2 values from mixed-effects linear models explaining variation in network metrics. *(See Tables S2.7 – S2.10 for details). Predictors were the prompt taxon, GPA tertile, and the interaction between the two (fixed effects), as well as student ID (random intercept). Marginal R^2 refers to the proportion of variance explained by the fixed effects alone, while the conditional R^2 quantifies the variance explained by the fixed and random effects together (Nakagawa & Schielzeth 2013). Bolded values are statistically significant (*** p < 0.001; ** p < 0.01; * p < 0.05; $^{\lambda}$ p < 0.1).*

Network Metric	Number of Vertices	Surface structure	Average degree of vertices	Web-like causality index
Intercept (mean of Cheetah-Tertile 1 group)	**4.88*** **	**3.93*** **	**1.55*** **	**0.11*** **
β_{Human}	-0.33	-0.38	-0.07*	-0.05**
$\beta_{GPA\ T2}$	**1.06****	**1.06****	0.06	0.001
$\beta_{GPA\ T3}$	0.63	0.47	-0.002	-0.03
$\beta_{Human:GPA\ T2}$	-0.12	-0.09	0.03	0.04
$\beta_{Human:GPA\ T3}$	0.01	0.13	**0.08***	0.04 $^{\lambda}$
Marginal R^2 (fixed effects alone) %	4.6	4.6	3.4	1.9

Table 2.8. (cont'd)

Conditional R^2 (fixed + random effects) %	64.6	64.0	49.3	50.8

The four network metrics we used in these analyses measured two main constructs: Size was measured by Surface structure and Number of vertices, and complexity was measured by Average degree of vertices and Web-like causality index (WCI). Model size showed consistent patterns with tertile and taxon across both metrics (Figs. 2.7a and 2.7b). For both metrics, the size of students' Cheetah models was slightly larger than their Human models across tertiles (but $p > 0.05$). Model size increased substantially between the low and mid-achievers ($p < 0.01$) and then decreased slightly between the mid and high-achievers.

With respect to complexity, the patterns were less consistent (Figs. 2.7c and 2.7d). Cheetah models were generally more complex than Human models, although estimated Average degree of vertices for high-achievers for Human models was marginally higher. We see a similar pattern as model size between tertiles – complexity tends to increase substantially between the low and mid-achievers and decrease between the mid and high-achievers. However, low-achievers had a WCI for Cheetah models that was as high as the mid-achievers. The low-achievers also showed large differences in complexity between the two taxa, while the mid and high-achievers showed small to negligible differences based on taxa. Across all metrics, the fixed effects alone explained less than 5% of the variance in the data; the random and fixed effects together accounted for 30-70% of the variance.

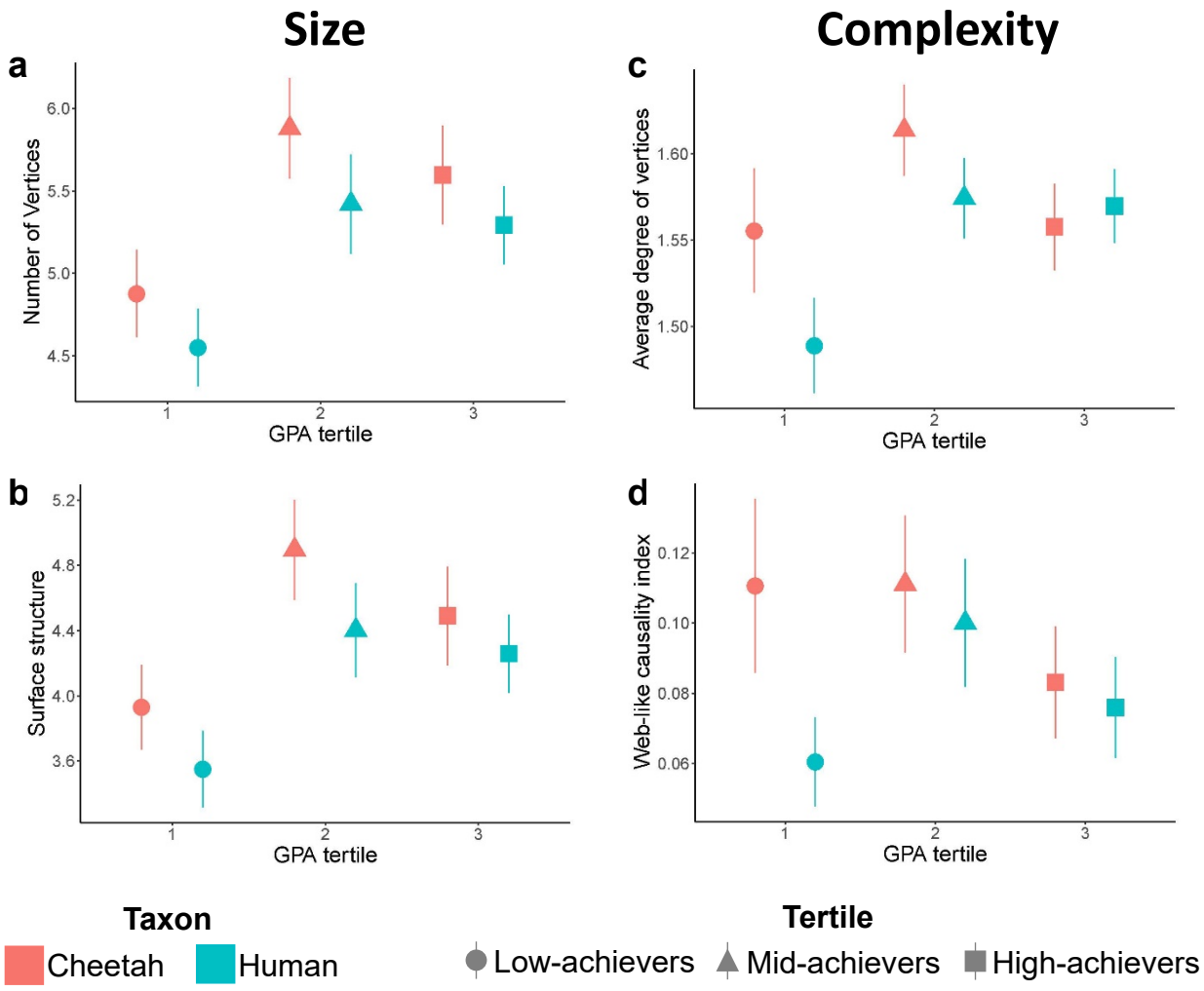


Figure 2.7. Association between network metrics and prior performance, for every combination of taxon and tertile for Size: a) Number of Vertices, b) Surface Structure; and Complexity: c) Average Degree of Vertices, and d) Web-like causality index*. Number of Vertices and Surface structure are measures of the size. Average degree of vertices and Web-like causality index are measures of complexity. Prior performance was determined on the basis of incoming GPA and students were binned into three tertiles: (1) Low-achievers, (2) Mid-achievers, and (3) High-achievers.*

DISCUSSION

These analyses provide new insights into the effect of item-feature context on students' model-based responses. Our results indicate that prompt taxon influenced both the architectural features and the conceptual content of student-generated models explaining evolution by natural selection. Additionally, we found that prior academic

achievement is associated with some of the differences we see in our data. In this section, we will discuss some plausible explanations for the patterns we see and the implications of our findings on instruction and assessment, especially with respect to using models for instruction and assessment.

Contextual Effects of the Prompt on the Content of Student-Constructed Models

A hallmark of understanding a concept is the ability to provide similar responses to conceptually equivalent questions testing that concept without being distracted by superficial factors in the prompt (Evans, 2008; Kampourakis & Zogza, 2009; Weston et al., 2015). The prompts in this study were designed to elicit the same conceptual information (evolution by natural selection) although they differed in surface features (taxon). Similar responses would indicate that students are using conceptual cues and not surface cues to access their cognitive structures and to retrieve relevant information while constructing their models. However, our results indicate that prompt context does influence the manner in which students access their CSs, build their mental models, and construct the required representation. The effect of prompt taxon was seen both in the content and the architecture of students' models.

Differences in the numbers of Key Concepts (KCs) and Naïve Ideas (NIs) in students' Human and Cheetah models show that the content of student-constructed models was affected by the context of the prompt. Students' Cheetah models had more KCs and fewer NIs than their Human models. Similar differences in response content due to prompt context have been shown by prior research including our own. Prompt context,

especially taxon, has affected students' narrative responses to questions about evolution by natural selection (Beggrow & Sbeglia, 2019; de Lima, 2020, p 12; Kampourakis & Zogza, 2008; Nehm & Ha, 2011; Prevost et al., 2013; Schurmeier et al., 2010). This shows that despite being required to represent their knowledge using a different mode of representation (model vs narrative), students demonstrate similar sensitivities to prompt context and are using surface features of the prompt as cues to access their CSs.

Constructing a response to a prompt based on surface as opposed to the conceptual features is indicative of a novice learner (Cheng et al., 2015; Hsu et al., 2012). Previous research has shown that when constructing Structure-Behaviour-Function (SBF) models, novices tended to include only the most salient structures instead of focusing on deeper relationships and emergent functions of the systems they were modelling (Hmelo-Silver & Pfeffer, 2004; Vattam et al., 2011). Our results could indicate that because these students are novice learners, their CSs for evolution are not particularly robust or intact, which could explain differences in the content they included in their models. A fragmented CS increases the challenge of retrieving relevant KCs without being distracted by irrelevant details and NIs (Dauer & Long, 2015; Hmelo-Silver et al., 2007). Previous research has shown that students have a hard time understanding evolution, especially human evolution, and that misconceptions and naïve ideas persist despite instruction (Beggrow & Sbeglia, 2019; Bishop & Anderson, 1990; Bray Speth et al., 2009; Nehm & Reilly, 2007). This is particularly supported by the finding from our study that students included fewer KCs and more NIs in their Human models as

compared to their Cheetah models. However, we did not detect any differences in the frequency of Threshold Concepts in students' responses based on taxa. Including a TC, by definition, is indicative of having an increased understanding of the subject matter and provides evidence of crossing a threshold of understanding which therefore would lead to decreased susceptibility to contextual influences.

Among the KCs, Variation appeared most frequently in student-constructed models. This was not surprising given that the students were enrolled in a course that was designed with Variation as a core theme. The course was organized around three themes focused on variation: how does biological variation arise? how is variation expressed? and, what are the consequences of variation among organisms? Variation is a central theme of evolution and is key to gaining a deeper understanding of evolution as a process (Emmons & Kelemen, 2015; Gregory, 2009; Halldén, 1988; Shtulman, 2006; Shtulman & Schulz, 2008). Previous studies have shown that students demonstrated a deeper understanding of evolution following instruction (Bray Speth et al., 2014). However, in our study, only 43% of the students (n=76) included ideas about variation at the genetic level. This could be because of an interaction between the affordances of the mode in which they were required to respond (construct a model) and because the prompt was framed at the organismal level. The spatial level referenced in the prompt could be another contextual cue that subsequently influenced the content of student-constructed models. Students could have started constructing their model at the organismal level and simply chosen not to include the genetic level either due to space constraints on the paper or because they assumed the prompt was

asking for a response at the organismal level. Alternatively, it is possible that they still do not have a keen understanding of the genetic basis of how variation originates and therefore did not include the genetic level. Previous studies have shown that understanding and articulating the genetic origin of variation is particularly difficult for students (Bray Speth et al., 2014). Additional analyses of classroom artifacts could shed insight onto whether this is a plausible explanation for this particular population.

In contrast, students were least likely to include the KC Limited Resources and Competition in their models (>60% of the students did not include the KC in either model). Of those that did include it, only 32 students included it in their Human model (as compared to 58 that included it in their Cheetah model). This could indicate that while students generally do not understand the role of Limited Resources and Competition with respect to evolution, they have an even greater difficulty thinking of it as relevant to Human evolution. This may be due to ingrained patterns of thinking (e.g., teleological and anthropomorphic thinking) that are notoriously resistant to instruction (Coley & Tanner, 2015; Inagaki & Hatano, 2006; Sinatra et al., 2008) or because they tend not to conceptualize humans as competing for resources in the ways other organisms do, particularly when that competition has consequences for fitness.

Our results also indicate that prior academic performance was associated with contextual susceptibility. Students with higher GPA's tended to include KCs and TCs in their models and were more consistent in the manner in which they included them – i.e., they included them in both models irrespective of context. Students who were

influenced by the context (and therefore included the KC/TC in only one of their models) and students who did not include KC/TCs in their models tended to have lower GPAs. This could indicate that students with higher GPAs have more robust or intact CSs. In a study in which students had to construct three types of models to illustrate the same phenomenon, Cheng & Lin (2015) found that students with higher science-learning performances also constructed models that were more coherent than their peers. Overall, this suggests that higher-performing students may be better at transferring concepts across contexts and less susceptible to contextual influences that distract from, or are unrelated to the underlying phenomenon.

Contextual Effects on the Architecture of Student-Constructed Models

Our study shows that the context of the prompt (taxon) affected the architecture of students' models. This is a unique contribution to the literature about the effect of context on students' reasoning and constructed representations. While previous studies, including our own, have shown the effect of context on students' responses, these findings are based primarily on analysis of written, text-based, narrative responses. This question has been relatively unexplored in model-based assessments, particularly for highly controlled prompt features as was the case in this study. Our findings show that even when we consider students' models sans content, we can see the effect of contextual influences. In this case, students' Cheetah models were both larger and more complex than their Human models, regardless of the specific content included in the models.

Our study also revealed differences in the ranges of model metrics for both size and complexity. Specifically, while the range for model size was large (Number of vertices, 2-14; Surface structure, 1-14) the range for complexity was relatively low (Average degree of vertices, 1-2.27; WCI, 0-1). This means that complexity was low even in larger models. This could be because students are used to linear thinking. Hay et al. (2008) postulate that our instructional choices have made students used to linear thinking. Because conventional lecturing often involves distilling down complex ideas into simple and easy-to-follow bullet points, this tends to emphasise linearity of thinking and rote-memorisation rather than engaging the complexity and intricacies of knowledge construction that is more reflective of reality, particularly in biological systems (Kinchin, 2006b, 2006a; Kinchin & Hay, 2007).

Alternatively, low model complexity could be attributed to the fact that students were required to construct these models using pen and paper. Royer & Royer (2004) reported that students using paper/pen constructed models that were lower in complexity than those that constructed models using a computer. Brandstädter et al. (2012) reported that students who used a computer constructed models that had higher propositional accuracy than those who constructed them using pen/paper. Lin et al. (2016) further showed that students not only learned better when they used computer mapping vs pen/paper, but found that computers were better for enabling students to collaborate and modify the model in real time as well as keep track of the various versions. It is therefore possible that because of the limitations of the medium (not being

able to reposition structures, concerns about making errors and cluttering the paper, unable to easily edit the model) lead to an overall lover complexity score.

If we consider prior achievement levels to be a proxy for progress on the novice to expert continuum, we see that lower-achievers have models that are smaller in size but lower in complexity than their higher achieving peers. Previous studies that have explored differences in expert-novice CSs based on differences in model architecture have similarly found that expert models differ significantly from novice models in both size and complexity (Hmelo-Silver & Pfeffer, 2004; Ifenthaler, 2011; Ifenthaler et al., 2011). This is to be expected as an expert would have both more information (larger model size) and would be better able to identify apparent as well as emergent relationships (greater complexity) within a study system/phenomenon. This could indicate that since our lower-achievers have fewer structures and propositions in their models, they also have smaller CSs on the topic of evolution (i.e., fewer concepts and links). Mid- and high-achievers may have better content preparation, more extensive prior experience, or more facile connection with material which then results in having a larger CS and therefore more concepts to draw from.

In our study, mid-achievers had the largest and most complex models. In a longitudinal study by Dauer et al. (2013), they observed that model complexity increased between assessments conducted early and mid-course, and then decreased between mid- and end-of-course, while model correctness continued to increase throughout. They attributed this to major restructuring of students' CSs that happened early in the

semester as students are accruing significant new information and identifying new relationships between concepts. In the second half of the semester, students' CSs then undergo minor restructuring as they fine-tune the information and shed unnecessary links, thereby making their CSs and models more parsimonious. Pearsall et al. (1997) reported similar findings. In their study the most "radical" restructuring of the CS takes place in the first part of the semester (first 4 weeks in most cases) and to a much lesser extent in the latter part of the semester. Therefore, it is plausible that our middle-achievers have accreted a lot of information and are on the path to fine-tuning it and developing a more parsimonious CS. Other model-based studies have shown that mid-achievers tend to be better at long-term conceptual retrieval (Bennett et al., 2020; Dauer & Long, 2015). This could indicate that the higher complexity of middle-achievers' models reflects a better-connected CS which then might facilitate their ability to retrieve relevant information after an extended period of time.

When associating prior performance with susceptibility to contextual effects, we saw that the general trend was that Human models were smaller and less complex than Cheetah models. The difference is most pronounced for low-achievers, particularly for complexity. This could indicate that because lower-achievers have less knowledge about the subject matter and/or modelling compared to their higher-achieving peers, they are most susceptible to contextual influences. It is also possible that these students have the greatest difficulty with the concept of human evolution, and it is this difficulty that is causing them to think in a more linear manner reflected in their lower complexity scores. Alternatively, it is also possible that they do not recognise that the two prompts

are analogous, and therefore feel compelled to produce distinctly different responses (Chi et al., 1981; Hmelo-Silver & Pfeffer, 2004; Nehm & Ridgway, 2011).

Implications for Instruction and Assessment

Models and model building are relevant not only in the scientific context (education or practice) but in a wide variety of daily life situations (Halloun, 2007). Recommendations from reports such as Vision & Change (AAAS, 2011) and NGSS (NGSS Lead States, 2013) have led many instructors in the life sciences to increase their use of models and modelling in classrooms (see Wilson et al., 2020). Our finding that context affects both the content and architecture of students' models has implications for both instruction and assessment. We hypothesize that some of the contextual influences we observed are due to the fact that students are novices in the subject matter. The fact that evolution, especially human evolution, is a notoriously difficult topic for students (Bishop & Anderson, 1990; Catley & Novick, 2009; Morabito et al., 2010; Nehm & Reilly, 2007; Smith, 2010b, 2010a), must be acknowledged by instructors while designing their curricula/instruction. Instructional strategies that are designed to improve students' understanding of evolution, particularly human evolution, have been proposed by multiple researchers (Alters & Nelson, 2002; Bray Speth et al., 2009, 2014; Kalinowski et al., 2010; Kampourakis & Zogza, 2009; Pobiner et al., 2018).

Students are also novices to modelling. Therefore, in order to develop expertise in modelling, students will have to understand not only how to construct a model, but what a model is, how to visualise it, and then how to represent it (Gilbert, 2004). Nicolaou &

Constantinou (2014) proposed a framework for developing modelling competence which stresses both modelling practises (e.g., creating, validating, etc.) and meta-modelling knowledge which includes understanding the nature and purpose of models and the modelling process. To facilitate this, students need practice in developing their own models in addition to working with provided models (van Driel et al., 2019). They need to understand not just what goes into the model, but why it belongs there. In SBF language, instruction should focus beyond the structures in a model to include the function of the model (Constantinou et al., 2019).

A major problem that multiple researchers have pointed out is that students tend to think that there is only one correct model for any system/phenomenon and find it difficult to accept the possibility of alternative models (Constantinou et al., 2019; Grosslight et al., 1991; Grünkorn et al., 2014). This belief reinforces a culture of learning by memorizing because there can be only one ‘right’ answer. Bennett et al. (2020) found that students who had memorised a model linking genetic variation to phenotype did well when asked to respond to the same prompt at a later date, but were unable to transfer concepts when asked to construct models of the same phenomenon in different contexts. Dauer & Long (2015) reported similar findings – students who relied on memorising models did not have a complete understanding of the system, and when asked to reproduce the model for a new context, faced difficulties. Grünkorn et al. (2014), suggest that this might be because the students have been taught about the historical development of models of systems (e.g., the atom) where a series of discoveries progressively

increased model accuracy until a ‘final’ model was achieved, and they therefore suggest that these examples be used with caution.

A major barrier to including student-constructed models in classrooms – especially on assessments – is the perception of increased difficulty in scoring them. However, there has been a lot of progress in the effectiveness of automated assessments, particularly when students construct models on the computer rather than on paper (Ifenthaler, 2010; Ifenthaler et al., 2011; Luckie et al., 2011). Researchers and educators have also developed resources and strategies for designing modelling-based instruction and assessments and delivering feedback at large scales (Upmeier zu Belzen et al., 2019b; Wilson et al., 2019).

Future Steps

Our findings pose some interesting questions that would be worth exploring in the future:

We have seen that context (taxon) influences both the content and the architecture of student-constructed models. Our previous research also showed that context influences the content of students’ narrative responses. However, this raises the question – does the context of the prompt predict the content of the response irrespective of the mode of response? What would be the effect of context on the content of students’ responses if they are asked to respond using two different modes of representation (e.g. narrative/model/drawing) at the same time? Will we see an interaction between the context of the prompt and the mode of representation?

The course in which this study was conducted, uses a lot of collaborative modelling activities. However, this particular assessment was conducted by individuals. There is evidence to show that when students constructed models collaboratively, they produce models that were of a higher quality (Kwon & Cifuentes, 2009), and perform better on biology tests (Brown, 2003) as compared to students who construct models individually. If this was done as a collaborative activity, we should expect to see an increase in the size of the model with respect to an individual model as they would be pooling their knowledge, however it would be interesting to explore the effect on the complexity of the model and the susceptibility to context. Would the complexity increase with because some students recognise relationships that others do not? Or will the models become more parsimonious because of an increase in the overall expertise of the group

The study also highlights some of the findings from previous studies that middle achievers seem to be interacting with models in a unique way. In our study, these students constructed models that were bigger in size and in complexity than their peers, and unlike the low-achievers they did not seem to be susceptible to context. It would be interesting to determine if they are using cues to access their CSs that are different from their peers. Or is their motivation while approaching modelling tasks different?

Answering these questions will pave the way to increasing our understanding of the way students engage with modelling and how context affects the way they access their CSs.

ACKNOWLEDGEMENTS

I thank Mitch Distin, Devin Babi and Hunter Hicks for help with data transcription; Socheatha Chan for logistical and technical support; and Mridul Thomas for help with data analysis; and Melanie Cooper, Amelia Gotwals, and Katherine Gross for comments on earlier versions of the manuscript. I express my deep gratitude to the world's best advisor, Tammy Long, for the mentorship and supervision provided during all stages of this project. Finally, I gratefully acknowledge all the anonymous students whose assessments we used as the data for this study.

Funding

This material is based in part upon research supported by the National Science Foundation under grant numbers DRL 1420492, DRL 0910278. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation. Additional funding was provided by the Michigan State University, Graduate School through the FAST (Future Academic Scholars in Teaching) Fellowship Program (2017-2019).

APPENDIX

Table S2.1. Odds ratios of multiple logistic regression for demographic analysis with lower and upper confidence intervals (95%) and p-values.

Predictors	*Odds Ratios*	*CI*	*p*
(Intercept)	0.29	0.05, 1.59	0.159
Gender [M]	1.03	0.68, 1.57	0.885
Ethnicity [Minority]	0.88	0.29, 2.75	0.827
Ethnicity [White (non-Hispanic)]	1.13	0.40, 3.27	0.816
First Generation Learner [yes]	0.85	0.52, 1.36	0.491
Start STEM credits	1.00	0.99, 1.01	0.785
Start GPA	1.40	0.99, 2.05	0.065
Observations	377		

Table S2.2. Additional network metrics used to analyse the architecture of student-constructed models. *Table adapted from Ifenthaler et al. (2011) and Dauer et al. (2013)*

Metric	How is it calculated?	Range	What does it indicate?	Metric value for Fig. 2.2
Graphical structure	Fewest number of edges between most distant vertices of spanning tree	0 – ∞	Higher values indicate broader understanding of the subject matter.	4
Connectedness	Mean probability of reaching every vertex from every other vertex	0 – 1	Higher values indicate deeper understanding of the subject matter	0.34
Ruggedness	Number of unlinked subgraphs	1 – ∞	Higher values indicate a greater lack of understanding of the subject matter	1

Table S2.3. Network metrics, their operationalisation, and the code used to calculate them for each model.

Metric	What does it indicate?	R *igraph* code used to calculate it
Number of vertices	Indicate of size of the CS	vcount(g)
Surface structure	Indicates development of the CS.	nrow(subdat)
Average degree of vertices	Indicates the complexity of the CS	mean(degree(g, mode = 'all'))
Web-like causality index	Another indicator of complexity	(length(degree(g, mode = 'in')[degree(g, mode = 'in') > 1]) / vcount(g)) + (length(degree(g, mode = 'out')[degree(g, mode = 'out') > 1]) / vcount(g))
Graphical structure	Indicates the breadth of subject matter understanding	diameter(mst(g))
Connectedness	Indicates depth of understanding of the subject matter	sum(distance_table(g, directed = TRUE)$res) / (sum(distance_table(g, directed = TRUE)$res) + distance_table(g, directed = TRUE)$unconnected)
Ruggedness	Increase in ruggedness indicates a lack of understanding of the subject matter	count_components(g)

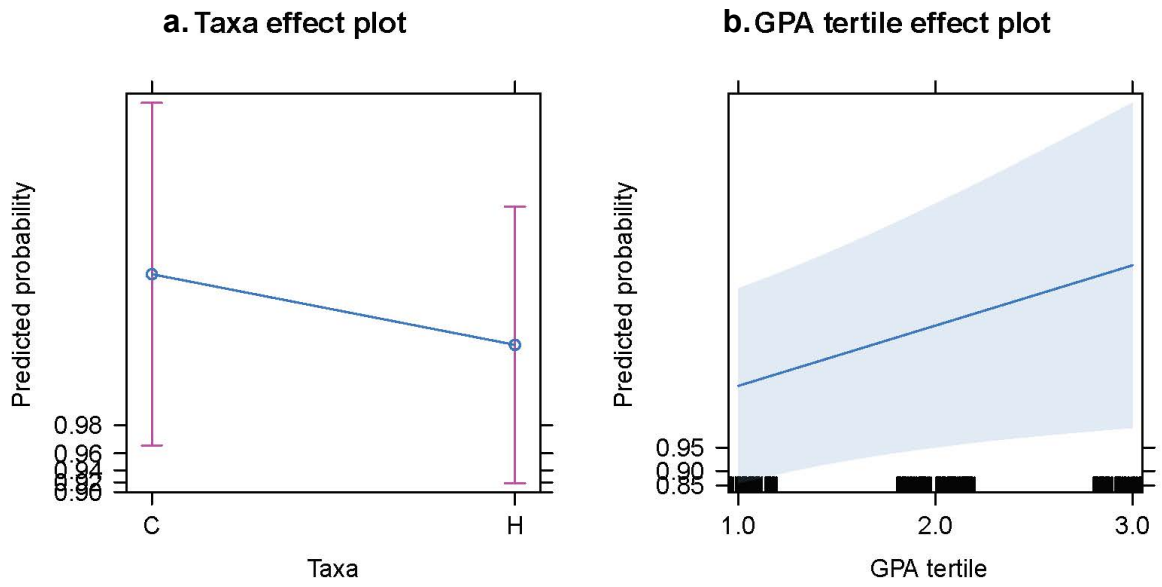


Figure S2.1. Model-inferred marginal probabilities for each of the fixed effects (a. taxa and b. tertile) on the presence of Variation in student-constructed models*. Model also included student identity as random intercept.*

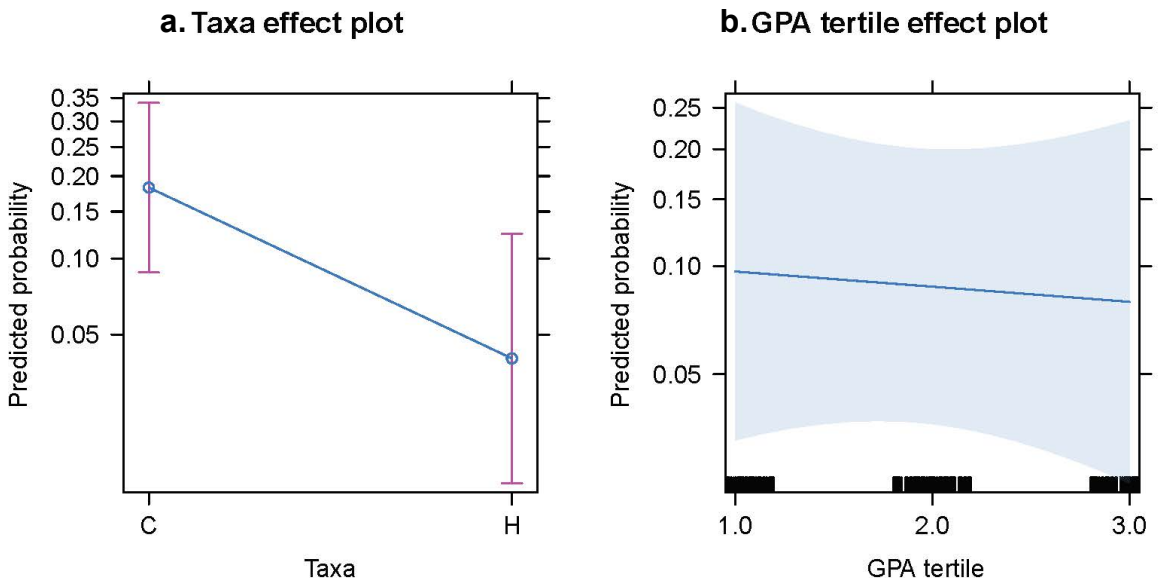


Figure S2.2. Model-inferred marginal probabilities for each of the fixed effects (a. taxa and b. tertile) on the presence of Limited Resources and Competition in student-constructed models. *Model also included student identity as random intercept.*

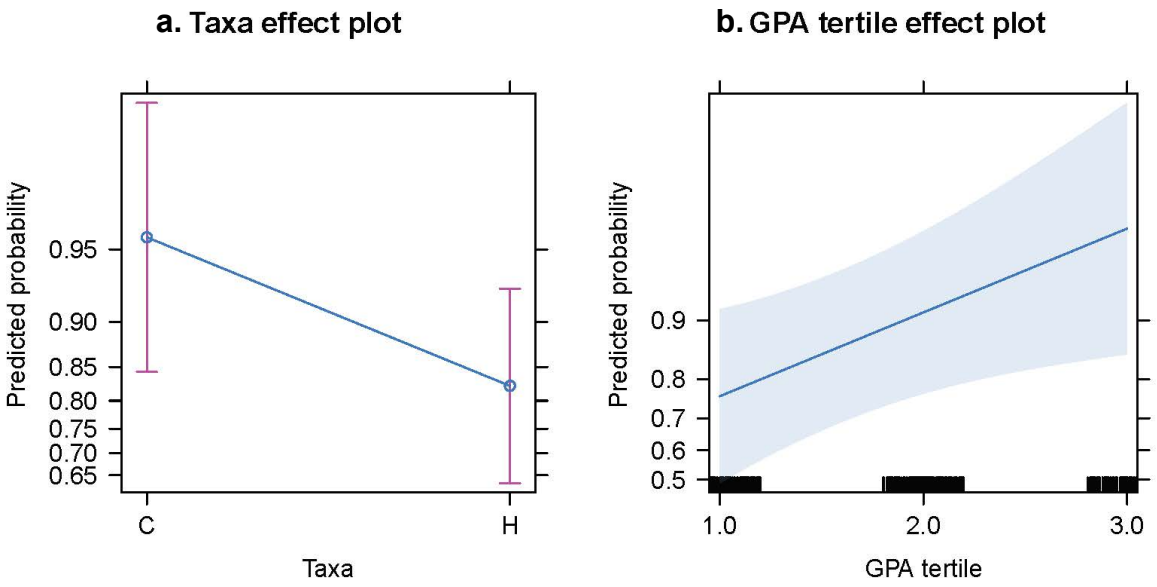


Figure S2.3. Model-inferred marginal probabilities for each of the fixed effects (a. taxa and b. tertile) on the presence of Differential Survival and Reproduction in student-constructed models. *Model also included student identity as random intercept.*

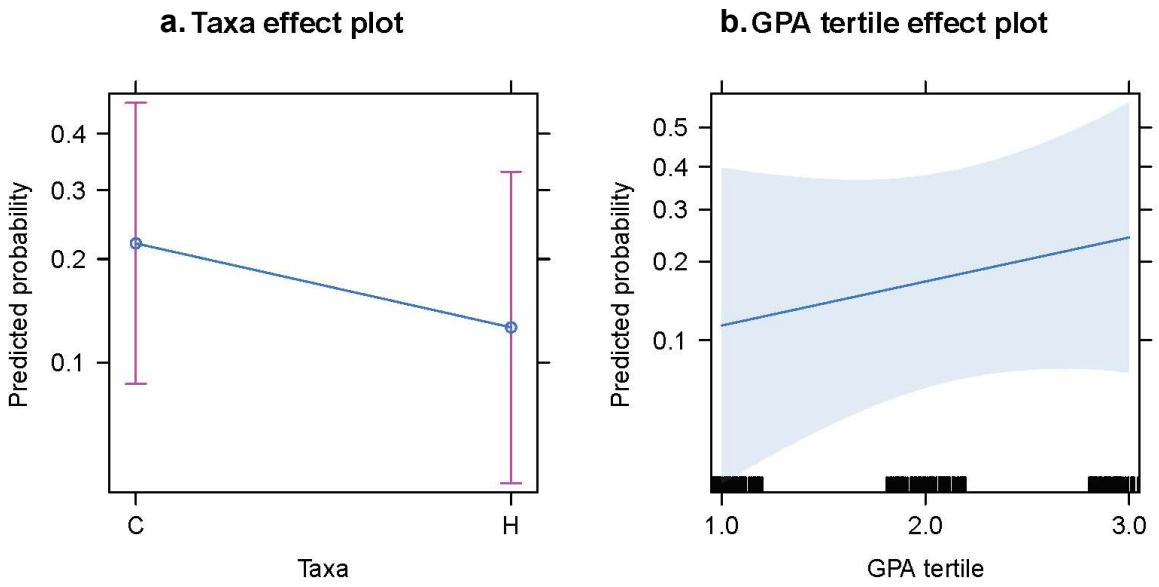


Figure S2.4. Model-inferred marginal probabilities for each of the fixed effects (a. taxa and b. tertile) on the presence of Heritability in student-constructed models. *Model also included student identity as random intercept.*

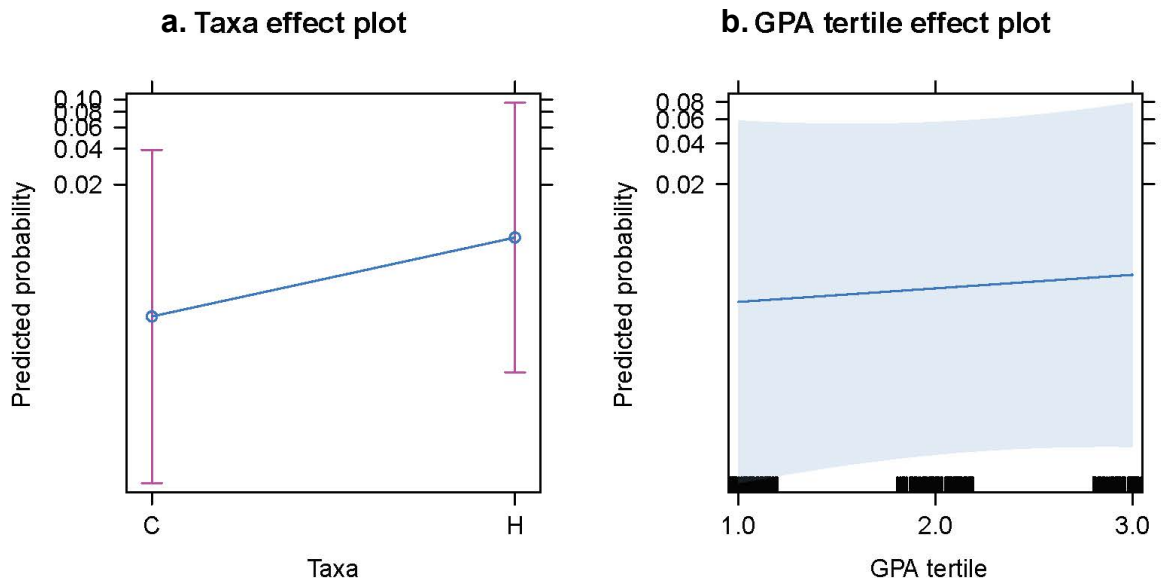


Figure S2.5. Model-inferred marginal probabilities for each of the fixed effects (a. taxa and b. tertile) on the presence of Naïve Ideas in student-constructed models*. Model also included student identity as random intercept.*

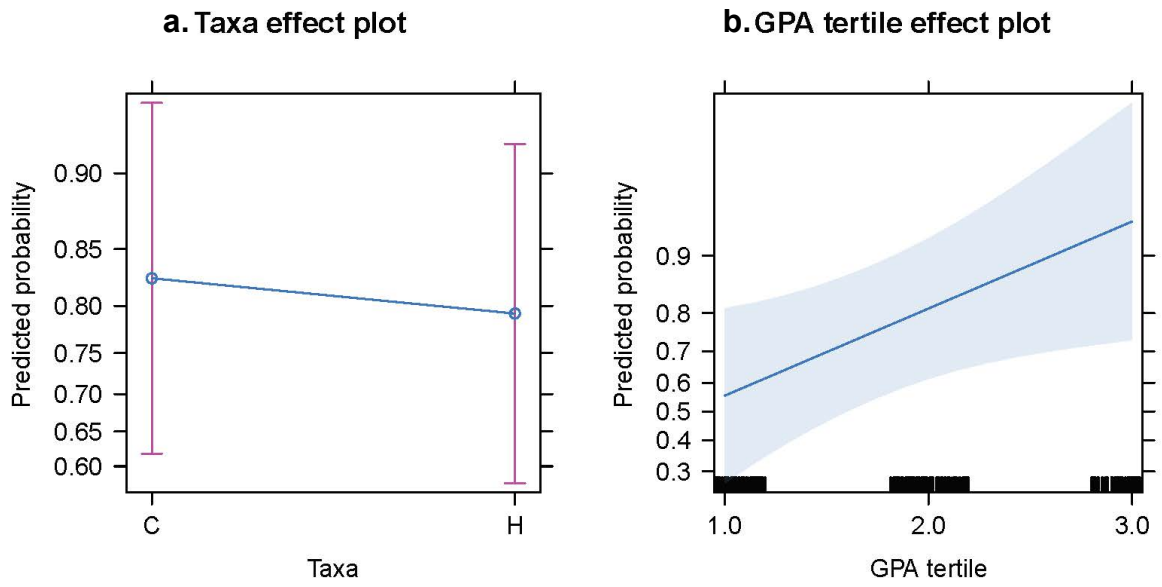


Figure S2.6. Model-inferred marginal probabilities for each of the fixed effects (a. taxa and b. tertile) on the presence of Threshold Concepts in student-constructed models. *Model also included student identity as random intercept.*

Table S2.4. Model-inferred marginal probability of concept use across the three tertiles.

	Key Concepts					
	Variation	**Limited Resources and Competition**	**Differential Survival and Reproduction**	**Heritability**	**Naïve Ideas**	**Threshold Concepts**
Tertile 1	0.99	0.09	0.75	0.11	0	0.56
Tertile 2	0.99	0.09	0.91	0.17	0	0.81
Tertile 3	0.99	0.08	0.97	0.24	0	0.94

Table S2.5. Paired t-tests comparing the mean value of each additional network metric in Cheetah and Human models.

Network Metric	Mean differences between Cheetah and Human Models	*CI*	*t*	*df*	*p*
Graphical structure	0.09	-0.09, 0.26	0.97	174	0.335
Connectedness	-0.01	-0.02, 0	-1.83	174	0.068
Ruggedness	0.04	0, 0.09	1.61	174	0.109

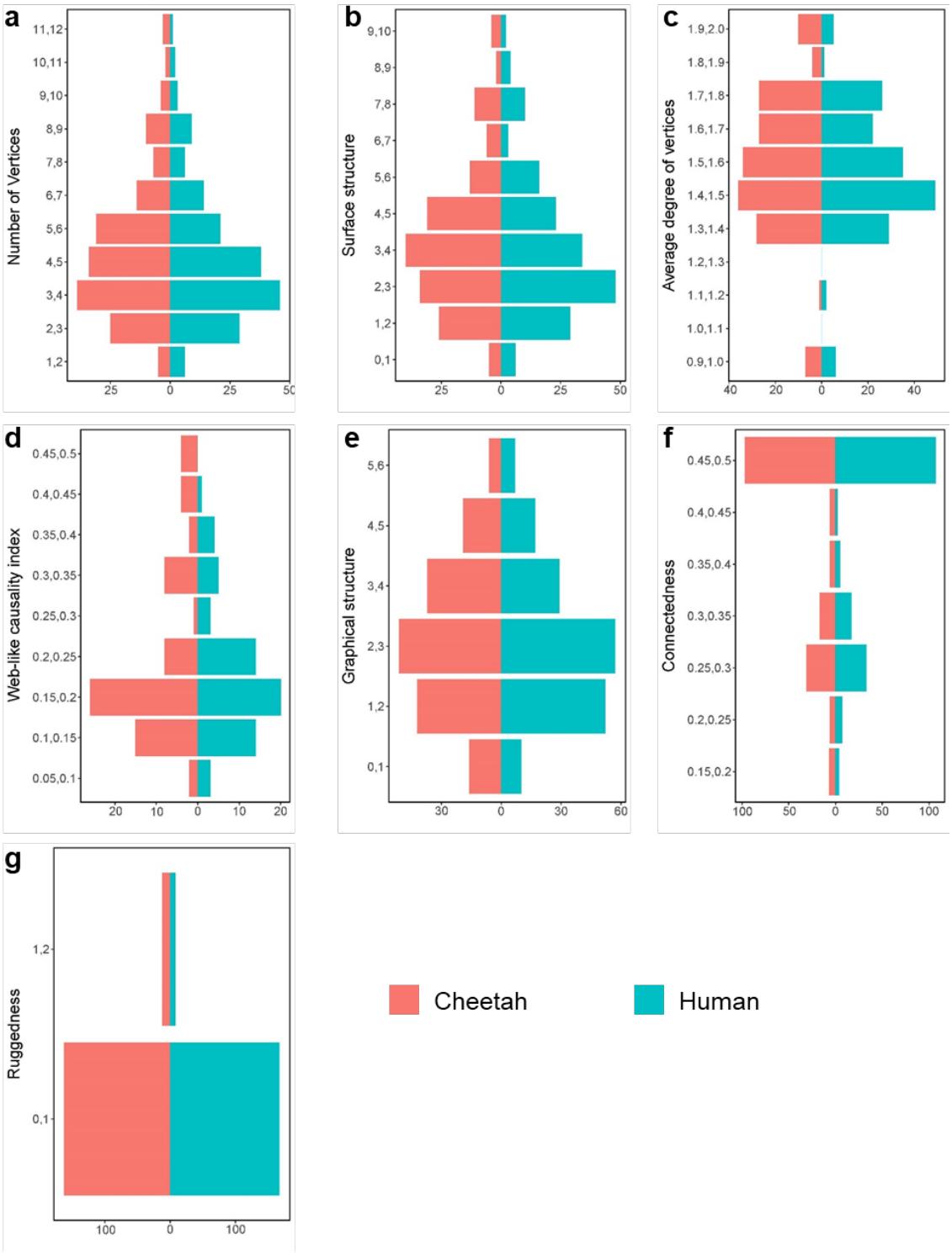


Figure S2.7. Distribution of network metrics for Cheetah vs Human models. a) Number of Vertices, b) Surface structure, c) Average degree of vertices, d) Web-like causality index, e) Graphical structure, f) Connectedness, g) Ruggedness

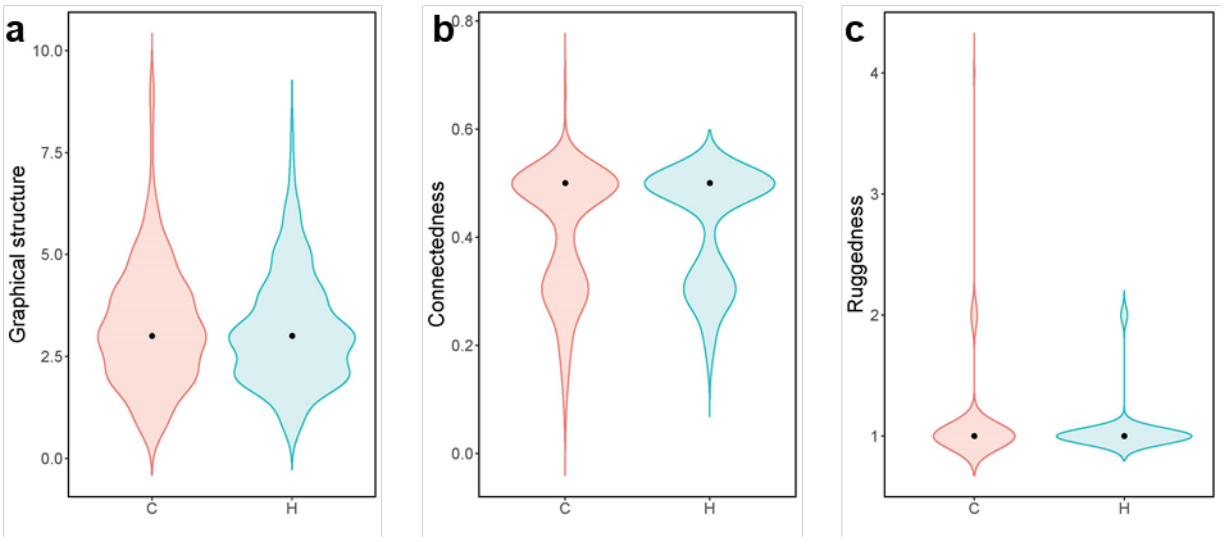


Figure S2.8. Violin plots showing distribution of network metrics for Cheetah vs Human models of a) Graphical Structure b) Connectedness, and c) Ruggedness. *The black dot indicates the mean.*

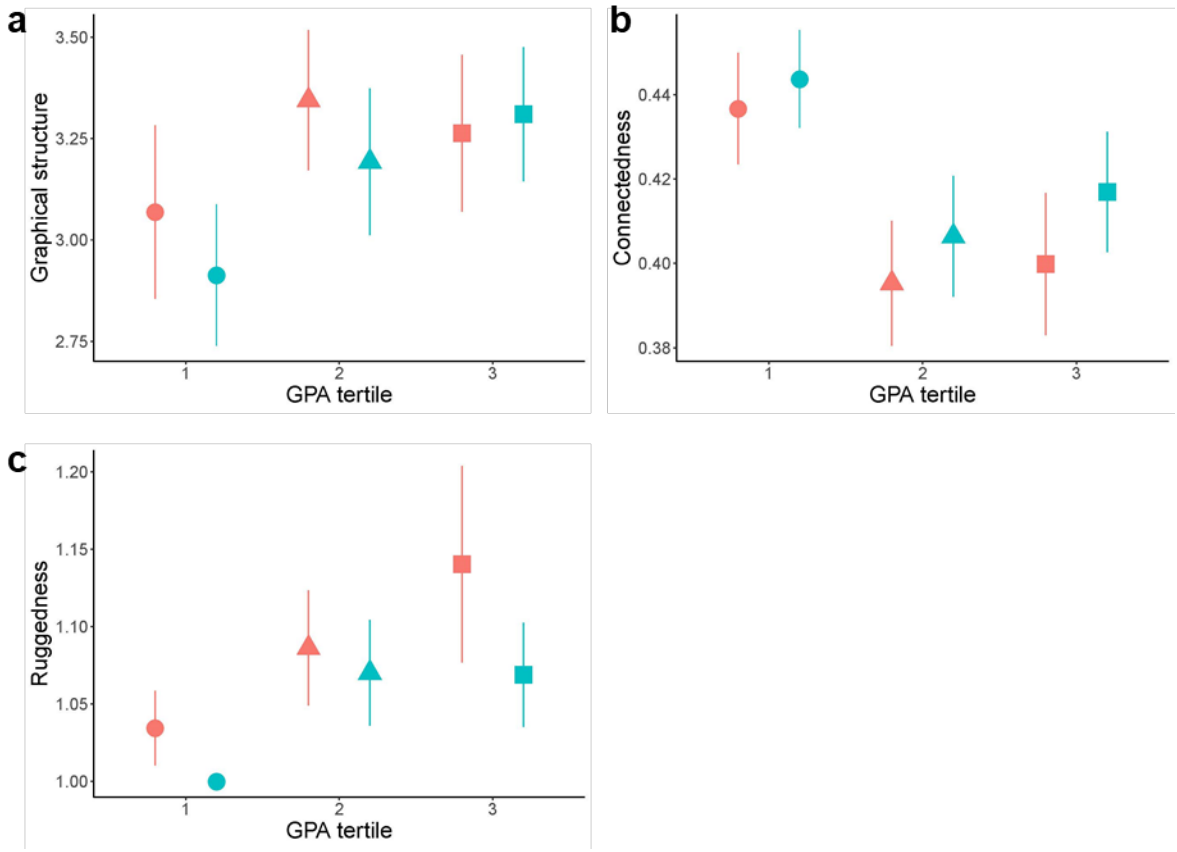


Figure S2.9. Plots showing mean (± SE) for each of the network metrics for every combination of taxon and tertile for a) Graphical Structure b) Connectedness, and c) Ruggedness.

Table S2.6. Regression coefficients, marginal R^2 and conditional R^2 values from a set of mixed-effects linear models explaining variation in additional network metrics. *(See tables S2.11 – S2.13). Predictors were the prompt taxon, GPA tertile, and the interaction between the two (fixed effects), as well as student ID (random intercept). Marginal R^2 refers to the proportion of variance explained by the fixed effects alone, while the conditional R^2 quantifies the variance explained by the fixed and random effects together (Nakagawa & Schielzeth 2013). Bolded values are statistically significant (*** p < 0.001; **p < 0.01; * p < 0.05; $^{\lambda}$ p <0.1).*

Network Metric	Graphical structure	Connectedness	Ruggedness
Intercept (mean of Cheetah-Tertile 1 group)	**3.069*** **	**0.44*** **	**1.03*** **
β_{Human}	-0.16	0.01	-0.03
$\beta_{GPA\ T2}$	0.32	**-0.04***	0.05
$\beta_{GPA\ T3}$	0.15	-0.03 $^{\lambda}$	**0.11***
$\beta_{Human:GPA\ T2}$	0.006	0.004	0.02
$\beta_{Human:GPA\ T3}$	0.01	0.13	**0.08***
Marginal R^2 (fixed effects alone) %	4.6	4.6	3.4
Conditional R^2 (fixed + random effects) %	64.6	64.0	49.3

Table S2.7. Mixed-effects linear model summary for Number of vertices. *Bolded values are statistically significant.*

	Number of vertices		
Predictors	*Estimates*	*CI*	*p*
(Intercept)	4.88	4.34 – 5.42	**<0.001**
Taxa [Human]	-0.33	-0.79 – 0.14	0.166
GPA tertile 2 [Mid-achievers]	1.09	0.33 – 1.84	**0.005**
GPA tertile 3 [High-achievers]	0.63	-0.13 – 1.39	0.105
Interaction 1: Taxa [H] * GPA tertile [2]	-0.12	-0.77 – 0.54	0.728
Interaction 2: Taxa [H] * GPA tertile [3]	0.01	-0.64 – 0.67	0.968
Random Effects			
σ^2	1.62		
$\tau_{00\ Deid}$	2.75		
ICC	0.63		
$N_{\ Deid}$	173		
Observations	346		
Marginal R^2 / Conditional R^2	0.047 / 0.647		

***Table S2.8. Mixed-effects linear model summary for Surface structure**. Bolded values are statistically significant.*

	Surface structure		
Predictors	*Estimates*	*CI*	*p*
(Intercept)	3.93	3.40 – 4.47	**<0.001**
Taxa [Human]	-0.38	-0.84 – 0.08	0.109
GPA tertile 2 [Mid-achievers]	1.06	0.30 – 1.81	**0.006**
GPA tertile 3 [High-achievers]	0.47	-0.29 – 1.22	0.227
Interaction 1: Taxa [H] * GPA tertile [2]	-0.09	-0.75 – 0.57	0.782
Interaction 2: Taxa [H] * GPA tertile [3]	0.13	-0.53 – 0.79	0.700
Random Effects			
σ^2	1.62		
$\tau_{00\ Deid}$	2.68		
ICC	0.62		
$N_{\ Deid}$	173		
Observations	346		
Marginal R^2 / Conditional R^2	0.046 / 0.640		

Table S2.9. Mixed-effects linear model summary for Average degree of vertices*. Bolded values are statistically significant.*

	Average degree of vertices		
Predictors	*Estimates*	*CI*	*p*
(Intercept)	1.56	1.50 – 1.61	**<0.001**
Taxa [Human]	-0.07	-0.12 – -0.01	**0.016**
GPA tertile 2 [Mid-achievers]	0.06	-0.01 – 0.14	0.100
GPA tertile 3 [High-achievers]	-0.00	-0.08 – 0.07	0.950
Interaction 1: Taxa [H] * GPA tertile [2]	0.03	-0.05 – 0.10	0.473
Interaction 2: Taxa [H] * GPA tertile [3]	0.08	0.00 – 0.15	**0.046**
Random Effects			
σ^2	0.02		
$\tau_{00\ Deid}$	0.02		
ICC	0.48		
$N_{\ Deid}$	173		
Observations	346		
Marginal R^2 / Conditional R^2	0.034 / 0.494		

Table S2.10. Mixed-effects linear model summary for Web-like causality index*. Bolded values are statistically significant.*

	Web-like causality index		
Predictors	*Estimates*	*CI*	*p*
(Intercept)	0.11	0.08 – 0.15	**<0.001**
Taxa [Human]	-0.05	-0.09 – -0.01	**0.005**
GPA tertile 2 [Mid-achievers]	0.00	-0.05 – 0.05	0.943
GPA tertile 3 [High-achievers]	-0.03	-0.08 – 0.02	0.257
Interaction 1: Taxa [H] * GPA tertile [2]	0.04	-0.01 – 0.09	0.124
Interaction 2: Taxa [H] * GPA tertile [3]	0.04	-0.01 – 0.09	0.093
Random Effects			
σ^2	0.01		
$\tau_{00\ Deid}$	0.01		
ICC	0.50		
$N_{\ Deid}$	173		
Observations	346		
Marginal R^2 / Conditional R^2	0.020 / 0.509		

Table S2.11. Mixed-effects linear model summary for Graphical structure. *Bolded values are statistically significant.*

	Graphical structure		
Predictors	*Estimates*	*CI*	*p*
(Intercept)	3.07	2.71 – 3.43	**<0.001**
Taxa [Human]	-0.16	-0.46 – 0.15	0.318
GPA tertile 2 [Mid-achievers]	0.32	-0.19 – 0.83	0.218
GPA tertile 3 [High-achievers]	0.15	-0.36 – 0.66	0.565
Interaction 1: Taxa [H] * GPA tertile [2]	0.01	-0.43 – 0.44	0.977
Interaction 2: Taxa [H] * GPA tertile [3]	0.20	-0.23 – 0.63	0.362
Random Effects			
σ^2	0.70		
$\tau_{00\ Deid}$	1.26		
ICC	0.64		
$N_{\ Deid}$	173		
Observations	346		
Marginal R^2 / Conditional R^2	0.012 / 0.647		

***Table S2.12. Mixed-effects linear model summary for Connectedness**. Bolded values are statistically significant.*

	Connectedness		
Predictors	*Estimates*	*CI*	*p*
(Intercept)	0.44	0.41 – 0.46	**<0.001**
Taxa [Human]	0.01	-0.01 – 0.03	0.527
GPA tertile 2 [Mid-achievers]	-0.04	-0.08 – -0.01	**0.026**
GPA tertile 3 [High-achievers]	-0.03	-0.07 – 0.01	0.091
Interaction 1: Taxa [H] * GPA tertile [2]	0.00	-0.03 – 0.03	0.796
Interaction 2: Taxa [H] * GPA tertile [3]	0.01	-0.02 – 0.04	0.528
Random Effects			
σ^2	0.00		
$\tau_{00\ Deid}$	0.01		
ICC	0.69		
$N_{\ Deid}$	173		
Observations	346		
Marginal R^2 / Conditional R^2	0.029 / 0.702		

***Table S2.13. Mixed-effects linear model summary for Ruggedness**. Bolded values are statistically significant.*

	Ruggedness		
Predictors	*Estimates*	*CI*	*p*
(Intercept)	1.03	0.96 – 1.11	**<0.001**
Taxa [Human]	-0.03	-0.12 – 0.05	0.428
GPA tertile 2 [Mid-achievers]	0.05	-0.05 – 0.15	0.319
GPA tertile 3 [High-achievers]	0.11	0.00 – 0.21	**0.044**
Interaction 1: Taxa [H] * GPA tertile [2]	0.02	-0.10 – 0.14	0.773
Interaction 2: Taxa [H] * GPA tertile [3]	-0.04	-0.16 – 0.08	0.556
Random Effects			
σ^2	0.05		
$\tau_{00\ Deid}$	0.02		
ICC	0.31		
$N_{\ Deid}$	173		
Observations	346		
Marginal R^2 / Conditional R^2	0.023 / 0.321		

REFERENCES

REFERENCES

Achér, A., Arcá, M., & Sanmartí, N. (2007). Modeling as a Teaching Learning Process for Understanding Materials: A Case Study in Primary Education. *Science Education*, 91(3), 398–418. https://doi.org/10.1002/sce

Alters, B. J., & Nelson, C. E. (2002). Perspective: Teaching Evolution in Higher Education. *Evolution*, 56(10), 1891–1901.

American Association for the Advancement of Science [AAAS]. (2011). *Vision and Change in Undergraduate Biology Education: a call to action*. http://visionandchange.org

Arnold, R. D., & Wade, J. P. (2015). A definition of systems thinking: A systems approach. *Procedia Computer Science*, 44(C), 669–678. https://doi.org/10.1016/j.procs.2015.03.050

Atran, S. (1998). Folk biology and the anthropology of science: cognitive universals and cultural particulars. *Behavioral and Brain Sciences*, 21(4), 547–569.

Atran, S., Medin, D., Lynch, E., Vapnarsky, V., Ucan Ek', E., & Sousa, P. (2001). Folkbiology doesn't Come from Folkpsychology: Evidence from Yukatek Maya in Cross-Cultural Perspective. *Journal of Cognition and Culture*, 1(1), 3–42. https://doi.org/10.1163/156853701300063561

Australian Curriculum Assessment and Reporting Authority [ACARA]. (2010). *The Australian Curriculum: Science*. https://www.australiancurriculum.edu.au/senior-secondary-curriculum/science/representation-of-general-capabilities/

Ausubel, D. G. (1963). Cognitive Structure and the Facilitation of Meaningful Verbal Learning. *Journal of Teacher Education*, 14(2), 217–222. https://doi.org/10.1177/002248716301400220

Barab, S. A., Hay, K. E., Barnett, M., & Keating, T. (2000). Virtual solar system project: Building understanding through model building. *Journal of Research in Science Teaching*, 37(7), 719–756. https://doi.org/10.1002/1098-2736(200009)37:7<719::AID-TEA6>3.0.CO;2-V

Barton, K. (2018). *MuMIn: Multi-Model Inference*. https://cran.r-project.org/package=MuMIn

Bates, D., Maechler, M., Bolker, B., & Walker, S. (2015). Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software*, 67(1), 1–48. https://doi.org/10.18637/jss.v067.i01.

Baze, C., & Gray, R. (2018). Two-Year Community: Modeling Tiktaalik: Using a Model-Based Inquiry Approach to Engage Community College Students in the Practices of Science During an Evolution Unit. *Journal of College Science Teaching*, 047(04). https://doi.org/10.2505/4/jcst18_047_04_12

Beggrow, E. P., & Sbeglia, G. C. (2019). Do disciplinary contexts impact the learning of evolution? Assessing knowledge and misconceptions in anthropology and biology students. *Evolution: Education and Outreach*, 12(1). https://doi.org/10.1186/s12052-018-0094-6

Ben-Zvi Assaraf, O., & Orion, N. (2005). Development of system thinking skills in the context of earth system education. *Journal of Research in Science Teaching*, 42(5), 518–560. https://doi.org/10.1002/tea.20061

Ben-Zvi Assaraf, O., & Orion, N. (2010). Four case studies, six years later: Developing system thinking skills in junior high school and sustaining them over time. *Journal of Research in Science Teaching*, 47(10), 1253–1280. https://doi.org/10.1002/tea.20383

Bennett, S., Gotwals, A. W., & Long, T. M. (2020). Assessing students' approaches to modelling in undergraduate biology. *International Journal of Science Education*, 1–18. https://doi.org/10.1080/09500693.2020.1777343

Bergan-Roller, H. E., Galt, N. J., Chizinski, C. J., Helikar, T., & Dauer, J. T. (2018). Simulated Computational Model Lesson Improves Foundational Systems Thinking Skills and Conceptual Knowledge in Biology Students. *BioScience*. https://doi.org/10.1093/biosci/biy054

Bierema, A. M.-K., Schwarz, C. V, & Stoltzfus, J. R. (2017). Engaging Undergraduate Biology Students in Scientific Modeling: Analysis of Group Interactions, Sense-Making, and Justification. *CBE—Life Sciences Education*, 16(4), ar68. https://doi.org/10.1187/cbe.17-01-0023

Bishop, B. A., & Anderson, C. W. (1990). Students conceptions of natural selection and its role in evolution. *Journal of Research in Science Teaching*, 27(5), 415–427.

Boulter, C. J., & Buckley, B. C. (2000). Constructing a Typology of Models for Science Education. *Developing Models in Science Education*, 41–57.

https://doi.org/10.1007/978-94-010-0876-1_3

Brandstädter, K., Harms, U., & Großschedl, J. (2012). Assessing System Thinking Through Different Concept-Mapping Practices. *International Journal of Science Education*, 34(14), 2147–2170. https://doi.org/10.1080/09500693.2012.716549

Bray Speth, E., Long, T. M., Pennock, R. T., & Ebert-May, D. (2009). Using Avida-ED for Teaching and Learning About Evolution in Undergraduate Introductory Biology Courses. *Evolution: Education and Outreach*, 2(3), 415–428. https://doi.org/10.1007/s12052-009-0154-z

Bray Speth, E., Shaw, N., Momsen, J., Reinagel, A., Le, P., Taqieddin, R., & Long, T. (2014). Introductory biology students' conceptual models and explanations of the origin of variation. *CBE Life Sciences Education*, 13(3), 529–539. https://doi.org/10.1187/cbe.14-02-0020

Brookhart, S. M. (1997). Effects of the classroom assessment environment on mathematics and science achievement. *Journal of Educational Research*, 90(6), 323–330. https://doi.org/10.1080/00220671.1997.10544590

Brown, D. S. (2003). High school biology: A group approach to concept mapping. *The American Biology Teacher*, 65(3), 192–197.

Brown, J. S., Collins, A., & Duguid, P. (2007). Situated Cognition and the Culture of Learning. *Educational Researcher*, 18(1), 32–42.

Bryce, C. M., Baliga, V. B., De Nesnera, K. L., Fiack, D., Goetz, K., Tarjan, L. M., Wade, C. E., Yovovich, V., Baumgart, S., Bard, D. G., Ash, D., Parker, I. M., & Gilbert, G. S. (2016). Exploring Models in the Biology Classroom. *The American Biology Teacher*, 8(1), 35–42. https://doi.org/10.1525/abt.2016.78.1.35

Casillas, A., Robbins, S., Allen, J., Kuo, Y. L., Hanson, M. A., & Schmeiser, C. (2012). Predicting Early Academic Failure in High School From Prior Academic Achievement, Psychosocial Characteristics, and Behavior. *Journal of Educational Psychology*, 104(2), 407–420. https://doi.org/10.1037/a0027180

Cassidy, S. (2012). Exploring individual differences as determining factors in student academic achievement in higher education. *Studies in Higher Education*, 37(7), 793–810. https://doi.org/10.1080/03075079.2010.545948

Catley, K. M., & Novick, L. R. (2009). Digging deep: Exploring college students' knowledge of macroevolutionary time. *Journal of Research in Science Teaching*,

46(3), 311–332. https://doi.org/10.1002/tea.20273

Cheng, M. F., Lin, J. L., Cheng, M. F., & Lin, J. L. (2015). Investigating the Relationship between Students' Views of Scientific Models and Their Development of Models. *International Journal of Science Education*, 37(15), 2453–2475. https://doi.org/10.1080/09500693.2015.1082671

Chi, M. T., Feltovich, P. J., & Glaser, R. (1981). Categorization and Representation of Physics Problems by Experts and Novices. *Cognitive Science*, 5(2), 121–152.

Christensen, R. H. B. (2019). *ordinal - Regression Models for Ordinal Data. R package version 2019.12-10*. https://cran.r-project.org/package=ordinal.

Clement, J. (1989). Learning via Model Construction and Criticism. In J. A. Glover, R. R. Ronning, & C. R. Reynolds (Eds.), *Handbook of Creativity* (pp. 341–381). Springer.

Clement, J. J., & Rea-Ramirez., M. A. (2008). *Model based learning and instruction in science, Volume 2*. Springer Science & Business Media.

Coley, J. D., & Tanner, K. (2015). Relations between intuitive biological thinking and biological misconceptions in biology majors and nonmajors. *CBE Life Sciences Education*, 14(1), 1–19. https://doi.org/10.1187/cbe.14-06-0094

Coll, R. K., France, B., & Taylor, I. (2005). The role of models/and analogies in science education: Implications from research. *International Journal of Science Education*, 27(2), 183–198. https://doi.org/10.1080/0950069042000276712

Constantinou, C. P., Nicolaou, C. T., & Papaevripidou, M. (2019). A Framework for Modeling-Based Learning, Teaching, and Assessment. In A. Upmeier zu Belzen, J. van Driel, & D. Krüger (Eds.), Towards a Competence-Based View on Models and Modeling in Science Education. *Models and Modeling in Science Education*, Vol 12. Springer, Cham. https://doi.org/10.1007/978-3-030-30255-9

Cooper, M. M., Caballero, M. D., Ebert-May, D., Fata-Hartley, C. L., Jardeleza, S. E., Krajcik, J. S., Laverty, J. T., Matz, R. L., Posey, L. A., & Underwood, S. M. (2015). Challenge faculty to transform STEM learning. *Science*, 350(6258), 281–282. https://doi.org/10.1126/science.aab0933

Csardi, G., & Nepusz, T. (2006). *The igraph software package for complex network research, InterJournal, Complex Systems 1695*. http://igraph.org.

Dauer, J. T., & Long, T. M. (2015). Long-term conceptual retrieval by college biology

majors following model-based instruction. *Journal of Research in Science Teaching*, 52(8), 1188–1206. https://doi.org/10.1002/tea.21258

Dauer, J. T., Momsen, J. L., Bray Speth, E., Makohon-Moore, S. C., & Long, T. M. (2013). Analyzing change in students' gene-to-evolution models in college-level introductory biology. *Journal of Research in Science Teaching*, 50(6), 639–659. https://doi.org/10.1002/tea.21094

de Lima, J. (2020). Contextual Influences on Undergraduate Biology Students' Reasoning and Representations of Evolutionary Concepts. [Unpublished doctoral dissertation]. Michigan State University.

diSessa, A. A., Gillespie, N. M., & Esterly, J. B. (2004). Coherence versus fragmentation in the development of the concept of force. *Cognitive Science*, 28(6), 843–900. https://doi.org/10.1016/j.cogsci.2004.05.003

Duschl, R. A., & Gitomer, D. H. (1997). Strategies and challenges to changing the focus of assessment and instruction in science classrooms. *Educational Assessment*, 4(1), 37–73. https://doi.org/10.1207/s15326977ea0401

Elias, S. M., & MacDonald, S. (2007). Using past performance, proxy efficacy, and academic self-efficacy to predict college performance. *Journal of Applied Social Psychology*, 37(11), 2518–2531. https://doi.org/10.1111/j.1559-1816.2007.00268.x

Emmons, N. A., & Kelemen, D. A. (2015). Young children's acceptance of within-species variation: Implications for essentialism and teaching evolution. *Journal of Experimental Child Psychology*, 139, 148–160. https://doi.org/10.1016/j.jecp.2015.05.011

Evagorou, M., Korfiatis, K., Nicolaou, C., & Constantinou, C. (2009). An investigation of the potential of interactive simulations for developing system thinking skills in elementary school: a case study with fifth-graders and sixth-graders. *International Journal of Science Education*, 31(5), 655–674.

Evans, E. M. (2008). Conceptual change and evolutionary biology: A developmental analysis. In S. Vosniadou (Ed.), *International Handbook of research on conceptual change* (pp. 263–294). Research in Science Education.

Fox, J. (2003). Effect Displays in R for Generalised Linear Models. *Journal of Statistical Software*, 8(15), 1–27. http://www.jstatsoft.org/v08/i15/

Freeman, S., Eddy, S. L., McDonough, M., Smith, M. K., Okoroafor, N., Jordt, H., &

Wenderoth, M. P. (2014). Active learning increases student performance in science, engineering, and mathematics. *Proceedings of the National Academy of Sciences*, 111(23), 8410–8415. https://doi.org/10.1073/pnas.1319030111

Freudenthaler, H. H., Spinath, B., & Neubauer, A. C. (2008). Predicting school achievement in boys and girls. *European Journal of Personality*, 22(3), 231–245. https://doi.org/10.1002/per.678

Gilbert, J. K. (2004). Models and Modelling: Routes To More Authentic Science Education. *International Journal of Science and Mathematics Education*, 2(2), 115–130. https://doi.org/10.1007/s10763-004-3186-4

Gilbert, J. K., Boulter, C. J., & Elmer, R. (2000). Positioning Models in Science Education and in Design and Technology Education. In John K. Gilbert & C. J. Boulter (Eds.), *Developing Models in Science Education*. Springer.

Gilbert, S. W. (1991). Model building and a definition of science. *Journal of Research in Science Teaching*, 28(1), 73–79. https://doi.org/10.1002/tea.3660280107

Gobert, J. D., O'Dwyer, L., Horwitz, P., Buckley, B. C., Levy, S. T., & Wilensky, U. (2011). Examining the relationship between students' understanding of the nature of models and conceptual learning in biology, physics, and chemistry. *International Journal of Science Education*, 33(5), 653–684. https://doi.org/10.1080/09500691003720671

Goel, A. K., & Stroulia, E. (1996). Functional device models and model-based diagnosis in adaptive design. *Artificial Intelligence for Engineering, Design, Analysis and Manufacturing*, 10(04), 355–370. https://doi.org/10.1017/S0890060400001669

Göransson, A., Orraryd, D., Fiedler, D., & Tibell, L. A. E. (2020). Conceptual Characterization of Threshold Concepts in Student Explanations of Evolution by Natural Selection and Effects of Item Context. *CBE Life Sciences Education*, 19(1), ar1. https://doi.org/10.1187/cbe.19-03-0056

Greca, I. M., & Moreira, M. A. (2000). Mental models, conceptual models, and modelling. *International Journal of Science Education*, 22(1), 11. https://doi.org/10.1080/095006900289976

Gregory, T. R. (2009). Understanding Natural Selection : Essential Concepts and Common Misconceptions. *Evolution Education Around the Globe*, 2(2), 156–175. https://doi.org/10.1007/s12052-009-0128-1

Gross, P., Buttrey, D., Goodenough, U., Koertge, N., Lerner, L. S., Schwartz, M., & Schwartz, R. (2013). *Final Evaluation of the Next Generation Science Standards*.

Grosslight, L., Unger, C., Jay, E., & Smith, C. L. (1991). Understanding models and their use in science: Conceptions of middle and high school students and experts. *Journal of Research in Science Teaching*, 28(9), 799–822. https://doi.org/10.1002/tea.3660280907

Grünkorn, J., zu Belzen, A. U., & Krüger, D. (2014). Assessing Students' Understandings of Biological Models and their Use in Science to Evaluate a Theoretical Framework. *International Journal of Science Education*, 36(10), 1651–1684. https://doi.org/10.1080/09500693.2013.873155

Haak, D. C., HilleRisLambers, J., Pitre, E., & Freeman, S. (2011). Increased Structure and Active Learning Reduce the Achievement Gap in Introductory Biology. *Science*, 332(6034), 1213–1216.

Halford, G. S. (1993). *Children's understanding: The development of mental models*. Lawrence Earlbaum Associates Inc.

Hall, R. (1996). Representation as Shared Activity : Situated Cognition and Dewey ' s Cartography of Experience. *The Journal of the Learning Sciences*, 5(3), 209–238.

Halldén, O. (1988). The evolution of the species : pupil perspectives and school perspectives. *International Journal of Science Education*, 10(5), 541–552.

Halloun, I. A. (2007). Mediated modeling in science education. *Science and Education*, 16(7–8), 653–697. https://doi.org/10.1007/s11191-006-9004-3

Hay, D., Kinchin, I., & Lygo-Baker, S. (2008). Making learning visible: The role of concept mapping in higher education. *Studies in Higher Education*, 33(3), 295–311. https://doi.org/10.1080/03075070802049251

Hmelo-Silver, C. E., & Azevedo, R. (2006). Understanding Complex Systems : Some Core Challenges. *The Journal of the Learning Sciences*, 15(1), 53–61.

Hmelo-Silver, C. E., Jordan, R., Eberbach, C., & Sinha, S. (2017). Systems learning with a conceptual representation : a quasi-experimental study. Instructional *Science,* 45(1), 53–72. https://doi.org/10.1007/s11251-016-9392-y

Hmelo-Silver, C. E., Marathe, S., & Liu, L. (2007). Fish Swim , Rocks Sit , and Lungs Breathe : Expert-Novice Understanding of Complex Systems. *The Journal of the*

Learning Sciences, 16(3), 307–331. https://doi.org/10.1080/10508400701413401

Hmelo-Silver, C. E., & Pfeffer, M. G. (2004). Comparing expert and novice understanding of a complex system from the perspective of structures, behaviors, and functions. *Cognitive Science*, 28(1), 127–138. https://doi.org/10.1016/S0364-0213(03)00065-X

Hsu, Y. S., Lin, L. F., Wu, H. K., Lee, D. Y., & Hwang, F. K. (2012). A Novice-Expert Study of Modeling Skills and Knowledge Structures about Air Quality. *Journal of Science Education and Technology*, 21(5), 588–606. https://doi.org/10.1007/s10956-011-9349-5

Hung, W. (2008). Enhancing systems-thinking skills with modelling. *British Journal of Educational Technology*, 39(6), 1099–1120. https://doi.org/10.1111/j.1467-8535.2007.00791.x

Ifenthaler, D. (2008). Practical solutions for the diagnosis of progressing mental models. In D. Ifenthaler, P. Pirnay-Dummer, & J. M. Spector (Eds.), *Understanding Models for Learning and Instruction* (pp. 43–61). Springer, Boston, MA. https://doi.org/10.1007/978-0-387-76898-4_3

Ifenthaler, D. (2010). Relational, structural, and semantic analysis of graphical representations and concept maps. *Educational Technology Research and Development*, 58(1), 81–97. https://doi.org/10.1007/s11423-008-9087-4

Ifenthaler, D. (2011). Identifying cross-domain distinguishing features of cognitive structure. *Educational Technology Research and Development*, 59(6), 817–840. https://doi.org/10.1007/s11423-011-9207-4

Ifenthaler, D., Masduki, I., & Seel, N. M. (2011). The mystery of cognitive structure and how we can detect it: Tracking the development of cognitive structures over time. *Instructional Science*, 39(1), 41–61. https://doi.org/10.1007/s11251-009-9097-6

Ifenthaler, D., Pirnay-Dummer, P., & Seel, N. (2007). The Role of Cognitive Learning Strategies and Intellectual Abilities in Mental Model Building Processes. *Technology, Instruction, Cognition and Learning*, 5(4), 353–366.

Inagaki, K., & Hatano, G. (2006). Young Children's Conception of the Biological World. *Current Directions in Psychological Science*, 15(4), 177–181. https://doi.org/10.1111/J.1467-8721.2006.00431.X

Jeppson, H., Hofmann, H., & Cook, D. (2018). *ggmosaic: Mosaic Plots in the "ggplot2"*

Framework. R package version 0.2.0. https://cran.r-project.org/package=ggmosaic

Jonassen, D., Strobel, J., & Gottdenker, J. (2005). Model building for conceptual change. *Interactive Learning Environments*, 13(1–2), 15–37. https://doi.org/10.1080/10494820500173292

Jones, M. G., Carter, G., & Rua, M. J. (2000). Exploring the development of conceptual ecologies: Communities of concepts related to convection and heat. *Journal of Research in Science Teaching*, 37(2), 139–159. https://doi.org/10.1002/(SICI)1098-2736(200002)37:2<139::AID-TEA4>3.0.CO;2-1

Kalinowski, S. T., Leonard, M. J., & Andrews, T. M. (2010). Nothing in Evolution Makes Sense Except in the Light of DNA. *CBE—Life Sciences Education*, 9(2), 87–97. https://doi.org/10.1187/cbe.09

Kampourakis, K., & Zogza, V. (2008). Students ' intuitive explanations of the causes of homologies and adaptations. *Science,* 17(1), 27–47. https://doi.org/10.1007/s11191-007-9075-9

Kampourakis, K., & Zogza, V. (2009). Preliminary evolutionary explanations: A Basic Framework for Conceptual Change and Explanatory Coherence in Evolution. *Science and Education*, 18(10), 1313–1340. https://doi.org/10.1007/s11191-008-9171-5

Kapteijn, M. (1990). The functions of organizational levels in biology for describing and planning biology education. In P. L. Lijnse, P. Licht, W. De Vos, & A. J. Waarlo (Eds.), *Relating Macroscopic Phenomena to Microscopic Particles: A Central Problem in Secondary Science Education* (pp. 139–150). CDB Press.

Kinchin, I. M. (2006a). Concept mapping, PowerPoint, and a pedagogy of access. *Journal of Biological Education*, 40(2), 79–83. https://doi.org/10.1080/00219266.2006.9656018

Kinchin, I. M. (2006b). Developing PowerPoint handouts to support meaningful learning. *British Journal of Educational Technology*, 37(4), 647–650. https://doi.org/10.1111/j.1467-8535.2006.00536.x

Kinchin, I. M., & Hay, D. B. (2007). The myth of the research-led teacher. *Teachers and Teaching: Theory and Practice*, 13(1), 43–61. https://doi.org/10.1080/13540600601106054

Kinchin, I. M., Hay, D. B., & Adams, A. (2000). How a qualitative approach to concept

map analysis can be used to aid learning by illustrating patterns of conceptual development. *Educational Research*, 42(1), 43–57. https://doi.org/10.1080/0013188800363908

Kohn, K. P., Underwood, S. M., & Cooper, M. M. (2018). Energy connections and misconnections across chemistry and biology. *CBE Life Sciences Education*, 17(1), 1–17. https://doi.org/10.1187/cbe.17-08-0169

Krell, M., Reinisch, B., & Krüger, D. (2015). Analyzing Students' Understanding of Models and Modeling Referring to the Disciplines Biology, Chemistry, and Physics. *Research in Science Education*, 45(3), 367–393. https://doi.org/10.1007/s11165-014-9427-9

Krell, M., Upmeier zu Belzen, A., & Krüger, D. (2012). Students ' Understanding of the Purpose of Models in Different Biological Contexts. *International Journal of Biology Education*, 2(2), 1–34.

Krell, M., Upmeier zu Belzen, A., & Krüger, D. (2014). Context-specificities in students`understanding of models and modelling: an issue of critical importance for both assessment and teaching. In C. P. Constantinou, N. Papadouris, & A. Hadjigeorgiou (Eds.), *Science Education Research for Evidence-based Teaching and Coherent Learning*. Proceedings of the ESERA 2013 Conference.

Kumar, S., Stecher, G., Suleski, M., & Hedges, S. B. (2017). TimeTree: A Resource for Timelines , Timetrees , and Divergence Times. *Molecular Biology and Evolution*, 34(7), 1812–1819. https://doi.org/10.1093/molbev/msx116

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2017). lmerTest Package: Tests in Linear Mixed Effects Models. *Journal of Statistical Software*, 82(13), 1–26. https://doi.org/10.18637/jss.v082.i13

Kwon, S. Y., & Cifuentes, L. (2009). The comparative effect of individually-constructed vs. collaboratively-constructed computer-based concept maps. *Computers and Education,* 52(2), 365–375. https://doi.org/10.1016/j.compedu.2008.09.012

Lee, S. W. Y., Chang, H. Y., & Wu, H. K. (2017). Students' Views of Scientific Models and Modeling: Do Representational Characteristics of Models and Students' Educational Levels Matter? *Research in Science Education*, 47(2), 305–328. https://doi.org/10.1007/s11165-015-9502-x

Lin, Y. T., Chang, C. H., Hou, H. T., & Wu, K. C. (2016). Exploring the effects of employing Google Docs in collaborative concept mapping on achievement, concept

representation, and attitudes. *Interactive Learning Environments*, 24(7), 1552–1573. https://doi.org/10.1080/10494820.2015.1041398

Lira, M. E., & Gardner, S. M. (2016). Structure-function relations in physiology education: Where's the mechanism? *Advances in Physiology Education*, 41(2), 270–278. https://doi.org/10.1152/advan.00175.2016

Liu, L., & Hmelo-Silver, C. E. (2009). Promoting complex systems learning through the use of conceptual representations in hypermedia. *Journal of Research in Science Teaching*, 46(9), 1023–1040. https://doi.org/10.1002/tea.20297

Long, T. M., Dauer, J. T., Kostelnik, K. M., Momsen, J. L., Wyse, S. A., Bray Speth, E., & Ebert-May, D. (2014). Fostering ecoliteracy through model- based instruction. *Frontiers in Ecology and the Environment*, 12(2), 138–139. https://doi.org/10.1890/1540-9295-12.2.138

Louca, L. T., & Zacharia, Z. C. (2012). Modeling-based learning in science education: Cognitive, metacognitive, social, material and epistemological contributions. *Educational Review*, 64(4), 471–492. https://doi.org/10.1080/00131911.2011.628748

Luckie, D., Harrison, S. H., & Ebert-May, D. (2011). Model-based reasoning: using visual tools to reveal student learning. *Advances in Physiology Education*, 35(1), 59–67. https://doi.org/10.1152/advan.00016.2010

Lüdecke, D. (2020). *sjPlot: Data Visualization for Statistics in Social Science*. https://cran.r-project.org/package=sjPlot

Manthey, S., & Brewe, E. (2013). Toward university modeling instruction-biology: Adapting curricular frameworks from physics to biology. *CBE Life Sciences Education*, 12(2), 206–214. https://doi.org/10.1187/cbe.12-08-0136

Meadows, D. H. (2008). *Thinking in Systems:A Primer*. Chelsea Green Publishing. https://doi.org/10.1017/CBO9781107415324.004

Moharreri, K., Ha, M., & Nehm, R. H. (2014). EvoGrader: an online formative assessment tool for automatically evaluating written evolutionary explanations. *Evolution: Education and Outreach*, 7(15), 1–14. https://doi.org/10.1186/s12052-014-0015-2

Morabito, N. P., Catley, K. M., & Novick, L. R. (2010). Reasoning about evolutionary history: Post-secondary students' knowledge of most recent common ancestry and

homoplasy. *Journal of Biological Education*, 44(4), 166–174. https://doi.org/10.1080/00219266.2010.9656217

Nakagawa, S., & Schielzeth, H. (2013). A general and simple method for obtaining R2 from generalized linear mixed-effects models. *Methods in Ecology and Evolution*, 4(2), 133–142. https://doi.org/10.1111/j.2041-210x.2012.00261.x

National Research Council [NRC]. (1996). *National Science Education Standards*. The National Academies Press. https://doi.org/https://doi.org/10.17226/4962

National Research Council [NRC]. (2012). *A Framework for K-12 Science Education: Practices, Crosscutting Concepts, and Core Ideas*. The National Academies Press.

Nehm, R. H., & Ha, M. (2011). Item feature effects in evolution assessment. *Journal of Research in Science Teaching*, 48(3), 237–256. https://doi.org/10.1002/tea.20400

Nehm, R. H., & Reilly, L. (2007). Biology majors' knowledge and misconceptions of natural selection. *Bioscience*, 57(3), 263–272. https://doi.org/10.1641/b570311

Nehm, R. H., & Ridgway, J. (2011). What Do Experts and Novices "See" in Evolutionary Problems? *Evolution: Education and Outreach*, 4(4), 666–679. https://doi.org/10.1007/s12052-011-0369-7

Nehm, R. H., & Schonfeld, I. S. (2008). Measuring knowledge of natural selection: A comparison of the CINS, an open-response instrument, and an oral interview. *Journal of Research in Science Teaching*, 45(10), 1131–1160. https://doi.org/10.1002/tea.20251

Nesbit, J. C., & Adesope, O. O. (2006). Learning with concept and knowledge maps: A meta-analysis. *Review of Educational Research*, 76(3), 413–448. https://doi.org/10.3102/00346543076003413

Nettle, D. (2010). Understanding of Evolution May Be Improved by Thinking about People. *Evolutionary Psychology*, 8(2), 205–228.

NGSS Lead States. (2013). *Next Generation Science Standards: For States, By States*. https://doi.org/10.17226/18290

Nicolaou, C. T., & Constantinou, C. P. (2014). Assessment of the modeling competence: A systematic review and synthesis of empirical research. *Educational Research Review*, 13, 52–73. https://doi.org/10.1016/j.edurev.2014.10.001

Novak, J. D., & Cañas, A. J. (2006). The origins of the concept mapping tool and the continuing evolution of the tool. *Information Visualization*, 5(3), 175–184. https://doi.org/10.1057/palgrave.ivs.9500126

Odenbaugh, J. (2005). Idealized, inaccurate but successful: A pragmatic approach to evaluating models in theoretical ecology. *Biology and Philosophy*, 20(2–3), 231–255. https://doi.org/10.1007/s10539-004-0478-6

Osbeck, L. M., & Nersessian, N. J. (2006). The distribution of representation. *Journal for the Theory of Social Behaviour*, 32(2), 141–160.

Ozdemir, G., & Clark, D. (2009). Knowledge structure coherence in Turkish students' understanding of force. *Journal of Research in Science Teaching*, 46(5), 570–596. https://doi.org/10.1002/tea.20290

Pearsall, N. R., Skipper, J. E. J., & Mintzes, J. J. (1997). Knowledge restructuring in the life sciences: A longitudinal study of conceptual change in biology. *Science* Education, 81(2), 193–215. https://doi.org/10.1002/(SICI)1098-237X(199704)81:2<193::AID-SCE5>3.0.CO;2-A

Plate, R. (2010). Assessing individuals' understanding of nonlinear causal structures in complex systems. *System Dynamics Review*, 26(1), 19–33. https://doi.org/10.1002/sdr

Pobiner, B., Beardsley, P. M., Bertka, C. M., & Watson, W. A. (2018). Using human case studies to teach evolution in high school A.P. biology classrooms. *Evolution: Education and Outreach*, 11(1). https://doi.org/10.1186/s12052-018-0077-7

Prevost, L. B., Knight, J., Smith, M. K., & Urban-Lurain, M. (2013). Student writing reveals their heterogeneous thinking about the origin of genetic variation in populations. *National Association for Research in Science Teaching.*

R Core Team. (2020). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. https://www.r-project.org/

Reinagel, A., & Bray Speth, E. (2016). Beyond the central dogma: Model-based learning of how genes determine phenotypes. *CBE Life Sciences Education*, 15(1), ar4. https://doi.org/10.1187/cbe.15-04-0105

Reiser, B. J., Black, J. B., & Abelson, R. P. (1985). Knowledge Structures in the Organization and Retrieval of Autobiographical Memories. *Cognitive Psychology*, 17(1), 89–137. https://doi.org/10.1016/0010-0285(85)90005-2

Royer, R., & Royer, J. (2004). Comparing hand drawn and computer generated concept mapping. *Journal of Computers in Mathematics and Science Teaching*, 23(1), 67–68.

Ruiz-Primo, M. A., & Shavelson, R. J. (1996). Problems and Issues in the Use of Concept Maps in Science Assessment. *Journal of Research in Science Teaching*, 33(6), 569–600. https://doi.org/10.1002/(SICI)1098-2736(199608)33:6<569::AID-TEA1>3.0.CO;2-M

Rumelhart, D. E., & Norman, D. A. (1978). Accretion, Tuning and Restructuring: three modes of learning. In R. L. Klatzky & J. W. Cotton (Eds.), *Semantic factors in cognition* (pp. 37–53). Lawrence Earlbaum.

Schurmeier, K. D., Atwood, C. H., Shepler, C. G., & Lautenschlager, G. J. (2010). Using item response theory to assess changes in student performance based on changes in question wording. *Journal of Chemical Education*, 87(11), 1268–1272. https://doi.org/10.1021/ed100422c

Schwarz, C. V., Reiser, B. J., Davis, E. A., Kenyon, L., Achér, A., Fortus, D., Shwartz, Y., Hug, B., & Krajcik, J. (2009). Developing a learning progression for scientific modeling: Making scientific modeling accessible and meaningful for learners. *Journal of Research in Science Teaching*, 46(6), 632–654. https://doi.org/10.1002/tea.20311

Schwarz, C. V. (2002). Is There a Connection? The Role of Meta-Modeling Knowledge in Learning with Models. Keeping Learning Complex: *The Proceedings of the Fifth International Conference of the Learning Sciences* (ICLS).

Seel, N. M. (2003). Model-centered learning and instruction. Technology, Instruction, Cognition and Learning, 1(1), 59–85.

Sekretariat der Ständigen Konferenz der Kultusminister der Länder in der Bundesrepublik Deutschland [KMK]. (2005a). *Bildungsstandards im Fach Biologie für den Mittleren Schulabschluss.* Wolters Kluwer. http://www.kmk.org/fileadmin/veroeffentlichungen_beschluesse/2004/2004_12_16-Bildungsstandards-Biologie.pdf

Sekretariat der Ständigen Konferenz der Kultusminister der Länder in der Bundesrepublik Deutschland [KMK]. (2005b). *Bildungsstandards im Fach Chemie für den Mittleren Schulabschluss*. Wolters Kluwer.

Sekretariat der Ständigen Konferenz der Kultusminister der Länder in der

Bundesrepublik Deutschland [KMK]. (2005c). *Bildungsstandards im Fach Physik für den Mittleren Schulabschluss.* Wolters Kluwer.

Shavelson, R. J. (1974). Methods for examining representations of a subject-matter structure in a student's memory. *Journal of Research in Science Teaching*, 11(3), 231–249. https://doi.org/10.1002/tea.3660110307

Shell, D. F., Brooks, D. W., Trainin, G., Wilson, K. M., Kauffman, D. F., & Herr, L. M. (2010). The Unified Learning Model. In *The Unified Learning Model*. Springer. https://doi.org/10.1007/978-90-481-3215-7

Shtulman, A. (2006). Qualitative differences between naïve and scientific theories of evolution. *Cognitive Psychology*, 52(2), 170–194. https://doi.org/10.1016/j.cogpsych.2005.10.001

Shtulman, A., & Schulz, L. (2008). The Relation Between Essentialist Beliefs and Evolutionary Reasoning. *Cognitive Science*, 32(8), 1049–1062. https://doi.org/10.1080/03640210801897864

Shute, V. J., & Zapata-Rivera, D. (2008). Using an Evidence-Based Approach to Assess Mental Models. In D. Ifenthaler, P. Pirnay-Dummer, & J. M. Spector (Eds.), *Understanding Models for Learning and Instruction* (pp. 23–41). Springer. https://doi.org/https://doi.org/10.1007/978-0-387-76898-4_2

Sinatra, G. M., Brem, S. K., & Evans, E. M. (2008). Changing Minds? Implications of Conceptual Change for Teaching and Learning about Biological Evolution. *Evolution: Education and Outreach*, 1(2), 189–195. https://doi.org/10.1007/s12052-008-0037-8

Sinatra, G. M., Southerland, S. A., McConaughy, F., & Demastes, J. W. (2003). Intentions and beliefs in students' understanding and acceptance of biological evolution. *Journal of Research in Science Teaching*, 40(5), 510–528. https://doi.org/10.1002/tea.10087

Smith, M. U. (2010a). Current status of research in teaching and learning evolution: I. Philosophical/epistemological issues. *Science and Education*, 19(6–8), 523–538. https://doi.org/10.1007/s11191-009-9215-5

Smith, M. U. (2010b). Current status of research in teaching and learning evolution: II. Pedagogical issues. *Science and Education*, 19(6–8), 539–571. https://doi.org/10.1007/s11191-009-9216-4

Sommer, C., & Lücken, M. (2010). System competence – Are elementary students able to deal with a biological system ? *Nordic Studies in Science Education*, 6(2), 125–143.

Spinath, B., Spinath, F. M., Harlaar, N., & Plomin, R. (2006). Predicting school achievement from general cognitive ability, self-perceived ability, and intrinsic value. *Intelligence*, 34(4), 363–374. https://doi.org/10.1016/j.intell.2005.11.004

Tripto, J., Ben Zvi Assaraf, O., & Amit, M. (2013). Mapping What They Know : Concept Maps as an Effective Tool for Assessing Students ' Systems Thinking. *American Journal of Operations Research*, 3(1), 245–258. https://doi.org/10.4236/ajor.2013.31A022

UK Department of Education. (2015). *National curriculum in England: science programmes of study.* https://www.gov.uk/government/publications/national-curriculum-in-england-science-programmes-of-study/national-curriculum-in-england-science-programmes-of-study

Upmeier zu Belzen, A., van Driel, J., & Krüger, D. (2019a). Introducing a Framework for Modeling Competence. In A. Upmeier zu Belzen, J. van Driel, & D. Krüger (Eds.), *Towards a Competence-Based View on Models and Modeling in Science Education. Models and Modeling in Science Education*, *Vol 12* (pp. 3–19). Springer, Cham. https://doi.org/10.1007/978-3-030-30255-9_1

Upmeier zu Belzen, A., van Driel, J., & Krüger, D. (Eds.). (2019b). *Towards a Competence-Based View on Models and Modeling in Science Education. Models and Modeling in Science Education, Vol 12*. Springer, Cham. https://doi.org/10.1007/978-3-030-30255-9

van Driel, J., Krüger, D., & Upmeier zu Belzen, A. (2019). Attainments and Challenges for Research on Modeling Competence. In A. Upmeier zu Belzen, J. van Driel, & D. Krüger (Eds.), *Towards a Competence-Based View on Models and Modeling in Science Education. Models and Modeling in Science Education, Vol 12* (pp. 311–321). Springer, Cham. https://doi.org/10.1007/978-3-030-30255-9_18

Vattam, S. S., Goel, A. K., Rugaber, S., Hmelo-Silver, C. E., Gray, S., & Sinha, S. (2011). Understanding Complex Natural Systems by Articulating Structure-Behavior- Function Models. *Journal of Educational Technology & Society*, 14(1), 66–81.

Verhoeff, R. P., Waarlo, A. J., & Boersma, K. T. (2008). Systems modelling and the

development of coherent understanding of cell biology. *International Journal of Science Education*, 30(4), 543–568. https://doi.org/10.1080/09500690701237780

Weston, M., Haudek, K. C., Prevost, L., Urban-lurain, M., & Merrill, J. (2015). Examining the Impact of Question Surface Features on Students ' Answers to Constructed-Response Questions on Photosynthesis. *CBE—Life Sciences Education*, 14(2), ar19. https://doi.org/10.1187/cbe.14-07-0110

White, B. Y. (1993). ThinkerTools: Causal Models, Conceptual Change, and Science Education. *Cognition and Instruction*, 10(1), 1–100. https://doi.org/10.1207/s1532690xci1001_1

White, B. Y., Collins, A., & Frederiksen, J. R. (2011). The Nature of Scientific Meta-Knowledge. In M. S. Khine & I. M. Saleh (Eds.), *Models and Modeling: Cognitive Tools for Scientific Enquiry (Vol. 6)* (pp. 41–76). Springer Science & Business Media.

Wickham, H. (2016). *ggplot2: Elegant Graphics for Data Analysis.* Springer-Verlag New York.

Wickham, Hadley. (2019). *stringr: Simple, Consistent Wrappers for Common String Operations.* R package version 1.4.0. https://cran.r-project.org/package=stringr

Wickham, Hadley, François, R., Henry, L., & Müller, K. (2020). *dplyr: A Grammar of Data Manipulation*. R package version 0.8.5. https://cran.r-project.org/package=dplyr

Wickham, Hadley, & Henry, L. (2020). *tidyr: Tidy Messy Data. R package version 1.0.3.* https://cran.r-project.org/package=tidyr

Williams, M. D., & Hollan, J. D. (1981). The process of retrieval from very long-term memory. *Cognitive Science*, 5(2), 87–119. https://doi.org/10.1207/s15516709cog0502_1

Wilson, K. J., Long, T. M., Momsen, J. L., & Bray Speth, E. (2019). Evidence Based Teaching Guide: Modeling in Classroom. *CBE Life Science Education.* https://lse.ascb.org/evidence-based-teaching-guides/modeling-in-the-classroom/

Wilson, K. J., Long, T. M., Momsen, J. L., & Bray Speth, E. (2020). Modeling in the Classroom: Making Relationships and Systems Visible. *CBE Life Sciences* Education, 19(1), fe1. https://doi.org/10.1187/cbe.19-11-0255

Windschitl, M., Thompson, J., & Braaten, M. (2008). Beyond the scientific method: Model-based inquiry as a new paradigm of preference for school science investigations. *Science Education*, 92(5), 941–967. https://doi.org/10.1002/sce.20259

CHAPTER THREE:

Model-based and narrative assessments elicit different ideas about evolution by natural selection.

ABSTRACT

Narrative responses are commonly used for assessing students' reasoning, but models are increasingly represented in college biology classrooms. Features of student-constructed models can provide insights into thinking and reasoning that are not captured in multiple choice or narrative responses. However, little is known about whether the two modes of response are equivalent in terms of eliciting students' ideas. In this study we explored the contextual influence of response mode on the content of students' explanations about evolution by natural selection.

We asked students in two sections of a large-enrolment introductory biology course to respond to prompts about evolution by natural selection by constructing both a model and written narrative. We used qualitative content analysis to develop a rubric for analysing the content of student responses. Responses were binned into levels of scientific plausibility that reflect inclusion of Key Concepts (KC), Naïve Ideas (NI), and Threshold Concepts (TC) that have been reported in research on evolution learning. We then used mixed-effects multiple ordinal logistic regressions and multiple logistic regressions to assess whether the mode of representation (model vs narrative) affected the probability of finding evolutionary concepts in student responses. Additionally, we assessed whether students similarly represented concepts across all their responses,

and if their prior academic performance was associated with consistency in the content of their responses.

We found that mode influenced the content of responses in various ways. Students' narratives were more likely to include the KCs Limited Resources and Competition ($p < 0.001$), and Differential Survival and Reproduction ($p < 0.01$), and the TC Randomness ($p < 0.001$), but were also more likely to contain NIs (Teleological ideas, $p < 0.001$). Students' models, however, were more likely to include the TC Probability ($p < 0.001$) and the KC Variation ($p = 0.69$). Other KCs, such as Heritability, were elicited no more frequently in narratives or models. We did not find any evidence to support the claim that mode of response influenced the consistency with which ideas were included in students' responses, however there was some association detected between prior achievement and consistency in the ideas included.

Our findings suggest that mode of response can influence the ideas elicited from students and in turn bias our interpretation of students' understanding of evolution by natural selection. Incorporating multiple modes of assessment has potential to generate a more holistic view of students' understanding and may promote greater transfer by requiring students to think and reason across contexts.

INTRODUCTION

Representing Knowledge

Knowledge is both internally stored and externally expressed through representations that organise and communicate knowledge about a particular concept (Daniel et al., 2018). The term 'representation' can have multiple meanings (Brachman et al., 2004; Pande & Chandrasekharan, 2017). It has been used to refer to internal representations (Ifenthaler, 2010), external representations (Gilbert, 2004; Ifenthaler, 2008), or both internal and external representations (Daniel et al., 2018; Paivio, 1990). Although researchers and educators are most interested in students' internal representations, these are not easily accessible. We therefore use students' external representations as windows into their internal representations (Daniel et al., 2018; Ifenthaler, 2008; Ifenthaler et al., 2011). Some researchers consider verbal representations an appropriate form of externalisation (Gilbert, 2004; Paivio, 1990; Tsui & Treagust, 2013) while others restrict their definition to 'pictorial and graphical descriptions of phenomena' (Schonborn & Anderson, 2009). In this study we will adopt a broad definition of external representation that includes verbal and text-based modes as well as graphical/pictorial.

Students' externalised responses to prompts give us a glimpse into their Cognitive Structure (CS) – an internal representation. CSs are features of long-term memory and act as repositories and hierarchical organisations of conceptual knowledge (Ausubel, 1963; Dauer & Long, 2015; Shell et al., 2010). As students learn more about a concept, their CS changes by adding new and relevant information about that concept, organising and reorganising links between pieces of information, and by pruning

irrelevant or erroneous information and links (Ifenthaler et al., 2011; Shavelson, 1974). Stable and well-organised CSs can facilitate learning, whereas the opposite type of CSs can inhibit learning (Ausubel, 1963). Information in the CS can be stored in the form of verbal representations (Clariana et al., 2014), visual representations, or as both visual and verbal representations (Paivio, 1990). A CS that has multiple types of representations will offer multiple access points and could enhance retrieval (Paivio, 1990; Schnotz & Bannert, 2003).

When responding to a prompt on an assessment, students access their long-term memory for a relevant CS by taking cues from a prompt and using these cues as specific access points. Dauer & Long (2015) identified three possible types of cues that students might use to access their CS: (i) the context of the prompt (i.e., the scenario or subject matter being described in the prompt, such as evolution of an insect population by natural selection), (ii) the specific task required (i.e., what the students are being asked to do, such as construct a model or write an essay), and (iii) specific words in the prompt (i.e., concepts that are included in the prompt, such as gene, allele, evolution).

Using these access points to their CS, students then build a relevant mental model (Dauer et al., 2013; Ifenthaler, 2008; Johnson-Laird, 1983). Given the same prompt, different people will produce different mental models based on both their existing CS and the cues they used to access it (Hmelo-Silver et al., 2007; Johnson-Laird, 1983). The mental model lives in short term memory, and unlike the CS, is ephemeral (Johnson-Laird, 1983; Shell et al., 2010). Additionally, the mental model is not a

complete representation of their CS and may not be entirely accurate in order to be useful to the task at hand (Johnson-Laird, 1983).

This mental representation (the mental model) can now be externally expressed as a response to a prompt through various modes (Dauer et al., 2013; Ifenthaler, 2008). However, during the process of building an external representation, students will continue to access their CS and modify/update their mental model (Lewis-Peacock & Postle, 2008). Gilbert (2004) describes five main modes of external representations: concrete, verbal, symbolic, visual, and gestural. Paivio (1990) used a more simplistic differentiation – they classified external representations as being 'language-like' or 'picture-like'. In this study, we consider two modes of external representations: Narrative responses ('verbal' according to the former classification and 'language-like' according to the latter), and model-based responses ('visual' according to the former and 'picture-like' according to the latter).

Multiple Modes of Representation in Learning and Assessment

Much research has been directed at understanding the effects of providing multiple modes of representations on student learning (Ainsworth, 2006; Goldman, 2003; Schnotz & Bannert, 2003; Schonborn & Anderson, 2009; Someren et al., 1998; Tsui & Treagust, 2013). In particular, various researchers have shown that learning is enhanced when the same information is presented using multiple modes (Cox, 1999; Jaipal, 2010; Mayer, 2003; Schnotz, 2002; Schnotz & Bannert, 2003; Wu & Puntambekar, 2012). By using multiple modes, students have the potential to make

new inferences by comparing modes (Gentner & Markman, 1997), enhance the robustness and flexibility of their existing CS (Ainsworth et al., 2002; Nesbit & Adesope, 2006), and facilitate problem solving (de Jong et al., 1998). Ainsworth (2006) argues that using multiple modes of representation allows for integration of information which, in turn, leads to a deeper understanding of the concept. This deeper understanding can then facilitate transfer of knowledge to other unknown but relevant contexts (Ainsworth, 2006). Chandler & Sweller (1992), however, argue that providing multiple modes might actually be detrimental by increasing cognitive load. This detriment can be ameliorated by ensuring that the modes are well-integrated, thereby reducing the cognitive load and enhancing the potential for learning and transfer.

The ability of a person to use multiple modes of representation “to make sense of and communicate understanding” is called representational competence (Daniel et al., 2018). A person who has representational competence will be able to both receive and convey equivalent information using different representations and/or representational modes (Kozma & Russell, 1997; Shafrir, 1999). Representational competence includes representational fluency, which is the ability to use multiple modes at the same time, and “to seamlessly move within and between” them (Daniel et al., 2018). Representational competence and fluency, i.e., being able to use, create, and manipulate multiple representations of the same concept, has been linked to expertise in the field (Ainsworth et al., 2002; Brenner et al., 1999; Kozma et al., 2000; Larkin et al., 1980; National Research Council [NRC], 2000; Pande & Chandrasekharan, 2017; Schonborn & Anderson, 2009).

Other studies have explored students' choices and preferences with respect to mode of representation in response to specific tasks. For example, students preferred using tables and graphs to respond to contextualised mathematical problems, but preferred equations for non-contextualised problems (Keller & Hirsch, 1998). The same study also found that personal preferences for a particular representation influenced students' choices. Çikla & Çakiroğlu (2006) similarly reported that while some students chose the same preferred mode of representation in response to all prompts, others chose modes depending on the context of the prompt. Negative preference for a particular mode (specifically, using number lines) to express fractions was reported by a majority of participants in a study conducted by Biber (2014). When given the option to use more than one mode, almost 90% of students in a study chose to use only one mode (Yerushalmy, 1991). Chi, Feltovich, & Glaser (1981) posit students' choices of representation for a concept can tell us about the way that particular concept is encoded in their CSs.

Existing literature, including that mentioned in this introduction, has added to our understanding of how students use multiple representations during learning and the affordances of different representations in learning and instruction. Researchers have also compared modes of representation provided during assessments (Ainsworth & Th Loizou, 2003) and discussed pros and cons of each. There is, however, a paucity of literature exploring the effect of representational mode used by the student on eliciting and assessing students' knowledge. This gap has been noted by some researchers who have called for investigations into the way students interact with representations,

including the ways students use and develop them. For example, Daniel et al. (2018) posit that increasing our understanding of how students interact with different modes of representation will help us to move them along the continuum of novice to expert. Ainsworth et al. (2002) specifically proposed using pairs of representations to increase our understanding about how different modes can affect what and how content is elicited in students' responses.

Research Questions

This study addresses the aforementioned gap in the literature by asking two questions: (1) How does representational mode influence the conceptual content of students' responses to a prompt about evolution by natural selection? And, (2) to what extent do students elicit and represent equivalent knowledge across modes? Here, we specifically compare two modes of representation – narratives and models – that have been extensively researched as ways to elicit students' knowledge and Cognitive Structures (O. R. Anderson & Demetrius, 1993; Dauer & Long, 2015; K. Anders Ericsson & Simon, 1998; Hmelo-Silver et al., 2007; Ifenthaler et al., 2011; Koubek & Mountjoy, 1991; Nesbit & Adesope, 2006; Tsai & Huang, 2002). In addition to exploring the influence of response mode, this study builds on previous findings to further examine influences of prompt contextual features (e.g., organismal taxon) as factors predicting the nature of students' responses to questions about evolution by natural selection (de Lima, 2020 p 12 and p 68). Our study therefore explores the way context – defined as both mode of response and prompt contextual features – influences the ideas elicited and represented in students' responses. Although we will classify specific

concepts as ‘key’ or ‘naïve’, our goal is not to characterize correctness or quantify students’ knowledge about evolution, but rather to describe the conceptual content of students’ constructed representations. We assume that both narratives and models are reflections of a student’s mental model and are therefore incomplete approximations of their CS (Daniel et al., 2018; Dauer et al., 2013).

METHODS

Setting and Participants

This study was conducted at a large, public university in the Midwest with highest research activity (The Carnegie Classification of Institutions of Higher Education, n.d.). Data for these analyses came from student responses in two sections of a large introductory biology course for majors (n = 384) that focused on content domains of genetics, evolution, and ecology. The course is second in a 2-course sequence required for life science majors; the first focused on cell and molecular biology. The course is designed for sophomores, but also has a high proportion of juniors (34%) and some freshman (3%) and seniors (8%). Data were collected from two different semesters (n = 190 and 194) and the two sections had different instructors. Both instructors use model-based instruction and students got practice in constructing and revising models based on feedback.

Assessment Design

We used a previously developed 'Human/Cheetah Assessment' (HCA) described in in (de Lima, 2020, p 12), which was developed based on the ACORNS instrument (Assessing COntextual Reasoning about Natural Selection; Nehm, Beggrow, Opfer, & Ha, 2012). ACORNS items were developed to elicit students' reasoning about evolution by natural selection using multiple contexts and have been shown to give valid and reliable insight into student thinking. HCA items follow the same structure as ACORNS items, but were designed to more narrowly explore the role of context. In this case, context varied as taxon of organism (humans vs non-human animals) and trait type (functional vs structural morphological traits).

For this study, students responded to two prompts which differed only in the taxon referenced (Humans vs Cheetahs). The following is an example of two of the prompts in the HCA. For each prompt, students were asked to construct their responses using two modes of representation (model and narrative) for a total of 4 responses per student (Table 3.1). In both sections, the HCA was administered toward the end of the semester during class as part of routine in-class assessment. Students received participation points on a scale of 1-3 based on perceived effort.

Table 3.1. Example Human/Cheetah Assessment prompts. *Pairs of prompts were constructed to examine the effect of taxon (human vs non-human animal) and mode (model vs narrative) on students' explanations of evolution by natural selection.*

Human Prompt	Cheetah Prompt
1. Modern humans have enlarged heels. How would biologists explain how a species of humans with enlarged heels evolved from an ancestral human species without enlarged heels? 1a. *Construct a model* that answers the question. 1b. *Construct a short essay* that answers the question.	2. A species of cheetah has long leg bones. How would biologists explain how a species of cheetah with long leg bones evolved from an ancestral cheetah species without long leg bones? 2a. *Construct a model* that answers the question. 2b. *Construct a short essay* that answers the question.

We tested for potential effects of prompt order (Federer et al., 2015; Schuman & Presser, 1996) by having multiple versions of the assessment (i.e., 'forms'). Forms differed in the order of the taxon (Human or Cheetah) or mode (model or essay) presented to students. Order of mode was kept the same in forms testing the effect of taxon order; taxon order was kept the same in forms testing the effect of mode order.

Data Processing

Selecting Data

Of a total of 384 students in the two class sections, 213 students attempted all four tasks (i.e., constructed one model and wrote one narrative for each of the two taxa) and were included in analyses. To ensure that the sub-population included in the analysis was representative of the entire population, we used a multiple logistic regression to assess how well demographic variables (Gender, Ethnicity, First Generation Learner,

Rank) and prior academic performance (GPA at the start of the semester) predicted whether students were included (Table S3.1). The sub-population included in the analysis was very similar to the sub-population excluded for most of the variables (Table 3.2). However, the sub-population excluded from the analysis had a lower mean GPA ($p < 0.05$), though the difference was driven by a few students with extremely low GPA and was marginal in terms of effect size.

Table 3.2. Demographic characteristics and prior academic achievement of the two student sub-populations (included and excluded from the study), and the total student population*. STEM credits are the number of STEM credits completed at the beginning of the semester. Start GPA is the cumulative GPA of the students (based on a 4-point system) at the beginning of the semester.*

	Students included in the analysis	**Students excluded from the analysis**	**Students in the course**
Gender (% Female)	60.1%	63.7%	61.7%
Ethnicity (% white non-Hispanic)	81.6%	68.4%	75.7%
First Generation Learner (%)	24.8%	28%	26.3%
Class Rank (% Sophomore/Junior)	54%/33%	52%/33%	53%/33%
STEM credits (mean)	46.29	48.29	47.15
Start GPA (mean)	3.3	3.2	3.3

Coding Responses

We developed a set of coding rubrics to assess and compare the conceptual content of students' responses related to their understanding of evolution by natural selection. As a guiding framework, we used 6 Key Concepts (KCs: Variation, Heritability, Competition, Limited Resources, Differential Survival, and Non-adaptive) and 3 Naïve Ideas (NIs: Adapt, Need, and Use/Disuse) described by Moharreri, Ha, & Nehm (2014). We coded responses in 2 phases: Emergent coding and Condensed coding.

Phase 1: Emergent Coding

We developed an emergent coding rubric using qualitative content analysis (Schreier, 2014) to capture qualitative attributes of students' responses related to their knowledge about natural selection. Table 3.3 shows the categories and guiding questions we used to unpack the content of responses and to develop codes.

Table 3.3. Categories and questions used in the process of qualitative content analysis to build the emergent coding rubric*. Key Concepts (KCs) related to students' understanding of evolution by natural selection informed our development of rubric categories. Naïve ideas (NIs) are included in our approach as they are applied to other categories.*

Category	Questions
Variation	• Is Variation pre-existing or caused? • What is the cause of the Variation? • What is the level at which the Variation occurs? • What are the consequences of the Variation • Other ideas related to Variation?

Table 3.3 (cont'd)

Differential Survival and Reproduction	• Is there a link between the trait and survival of the organism/population? • Is there a link between the trait and reproduction of the organism/population? • Does selection act on the trait?
Limited Resources and Competition	• Does the trait influence competition? • Does the trait lead to differential access to food/resources? • Does the trait lead to differential interactions with predators/prey? • Does the trait offer any additional benefits (e.g. abilities)?
Heritability	• What is the unit of inheritance? • Who are the inheritors? • What is the mechanism of inheritance? • Does the response refer to differential probabilities with respect to inheritance? • Any other ideas with respect to inheritance?
Holistic	• Does the response have an accurate explanation for the origin of variation, including the mechanism and the level? • Other ideas not captured in the previous categories?

A subset of responses (n = 94) was used to develop the emergent coding rubric. To ensure a representative sample of student responses across achievement levels, we randomly selected responses across tertile bins based on GPA at the start of class. We also ensured that the responses equally represented both taxa. Two independent raters reviewed 10 responses and assigned tags to specific conceptual ideas reflective of categories described in Table 3.3. The two coders then met and discussed each tag and the relevant information it encoded. After reaching consensus, relevant tags became codes. Iterative bouts of independent coding and discussion continued until no new codes emerged and an acceptable IRR had been achieved (> 75% agreement and > 0.7 for Cohens Kappa; Table 3.4).

Table 3.4. Inter-Rater Reliability values during the last round of iterative independent coding by two raters indicate achievement of an acceptable IRR. *The table gives the range of IRR values for the whole category (each category had multiple codes), as well as the mean IRR for the whole category. * Cohens K is not meaningful when the data is highly skewed.*

Category	% Agreement		Cohens K	
	Range	Mean	Range	Mean
Variation	81.25 - 100	95.39	0.38* - 1	0.81
Differential Survival and Reproduction	81.25 - 100	95.71	0.33* - 1	0.82
Limited Resources and Competition	90.32 - 100	97.02	0.61 - 1	0.93
Heritability	75 - 100	92.14	0.30* - 1	0.74
Holistic	75 - 100	95.71	0.45* - 1	0.89

The final codebook for the emergent phase includes 59 codes in 5 categories (Tables S3.2 – S3.6). Responses were coded for presence (1) or absence (0) of each code. Two raters used the codebook to code 852 responses where a 'response' represents a unique combination of taxon and mode. Only students who attempted all 4 tasks (one model and one narrative for each of the two taxa) were included in analyses (n = 213 students). Data obtained from the emergent phase were then analysed in the condensed coding phase.

Phase 2: Condensed Coding

In this phase, we developed a rubric to quantify the scientific plausibility of students' ideas. Each of the four KCs (Variation, Differential Survival and Reproduction, Limited Resources and Competition, and Heritability) was ranked on a scale from scientifically implausible/inaccurate to scientifically plausible/accurate. The number of levels varies among KCs and is based on the variation we observed in students' responses (Table 3.5).

Logic Statements:

To bridge the emergent coding rubric (59 distinct presence/absence codes) and the condensed coding rubric (levels of scientific plausibility), we developed a series of logic statements that used the codes to bin students into levels of scientific plausibility for each of the KCs. Similar logic statements were also developed to determine the presence/absence of the 3 NIs (Need, Use, Adapt) and to assess the presence of Threshold Concepts (TCs) related to evolution (Göransson et al., 2020). For example,

we used logic statements to determine the presence/absence of ideas related to Probability, Randomness, and Level of Biological Organisation (e.g., Genetic, Organismal, and Population).

Table 3.5. Condensed coding rubric with examples of student model and narrative responses.

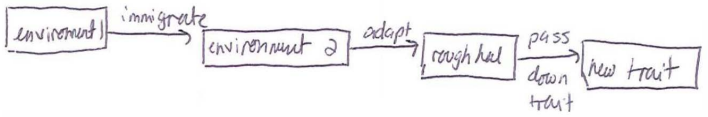
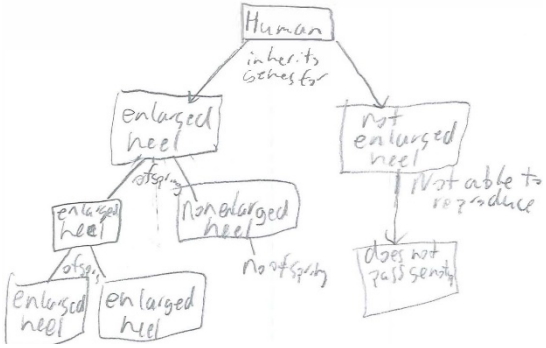
Category	Levels	Description	Example
Variation	Absent	No evidence of Variation	*The phenotype traits expressed by long leg boned cheetahs shows evolution. The ancestors without long leg bones need to adapt to their environment and so the offspring of the ancestors evolved to longer legs.* 
	Level 1	Criteria: Response includes variation at the genotypic or phenotypic level. Evidence: • Reference to variation in traits • Reference to variation in genes • Reference to variation in the organisms	*The switch from slow cheetahs to fast cheetahs took many thousands of years. The switch would of began by a mutation. Then members of the population would select sexually for that mutation. Eventually the entire population would be fast.* 

Table 3.5. (Cont'd)

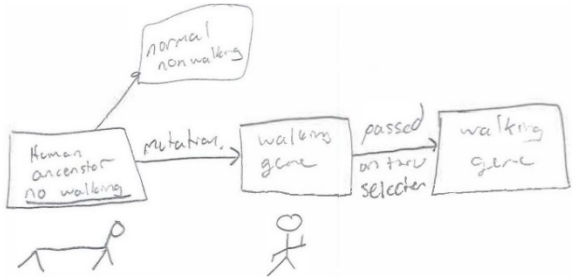
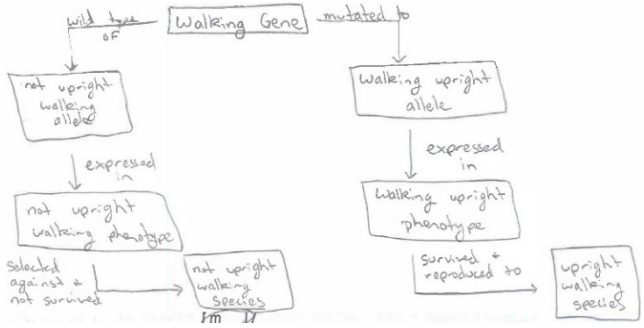
Variation	Level 2	Criteria: • Variation originates through a **mutation/ change.** • This mutation/change originates at the **genetic level** • Mutation/change causes **phenotypic variation or genotypic variation**	*Biologists would argue that long leg cheetahs evolved through selection. A variation in cheetah genetic code may have originally resulted in long legs. This variation may have proved beneficial to the species because longer legs helps them run faster. As a result, more long leg cheetahs are born because it is more fit to survive.* 
	Level 3	Criteria: • Mutation leads to **the formation of a genetic entity (that can be inherited)** • The **new genetic entities lead to phenotypic variation**	*Ancestral cheetahs had short lef bones. A mutation occurs to bring about an allele that causes the long leg bone phenotype. These cheetahs had a higher fitness than their short legged counterparts. Over time, short legged cheetas were removed from the gene pool by natural selection.* 

Table 3.5. (Cont'd)

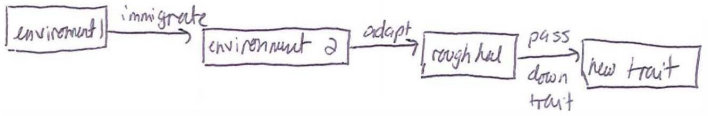
Differential Survival and Reproduction	Absent	No evidence of differential survival and/or reproduction	*The phenotype traits expressed by long leg boned cheetahs shows evolution. The ancestors without long leg bones need to adapt to their environment and so the offspring of the ancestors evolved to longer legs.* 
	Level 1	Criteria: Response indicates trait leads to some sort of benefit Evidence: • Differential ability to inhabit the environment • Possibility of selection pressure	*A mutation that was once present allowed for some individual cheetahs to run faster than others. They were better fit for the environment than most others and were able to survive the best. This trait was then inherited by the offspring and passed down as this was a favored characteristic that benefitted the cheetahs.* 

Table 3.5. (Cont'd)

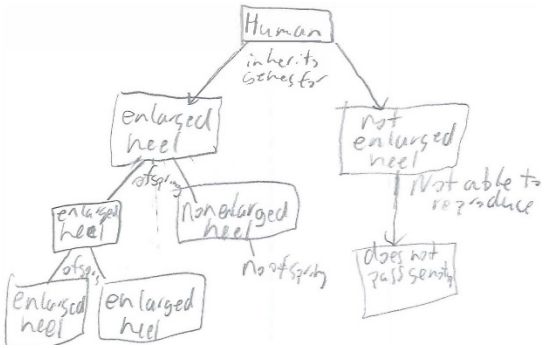
Differential Survival and Reproduction	Level 2	Criteria: Response identifies what benefit is afforded by the trait (with respect to survival and reproduction) Evidence: • Response indicates that the trait leads to differential survival and/or reproduction either for the animal and/or for their progeny • Correctly use and unpacking of terminology (e.g. fitness, selection)	*A random mutation gives an individual the ability to function on two legs. This makes getting food easier, giving them a higher fitness. They pass this trait on to their offspring who will have an advantage. This will allow them to better survive and reproduce causing the frequency of the trait in the population.* 

Table 3.5. (Cont'd)

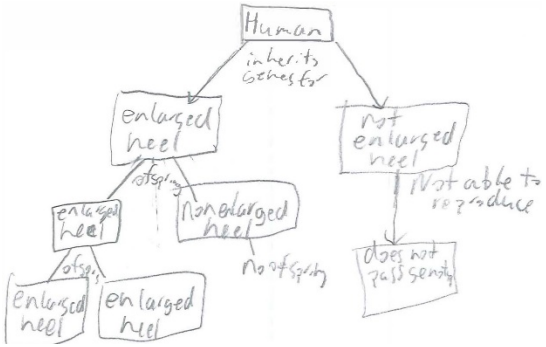
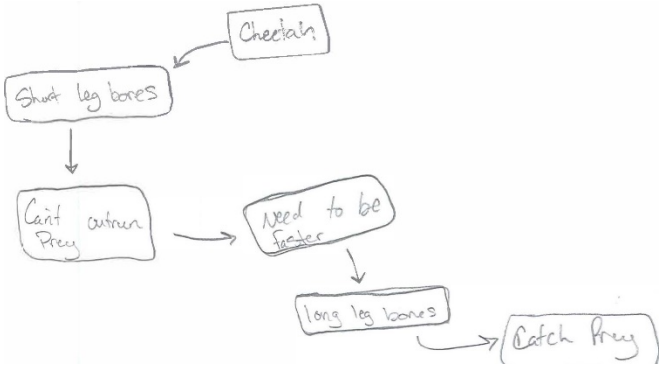
Limited Resources and Competition	Absent	No evidence of Competition and Limited Resources	*The phenotype traits expressed by long leg boned cheetahs shows evolution. The ancestors without long leg bones need to adapt to their environment and so the offspring of the ancestors evolved to longer legs.* 
	Level 1	Criteria: Responses refer to ideas with respect to Competition and Limited Resources. Evidence: • Differential access to food • Differential access to other resources • Differential interactions with predators/prey	*A mutation caused long leg bones. This made it so the cheetahs could run faster and catch all the prey. They outcompeted the competition.* 

Table 3.5. (Cont'd)

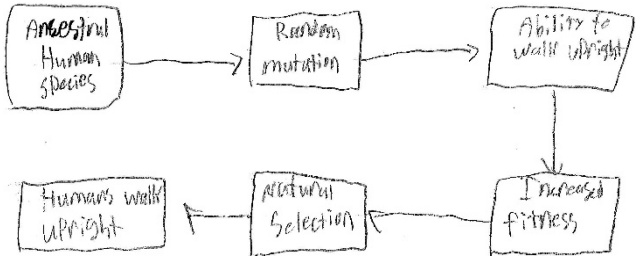
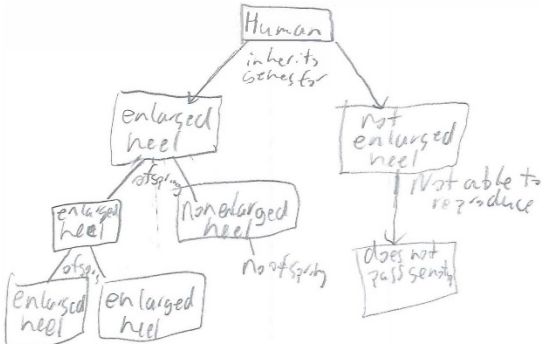
Heritability	Absent	No evidence of Heritability	*The phenotype traits expressed by long leg boned cheetahs shows evolution. The ancestors without long leg bones need to adapt to their environment and so the offspring of the ancestors evolved to longer legs.* 
	Level 1	Criteria: Response indicates transfer of something from parent to offspring, or across generations Evidence: • Reference to inheritance of traits or mutations	*Cheetahs with the ability to run fast evolved because they were more fit than the slow cheetahs. Their speed allowed them to catch prey and survive longer, allowing them to reproduce. The slow cheetahs couldn't catch food and eventually starved to death. The fast cheetahs passed along their speed trait to their offspring.* 

Table 3.5. (Cont'd)

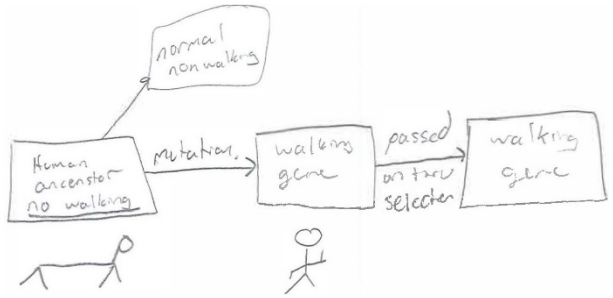
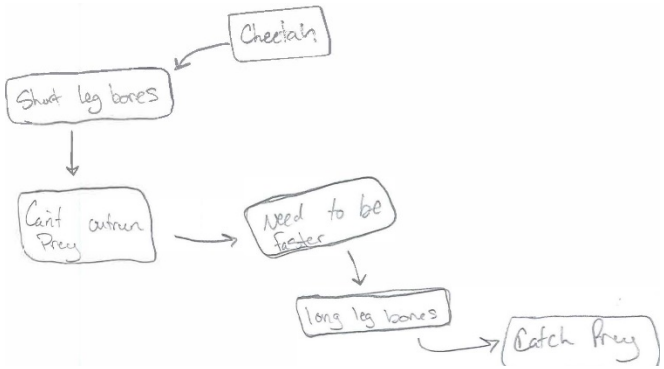
Heritability	Level 2	Criteria: Response indicates transfer of information or material at the genetic level from parent to offspring, or across generations Evidence: • Reference to inheritance of alleles or genes	*Cheetahs leg size represents the theory of selection. They have evolved from short legs to long legs. In the past, there must have been a mutation that occurred allwoing a cheetah to have longer legs. With longer legs, that cheetah can run faster to his prey and therefor have a better chance of survival and to pass down his long legged alleles rather than that short legged cheetah.* 
Naïve Ideas	Need	Responses demonstrate teleological thinking. (Teleological ideas were mostly associated with variation and/or heritability)	*Humans began to walk upright rather than being hunch-backed due to environmental influences of being and looking civilized, and having to do activities that required them to model good posture.* 

Table 3.5. (Cont'd)

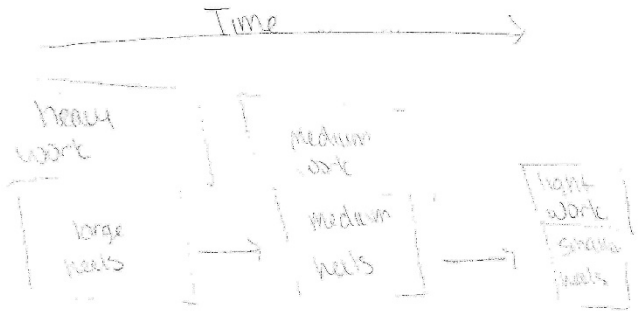
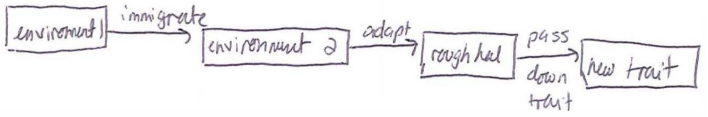
Naïve Ideas	Use	Responses that have Lamarckian ideas about evolution.	*Humans began walking upright which caused our heels to get bigger.* 
	Adapt	Responses that referenced adaptation happening in the lifetime of an organism. (These beneficial adaptations could then be passed on to offspring)	*Biologist's first explanation would be that a random mutation occurred within one of the ancestral cheetah that caused the long leg bones. This could have been enabled the cheetah to be better able to adapt in the environment. As the species that the cheetahs were hunting became faster and faster, the cheetah needed a way to select for a trait that also makes them faster in order to better adapt in their environment. The long bones were selected for and whoever possessed the alleles for it was more able to make & produce offspirng. Evolution occurred when the frequency of the allele increased.* 

Table 3.5. (Cont'd)

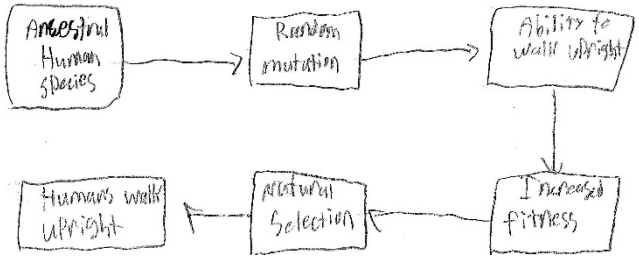
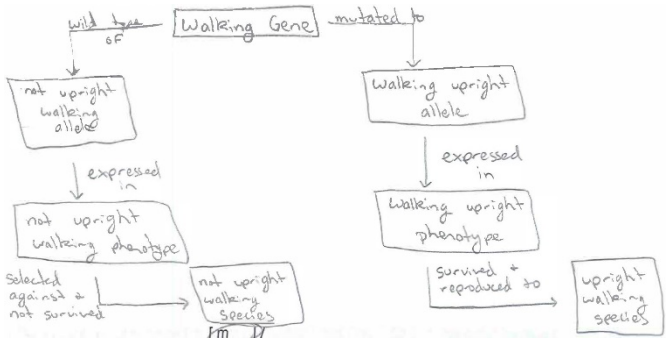
Threshold Concepts	Randomness	Responses specifically referred to randomness.	*A random mutation occurred giving a cheetah long leg bones. This trait made it easier for the cheetah to survive and reproduce. Since this trait was advantagous to the cheetah's survival and reproduction, the trait was passed on.* 
	Probability	Responses included some reference to probability in the content. This could be related to the probability of survival, reproduction, access to resources etc.	*Cheetah's overtime would evolve long leg bones because this will increase the cheetah's stride length. An increased stride length leads to increased ability to successfully hunt. Better hunters are more likely to survive and more likely to pas on their genes.* 

Table 3.5. (Cont'd)

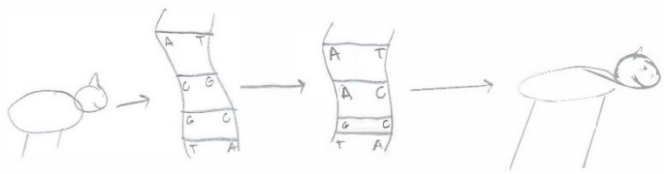
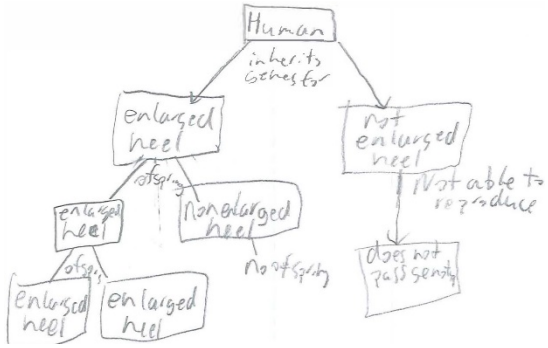
Threshold Concepts	Level of Biological Organisation	Genetic	Responses referenced genetic level entities. E.g. DNA, genes, alleles etc.	*This trait started with genetic variation. As time went on, humans with the ability to walk upright wre more fit and so had more successful offspring to pass on this genetic information. This resulted in a growing population of upright humans and shirinking population of non- upright humans.* 
		Organismal	Responses referenced organismal level entities. E.g. behavioural and structural traits, and organisms.	*The switch from slow cheetahs to fast cheetahs took many thousands of years. The switch would of began by a mutation. Then members of the population would select sexually for that mutation. Eventually the entire population would be fast.* 

Table 3.5. (Cont'd)

Threshold Concepts	Level of Biological Organisation	Population	Responses referenced population level entities. E.g. reference to communities, the entire population, future generations.	*The ancestral species had some sort of mutation (genetic variation, crossing over, etc.) that was passed onto the next generations. This mutation caused some of the offspring to have enlarged heels. These enlarged heels held some sort of survival value and was therefore a trait passed on to future generations. This is a case for evolution.* 

Data Analysis

A total of 852 student responses were included in analyses (n=213 students; 4 responses per student). Each student provided one narrative and one model-based response to each of two prompts that differed in taxon of organism (Human or Cheetah). We conducted analyses to test two related research questions:

1. What is the effect of the representational mode on the content of student responses?

We assessed whether the mode of representation (model vs narrative) affected the probability of finding KCs, NIs, and TCs in student responses. Prior studies that have used variations of the HCA prompt have shown that 'taxon' as a contextual feature of prompts affects the content of student responses (de Lima, 2020, p 12 and p 68; Göransson et al., 2020; Nehm & Ha, 2011). We therefore also assessed the influence of 'taxon' in this study, as well as its interaction with mode of representation.

We used mixed-effects multiple ordinal logistic regressions and multiple logistic regressions to test our questions. We use ordinal models because some of the KCs can exist at different levels. We verified results by also using non-ordinal logistic models comparing pairs of successive levels. Results are presented as odds ratios and model-predicted conditional probabilities. All models included random intercepts to account for student-level variation in the data.

2. Are students' ideas represented consistently across assessments that vary in mode and taxon?

We assessed whether students similarly represented KCs, NIs, and TCs across their 4 responses explaining evolution by natural selection. Additionally, we assessed differences between groups of individuals ('subpopulations') who varied in their consistency. To do this, we defined a set of subpopulations based on patterns of presence and absence of KCs, NIs, and TCs in their 4 responses (Narrative Human, Narrative Cheetah, Model Human, and Model Cheetah). We then compared these subpopulations to see whether they differed in their prior academic performance (measured by incoming GPA).

Software

All analyses were done using the R statistical environment v 3.6.3 (R Core Team, 2020). We used the *dplyr* (Wickham, François, Henry, & Müller, 2020) and *tidyr* (Wickham & Henry, 2020) packages for data processing, *lme4* (Bates et al., 2015) for mixed effects logistic regressions, *ordinal* (Christensen, 2019) for mixed effects ordinal logistic regressions, *effects* (Fox, 2003) for calculating and plotting model output, *ggplot2* (Wickham, 2016) for plotting, and *sjPlot* (Lüdecke, 2020) for generating tables.

RESULTS

1. Effect of Context (Mode and Taxa) on Content and Level of Student Responses.

Our data showed that almost all the students included Key Concepts (KCs) in their responses, most included Threshold Concepts (TCs), and a few included Naïve Ideas

(NIs) (Fig. 3.1). While KCs and NIs were represented almost equally across mode of representation, narrative responses elicited more NIs than model-based responses.

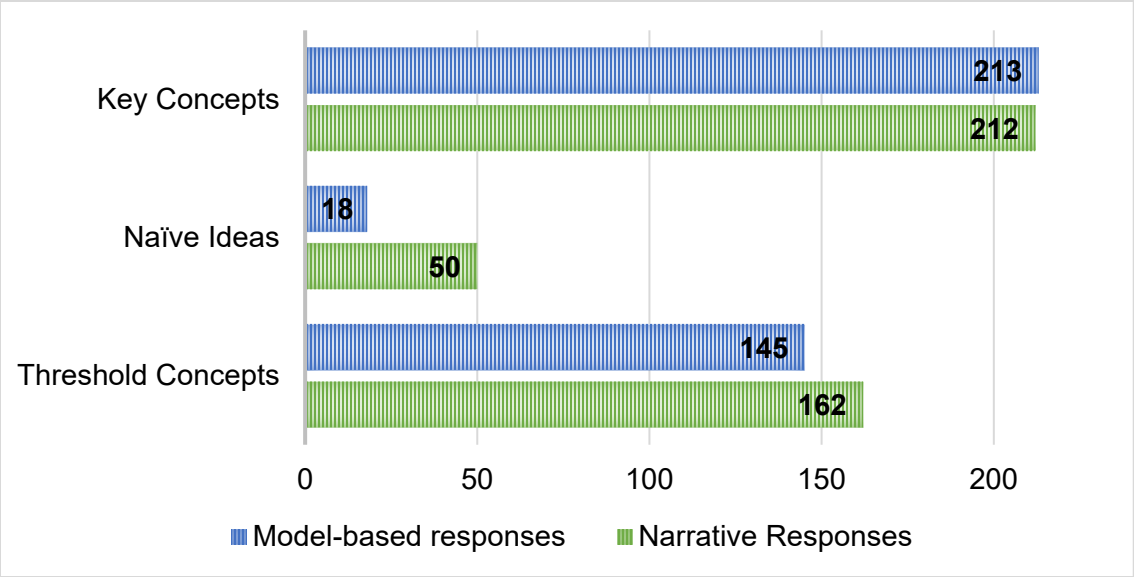


Figure 3.1. The number of students who included KCs (at least 1), NIs (at least 1), and TCs (Probability, or Randomness, or at least 2 Levels of Biological Organisation) in their narrative and model-based responses. *These are not exclusive categories.*

The results of our ordinal logistic regressions show that the mode of representation was a predictor for the presence of the KCs Limited Resources and Competition ($p < 0.001$), and Differential Survival and Reproduction ($p < 0.01$) (Table 3.6, Figs. 3.2b and 3.2c). Students were twice as likely to include ideas about Limited Resources and Competition in their narrative responses as compared to their model-based responses with minor differences between taxa (Fig. 3.2b). In the case of Differential Survival and Reproduction, the logistic regressions and ordinal logistic regressions differ in whether the mode is statistically significant (Tables 3.6, 3.7). The general pattern, however, is consistent: students are more likely to include Differential Survival and Reproduction in their narratives than in their models. In contrast, the odds of students including ideas

about Variation were higher in their model-based responses, although this difference was not statistically significant (Table 3.6).

For Heritability, the mode did not affect the probability of responses being in the higher or lower level of scientific plausibility for Human responses (Table 3.7, Fig. 3.2d). However, for Cheetah responses, narratives were 1.5 times more likely to be more scientifically plausible as compared to their model-based responses (i.e., students were more likely to include references to genetic material being inherited in their narrative responses than in their model-based responses). We did not find any significant interactions between the mode of representation and the prompt taxon for any of the other KCs.

Table 3.6. Odds ratios of ordinal logistic regression analysis for Key Concepts. *Bolded values are statistically significant (*** p < 0.001; ** p < 0.01; * p < 0.05; $^{\lambda}$ p < 0.1). Lower and Upper Confidence intervals are provided in the brackets.*

	Mode [Narrative]	**Taxon [Human]**	**Interaction [Taxa/mode]**
Variation	0.69 [0.42, 1.12]	**0.58*** [0.35, 0.94]	1.11 [0.56, 2.20]
Limited Resources and Competition	**2.84***** [1.63, 5.05]	**0.21***** [0.11, 0.39]	1.01 [0.44, 2.28]
Differential Survival and Reproduction	**2.16**** [1.34, 3.48]	**0.42***** [0.27, 0.66]	0.82 [0.42, 1.55]
Heritability	1.09 [0.7, 1.71]	0.75 [0.48, 1.18]	1.09 [0.58, 2.07]

Similar to previous studies (de Lima, 2020, p 12 and p 68; Göransson et al., 2020; Nehm & Ha, 2011) we saw that item-feature context (i.e., taxon) influenced the presence/absence of KCs in responses (Table 3.6, Figs. 3.2a, 3.2b, and 3.2c). Students were more likely to include Limited Resources and Competition (~3 times as likely, $p < 0.001$), when responding to prompts about Cheetahs as compared to Humans (Table 3.7, Fig. 3.2b). They were also slightly more likely (~3% more, $p < 0.001$) to include Variation in their Cheetah responses, but since Variation was present in almost every response, this difference is not very meaningful (Table 3.7, Fig. 3.2a). Students also include ideas about Differential Survival and Reproduction to a greater extent in their Cheetah responses, and the difference between the taxa is higher in models (~3.5 times as likely) than in narratives (~1.7 times as likely) ($p < 0.05$, Table 3.7, Fig. 3.2c). When we analysed different levels of scientific plausibility, we saw that the significant differences were in the Level 1 vs Absent comparison (Table 3.7). We did not see significant differences when comparing higher levels.

Table 3.7. Odds ratios of logistic regression analysis for Key Concepts. *Bolded values are statistically significant (*** p < 0.001; ** p<0.01; * p<0.05; λ p<0.1). Lower and Upper Confidence intervals are provided in the brackets.*

		Mode [Narrative]	***Taxon [Human]***	***Interaction Taxa/mode***
Variation	Level 1 v Absent	0.96 [0.33, 2.78]	**0.16*** ** [0.05, 0.39]	1.43 [0.38, 5.34]
	Level 2 v Level 1	0.99 [0.32, 2.99]	1.83 [0.66, 5.29]	1.04 [0.24, 4.47]
	Level 3 V Level 2	1.41 [0.32, 7.67]	0.56 [0.13, 2.11]	0.57 [0.06, 4.05]
Limited Resources and Competition	Level 1 V Absent	**2.84*** ** [1.63, 5.05]	**0.21*** ** [0.11, 0.39]	1.01 [0.44, 2.28]
Differential Survival and Reproduction	Level 1 V Absent	2.37 [0.64, 9.62]	**0.26*** [0.06, 0.83]	2.04 [0.36, 1.21]
	Level 2 V L 1	1.34 [0.60, 3.04]	0.77 [0.33, 1.79]	0.41 [0.12, 1.26]
Heritability	Level 1 V Absent	0.63 [0.34, 1.16]	0.63 [0.34, 1.14]	1.95 [0.82, 4.65]
	Level 2 v Level 1	**3.07*** [1.20, 8.37]	1.64 [0.62, 4.47]	0.29 λ [0.07, 1.10]

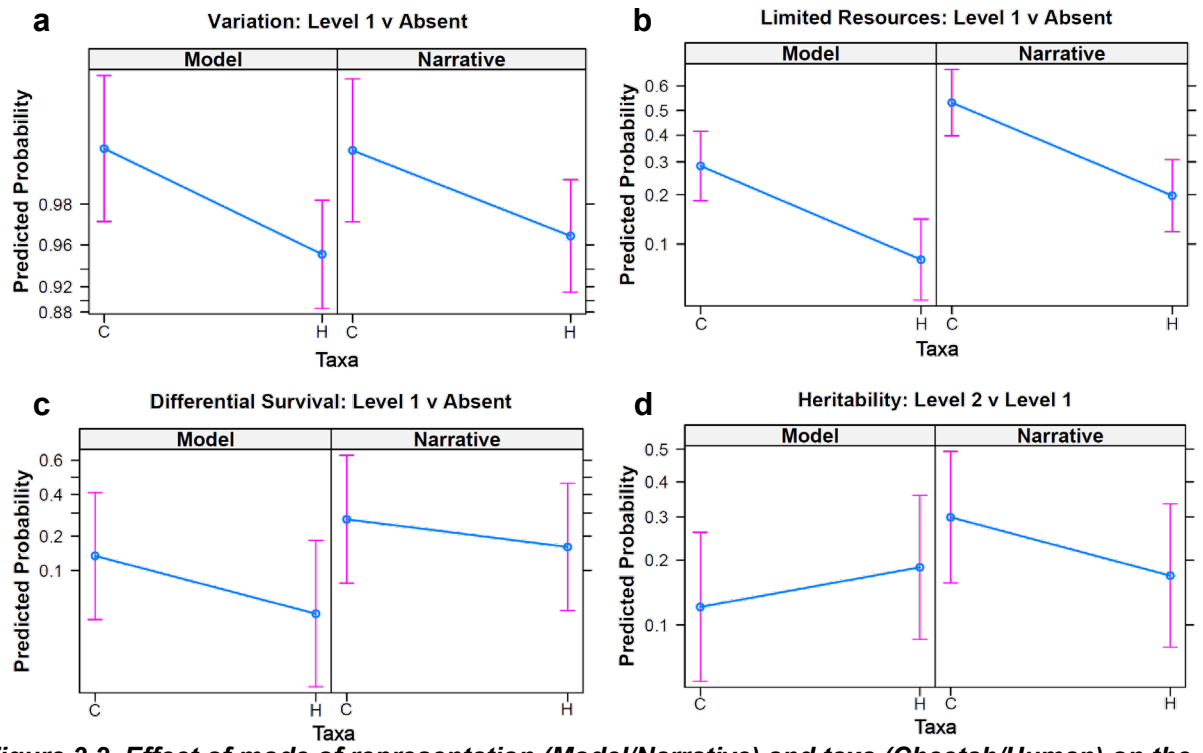


Figure 3.2. Effect of mode of representation (Model/Narrative) and taxa (Cheetah/Human) on the probability of Key Concepts - (a) Variation, (b) Differential Survival and Reproduction, (c) Heritability, and (d) Limited Resources and Competition - occurring in student responses based on presence and levels of scientific plausibility*. Note that in this and all subsequent figures, the confidence intervals appear very wide, but inference based on visualisation of the error bars is not very reliable. This is especially true because of the presence of additional terms in the model (student identity as random intercept). Refer to odds ratio tables for additional information relevant to statistical inference.*

As a whole, students were far less likely to include Naïve Ideas in their responses as compared to the Key Concepts (Figs. 3.2 and 3.3). However, when they did include NIs, they were much more likely to include them in their narrative responses as compared to their model-based responses. Most of this was driven by the NI Need (Table 3.8, Fig. 3.3). Students' included Need ~8.5 times more in Cheetah narratives as compared to Cheetah models, and ~4.5 times more in Human narratives as compared to Human models ($p < 0.001$). In general, they also tended to use Need more when they were responding to prompts about Humans as compared to prompts about Cheetahs. Need was ~3 times more frequent in Human models as compared to Cheetah models and

~1.3 times more in Human narratives as compared to Cheetah narratives ($p < 0.05$). The frequency of occurrence for the other two NIs – Use and Adapt – was very low, and we did not find significant differences in their occurrence. We did not find any interactions between the mode of representation and the prompt taxon affecting the predicted probabilities for Naïve Ideas.

Table 3.8. Odds ratios of logistic regression analysis for Naïve Ideas**. Bolded values are statistically significant (p < 0.001; ** p < 0.01; * p < 0.05; ^λ p < 0.1). Lower and Upper Confidence intervals are provided in the brackets.*

	Mode [Narrative]	***Taxon [Human]***	***Interaction Taxa/mode***
Need	**9.74*** **** [3.17, 38.7]	**4.02*** [1.22, 16.3]	0.60 [0.12, 2.37]
Use	3.03 [0.38, 61.4]	1.00 [0.03, 25.6]	1.34 [0.04, 45.01]
Adapt	2.23 [0.38, 17.88]	3.65 [0.73, 27.9]	0.75 [0.07, 6.21]

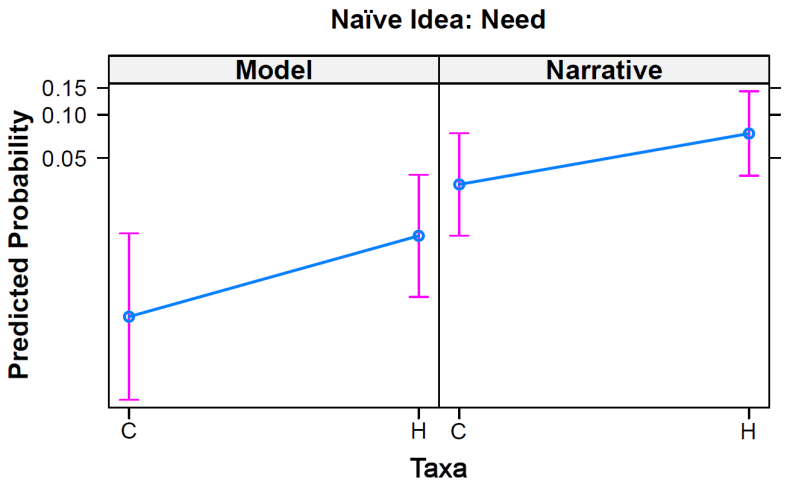


Figure 3.3. Effect of mode of representation (Model/Narrative) and taxa (Cheetah/Human) on the probability of Naïve Ideas occurring in student responses.

Our results indicate that mode of response highly influences the presence of Threshold Concepts in students' responses (Table 3.9, Fig. 3.4). Ideas relating to Probability were more than 3 times as likely to be present in their model-based responses as compared to their narratives (Fig. 3.4a). In contrast, students were more likely to talk about Randomness in their narrative responses (>4.5 times as likely; Fig. 3.4b). Narrative responses were also >1.3 times as likely to refer to entities at all three Levels of Biological Organisation (Genetic, Organismal, and Population), whereas model-based responses were ~4% more likely to refer to only two Levels of Biological Organisation - either Genetic and Organismal or Organismal and Population (Figs.3. 4c and 3.4d). Prompt taxon and interactions between mode of representation and prompt taxon did not affect predicted probabilities for TCs

Table 3.9. Odds ratios of logistic regression analysis for Threshold Concepts*. Bolded values are statistically significant (*** p < 0.001; ** p < 0.01; * p < 0.05; $^{\lambda}$ p < 0.1). Lower and Upper Confidence intervals are provided in the brackets.*

	Mode [Narrative]	***Taxon [Human]***	***Interaction Taxa/mode***
Probability	**0.30**** [0.14, 0.59]	0.71 [0.38, 1.31]	0.68 [0.23, 1.91]
Randomness	**6.93***** [2.95, 17.3]	1.73 [0.74, 4.10]	0.67 [0.23, 2.06]
Any 2 Levels of Biological Organisation	**0.42*** [0.21, 0.82]	1.07 [0.51, 2.24]	1.70 [0.64, 4.60]
All 3 Levels of Biological Organisation	**2.37*** [1.21, 4.78]	0.93 [0.44, 1.94]	0.59 [0.21, 1.57]

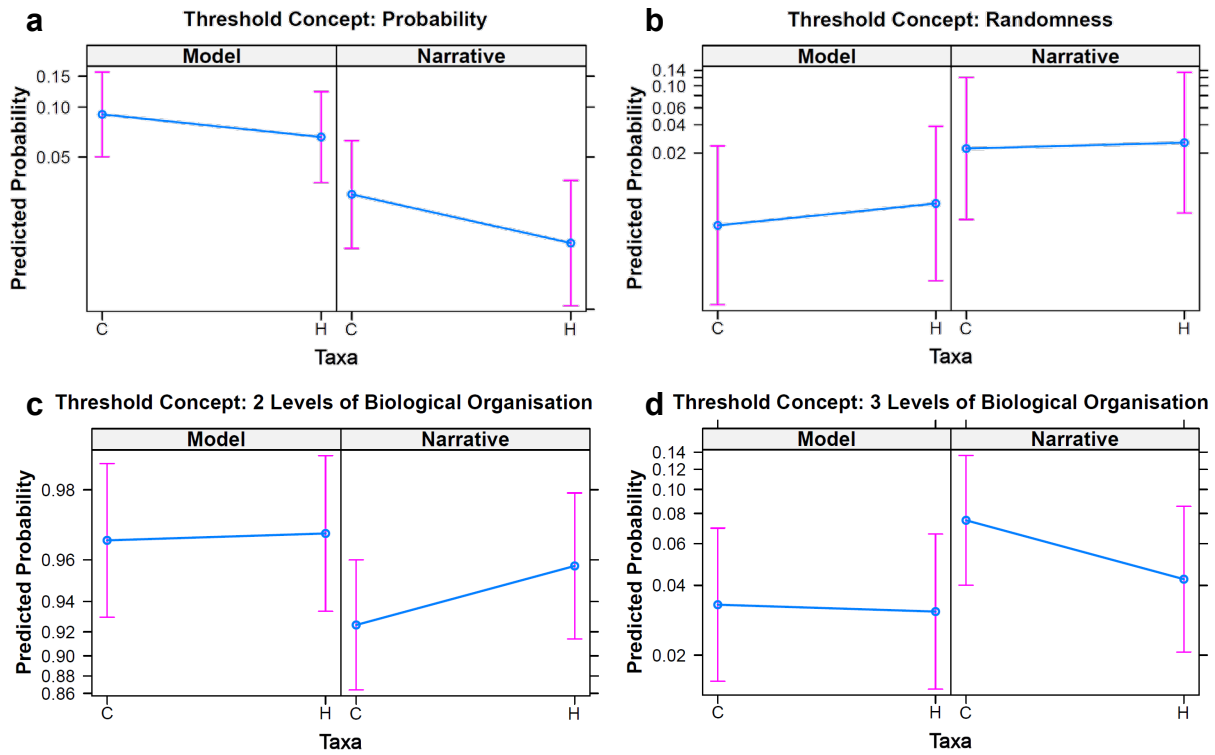


Figure 3.4. Effect of mode of representation (Model/Narrative) and taxa (Cheetah/Human) on the probability of Threshold Concepts - (a) Probability, (b) Randomness, (c) 2 Levels of Biological Organisation, and (d) 3 Levels of Biological Organisation - occurring in student responses.

2. Consistency of Key Concepts (KCs), Naïve Ideas (NIs) and Threshold Concepts (TCs) in Students' Responses

We defined a set of subpopulations based on patterns of presence and absence of KCs, NIs, and TCs in students' 4 responses (Narrative Human, Narrative Cheetah, Model Human, and Model Cheetah). Students were 'consistent' if they expressed an idea (by virtue of its presence) consistently in all four responses. Students could also be 'consistent within taxon' if an idea was expressed for only one of the taxa, but included across modes, or 'consistent within mode' if an idea was expressed only in model-based or narrative responses regardless of taxon. Because subpopulations were significantly skewed in size (Table 3.10) we did not try to make any statistical inferences but looked at trends in the data.

The most common and consistent KC found in student responses was Variation; 80% (n=169) of students included Variation in all their responses and only 3 students did not include it in any of their responses. Similarly, almost 50% of students (n=104) consistently included ideas about Differential Survival and Reproduction in their responses. In contrast, <16% of students were consistent in including ideas about Limited Resources and Competition (n=31) and Heritability (n=34) in all their responses. Most students (73%, n=155) did not include NIs in their responses and only 2 students had an NI in all 4 of their responses. Almost a quarter of the students (n=49) had at least one TC in all 4 responses.

We did not find evidence that any of the KCs, NIs, or TCs that we coded were exclusive to any mode of representation. Among the students who were consistent only within mode or within taxon, mode elicited a greater degree of consistency for TCs, NIs, and KCs (except for Limited Resources and Competition).

In general, high-achieving students (highest mean GPA) were most consistent in including KCs and not including NIs in their responses. Students with lower mean GPAs tended to include NIs and were more inconsistent with including KCs in their responses.

Table 3.10. Comparison of the number and GPA (mean ± Standard Error) of students who expressed an idea consistently in all four responses, or responses to the same taxa, or in responses using the same mode. *Since the n for Naïve Ideas and Threshold Concepts was low, we aggregated the data. Responses with any NI were counted as having an NI, and responses that had either Probability, Randomness, or at least two Levels of Biological Organisation were counted as having a TC. We used incoming GPA to calculate the mean. Students whose responses had a concept in all 4 responses were counted in the 'Consistently included the concept' group'. Students who used the concept only in both the responses in the same mode or for the same taxa were counted in the 'Consistent within mode' and 'Consistent within taxa' groups. Students who were inconsistent in their usage of concepts (had it in 1 or 3 of their responses) are not counted in this table. Within each concept, the groups are exclusive.*

Concept		**Consistently included the concept** (concept present in all 4 responses)	**Consistent within mode** (concept present only in both taxa for one of the modes)	**Consistent within taxa** (concept present only in both modes for one of the taxa)	**Concept absent from all responses**
Variation	N	169	10	7	3
	GPA	3.44 (± 0.03)	2.71 (± 0.34)	3.03 (± 0.30)	2.47 (± 0.18)
Limited Resources and Competition	N	31	16	22	81
	GPA	3.41 (± 0.09)	3.28 (± 0.23)	3.36 (± 0.11)	3.32 (± 0.05)
Differential Survival and Reproduction	N	104	29	16	17
	GPA	3.44 (± 0.05)	3.36 (± 0.08)	3.38 (± 0.13)	3.18 (± 0.10)
Heritability	N	34	34	7	78
	GPA	3.41 (± 0.06)	3.46 (± 0.07)	3.61 (± 0.15)	3.16 (± 0.07)
Naïve Ideas	N	2	13	6	155
	GPA	3.24 (± 0.44)	2.95 (± 0.14)	3.45 (± 0.16)	3.41 (± 0.04)
Threshold Concepts	N	49	55	18	37
	GPA	3.48 (± 0.05)	3.48 (± 0.05)	3.46 (± 0.10)	3.02 (± 0.12)

DISCUSSION

Our results indicate that the content of students' responses is influenced by the mode of representation students use to convey their knowledge and reasoning. In this section, we compare our findings with previous studies and offer some possible explanations for the patterns we see. Additionally, we will discuss the implications of our findings on instruction and assessment.

Contextual Effects of Mode of Representation

Our results indicate that while mode of representation did not seem to have an effect in eliciting Key Concepts (KCs) or Threshold Concepts (TCs) overall, mode did affect particular KCs, TCs, and Naïve Ideas (NIs). We posit that some differences could be because of the affordances of the particular mode. For example, students' model-based responses were more likely to include ideas relating to Probability. This could be because students often use branching in their modes, and this directly affords a comparison between one branch leading to something that has a higher probability of occurring than the other branch. Similar affordances of the mode could contribute to responses being coded for the presence of certain KCs. The branching in the model inherently leads a student to think of variation in the organism/trait, which logically leads to representing ideas about Differential Survival and Reproduction. Additionally, arrows leading to successive generations implicitly indicate Heritability. We did not detect any differences for the presence of Variation (although the trend in the data was to show that students included Variation to a higher degree in their models) or Heritability based on mode of representation. However, this could either be because students intentionally

included both the KCs in their narratives and models, or because architectural affordances of the model allowed for the coding of the KC in their model based response when they had only included it in their narrative.

However, such architectural affordances are not provided to the TC, Random, and the KC, Limited Resources and Competition. Since there is no obvious shortcut for these terms (such as an arrow or branch), students would have to explicitly include the term ‘random’ or write about limited resources and competition in their model (which some students did do). Students’ are used to writing the phrase ‘a random mutation’ and therefore could have used it in their narratives as a matter of rote and not really by intent. Additionally, for this particular KC, the prompt (about heel/leg bones) may not directly trigger thoughts about limited resources – unlike, for example, a plate of bacteria which have a limited medium to grow on. The relative benefit of the trait with respect to the survival and reproduction may not be obvious.

Naïve ideas – especially with respect to teleology - were overrepresented in students’ narrative responses. This could be because the affordances of a narrative make it is easier to be verbose and descriptive while models are intended to be parsimonious and condensed.

For the Threshold Concept, Levels of Biological Organisation, we noticed an interesting trend in the number of levels elicited by each mode of representation. While all three levels (Genetic, Organismal, and Population) were found most often in students’

narratives, their model-based responses tended to include two levels. We found that while almost all models included the Organismal level, students were split in their inclusion of Genetic (97 models) or Population (124 models) as their second level. As the prompt is directed at the organismal level, it is plausible that students started their models at the organismal level and then chose to either move up or down a level (Population vs Genetic levels). This could also be due to a limitation of the physical space in which they had to construct their model. Repeating the same set of prompts in a virtual environment could reveal whether space is a factor since this would not be a constraint.

Another consequence of not including the Genetic level in model-based responses was that those particular models were then coded as Level 1 (for scientific plausibility) for the KC Heritability. In order to get coded as Level 2 for Heritability, the response had to indicate transfer of information or material at the genetic level. This could explain why we found that the students' narrative responses were more likely to be at Level 2 for Heritability as compared to their model-based responses.

Consistency in the Occurrence of KCs, NIs, and TCs

We propose that consistency in the occurrence of concepts is indicative of a stable Cognitive Structure. We think of consistency in terms of the same concept being expressed at the same level in all 4 responses. Students with a fragile or piecemeal understanding of a phenomenon (in this case evolution by natural selection) do not have a stable CS, are unable to build a robust mental model of the phenomenon, and

this is reflected in the inconsistency with which they include concepts in their representations.

In this study, we saw that Variation is the most stable KC, with most students including it in all 4 responses. This could be because the course they were taking was designed to emphasise thinking about variation with respect to evolution, especially in terms of how variation arises and is maintained in populations. However, we also saw that an increase in the level (plausibility) at which the response was coded does not translate to an increase in consistency of KCs. For both Variation and Heritability, consistency is highest at the lowest level of plausibility. However, it should be noted that Variation and Heritability potentially interact with each other and with additional concepts, namely Level of Biological Organisation. We previously noted that students tended to include either the population or genetic level in their model-based responses, but infrequently included both. Those that did not include the genetic level (by choice or as a consequence of spatial constraints on a page) could not be classified at higher levels for Heritability and Variation since these explicitly require the response to refer to the genetic level. Therefore, allowances should be made while making inferences about the stability of students' CS with respect to consistency and level of response. For example, we see the opposite finding for Differential Survival and Reproduction, where higher level responses are most consistent across modes. We think that this is because the threshold between Level 1 and Level 2 is not a difficult one to cross, so the higher level is both more prevalent and more consistent.

Additional clues about the robustness of the CS can be gained by the presence/ absence of Naïve Ideas and Threshold Concepts. Presence of NIs could indicate an unsophisticated understanding of evolution. This is supported by our results. Although NIs were present to a much greater extent in students' narrative responses, they were not consistently expressed. Within the two students' narratives, NIs were often present in one response or the other, but rarely in both. Additionally, presence and consistency of NIs corresponded with students having lower academic achievement. A study by Stenning, Cox, & Oberlander (1995) similarly showed that prior performance had a disproportionate influence on students' performance using multiple representations. Higher-achieving students did better when provided with instruction using a novel mode of representation, while the performance of lower-achieving students decreased. In contrast, Dauer, Momsen, Speth, Makohon-Moore, & Long (2013) showed that given repeated practice and feedback, lower-performing students reduced the achievement gap with higher-performing students on tasks that involved creation of models that were similar to the models used in our study.

It is to be noted that the assessments used in this study were administered at the end of a semester of instruction on natural selection. Prior research has shown that Naïve Ideas are notoriously difficult to eradicate even after prolonged instruction (Bishop & Anderson, 1990; Bray Speth et al., 2009; Nehm & Reilly, 2007). This could explain that while students can generate explanations without using NIs such as Need, Use, and Adapt, it is at times difficult to overcome this deeply entrenched and intuitive way of thinking (Coley & Tanner, 2015). Additionally, it is also possible that students are using

these terms as shorthand or as a more colloquial way of speaking. It is possible that in interview contexts, further probing might have yielded a more canonical response.

As with Key Concepts, presence and consistency of Threshold Concepts could indicate a well-developed CS with respect to evolution. By definition, TCs indicate that a student has mastered a particular baseline level of a concept, and having crossed that threshold to a higher level of understanding, is unlikely to go back (Batzli et al., 2016; Harms & Fiedler, 2019; Meyer & Land, 2003). With respect to evolution, researchers have identified the abstract concepts of Randomness and Probability (Garvin-Doxas & Klymkowsky, 2008; Mead & Scott, 2010) and the ability to think at various Levels of Biological Organisation (Ross et al., 2010; Tibell & Harms, 2017). While we have discussed potential of response mode to confound our interpretation about consistency of TCs (specifically, Level of Biological Organisation) in student responses, we still have some evidence allowing us to gauge if a student is using a TC just because of a modal affordance or because the student has truly crossed a threshold with respect to evolutionary understanding. As discussed earlier, Randomness is a TC that is present most commonly in narratives. However, when it is present in a student's models, it is also present in their narratives. Additionally, Randomness appears not to be biased by taxon (consistent between taxa) and therefore could indicate that consistency in the TC Randomness could be indicative of a more developed CS.

Possible Explanations

Students' whose responses were unaffected by representational mode and consistent in the ideas elicited in their responses could be said to have representational competence. Conversely, students whose responses were influenced by mode and therefore inconsistent in the assemblage of ideas conveyed in their responses could be said to be lacking in representational competence. However, it is important to consider a few caveats. Many studies have indicated that novices face significant challenges with respect to both learning how to use multiple modes and gaining representational competence (Ainsworth, 2006; Chi et al., 1981; Kozma & Russell, 1997; Petre & Green, 1993).

When presented with a novel mode of representation, novices need to understand multiple facets of the mode in order to effectively use it as a representation. These include understanding how information is encoded and communicated in the representation, linking the mode with the domain that it is being used in, being able to identify which situations are appropriate for using it, and knowing how to leverage a representation to communicate knowledge using the appropriate level of abstraction (Ainsworth, 2006). These are challenging requirements. Research has shown that novices experience difficulty using new modes of representation as tools to further their learning (Cavallo, 1996) and struggle to select appropriate representations for given contexts (Kozma & Russell, 1997) (Chi et al., 1981). In addition, novices typically fail to discern patterns, notice discrepancies (Dufour-Janvier et al., 1987), or perceive and

understand perceptual cues not made explicit in a representation (Petre & Green, 1993).

Novices also face challenges developing representational competence (K. C. Anderson & Leinhardt, 2002; Cooper et al., 2010; Daniel et al., 2018; Treagust et al., 2003). Learners find it difficult to translate between different modes of representation (Anzai, 1991; Ferk et al., 2003; Hinton & Nakhleh, 1999; Kozma et al., 2000; Kozma & Russell, 1997; Wilder & Brinkerhoff, 2007) which could potentially be due to an unstable CS (Anzai, 1991; Pande & Chandrasekharan, 2017). Meir, Perry, Herron, & Kingsolver (2007) showed that students had misconceptions about a mode of representation (evolutionary trees) despite explicit instruction. Nitz et al. (2014) measured representational competence before and after a unit (~15 lessons) on photosynthesis and found minimal gains in students' representational competence (5%). Although Wilder & Brinkerhoff (2007) observed modest gains in representational competence after a 10-week module that involved explicit instruction about a new mode of representation (computer-based bimolecular visualisations), students did poorly when asked to translate between the new representation and one they were already familiar with (graphs). Other researchers have also indicated that developing representational competence involves a steep learning curve and requires more time and effort than can be fitted into a semester of instruction in one class (Kozma & Russell, 2005).

Our findings are consistent with prior research on novice learners using representations and developing representational competence. In our study, students had been taught to

use and develop models during the same semester in which our assessment was conducted. The differences we noticed in the degree to which narratives and models elicited different KCs can be explained by the fact that our population consists of novice learners that are on the path to developing representational competence, but are not there yet. Additionally, the inherent complexity of biology as discipline (i.e., multi-level organisation of systems characterised by dynamic and emergent interactions) makes achieving representational competence potentially more challenging than in other STEM domains (Pande & Chandrasekharan, 2017).

As novices, it is also likely that students in our study relied on surface features of the prompt as cues for accessing their CS and failed to recognize deeper conceptual aspects that the prompt was intended to elicit. Bennett, Gotwals, & Long (2020) showed that students cue into superficial contextual features of prompts and this changes the way they approach constructing required representations. Other researchers have similarly shown that novices are more likely to be influenced by surface, rather than deeper conceptual features of problems (Chi et al., 1981; Kozma & Russell, 1997; R. K. Lowe, 2003; Richard K. Lowe, 1996). These contextual influences are additionally influenced by the fact that students use such surface features when constructing their CSs in the first place (Chi et al., 1981; Larkin et al., 1980). If students are then using contextual cues to access whole or fragmented bits of their CSs that were encoded and influenced by different sets of cues, then it is not surprising that varying the contextual features of prompts will elicit externalised representations of students' mental models that are inconsistent in the content they convey.

Implications for Learning and Assessment

There is extensive literature on benefits of including multiple representations during instruction (Ainsworth, 2006; Ainsworth et al., 2011; Biber, 2014; Bransford & Schwartz, 1999; Pande & Chandrasekharan, 2017; Schnotz & Bannert, 2003; Spiro & Jehng, 1990; Tsui & Treagust, 2013). Best practice recommendations to support students' learning with multiple representations include giving explicit instruction about representations used (both mode and content), providing sufficient scaffolding to enable students to connect representations across scales and domains, and iterating bouts of feedback with opportunities for students to revise representations based on the feedback provided (Anzai, 1991; Chi et al., 1981; Cooper et al., 2012; Cox, 1999; Dauer et al., 2013; Long et al., 2014; Mayer, 2003; Nesbit & Adesope, 2006; Schnotz, 2002; Wu & Puntambekar, 2012; Yerushalmy, 1991).

A goal of education is to develop expertise, and one part of increasing expertise is being able to communicate using representations that are common to a discipline (NRC, 2015). Practicing scientists use multiple modes of representation not only to gain knowledge, but also to communicate their findings to others – both peers and non-scientists. Therefore, as part of building representational competence and disciplinary expertise in our students, we should not only instruct using multiple modes of representation but also assess their ability to communicate their knowledge in more than one mode. For example, in the course in which this study was conducted, students are expected to construct arguments, graphs, diagrams, models, and narratives that build on and support each other. Students must not only construct, but must interpret

and reason about diverse representations of the same or related concepts. Ideally, allowing and encouraging students to use more than one mode to respond to a prompt might better enable them to communicate their thinking more effectively and support their development of representational competence.

Using multiple modes of representation can support students in considering multiple perspectives (NRC, 1996) which can then enrich their responses and deepen their understanding. Requiring students to access their CSs in diverse ways will help build linkages within and among their CSs. This, in turn, will help stabilize connections and promote longer-term storage of content as well as subsequent retrieval (Paivio, 1990; Verdi & Kulhavy, 2002). From an assessment perspective, if we use the analogy that a representation is a window into the student's CS, having them construct multiple representations for the same content will provide multiple windows into their CS, and therefore a more holistic idea about its content and organisation.

Next Steps

diSessa, Hammer, Sherin, & Kolpakowski (1991) explored the way students 'invented, critiqued, improved, applied, and moved fluidly' between multiple modes of representation, and called these abilities meta-representational competence. A logical step forward for this study would therefore observe and quantify meta-representational competence in the students completing this assessment. For example, one could have students compare their representations constructed for different contexts and critique their own work. Additionally, students could be asked to modify their representations

based on their own critique and asked to explain their choices. What do they choose to include or delete? From which representation? And, what is the subsequent effect on the accuracy of the representations? Responses to such questions will provide insights about whether the content of a student's representation is an affordance of the mode or a part of their CS, as well as reveal how stable their CS is with respect to a concept of interest.

ACKNOWLEDGEMENTS

I thank Mitch Distin, Devin Babi and Hunter Hicks for help with data transcription; Socheatha Chan for logistical and technical support; Mridul Thomas for help with data analysis; and Melanie Cooper, Amelia Gotwals, and Katherine Gross for comments on earlier versions of the manuscript . I am grateful to Thomas Kiørboe and Ken Haste Andersen at the Centre for Ocean Life, Danish Technical University, Denmark, for providing me space and resources while writing this manuscript. I express my deep gratitude to the world's best advisor, Tammy Long, for the mentorship and supervision provided during all stages of this project. Finally, I gratefully acknowledge all the anonymous students whose assessments we used as the data for this study.

Funding

This material is based in part upon research supported by the National Science Foundation under grant numbers DRL 1420492 and DRL 0910278. Any opinions, findings, and conclusions or recommendations expressed in this material are those of

the author(s) and do not necessarily reflect the views of the National Science Foundation.

APPENDIX

Table S3.1. Odds ratios of multiple logistic regression for demographic analysis*. Bolded are statistically significant (* p < 0.05).*

Predictors	**Odds Ratios**	***CI***	***p***
(Intercept)	0.25	0.04 – 1.42	0.121
Gender [F]	0.89	0.58 – 1.37	0.611
Ethnicity [Minority]	0.80	0.26 – 2.44	0.695
Ethnicity [White (non-Hispanic)]	1.49	0.53 – 4.22	0.442
First Generation Learner [yes]	0.97	0.60 – 1.57	0.895
Start STEM credits	1.00	0.99 – 1.01	0.791
Start GPA	**1.59***	1.12 – 2.32	0.012
Observations		377	

Table S3.2. Emergent codebook section for Variation

VARIATION (20 codes)

Description	Code
Section 1: Linking statement: Cause and consequence of variation	
This set of codes captures the link between the cause and the consequence of variation, they only apply when there is a link established and capture the two (or more) components of the link. The code is written in two stages (or more). These codes are applied when the response specifically talks about the cause of the variation. And not just that variation exists. Do not worry about where the mutation/change is occurring. If the response is not referring to the cause of variation but rather just the presence of variation - look at Section 3	
Stage 1: Cause of the variation	
The response specifically talks about mutation being the cause of variation.	**Mut ->**
The response talks about a change or difference occurring and being the cause of variation but does not mention the word mutation.	**Chg ->**
The response refers to variation occurring due to mating or reproduction.	**Mate ->**
Any other cause of variation. Please specify.	**other-cau-var ->**
Stage 2: Consequence of variation	
The cause of variation leads to a genotypic change which in turn leads to a phenotypic change	**-> GV -> PV**
A cause of variation leads to genetic variation. Code this if previous code does not apply	**-> GV**
The cause of variation leads to a phenotypic change	**-> PV**
This applies only when the genetic variation has no cause of its own - these responses will be coded in Section 3.	**GV -> PV**

Table S3.2 (Cont'd)

Section 2: Level at which the cause of variation occurs	
Use this set of codes only if something has been captured in stage 1 of Section 1. These codes capture the level at which the previously coded cause of variation is occurring. These might be most relevant if the response mentions mutations or changes.	
The change/mutation is occurring at a genetic level. (DNA/gene/ genotype/allele).	**gen**
The change/mutation is occurring at the phenotypic or the trait level.	**trait**
The change/mutation is at the organismal level	**org**
The level of the change is not specified	**level ns**

Section 3: Variation mentioned	
These codes apply only if variation is mentioned as something that already exists. And there is no cause attributed to the presence of variation.	
Genetic variation exists in the population or it is mentioned without linking it to a cause or consequence.	**GV exists**
Phenotypic variation already exists in the population or it is mentioned without linking it to a cause or consequence.	**PV exists**

Section 4: Other ideas	
This set of codes looks at other ideas that students might have about the causes and consequences of variation. More than one of these ides might occur in the same response.	
Mutation or genetic variation specifically referred to as a random event. Absence of this code indicates that the response did not have this word.	**Rand**
There is a difference in allele frequency. This can be a pre-existing condition or can be the result of mutation/evolution	**alle freq_var**
Evolution is directed towards the phenotype in question	**evo-dir**

Table S3.2 (Cont'd)

Variation is a consequence of activity/behaviour or an organism changes within its lifetime to adapt to the environment	**lamk**
Variation exists because it was needed or because it enabled some function	**teleo-ori_var**
Any other ideas related to the causes or consequences of variation. Please write the idea in addition to the code.	**other-var**

Table S3.3. Emergent codebook section for Differential Survival and Reproduction

DIFFERENTIAL SURVIVAL AND REPRODUCTION (12 Codes)

Description	Code
Section 1: Causal/mechanistic	
This pair of codes looks for differences in causal and mechanistic reasoning with respect to survival. Code only if the response talks about survival. (the difference in these 2 could be because of the differences in structural and functional traits)	
The trait leads to increased chances of survival. There is a direct link between the trait and survival	**surv-dir**
The trait leads to something that in turn might lead to survival. Indirect link between trait and survival. If this code is given, also make sure to look at section 3.	**surv-indir**
Section 2: Directionality	
This pair of codes looks at the directionality between survival and reproduction	
Differential survival (of the individual or the population) leads to (differential) reproduction	**surv->repr**
Differential reproduction (of the individual or the population) leads to (differential) survival (probably of the population)	**repr->surv**
Section 3: Trait leads to…	
This set of codes captures the other consequences of variation that have not been captured under the main codes of limited resources and completion (i.e. Differential abilities, differential access to food and resources, and differential interactions with predators/prey.) It could be all the indirect reasons for a trait leading to differential survival (The trait leads to the coded feature which in turn leads to survival).	
The trait leads to an increase in reproduction or an increase in the rate of reproduction (including changes in the number of offspring)	**diff repr**
The trait leads to the ability to better fit the environment - not to be confused with fitness in general	**diff envt fit**
The trait leads to increased fitness (This is when the term fitness is specifically mentioned)	**diff fit**

Table S3.3 (Cont'd)

Section 4: Consequences for Offspring	
This pair of codes considers the differential consequences for the offspring. This could be the F1 generation or any subsequent generation. Having more offspring does not count - it is captured under the previous section.	
The trait leads to increased survival of the offspring.	**off-surv**
The trait leads to increased reproduction in the offspring	**off - repr**

Section 5: Selection	
This set of codes captures ideas related to selection acts on the trait.	
Selection acts on the trait - natural selection not specified. Also, not sexual selection.	**sel**
Natural selection acts on the trait	**nat sel**
Sexual selection acts on the trait	**sex sel**

Table S3.4. Emergent codebook section for Limited Resources and Competition

LIMITED RESOURCES AND COMPETITION (6 codes)

Description	Code
Section 1: Competition	
This code is applied when competition is specifically mentioned	
is competition mentioned in the response?	**comp**
Section 2: Consequences of Limited Resources	
This set of codes looks at the consequences of Limited resources. This could be directly or indirectly. These codes also apply when only one of the phenotypes is referenced in the response but it is clear that the phenotype leads to some sort of a difference.	
The trait leads to differential access to food specifically. If the response mentions prey - that is not to be coded here. It has a code of its own.	**diff food**
The trait leads to differential access to all non-food-based resources and in cases where just the word 'resources' has been used. (one or more resources - coded just once)	**diff res**
The trait leads to differential interactions with prey. Here the assumption is that the animal under consideration is the predator.	**diff prey**
The trait leads to differential interactions with predators. Here the assumption is that the animal under consideration is the prey.	**diff pred**
The trait leads to one or more different abilities. Before coding this look at the column that tells if the question that was asked was a structural or a functional question. do not code if it was a functional question and the student mentions: > "ability to walk upright" for a Human question, or > "ability to run fast" for the Cheetah question. If these abilities are mentioned for the opposite animal (i.e. run fast for a Human - they get this code. Additionally, if these abilities are mentioned for the structural questions, they also get a code. If there is any ambiguity (i.e. variations of these phrases) make a note of them and then they will be sorted out through discussion. Any and all other abilities are to be coded.	**diff ab**

Table S3.5. Emergent codebook section for Heritability

HERITABILITY (13 codes)	
Description	**Code**
Section 1: What is inherited?	
This set of codes describes what the response says is being passed down or inherited. More than one of these codes can apply to each response.	
Genetic material or DNA is passed down/inherited	**what-gen**
A specific allele is inherited	**what-all**
The trait in question, or an ability is passed down / inherited. Before coding this look at the next code	**what-trait**
The/a mutation is inherited/passed down	**what-mut**
Section 2: Who Inherits?	
This set of codes describes the inheritor(s). These codes are used only when the inheritor is specified. If no inheritor is specified, do not code this section. Both the codes can be used in one response.	
The offspring inherit. Synonyms like babies, children, F1 generation all included here.	**who-off**
Future generations inherit.	**who-fut**
Section 3: How does inheritance take place?	
How does inheritance take place, or how is 'it' passed down? Both the codes can be used in one response.	
The material is passed down through reproduction.	**how-rep**
The material is passed down by producing offspring. This is to be used when the response specifically mentions producing offspring	**how-off**
The material is passed down through natural selection	**how-natsel**

Table S3.5 (Cont'd)

Section 4: Differential Probabilities/Quantities	
These two codes are used when the response quantifies inheritance or what is being inherited.	
The response talks about a difference in the likelihood or the chance that material is inherited	**quant-chance**
The trait itself increases or decreases	**quant-trait**

Section 5: Other ideas	
This set of codes looks at other ideas that students might have about heritability. More than one of these ides might occur in the same response	
Response specifically mentions the phrase 'heritable trait/mutation/material'	**heritable**
Teleological ideas about heritability	**teleo-her**

Table S3.6. Emergent codebook section for the Holistic section

HOLISTIC (8 codes)	
Description	**Code**
Section 1: Presence of Genetic variation	
Presence of genetic variation. Since an accurate response will account for variation at the genetic level, this set of codes explores if GV has been included in the response and if it has been included accurately.	
GV included accurately (plausible cause).	**GV +**
GV included inaccurately. - Response mentions genetic components but it is used inaccurately. - the response does not respond to the question. - Variation is shown as a consequence	**GV -**
GV is ambiguous/vague. - This includes where responses refer to mutations *but do not specify the level.* - Could also include responses where 'mutation' is a noun rather than a verb.	**GV 0**
Section 2: Other ideas	
This section looks at other ideas that are present in the explanation. These are other ideas related to variation/ LR or comp/ diff survival or reproduction/ heritability.	
Other ideas included and they are all accurate/plausible	**idea +**
Other ideas included and some are plausible or accurate	**idea -**
Section 3: Other general impressions	
This section looks at general impressions of the response	
The response is teleological	**teleo-hol**
The response attributes agency to the organism. The organism can bring about the desired change by willing/wanting it	**agency_hol**
The response has some other idea that will not be captured in any of the other codes. Please specify.	**other-hol**

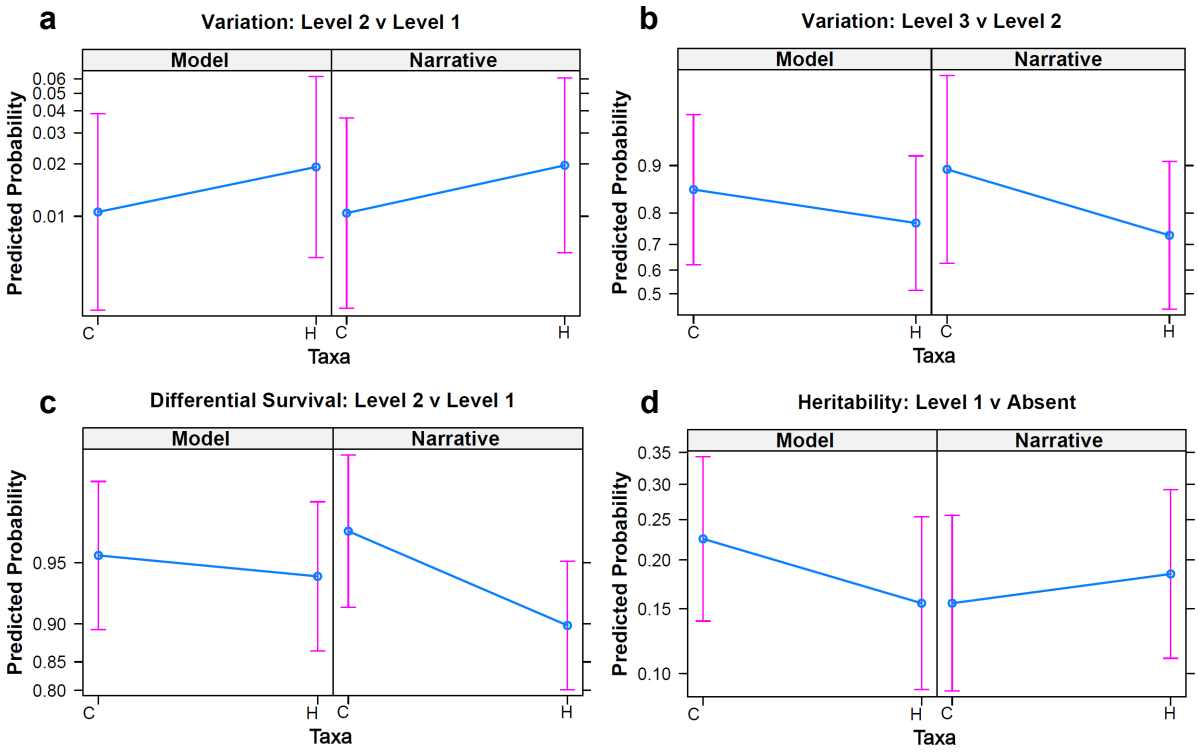


Figure S3.1. Effect of mode of representation (Model/Narrative) and taxa (Cheetah/Human) on the probability of Key Concepts - (a) Variation: Level 2 vs Level 1, (b) Variation: Level 3 vs Level 2, (c) Differential Survival and Reproduction: Level 2 vs Level 1, and (d) Heritability: Level 1 vs Absent - occurring in student responses.

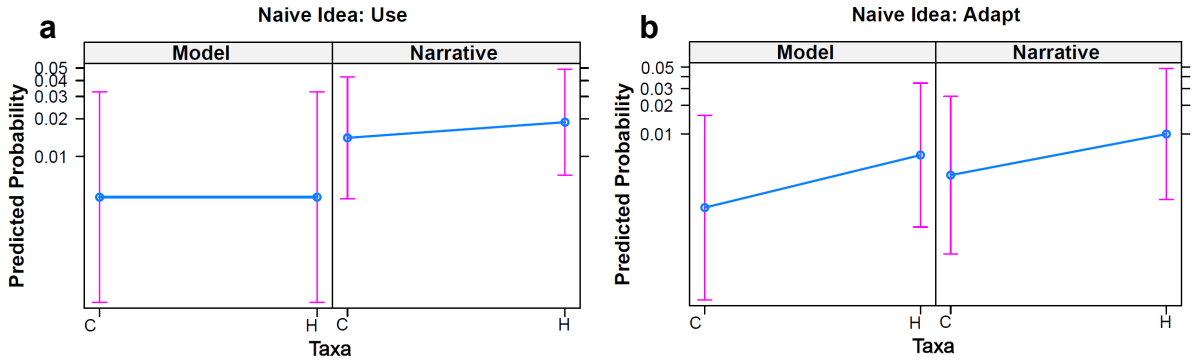


Figure S3.2. Effect of mode of representation (Model/Narrative) and taxa (Cheetah/Human) on the probability of Naïve Ideas - (a) Use, (b) Adapt - occurring in student responses.

REFERENCES

REFERENCES

Ainsworth, S. (2006). DeFT: A conceptual framework for considering learning with multiple representations. *Learning and Instruction*, 16(3), 183–198. https://doi.org/10.1016/j.learninstruc.2006.03.001

Ainsworth, S., Bibby, P., & Wood, D. (2002). Examining the Effects of Different Multiple Representational Systems in Learning Primary Mathematics. *The Journal of the Learning Sciences*, 11(1), 25–61.

Ainsworth, S., Prain, V., & Tytler, R. (2011). Drawing to learn in science. *Science*, 333(6046), 1096–1097. https://doi.org/10.1126/science.1204153

Ainsworth, S., & Th Loizou, A. (2003). The effects of self-explaining when learning with text or diagrams. *Cognitive Science*, 27(4), 669–681. https://doi.org/10.1016/S0364-0213(03)00033-8

Anderson, K. C., & Leinhardt, G. (2002). Maps as Representations : Expert Novice of Projection Comparison Understanding. *Cognition and Instruction*, 20(3), 283–321. https://doi.org/10.1207/S1532690XCI2003

Anderson, O. R., & Demetrius, O. J. (1993). A flow-map method of representing cognitive structure based on respondents' narrative using science content. *Journal of Research in Science Teaching*, 30(8), 953–969. https://doi.org/10.1002/tea.3660300811

Anzai, Y. (1991). Learning and use of representations for physics expertise. In Karl Anders Ericsson & J. Smith (Eds.), *Towards a General Theory of Expertise: Prospects and Limits*. (pp. 64–92). Cambridge University Press.

Ausubel, D. G. (1963). Cognitive Structure and the Facilitation of Meaningful Verbal Learning. *Journal of Teacher Education*, 14(2), 217–222. https://doi.org/10.1177/002248716301400220

Bates, D., Maechler, M., Bolker, B., & Walker, S. (2015). Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software*, 67(1), 1–48. https://doi.org/10.18637/jss.v067.i01.

Batzli, J. M., Knight, J. K., Hartley, L. M., Maskiewicz, A. C., & Desy, E. A. (2016). Crossing the threshold: Bringing biological variation to the foreground. *CBE Life Sciences Education*, 15(4), es9. https://doi.org/10.1187/cbe.15-10-0221

Bennett, S., Gotwals, A. W., & Long, T. M. (2020). Assessing students' approaches to modelling in undergraduate biology. *International Journal of Science Education*, 1–18. https://doi.org/10.1080/09500693.2020.1777343

Biber, A. Ç. (2014). Mathematics teacher candidates skills of using multiple representations for division of fractions. *Educational Research and Reviews*, 9(8), 237–244. https://doi.org/10.5897/err2013.1703

Bishop, B. A., & Anderson, C. W. (1990). Students conceptions of natural selection and its role in evolution. *Journal of Research in Science Teaching*, 27(5), 415–427.

Brachman, R. J., Levesque, H. J., & Pagnucco, M. (2004). *Knowledge Representation and Reasoning*. Morgan Kaufmann Publishers.

Bransford, J. D., & Schwartz, D. L. (1999). Rethinking Transfer : A Simple Proposal with Multiple Implications. *Review of Research in Education*, 24, 61–100.

Bray Speth, E., Long, T. M., Pennock, R. T., & Ebert-May, D. (2009). Using Avida-ED for Teaching and Learning About Evolution in Undergraduate Introductory Biology Courses. *Evolution: Education and Outreach*, 2(3), 415–428. https://doi.org/10.1007/s12052-009-0154-z

Brenner, M. E., Herman, S., Ho, H.-Z., & Zimmer, J. M. (1999). Cross-national comparison of representational competence. *Journal for Research in Mathematics Education*, 30(5), 541–557.

Cavallo, A. M. L. (1996). Meaningful Learning, Reasoning Ability, and Students' Understanding and Problem Solving of Topics in Genetics. *Journal of Research in Science Teaching*, 33(6), 625–656. https://doi.org/10.1002/(SICI)1098-2736(199608)33:6<625::AID-TEA3>3.0.CO;2-Q

Chandler, P., & Sweller, J. (1992). The split-attention effect as a factor in the design of instruction. *British Journal of Educational Psychology*, 62(2), 233–246.

Chi, M. T., Feltovich, P. J., & Glaser, R. (1981). Categorization and Representation of Physics Problems by Experts and Novices. *Cognitive Science*, 5(2), 121–152.

Christensen, R. H. B. (2019). *ordinal - Regression Models for Ordinal Data*. R package version 2019.12-10. https://cran.r-project.org/package=ordinal.

Çikla, O. A., & Çakiroğlu, E. (2006). Seventh grade students' use of multiple representations in pattern related algebra tasks. *Hacettepe Üniversitesi Eğitim*

Fakültesi Dergisi, 31(31), 13–24.

Clariana, R. B., Wolfe, M. B., & Kim, K. (2014). The influence of narrative and expository lesson text structures on knowledge structures: alternate measures of knowledge structure. *Educational Technology Research and Development*, 62(5), 601–616. https://doi.org/10.1007/s11423-014-9348-3

Coley, J. D., & Tanner, K. (2015). Relations between intuitive biological thinking and biological misconceptions in biology majors and nonmajors. *CBE Life Sciences Education*, 14(1), 1–19. https://doi.org/10.1187/cbe.14-06-0094

Cooper, M. M., Grove, N., Underwood, S. M., & Klymkowsky, M. W. (2010). Lost in lewis structures: An investigation of student difficulties in developing representational competence. *Journal of Chemical Education*, 87(8), 869–874. https://doi.org/10.1021/ed900004y

Cooper, M. M., Underwood, S. M., Hilley, C. Z., & Klymkowsky, M. W. (2012). Development and assessment of a molecular structure and properties learning progression. *Journal of Chemical Education*, 89(11), 1351–1357. https://doi.org/10.1021/ed300083a

Cox, R. (1999). Representation construction, externalised cognition and individual differences. *Learning and Instruction*, 9(4), 343–363. https://doi.org/10.1016/S0959-4752(98)00051-6

Daniel, K. L., Bucklin, C. J., Austin Leone, E., & Idema, J. (2018). Towards a Definition of Representational Competence. In K. L. Daniel (Ed.), *Towards a Framework for Representational Competence in Science Education. Models and Modeling in Science Education, vol 11*. Springer, Cham. https://doi.org/10.1007/978-3-319-89945-9_1

Dauer, J. T., & Long, T. M. (2015). Long-term conceptual retrieval by college biology majors following model-based instruction. *Journal of Research in Science Teaching*, 52(8), 1188–1206. https://doi.org/10.1002/tea.21258

Dauer, J. T., Momsen, J. L., Bray Speth, E., Makohon-Moore, S. C., & Long, T. M. (2013). Analyzing change in students' gene-to-evolution models in college-level introductory biology. *Journal of Research in Science Teaching*, 50(6), 639–659. https://doi.org/10.1002/tea.21094

de Jong, T., Ainsworth, S., Dobson, M., van der Hulst, A., Levonen, J., Reimann, P., Sime, J.-A., van Someren, M. W., Spada, H., & Swaak, J. (1998). Acquiring

knowledge in science and mathematics: The use of multiple representations in technology-based learning environments. In M. W. van Someren, P. Reimann, H. P. A. Boshuizen, & T. de Jong (Eds.), *Learning with Multiple Representations* (pp. 9–40). Pergamon Press.

de Lima, J. (2020). Contextual Influences on Undergraduate Biology Students' Reasoning and Representations of Evolutionary Concepts. [Unpublished doctoral dissertation]. Michigan State University.

diSessa, A. A., Hammer, D., Sherin, B., & Kolpakowski, T. (1991). Inventing graphing: Meta-representational expertise in children. *The Journal of Mathematical Behavior*, 10(2), 117–160.

Dufour-Janvier, B., Bednarz, N., & Belanger, M. (1987). Pedagogical considerations concerning the problem of representation. In C. Janvier (Ed.), *Problems of representation in the teaching and learning of mathematics* (pp. 109–122). Lawrence Earlbaum Associates.

Ericsson, K. Anders, & Simon, H. A. (1998). Thinking and Speech and Protocol Analysis Thinking and Speech and Protocol Analysis. *Mind, Culture, and Activity*, 5(3), 178–186. https://doi.org/10.1207/s15327884mca0503_3

Federer, M. R., Nehm, R. H., Opfer, J. E., & Pearl, D. (2015). Using a constructed-response instrument to explore the effects of item position and item features on the assessment of students' written scientific explanations. *Research in Science Education*, 45(4), 527–553. https://doi.org/10.1007/s11165-014-9435-9

Ferk, V., Vrtacnik, M., Blejec, A., & Gril, A. (2003). Student's understanding of molecular structure representations. *International Journal of Science Education*, 25(10), 1227–1245. https://doi.org/10.1080/0950069022000038231

Fox, J. (2003). Effect Displays in R for Generalised Linear Models. *Journal of Statistical Software*, 8(15), 1–27. http://www.jstatsoft.org/v08/i15/

Garvin-Doxas, K., & Klymkowsky, M. W. (2008). Understanding Randomness and its Impact on Student Learning: Lessons Learned from Building the Biology Concept Inventory (BCI). *CBE—Life Sciences Education*, 7(2), 227–233. https://doi.org/https://doi.org/10.1187/cbe.07-08-0063

Gentner, D., & Markman, A. B. (1997). Structure Mapping in Analogy and Similarity. *American Psychologist*, 52(1), 45–56. https://doi.org/10.7551/mitpress/4631.003.0008

Gilbert, J. K. (2004). Models and Modelling: Routes To More Authentic Science Education. *International Journal of Science and Mathematics Education*, 2(2), 115–130. https://doi.org/10.1007/s10763-004-3186-4

Goldman, S. R. (2003). Learning in complex domains: When and why do multiple representations help? *Learning and Instruction*, 13(2), 239–244. https://doi.org/10.1016/s0959-4752(02)00023-3

Göransson, A., Orraryd, D., Fiedler, D., & Tibell, L. A. E. (2020). Conceptual Characterization of Threshold Concepts in Student Explanations of Evolution by Natural Selection and Effects of Item Context. *CBE Life Sciences Education*, 19(1), ar1. https://doi.org/10.1187/cbe.19-03-0056

Harms, U., & Fiedler, D. (2019). Improving Student Understanding of Randomness and Probability to Support Learning About Evolution. In U. Harms & M. J. Reiss (Eds.), *Evolution Education Re-considered* (pp. 271–283). Springer. https://doi.org/10.1007/978-3-030-14698-6

Hinton, M. E., & Nakhleh, M. B. (1999). Students' Microscopic, Macroscopic, and Symbolic Representations of Chemical Reactions. *The Chemical Educator*, 4(5), 158–167. https://doi.org/10.1007/s00897990325a

Hmelo-Silver, C. E., Marathe, S., & Liu, L. (2007). Fish Swim , Rocks Sit , and Lungs Breathe : Expert-Novice Understanding of Complex Systems. *The Journal of the Learning Sciences*, 16(3), 307–331. https://doi.org/10.1080/10508400701413401

Ifenthaler, D. (2008). Practical solutions for the diagnosis of progressing mental models. In D. Ifenthaler, P. Pirnay-Dummer, & J. M. Spector (Eds.), *Understanding Models for Learning and Instruction* (pp. 43–61). Springer, Boston, MA. https://doi.org/10.1007/978-0-387-76898-4_3

Ifenthaler, D. (2010). Relational, structural, and semantic analysis of graphical representations and concept maps. *Educational Technology Research and Development*, 58(1), 81–97. https://doi.org/10.1007/s11423-008-9087-4

Ifenthaler, D., Masduki, I., & Seel, N. M. (2011). The mystery of cognitive structure and how we can detect it: Tracking the development of cognitive structures over time. *Instructional Science*, 39(1), 41–61. https://doi.org/10.1007/s11251-009-9097-6

Jaipal, K. (2010). Meaning making through multiple modalities in a biology classroom: A multimodal semiotics discourse analysis. *Science Education*, 94(1), 48–72. https://doi.org/10.1002/sce.20359

Johnson-Laird, P. N. (1983). *Mental Models: Towards a Cognitive Science of Language, Inference and Consciousness*. Cambridge University Press.

Keller, B. A., & Hirsch, C. R. (1998). Student preferences for representations of functions. *International Journal of Mathematical Education in Science and Technology*, 29(1), 1–17. https://doi.org/10.1080/0020739980290101

Koubek, R. J., & Mountjoy, D. N. (1991). *Toward a Model of Knowledge Structure and a Comparative Analysis of Knowledge Structure Measurement Techniques*.

Kozma, R. B., & Russell, J. (1997). Multimedia and Understanding: Expert and Novice Responses to Different Representations of Chemical Phenomena. *Journal of Research in Science Teaching*, 34(9), 949–968. https://doi.org/10.1002/(SICI)1098-2736(199711)34:9<949::AID-TEA7>3.0.CO;2-U

Kozma, R., Chin, E., Russell, J., & Marx., N. (2000). The roles of representations and tools in the chemistry laboratory and their implications for chemistry learning. *Journal of the Learning Sciences*, 9(2), 105–143. https://doi.org/10.1207/s15327809jls0902

Kozma, R., & Russell, J. (2005). Students Becoming Chemists: Developing Representationl Competence. In John K. Gilbert (Ed.), *Visualization in Science Educatio. Models and Modeling in Science Education, vol 1* (pp. 121–145). Springer. https://doi.org/10.1007/1-4020-3613-2_8

Larkin, J., Mcdermott, J., & Simon, D. P. (1980). Expert and Novice Performance in Solving Physics Problems. *Science*, 208(4450), 1335–1342.

Lewis-Peacock, J. A., & Postle, B. R. (2008). Temporary activation of long-term memory supports working memory. *Journal of Neuroscience*, 28(35), 8765–8771. https://doi.org/10.1523/JNEUROSCI.1953-08.2008

Long, T. M., Dauer, J. T., Kostelnik, K. M., Momsen, J. L., Wyse, S. A., Bray Speth, E., & Ebert-May, D. (2014). Fostering ecoliteracy through model- based instruction. *Frontiers in Ecology and the Environment*, 12(2), 138–139. https://doi.org/10.1890/1540-9295-12.2.138

Lowe, R. K. (2003). Animation and learning: selective processing of information in dynamic graphics. *Learning and Instruction*, 13(2), 157–176. https://doi.org/10.1016/S0959-4752(02)00018-X

Lowe, Richard K. (1996). Background knowledge and the construction of a situational

representation from a diagram. *European Journal of Psychology of Education*, 11(4), 377–397. https://doi.org/10.1007/bf03173279

Lüdecke, D. (2020). *sjPlot: Data Visualization for Statistics in Social Science*. https://cran.r-project.org/package=sjPlot

Mayer, R. E. (2003). The promise of multimedia learning: using the same instructional design methods across different media. *Learning and Instruction*, 13(2), 125–139. https://doi.org/10.1016/s0959-4752(02)00016-6

Mead, L. S., & Scott, E. C. (2010). Problem Concepts in Evolution Part II: Cause and Chance. *Evolution: Education and Outreach*, 3(2), 261–264. https://doi.org/10.1007/s12052-010-0231-3

Meir, E., Perry, J., Herron, J. C., & Kingsolver, J. (2007). College Students' Misconceptions About Evolutionary Trees. *The American Biology Teacher*, 69(7), e71–e76. https://doi.org/10.1662/0002-7685(2007)69[71:csmaet]2.0.co;2

Meyer, J. H. F., & Land, R. (2003). Threshold Concepts and Troublesome Knowledge : linkages to ways of thinking and practising within the disciplines. In C. Rust (Ed.), *Improving Student Learning – Ten Years On* (pp. 412–424). Oxford Centre for Staff and Learning Development.

Moharreri, K., Ha, M., & Nehm, R. H. (2014). EvoGrader: an online formative assessment tool for automatically evaluating written evolutionary explanations. *Evolution: Education and Outreach*, 7(15), 1–14. https://doi.org/10.1186/s12052-014-0015-2

National Research Council [NRC]. (1996). *National Science Education Standards*. The National Academies Press. https://doi.org/https://doi.org/10.17226/4962

National Research Council [NRC]. (2000). How Experts Differ from Novices People. In *How People Learn: Brain, Mind, Experience, and School: Expanded Edition*. National Academies Press. https://doi.org/10.17226/9853

National Research Council [NRC]. (2015). *Reaching Students: What Research Says About Effective Instruction in Undergraduate Science and Engineering*. The National Academies Press. https://doi.org/https://doi.org/10.17226/18687

Nehm, R. H., Beggrow, E. P., Opfer, J. E., & Ha, M. (2012). Reasoning About Natural Selection: Diagnosing Contextual Competency Using the ACORNS Instrument. *The American Biology Teacher*, 74(2), 92–98. https://doi.org/10.1525/abt.2012.74.2.6

Nehm, R. H., & Ha, M. (2011). Item feature effects in evolution assessment. *Journal of Research in Science Teaching*, 48(3), 237–256. https://doi.org/10.1002/tea.20400

Nehm, R. H., & Reilly, L. (2007). Biology majors' knowledge and misconceptions of natural selection. *Bioscience*, 57(3), 263–272. https://doi.org/10.1641/b570311

Nesbit, J. C., & Adesope, O. O. (2006). Learning with concept and knowledge maps: A meta-analysis. *Review of Educational Research*, 76(3), 413–448. https://doi.org/10.3102/00346543076003413

Nitz, S., Ainsworth, S. E., Nerdel, C., & Prechtl, H. (2014). Do student perceptions of teaching predict the development ofrepresentational competence and biological knowledge? *Learning and Instruction*, 31, 13–22. https://doi.org/10.1016/j.learninstruc.2013.12.003

Paivio, A. (1990). *Mental representations: A dual coding approach*. Oxford University Press.

Pande, P., & Chandrasekharan, S. (2017). Representational competence: towards a distributed and embodied cognition account. *Studies in Science Education*, 53(1), 1–43. https://doi.org/10.1080/03057267.2017.1248627

Petre, M., & Green, T. R. G. (1993). Learning to read graphics: Some evidence that "seeing" an information display is an acquired skill. In *Journal of Visual Languages and Computing* (Vol. 4, Issue 1, pp. 55–70). https://doi.org/10.1006/jvlc.1993.1004

R Core Team. (2020*). R: A language and environment for statistical computing*. R Foundation for Statistical Computing. https://www.r-project.org/

Ross, P. M., Taylor, C. E., Hughes, C., Kofod, M., Whitaker, N., Lutze-Mann, L., & Tzioumis, V. (2010). Threshold Concepts: Challenging the Way We Think, Teach and Learn In Biology. In J. H. F. Meyer, R. Land, & C. Baillie (Eds.), *Threshold Concepts and Transformational Learning* (pp. 165–177). Sense Publishers. https://doi.org/10.1037//0003-066X.46.5.506

Schnotz, W. (2002). Towards an integrated view of learning from text and visual displays. *Educational Psychology Review*, 14(1), 101–120. https://doi.org/10.1023/A:1013136727916

Schnotz, W., & Bannert, M. (2003). Construction and interference in learning from multiple representation. *Learning and Instruction*, 13(2), 141–156. https://doi.org/10.1016/S0959-4752(02)00017-8

Schonborn, K. J., & Anderson, T. R. (2009). A model of factors determining students' ability to interpret external representations in biochemistry. *International Journal of Science Education*, 31(2), 193–232. https://doi.org/10.1080/09500690701670535

Schreier, M. (2014). Qualitative Content Analysis. In U. Flick (Ed.), *The SAGE Handbook of Qualitative Data Analysis* (pp. 170–183). SAGE Publications, Inc. https://doi.org/http://dx.doi.org/10.4135/9781446282243.n12

Schuman, H., & Presser, S. (1996). *Questions and answers in attitude surveys: Experiments on question form, wording, and context*. SAGE Publications, Inc.

Shafrir, U. (1999). Representational Competence. In I. E. Sigel (Ed.), *Development of Mental Representation: Theories and Applications* (pp. 371–389). Lawrence Earlbaum Associates.

Shavelson, R. J. (1974). Methods for examining representations of a subject-matter structure in a student's memory. *Journal of Research in Science Teaching,* 11(3), 231–249. https://doi.org/10.1002/tea.3660110307

Shell, D. F., Brooks, D. W., Trainin, G., Wilson, K. M., Kauffman, D. F., & Herr, L. M. (2010). *The Unified Learning Model. In The Unified Learning Model*. Springer. https://doi.org/10.1007/978-90-481-3215-7

Someren, M. W. van, Reimann, P., Boshuizen, H. P. A., & de Jong, T. (Eds.). (1998). *Learning with Multiple Representations*. Pergamon Press.

Spiro, R. J., & Jehng, J.-C. (1990). Cognitive Flexibility ond Hypertext: Theory and Technology for the Nonlinear and Multidimensional Traversal of Complex Subject Matter. In D. Nix & R. Spiro (Eds.), *Cognition, Education, and Multimedia: Exploring Ideas in High Technology.* Lawrence Earlbaum Associates. https://doi.org/10.4324/9780203052174

Stenning, K., Cox, R., & Oberlander, J. (1995). Contrasting the Cognitive Effects of Graphical and Sentential Logic Teaching: Reasoning, Representation and Individual Differences. *Language and Cognitive Processes*, 10(3–4), 333–354. https://doi.org/10.1080/01690969508407099

The Carnegie Classification of Institutions of Higher Education. (n.d.). Retrieved March 23, 2018, from http://carnegieclassifications.iu.edu/

Tibell, L. A. E., & Harms, U. (2017). Biological Principles and Threshold Concepts for Understanding Natural Selection: Implications for Developing Visualizations as a

Pedagogic Tool. *Science & Education*, 26(7–9), 953–973.

Treagust, D. F., Chittleborough, G., & Mamiala, T. L. (2003). The role of submicroscopic and symbolic representations in chemical explanations. *International Journal of Science Education*, 25(11), 1353–1368. https://doi.org/10.1080/0950069032000070306

Tsai, C. C., & Huang, C. M. (2002). Exploring students' cognitive structures in learning science: A review of relevant methods. *Journal of Biological Education*, 36(4), 163–169. https://doi.org/10.1080/00219266.2002.9655827

Tsui, C.-Y., & Treagust, D. F. (2013). Introduction to Multiple Representations: Their Importance in Biology and Biological Education. In C.-Y. Tsui & D. F. Treagust (Eds.), *Multiple Representations in Biological Education. Models and Modeling in Science Education* (Vol. 7, pp. 3–18). Springer, Dordrecht. https://doi.org/10.1007/978-94-007-4192-8

Verdi, M. P., & Kulhavy, R. W. (2002). Learning with maps and texts: An overview. *Educational Psychology Review*, 14(1), 27–46. https://doi.org/10.1023/A:1013128426099

Wickham, H. (2016). *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York.

Wickham, Hadley, François, R., Henry, L., & Müller, K. (2020). *dplyr: A Grammar of Data Manipulation*. R package version 0.8.5. https://cran.r-project.org/package=dplyr

Wickham, Hadley, & Henry, L. (2020). *tidyr: Tidy Messy Data*. R package version 1.0.3. https://cran.r-project.org/package=tidyr

Wilder, A., & Brinkerhoff, J. (2007). Supporting Representational Competence in High School Biology With Computer-Based Biomolecular Visualizations. *Journal of Computers in Mathematics and Science Teaching,* 26(1), 5–26. https://doi.org/10.1007/978-90-481-9449-0

Wu, H. K., & Puntambekar, S. (2012). Pedagogical Affordances of Multiple External Representations in Scientific Processes. *Journal of Science Education and Technology*, 21(6), 754–767. https://doi.org/10.1007/s10956-011-9363-7

Yerushalmy, M. (1991). Student perceptions of aspects of algebraic function using multiple representation software. *Journal of Computer Assisted Learning*, 7(1), 42–

57. https://doi.org/10.1111/j.1365-2729.1991.tb00223.x

CONCLUSIONS AND IMPLICATIONS

This dissertation contributes to our understanding of how context influences the way students both reason about and represent evolutionary concepts. In this chapter I briefly summarise my findings and discuss implications for instruction and assessment.

An idea, event, or phenomenon is fully understood only when considered in context, wherein context is the background or setting in which the idea, event or phenomenon occurs. In education, contextual influences are of great relevance because while learning happens in one context (e.g., instruction in the classroom), the application of the skills and knowledge learnt happens in a different context (e.g., in real life or on an assessment). When students learn a new concept or a skill, this information is encoded into their Cognitive Structures (CS). The process of encoding information into the CS (i.e., learning) is affected by the context in which the learning occurs. When students are later asked to apply the knowledge or skills, they access their CS to retrieve the relevant information. This retrieval process is similarly affected by the context in which the information is retrieved.

In this dissertation, I have explored how the context in which information is retrieved, affects what is retrieved. In education research, context has been defined in multiple ways, from the broad societal/cultural setting (National Academies of Sciences, Engineering, and Medicine [NASEM], 2016), to the discipline in question (Nehring et al., 2012; Topcu, 2013), to the types of questions (Driver et al., 1994; Watkins & Elby,

2013), the text used in the question (Brown et al., 2011), and even the order in which the questions are asked (Federer et al., 2015). For the purpose of this dissertation, context has been defined as the words used in the prompt (i.e., item feature context) and the type of representation required in response to the prompt (i.e., mode of response). Additionally, I explored the effects of instruction on, and the association of prior achievement with, contextual susceptibility.

To answer these questions, I asked students to respond to two isomorphic prompts which contained the following basal structure: "(Taxon) has (trait). How would biologists explain how a (taxon) with (trait) evolved from an ancestral (taxon) without (trait)?" I tested for the effect of item-feature by varying the taxon in the prompts (Cheetah v Human) and tested the effect of mode of response by requiring students to respond to each prompt by writing a narrative and by constructing a model. All the responses were coded for the concepts that were included and the architecture of model-based responses was also graphically analysed using network metrics.

I found that taxon – specifically, Cheetah v Human - was a strong predictor of the content in students' responses. Students included more Key Concepts (KCs; concepts that are key towards understanding evolution) and fewer Naïve Ideas (NIs; intuitive and inaccurate ways of thinking about evolution) in responses to Cheetah prompts than Human prompts in both narrative and model-based responses. Cheetah models were also more likely to have certain KCs (specifically, Limited Resources, Competition, and Differential Survival and Reproduction) in their Cheetah models as compared to their

Human models. Additionally, the odds of students including NIs was greater when responding to prompts about Humans. Prompt taxon also influenced the architecture of students' model-based responses. Cheetah models were significantly larger and more complex than models constructed in response to the Human prompt.

Instruction and prior academic achievement were associated with some decrease in contextual susceptibility to the prompt taxon. While instruction had a significant effect on the average number of KCs (increased by 30%) and NIs (decreased by 40%) in students' narrative responses, taxon-specific differences in KCs decreased only moderately with instruction (e.g., 4.4% for the KC Heritability and 3.7% for Limited Resources). With respect to students' model-based responses, the odds of students including certain KCs (Variation and Differential Survival and Reproduction) increased with increasing levels of prior academic achievement. Additionally, higher achieving students showed decreased susceptibility to prompt taxon and were more consistent in including KCs in both models, regardless of taxon of the prompt. However, with regards to model architecture, middle achievers in general had models that were larger in size and more complex than their peers.

Mode of response was also a significant predictor of the content in students' responses. Students' narrative responses had more KCs and NIs than their model-based responses. However, when I analysed students' responses for the presence of evolutionary Threshold Concepts (TCs; concepts that indicate mastery of the baseline level of a concept), we did not see any contextual effects of mode of response.

However, presence of individual KCs, NIs, and TCs was influenced by mode of response. Students' narrative responses were more likely to include the KCs, Limited Resources, Competition, and Differential Survival and Reproduction, the NI, Need, and the TCs, Randomness and Three Levels of Biological Organisation. Conversely, the TCs, Probability and Two Levels of Biological Organisation, were more likely to be present in their model-based responses.

These results show that students are susceptible to context – both when context is defined as item features of the prompt and mode of response. This indicates that students face difficulties transferring knowledge and skills from the context in which they acquired them to a different context in which they are required to retrieve and apply them. Reasons for students' susceptibility to context are likely to be complex and multifaceted. For example, the contextual susceptibility that students demonstrated due to prompt taxon could be attributed to the specific taxa that were contrasted in these studies (i.e., Cheetah and Human). Acceptance of human evolution by natural selection has been historically problematic (Mayr, 1982). Even now, many people, including college educated adults, are less accepting of natural selection in humans as compared to other animals; acceptance increases however with increasing evolutionary distance from humans (Brenan, 2019; Evans, 2008; Miller et al., 2006; Sinatra et al., 2003). It is therefore possible that students' susceptibility to the contextual influences of taxon observed in this study are related to their acceptance of human evolution and perceptions that humans are taxonomically unique and apart from the evolutionary tree

(American Association for the Advancement of Science [AAAS], 2018; Coley, 2007; Coley & Tanner, 2015).

Contextual susceptibility attributed to the mode of response could be explained by the affordances of the modes in question. For example, students may have included the phrase ‘random mutation’ in their narratives out of habit because they are used to saying, hearing, or writing that phrase. However, this lexical affordance is not as easily accessible in models where its inclusion might be less reflexive and require more purposeful intent. Models similarly have affordances that may explain the occurrence of certain ideas that are less prevalent in narratives. For example, most students’ models included branches which lend themselves to notions of alternative pathways (i.e., Variation) that may differ in their likelihood of occurring (i.e., Probability). Therefore, some ideas may be easier to express in models merely on account of the model’s architectural features.

Students contextual susceptibility could also be due to the fact that they are novice learners. Novices are more likely to have fragmented and fragile knowledge structures (diSessa, 1988, 2013; diSessa et al., 2004; Ifenthaler et al., 2011), which increases the challenge of retrieving relevant information (e.g., KCs) without being distracted by irrelevant details in the prompt or in their CS (e.g., NIs; Dauer & Long, 2015; Hmelo-Silver et al., 2007). In addition, research has shown that novice learners are unable to recognise the deeper conceptual features of prompts and tend to use surface cues while accessing their CS to retrieve information (Cheng et al., 2015; Hmelo-Silver &

Pfeffer, 2004; Hsu et al., 2012; Vattam et al., 2011). In my studies, contextual susceptibility to taxon could be because these students are novices to evolution learning. Evolution is a topic that students have great difficulty understanding and multiple studies have shown that despite explicit instruction, students' explanations of evolution often include misconceptions (Bishop & Anderson, 1990; Bray Speth et al., 2009; Morabito et al., 2010; Nehm & Ridgway, 2011; M. U. Smith, 2010). Similarly, contextual susceptibility to the mode of response could be due to the fact that these students are also novices to modelling. As undergraduates, most students will have had many opportunities to represent their knowledge via narratives, but model-based assessments are far less common in both K-12 and undergraduate classrooms. Novices require both time and practice to become competent in expressing their ideas through novel modes of representation, such as models (Ainsworth, 2006; Chi et al., 1981; Constantinou et al., 2019; Kozma & Russell, 1997; Petre & Green, 1993).

Understanding students' contextual susceptibility has implications for both instruction and assessment. Instruction aims to help students progress along the continuum from being novices to becoming experts. In order to facilitate the development of robust CSs which are less susceptible to contextual influences, our instruction should be tailored to help students develop deeper conceptual understanding (Chin & Brown, 2000; McNeill et al., 2006; Sedikides & Skowronski, 1991; Smith & Colby, 2007; Warburton, 2003). For example, one strategy to increase the efficiency of instruction could be to teach evolution using humans as examples. Previous studies have shown that students are interested in learning about themselves and their development (Pobiner et al., 2018)

and that students' learn best when taught using relatable and relevant examples (National Research Council [NRC], 2009).

Instruction that promotes fluency with representations commonly used in the discipline can also help advance novices along the continuum toward expertise (NRC, 2015). Since models are commonly used by scientists when generating, evaluating, and communicating science (Gilbert, 2004; Halloun, 2007; Long et al., 2014; Upmeier zu Belzen et al., 2019a), science instruction should help students develop competence in modelling as well as in using other representations common to science (S. W. Gilbert, 1991; Liu & Hmelo-Silver, 2009; Louca & Zacharia, 2012; Verhoeff et al., 2008). Fortunately, there is extensive literature to support instruction using modelling and other representations (Bierema et al., 2017; Hobbs et al., 2013; Nicolaou & Constantinou, 2014; Offerdahl et al., 2017; Upmeier zu Belzen et al., 2019b; Wilson et al., 2019; Windschitl et al., 2008).

Additionally, instructors are encouraged to teach science the way it is practised (AAAS, 2011; Cooper et al., 2015). Instruction that encourages students to construct their knowledge by engaging in scientific practises has been shown to improve student achievement (Armbruster et al., 2009; Blasco-Arcas et al., 2013; Freeman et al., 2007, 2014; Haak et al., 2011; Jensen et al., 2015; Martin et al., 2007; Oliver-Hoyo et al., 2004; Pearsall et al., 1997; Pierce & Fox, 2012; Yoder & Hochevar, 2005; Yuretich et al., 2001). Other best practise recommendations include providing scaffolding, giving explicit instruction about the content and representations used during instruction, and

ensuring that students are able to iteratively revise their representations based on feedback (Anzai, 1991; Chi et al., 1981; Cooper et al., 2012; Cox, 1999; Dauer et al., 2013; Hattie & Timperley, 2007; Long et al., 2014; Mayer, 2003; Nesbit & Adesope, 2006; Schnotz, 2002; Wu & Puntambekar, 2012; Yerushalmy, 1991)

Considering context is also of vital importance when designing assessments. The context in which information is retrieved from the CS can affect the transfer of knowledge and its subsequent retrieval (Gentner et al., 2003; Jacobson & Spiro, 1995; Loewenstein & Gentner, 2001; Vosniadou, 1989). Therefore, if the context used on an assessment acts as a barrier to knowledge transfer, the assessment will be an invalid measure of students' knowledge and skills. While designing prompts that are equivalent in terms of difficulty levels is, difficult (Hamp-Lyons & Mathias, 1994; Lee & Anderson, 2007; Li, 2018; Sydorenko, 2011), our study, as well as others (Göransson et al., 2020; Nehm & Ha, 2011), shows that even prompts that are truly isomorphic do not elicit the same information.

Using multiple contexts, both in terms of the words used in the prompt and the mode of response required, could make assessments more holistic and provide better insights into students' knowledge and skills. Additionally, since students' preferences for mode of representation vary, giving them the opportunity to represent their reasoning using multiple modes might make instruction more equitable and inclusive. And although the task of scoring all these assessments, especially for large-enrolment classes might seem daunting, researchers and educators have made tremendous strides in

developing automated assessments, even of models (Gray et al., 2013; Ifenthaler, 2010; Ifenthaler et al., 2011; Luckie et al., 2011; Zhai et al., 2020)

While it is known that context affects the way students reason and represent their thinking, my studies contribute to our understanding about how and to what degree students are influenced by contextual susceptibility. It is my hope that my findings can benefit college biology education by making it more efficient, effective, and inclusive.

REFERENCES

REFERENCES

Ainsworth, S. (2006). DeFT: A conceptual framework for considering learning with multiple representations. *Learning and Instruction*, 16(3), 183–198. https://doi.org/10.1016/j.learninstruc.2006.03.001

American Association for the Advancement of Science [AAAS]. (2011). *Vision and Change in Undergraduate Biology Education: a call to action.* http://visionandchange.org

American Association for the Advancement of Science [AAAS]. (2018). *Project 2061: Evolution and Natural Selection. AAAS Science Assessment.* http://assessment.aaas.org/topics/1/EN#/0

Anzai, Y. (1991). Learning and use of representations for physics expertise. In K. A. Ericsson & J. Smith (Eds.), *Towards a General Theory of Expertise: Prospects and Limits*. (pp. 64–92). Cambridge University Press.

Armbruster, P., Patel, M., Johnson, E., & Weiss, M. (2009). Active Learning and Student-centered Pedagogy Improve Student Attitudes and Performance in Introductory Biology. *CBE-Life Sciences Education*, 8(3), 203–213. https://doi.org/10.1187/cbe.09-03-0025

Bierema, A. M.-K., Schwarz, C. V, & Stoltzfus, J. R. (2017). Engaging Undergraduate Biology Students in Scientific Modeling: Analysis of Group Interactions, Sense-Making, and Justification. *CBE—Life Sciences Education*, 16(4), ar68. https://doi.org/10.1187/cbe.17-01-0023

Bishop, B. A., & Anderson, C. W. (1990). Students conceptions of natural selection and its role in evolution. *Journal of Research in Science Teaching*, 27(5), 415–427.

Blasco-Arcas, L., Buil, I., Hernández-Ortega, B., & Sese, F. J. (2013). Using clickers in class. the role of interactivity, active collaborative learning and engagement in learning performance. *Computers and Education*, 62, 102–110. https://doi.org/10.1016/j.compedu.2012.10.019

Bray Speth, E., Long, T. M., Pennock, R. T., & Ebert-May, D. (2009). Using Avida-ED for Teaching and Learning About Evolution in Undergraduate Introductory Biology Courses. *Evolution: Education and Outreach*, 2(3), 415–428. https://doi.org/10.1007/s12052-009-0154-z

Brenan, M. (2019). *40% of Americans Believe in Creationism*. Gallup. https://news.gallup.com/poll/261680/americans-believe-creationism.aspx

Brown, S., Lewis, D., Montfort, D., & Borden, R. (2011). The Importance of Context in Students ' Understanding of Normal and Shear Stress in Beams. *118th ASEE Annual Conference & Exposition*, Session W522B.

Cheng, M. F., Lin, J. L., Cheng, M. F., & Lin, J. L. (2015). Investigating the Relationship between Students' Views of Scientific Models and Their Development of Models. *International Journal of Science Education*, 37(15), 2453–2475. https://doi.org/10.1080/09500693.2015.1082671

Chi, M. T., Feltovich, P. J., & Glaser, R. (1981). Categorization and Representation of Physics Problems by Experts and Novices. *Cognitive Science*, 5(2), 121–152.

Chin, C., & Brown, D. E. (2000). Learning in science: A comparison of deep and surface approaches. *Journal of Research in Science Teaching*, 37(2), 109–138. https://doi.org/10.1002/(SICI)1098-2736(200002)37:2<109::AID-TEA3>3.0.CO;2-7

Coley, J. D. (2007). The Human Animal: Developmental Changes in Judgments of Taxonomic and Psychological Similarity Among Humans and Other Animals. *Cognition, Brain, Behavior*, 11(4), 733–756.

Coley, J. D., & Tanner, K. (2015). Relations between intuitive biological thinking and biological misconceptions in biology majors and nonmajors. *CBE Life Sciences Education*, 14(1), 1–19. https://doi.org/10.1187/cbe.14-06-0094

Constantinou, C. P., Nicolaou, C. T., & Papaevripidou, M. (2019). A Framework for Modeling-Based Learning, Teaching, and Assessment. In A. Upmeier zu Belzen, J. van Driel, & D. Krüger (Eds.), *Towards a Competence-Based View on Models and Modeling in Science Education.* Models and Modeling in Science Education, Vol 12. Springer, Cham. https://doi.org/10.1007/978-3-030-30255-9

Cooper, M. M., Caballero, M. D., Ebert-May, D., Fata-Hartley, C. L., Jardeleza, S. E., Krajcik, J. S., Laverty, J. T., Matz, R. L., Posey, L. A., & Underwood, S. M. (2015). Challenge faculty to transform STEM learning. *Science*, 350(6258), 281–282. https://doi.org/10.1126/science.aab0933

Cooper, M. M., Underwood, S. M., Hilley, C. Z., & Klymkowsky, M. W. (2012). Development and assessment of a molecular structure and properties learning progression. *Journal of Chemical Education*, 89(11), 1351–1357. https://doi.org/10.1021/ed300083a

Cox, R. (1999). Representation construction, externalised cognition and individual differences. *Learning and Instruction*, 9(4), 343–363. https://doi.org/10.1016/S0959-4752(98)00051-6

Dauer, J. T., & Long, T. M. (2015). Long-term conceptual retrieval by college biology majors following model-based instruction. *Journal of Research in Science Teaching*, 52(8), 1188–1206. https://doi.org/10.1002/tea.21258

Dauer, J. T., Momsen, J. L., Bray Speth, E., Makohon-Moore, S. C., & Long, T. M. (2013). Analyzing change in students' gene-to-evolution models in college-level introductory biology. *Journal of Research in Science Teaching*, 50(6), 639–659. https://doi.org/10.1002/tea.21094

diSessa, A. A. (1988). Knowledge in pieces. In G. Forman & P. B. Pufall (Eds.), *Constructivism in the computer age* (pp. 49–70). Lawrence Earlbaum Associates Inc.

diSessa, A. A. (2013). A bird's-eye view of the "pieces" vs. "coherence" controversy (from the "pieces" side of the fence). In S. Vosniadou (Ed.), *International Handbook of Research on Conceptual Change*, Second Edition (pp. 31–48). Routledge. https://doi.org/10.4324/9780203154472

diSessa, A. A., Gillespie, N. M., & Esterly, J. B. (2004). Coherence versus fragmentation in the development of the concept of force. *Cognitive Science*, 28(6), 843–900. https://doi.org/10.1016/j.cogsci.2004.05.003

Driver, R., Asoko, H., Leach, J., Scott, P., & Mortimer, E. (1994). Constructing Scientific Knowledge in the Classroom. *Educational Researcher*, 23(7), 5–12.

Evans, E. M. (2008). Conceptual change and evolutionary biology: A developmental analysis. In S. Vosniadou (Ed.), *International Handbook of research on conceptual change* (pp. 263–294). Research in Science Education.

Federer, M. R., Nehm, R. H., Opfer, J. E., & Pearl, D. (2015). Using a constructed-response instrument to explore the effects of item position and item features on the assessment of students' written scientific explanations. *Research in Science Education*, 45(4), 527–553. https://doi.org/10.1007/s11165-014-9435-9

Freeman, S., Eddy, S. L., McDonough, M., Smith, M. K., Okoroafor, N., Jordt, H., & Wenderoth, M. P. (2014). Active learning increases student performance in science, engineering, and mathematics. *Proceedings of the National Academy of Sciences*, 111(23), 8410–8415. https://doi.org/10.1073/pnas.1319030111

Freeman, S., O'Connor, E., Parks, J. W., Cunningham, M., Hurley, D., Haak, D., Dirks, C., & Wenderoth, M. P. (2007). Prescribed active learning increases performance in introductory biology. *CBE - Life Sciences Education*, 6(2), 132–139.

Gentner, D., Loewenstein, J., & Thompson, L. (2003). Learning and transfer: A general role for analogical encoding. *Journal of Educational Psychology*, 95(2), 393–408. https://doi.org/10.1037/0022-0663.95.2.393

Gilbert, J. K. (2004). Models and Modelling: Routes To More Authentic Science Education. *International Journal of Science and Mathematics Education*, 2(2), 115–130. https://doi.org/10.1007/s10763-004-3186-4

Gilbert, S. W. (1991). Model building and a definition of science. *Journal of Research in Science Teaching*, 28(1), 73–79. https://doi.org/10.1002/tea.3660280107

Göransson, A., Orraryd, D., Fiedler, D., & Tibell, L. A. E. (2020). Conceptual Characterization of Threshold Concepts in Student Explanations of Evolution by Natural Selection and Effects of Item Context. *CBE Life Sciences Education*, 19(1), ar1. https://doi.org/10.1187/cbe.19-03-0056

Gray, S. A., Gray, S., Cox, L. J., & Henly-Shepard, S. (2013). Mental Modeler: A fuzzy-logic cognitive mapping modeling tool for adaptive environmental management. *Proceedings of the Annual Hawaii International Conference on System Sciences*, 965–973. https://doi.org/10.1109/HICSS.2013.399

Haak, D. C., HilleRisLambers, J., Pitre, E., & Freeman, S. (2011). Increased Structure and Active Learning Reduce the Achievement Gap in Introductory Biology. *Science*, 332(6034), 1213–1216.

Halloun, I. A. (2007). Mediated modeling in science education. *Science and Education*, 16(7–8), 653–697. https://doi.org/10.1007/s11191-006-9004-3

Hamp-Lyons, L., & Mathias, S. P. (1994). Examining expert judgments of task difficulty on essay tests. *Journal of Second Language Writing*, 3(1), 49–68. https://doi.org/10.1016/1060-3743(94)90005-1

Hattie, J., & Timperley, H. (2007). The power of feedback. *Review of Educational Research*, 77(1), 81–112. https://doi.org/10.3102/003465430298487

Hmelo-Silver, C. E., Marathe, S., & Liu, L. (2007). Fish Swim , Rocks Sit , and Lungs Breathe : Expert-Novice Understanding of Complex Systems. *The Journal of the Learning Sciences*, 16(3), 307–331. https://doi.org/10.1080/10508400701413401

Hmelo-Silver, C. E., & Pfeffer, M. G. (2004). Comparing expert and novice understanding of a complex system from the perspective of structures, behaviors, and functions. *Cognitive Science*, 28(1), 127–138. https://doi.org/10.1016/S0364-0213(03)00065-X

Hobbs, F. C., Johnson, D. J., & Kearns, K. D. (2013). A deliberate practice approach to teaching phylogenetic analysis. *CBE Life Sciences Education*, 12(4), 676–686. https://doi.org/10.1187/cbe-13-03-0046

Hsu, Y. S., Lin, L. F., Wu, H. K., Lee, D. Y., & Hwang, F. K. (2012). A Novice-Expert Study of Modeling Skills and Knowledge Structures about Air Quality. *Journal of Science Education and Technology*, 21(5), 588–606. https://doi.org/10.1007/s10956-011-9349-5

Ifenthaler, D. (2010). Relational, structural, and semantic analysis of graphical representations and concept maps. *Educational Technology Research and Development*, 58(1), 81–97. https://doi.org/10.1007/s11423-008-9087-4

Ifenthaler, D., Masduki, I., & Seel, N. M. (2011). The mystery of cognitive structure and how we can detect it: Tracking the development of cognitive structures over time. *Instructional Science*, 39(1), 41–61. https://doi.org/10.1007/s11251-009-9097-6

Jacobson, M. J., & Spiro, R. J. (1995). Hypertext Learning Environments, Cognitive Flexibility, and the Transfer of Complex Knowledge: An Empirical Investigation. *Journal of Educational Computing Research*, 12(4), 301–333. https://doi.org/10.2190/4T1B-HBP0-3F7E-J4PN

Jensen, J. L., Kummer, T. A., & Godoy, P. D. D. M. (2015). Improvements from a Flipped Classroom May Simply Be the Fruits of Active Learning. *CBE Life Sciences Education*, 14(Spring), 1–12. https://doi.org/10.1187/10.1187/cbe.14-08-0129

Kozma, R. B., & Russell, J. (1997). Multimedia and Understanding: Expert and Novice Responses to Different Representations of Chemical Phenomena. *Journal of Research in Science Teaching*, 34(9), 949–968.

Lee, H. K., & Anderson, C. (2007). Validity and topic generality of a writing performance test. *Language Testing*, 24(3), 307–330. https://doi.org/10.1177/0265532207077200

Li, J. (2018). Establishing Comparability Across Writing Tasks With Picture Prompts of Three Alternate Tests. *Language Assessment Quarterly*, 15(4), 368–386. https://doi.org/10.1080/15434303.2017.1405422

Liu, L., & Hmelo-Silver, C. E. (2009). Promoting complex systems learning through the use of conceptual representations in hypermedia. *Journal of Research in Science Teaching*, 46(9), 1023–1040. https://doi.org/10.1002/tea.20297

Loewenstein, J., & Gentner, D. (2001). Spatial Mapping in Preschoolers: Close Comparisons Facilitate Far Mappings. *Journal of Cognition and Development*, 2(2), 189–219. https://doi.org/10.1207/S15327647JCD0202_4

Long, T. M., Dauer, J. T., Kostelnik, K. M., Momsen, J. L., Wyse, S. A., Bray Speth, E., & Ebert-May, D. (2014). Fostering ecoliteracy through model- based instruction. *Frontiers in Ecology and the Environment*, 12(2), 138–139. https://doi.org/10.1890/1540-9295-12.2.138

Louca, L. T., & Zacharia, Z. C. (2012). Modeling-based learning in science education: Cognitive, metacognitive, social, material and epistemological contributions. *Educational Review*, 64(4), 471–492. https://doi.org/10.1080/00131911.2011.628748

Luckie, D., Harrison, S. H., & Ebert-May, D. (2011). Model-based reasoning: using visual tools to reveal student learning. *Advances in Physiology Education*, 35(1), 59–67. https://doi.org/10.1152/advan.00016.2010

Martin, T., Rivale, S. D., & Diller, K. R. (2007). Comparison of student learning in challenge-based and traditional instruction in biomedical engineering. *Annals of Biomedical Engineering*, 35(8), 1312–1323. https://doi.org/10.1007/s10439-007-9297-7

Mayer, R. E. (2003). The promise of multimedia learning: using the same instructional design methods across different media. *Learning and Instruction*, 13(2), 125–139. https://doi.org/10.1016/s0959-4752(02)00016-6

Mayr, E. (1982). *The Growth of Biological Thought*. The Belknap Press of Harvard University Press Cambridge.

McNeill, K. L., Lizotte, D. J., Krajcik, J., & Marx, R. W. (2006). Supporting students' construction of scientific explanations by fading scaffolds in instructional materials. *Journal of the Learning Sciences*, 15(2), 153–191. https://doi.org/10.1207/s15327809jls1502_1

Miller, J. D., Scott, E. C., & Okamoto, S. (2006). Public Acceptance of Evolution. *Science*, 313(5788), 765–766. https://doi.org/10.1126/science.1126746

Morabito, N. P., Catley, K. M., & Novick, L. R. (2010). Reasoning about evolutionary history: Post-secondary students' knowledge of most recent common ancestry and homoplasy. *Journal of Biological Education*, 44(4), 166–174. https://doi.org/10.1080/00219266.2010.9656217

National Academies of Sciences, Engineering, and Medicine [NASEM]. (2016). *Science Literacy: Concepts, Contexts, and Consequences*. The National Academies Press. https://doi.org/10.17226/23595

National Research Council [NRC]. (2009). *Learning Science in Informal Environments: People, Places, and Pursuits*. The National Academies Press.

National Research Council [NRC]. (2015). *Reaching Students: What Research Says About Effective Instruction in Undergraduate Science and Engineering*. The National Academies Press. https://doi.org/https://doi.org/10.17226/18687

Nehm, R. H., & Ha, M. (2011). Item feature effects in evolution assessment. *Journal of Research in Science Teaching*, 48(3), 237–256. https://doi.org/10.1002/tea.20400

Nehm, R. H., & Ridgway, J. (2011). What Do Experts and Novices "See" in Evolutionary Problems? *Evolution: Education and Outreach*, 4(4), 666–679. https://doi.org/10.1007/s12052-011-0369-7

Nehring, A., Nowak, K. H., Upmeier zu Belzen, A., & Tiemann, R. (2012). Doing Inquiry in Chemistry and Biology. The Context's Influence on the Students' Cognitive Load. *La Chimica Nella Scuola*, XXXIV–3(January), 253–258.

Nesbit, J. C., & Adesope, O. O. (2006). Learning with concept and knowledge maps: A meta-analysis. *Review of Educational Research*, 76(3), 413–448. https://doi.org/10.3102/00346543076003413

Nicolaou, C. T., & Constantinou, C. P. (2014). Assessment of the modeling competence: A systematic review and synthesis of empirical research. *Educational Research Review*, 13, 52–73. https://doi.org/10.1016/j.edurev.2014.10.001

Offerdahl, E. G., Arneson, J. B., & Byrne, N. (2017). Lighten the load: Scaffolding visual literacy in biochemistry and molecular biology. *CBE Life Sciences Education*, 16(1), 1–11. https://doi.org/10.1187/cbe.16-06-0193

Oliver-Hoyo, M. T., Allen, D. D., Hunt, W. F., Hutson, J., & Pitts, A. (2004). Effects of an active learning environment: Teaching innovations at a research I institution. *Journal of Chemical Education*, 81(3), 441–448. https://doi.org/10.1021/ed081p441

Pearsall, N. R., Skipper, J. E. J., & Mintzes, J. J. (1997). Knowledge restructuring in the life sciences: A longitudinal study of conceptual change in biology. *Science* Education, 81(2), 193–215.

Petre, M., & Green, T. R. G. (1993). Learning to read graphics: Some evidence that "seeing" an information display is an acquired skill. *Journal of Visual Languages and Computing*, 4 (1), 55–70. https://doi.org/10.1006/jvlc.1993.1004

Pierce, R., & Fox, J. (2012). Vodcasts and active-learning exercises in a "flipped classroom" model of a renal pharmacotherapy module. *American Journal of Pharmaceutical Education*, 76(10). https://doi.org/10.5688/ajpe7610196

Pobiner, B., Beardsley, P. M., Bertka, C. M., & Watson, W. A. (2018). Using human case studies to teach evolution in high school A.P. biology classrooms. *Evolution: Education and Outreach*, 11(1). https://doi.org/10.1186/s12052-018-0077-7

Schnotz, W. (2002). Towards an integrated view of learning from text and visual displays. *Educational Psychology Review*, 14(1), 101–120. https://doi.org/10.1023/A:1013136727916

Sedikides, C., & Skowronski, J. J. (1991). The Law of Cognitive Structure Activation. *Psychological Inquiry*, 2(2), 169–184.

Sinatra, G. M., Southerland, S. A., McConaughy, F., & Demastes, J. W. (2003). Intentions and beliefs in students' understanding and acceptance of biological evolution. *Journal of Research in Science Teaching*, 40(5), 510–528. https://doi.org/10.1002/tea.10087

Smith, M. U. (2010). Current status of research in teaching and learning evolution: I. Philosophical/epistemological issues. *Science and Education*, 19(6–8), 523–538. https://doi.org/10.1007/s11191-009-9215-5

Smith, T. W., & Colby, S. A. (2007). Teaching for Deep Learning. The Clearing House: *A Journal of Educational Strategies, Issues and Ideas*, 80(5), 205–210. https://doi.org/10.3200/tchs.80.5.205-210

Sydorenko, T. (2011). Item writer judgments of item difficulty versus actual item difficulty: A case study. *Language Assessment Quarterly*, 8(1), 34–52. https://doi.org/10.1080/15434303.2010.536924

Topcu, M. S. (2013). Preservice teachers' epistemological beliefs in physics, chemistry, and biology: A mixed study. *International Journal of Science and Mathematics*

Education, 11(2), 433–458. https://doi.org/10.1007/s10763-012-9345-0

Upmeier zu Belzen, A., van Driel, J., & Krüger, D. (2019a). Introducing a Framework for Modeling Competence. In A. Upmeier zu Belzen, J. van Driel, & D. Krüger (Eds.), *Towards a Competence-Based View on Models and Modeling in Science Education*. Models and Modeling in Science Education, Vol 12 (pp. 3–19). Springer, Cham. https://doi.org/10.1007/978-3-030-30255-9_1

Upmeier zu Belzen, A., van Driel, J., & Krüger, D. (Eds.). (2019b). *Towards a Competence-Based View on Models and Modeling in Science Education*. Models and Modeling in Science Education, Vol 12. Springer, Cham. https://doi.org/10.1007/978-3-030-30255-9

Vattam, S. S., Goel, A. K., Rugaber, S., Hmelo-Silver, C. E., Gray, S., & Sinha, S. (2011). Understanding Complex Natural Systems by Articulating Structure-Behavior- Function Models. *Journal of Educational Technology & Society*, 14(1), 66–81.

Verhoeff, R. P., Waarlo, A. J., & Boersma, K. T. (2008). Systems modelling and the development of coherent understanding of cell biology. *International Journal of Science Education*, 30(4), 543–568. https://doi.org/10.1080/09500690701237780

Vosniadou, S. (1989). Analogical reasoning as a mechanism in knowledge acqisition: a developmental perspective. *Similarity and Analogical Reasoning, Technical*, 413–437.

Warburton, K. (2003). Deep learning and education for sustainability. International *Journal of Sustainability in Higher Education*, 4(1), 44–56. https://doi.org/10.1108/14676370310455332

Watkins, J., & Elby, A. (2013). Context dependence of students' views about the role of equations in understanding biology. *CBE Life Sciences Education*, 12(2), 274–286. https://doi.org/10.1187/cbe.12-11-0185

Wilson, K. J., Long, T. M., Momsen, J. L., & Bray Speth, E. (2019). Evidence Based Teaching Guide: Modeling in Classroom. *CBE Life Science Education*. https://lse.ascb.org/evidence-based-teaching-guides/modeling-in-the-classroom/

Windschitl, M., Thompson, J., & Braaten, M. (2008). Beyond the scientific method: Model-based inquiry as a new paradigm of preference for school science investigations. *Science Education*, 92(5), 941–967. https://doi.org/10.1002/sce.20259

Wu, H. K., & Puntambekar, S. (2012). Pedagogical Affordances of Multiple External Representations in Scientific Processes. *Journal of Science Education and Technology*, 21(6), 754–767. https://doi.org/10.1007/s10956-011-9363-7

Yerushalmy, M. (1991). Student perceptions of aspects of algebraic function using multiple representation software. *Journal of Computer Assisted Learning*, 7(1), 42–57. https://doi.org/10.1111/j.1365-2729.1991.tb00223.x

Yoder, J. D., & Hochevar, C. M. (2005). Encouraging Active Learning Can Improve Students' Performance on Examinations. *Teaching of Psychology*, 32(2), 91–95. https://doi.org/10.1207/s15328023top3202_2

Yuretich, R. F., Khan, S. A., Leckie, R. M., & Clement, J. J. (2001). Active-learning methods to improve student performance and scientific interest in a large introductory oceanography course. *Journal of Geoscience Education*, 49(2), 111–119. https://doi.org/10.5408/1089-9995-49.2.111

Zhai, X., C. Haudek, K., Shi, L., H. Nehm, R., & Urban-Lurain, M. (2020). From substitution to redefinition: A framework of machine learning-based science assessment. *Journal of Research in Science Teaching*, 57(9), 1430–1459. https://doi.org/10.1002/tea.21658