Winter 2017

Validation of an Instrument for Assessing Conceptual Change with Respect to The Theory of Evolution By Secondary Biology Students

Kevin David Goff
College of William and Mary - School of Education, kdgoff@email.wm.edu

Recommended Citation

Goff, Kevin David, "Validation of an Instrument for Assessing Conceptual Change with Respect to The Theory of Evolution By Secondary Biology Students" (2017). *Dissertations, Theses, and Masters Projects.* Paper 1499449907.
http://doi.org/10.21220/W43M27

VALIDATION OF AN INSTRUMENT FOR ASSESSING CONCEPTUAL CHANGE WITH RESPECT TO THE THEORY OF EVOLUTION BY SECONDARY BIOLOGY STUDENTS

A Dissertation
Presented to
The Faculty of the School of Education
The College of William and Mary in Virginia

In Partial Fulfillment
Of the Requirements for the Degree
Doctor of Education

by
Kevin David Goff
May 2017

VALIDATION OF AN INSTRUMENT FOR ASSESSING CONCEPTUAL CHANGE WITH RESPECT TO THE THEORY OF EVOLUTION BY SECONDARY BIOLOGY STUDENTS

by

Kevin David Goff

Approved May 2017 by

Juanita Jo Matkins, Ph.D.

Chairperson of Doctoral Committee

Paul D. Heideman, Ph.D.

Lee A. Kirkpatrick, Ph.D.

Carol L. Tieso, Ph.D.

Dedication

3

Table of Contents

Acknowledgments

Thank you, Carol, Lee, Paul, and Juanita Jo, for seeing me through. Your wisdom and influence are woven everywhere throughout these pages, and forever scratched onto the lenses through which I take in the world. From the very start I was resolved to make my doctoral study a crossroads where the human sciences would intersect with the hard sciences, where genuinely theory-driven research and real world application could converge, and where such ordinarily disparate fields as science education, statistics, curriculum and instruction, biology, anthropology, cognitive science, and evolutionary psychology could rendezvous around a common question. What hope could there be of drawing together a doctoral committee of good and giving souls with expertise spanning all those domains? Somehow it happened. I got lucky, I know, and I'm deeply grateful.

Thank you, Ann, Anne, Carla, Christy, Chuck, Jean, John, Kristi, Liz, Martha, and Tony (last names best left confidential), for helping me through the formidable challenge of accessing students and collecting data. And thanks, all you incognito test takers out there, for volunteering to participate in my research.

Thank you, Laura Bitto, Michael DiPaola, Chris Gareis, Leslie Grant, Judi Harris, Paul Heideman, Mark Hofer, Kyung Hee Kim, Lee Kirkpatrick, Heather MacDonald, Juanita Jo Matkins, Jackie McDonnough, Steve Staples, Carol Tieso, Megan Tschannen-Moran, Kristi Wagner, and Tom Ward – my teachers and colleagues in the School of Ed and beyond – for stretching me, changing me, and practicing the very pedagogy that is preached. You, too, Greg Capelli: hands down the most fearless thinker and teacher I've ever known …and I've been around. Sure wish you could read this.

Thank you, Robert Boyd, Pascal Boyer, David Buss, Susan Carey, Leda Cosmides, Richard Dawkins, Dan Dennett, David Geary, Robert Frank, Susan Gelman, Judith Rich Harris, Annette Karmiloff-Smith, Frank Keil, Rob Kurzban, Scott Mithen, Steven Pinker, John Tooby, Paul Richerson, Matt Ridley, Michael Tomasello, Dan Sperber, Bob Trivers, Frans de Waal, George Williams, E. O. Wilson, Robert Wright, and, of course, Chuck D – and many others lurking in these pages and on my bookshelf – for ushering me through one tectonic, paradigm-shifting conceptual change after another, most of them welcome, all of them transformative. Hope to meet you one day.

Thank you, Mom and Dad, for supporting me once again in countless ways, with boundless love and beneficence, during this latest pilgrimage. Same as you always have.

And most of all, thank you, Page. Thank you. No words will do...

List of Tables

List of Figures

VALIDATION OF AN INSTRUMENT FOR ASSESSING CONCEPTUAL CHANGE WITH RESPECT TO THE THEORY OF EVOLUTION BY SECONDARY BIOLOGY STUDENTS

ABSTRACT

This pilot study evaluated the validity of a new quantitative, closed-response instrument for assessing student conceptual change regarding the theory of evolution. The instrument has two distinguishing design features. First, it is designed not only to gauge student mastery of the scientific model of evolution, but also to elicit a trio of deeply intuitive tendencies that are known to compromise many students' understanding: the projection of intentional agency, teleological directionality, and immutable essences onto biological phenomena. Second, in addition to a section of conventional multiple choice questions, the instrument contains a series of items where students may simultaneously endorse both scientifically normative propositions and intuitively appealing yet unscientific propositions, without having to choose between them. These features allow for the hypothesized possibility that the three intuitions are partly innate, themselves products of cognitive evolution in our hominin ancestors, and thus may continue to inform students' thinking even after instruction and conceptual change. The test was piloted with 340 high school students from diverse schools and communities. Confirmatory factor analysis and other statistical methods provided evidence that the instrument already has strong potential for validly distinguishing students who hold a correct scientific understanding from those who do not, but that revision and retesting are needed to render it valid for gauging students' adherence to intuitive misconceptions.

Ultimately the instrument holds promise as a tool for classroom intervention studies by conceptual change researchers, for diagnostic testing and data gathering by instructional leaders, and for provoking classroom dialogue and debate by science teachers.

KEVIN DAVID GOFF

EDUCATIONAL POLICY, PLANNING, AND LEADERSHIP (EPPL)
CURICULUM LEADERSHIP

THE COLLEGE OF WILLIAM AND MARY IN VIRGINIA

VALIDATION OF AN INSTRUMENT FOR ASSESSING CONCEPTUAL CHANGE WITH RESPECT TO THE THEORY OF EVOLUTION BY SECONDARY BIOLOGY STUDENTS

Chapter 1: Introduction

Even after three decades of intense research, a wealth of scholarly dialogue and debate, and the evaluation of an astonishing variety of instructional interventions, the mission to help students undergo *conceptual change* with respect to difficult scientific concepts remains a daunting, plaguing challenge in science education. The problem is that students are not blank slates. They come to the science classroom already with intuitive understandings about natural phenomena, often deeply held and resistant to revision (Chinn & Brewer, 1993; Driver, Squires, Rushworth, & Wood-Robinson, 1994). At the same time, the concepts, principles, models, and theories of contemporary science are often less than intuitive, even counterintuitive; they are not easy to adopt or master. While it is true, as constructivist learning theory dictates, that science educators can often fruitfully build upon students' pre-instructional ideas (Bransford, Brown, & Cocking, 1999; Smith, diSessa, & Roschelle, 1993), it is also true that some prior ideas can actually impede their acquisition of scientifically normative concepts (e.g., Chi, 1992; Chinn & Brewer, 1993; Vosniadou, Vamvakoussi, & Skopetliti, 2008). Consequently, a teacher must sometimes design instruction to help students restructure, replace, or at least override their working conceptions within a content area; that is, to provoke conceptual change (Carey, 2000; Driver et al., 1994; Posner, Strike, Hewson, & Gertzog, 1982; Vosniadou & Brewer, 1987).

The challenge of leading students through conceptual change has not been unyielding, and effective educational strategies have been developed for particular scientific ideas. Researchers are far from consensus, however, on any single, widely applicable model of the *cognitive process* and *psychological mechanisms* at work whenever successful conceptual change does occur (Clement, 2008). I hope to contribute to the quest for an adequate cognitive model of conceptual change by taking a novel approach, heretofore unexplored in the science education research community: a Darwinian approach. The guiding theory here is that those cognitive mechanisms may be products of evolution by natural selection. That is, the capacity for conceptual change may be a psychological *adaptation* – subserved by innate neural architecture – which evolved in early humans because it enabled them to solve certain adaptive problems in their physical, ecological, and/or social environments. If we can determine what those ancient adaptive problems were, I believe we may gain useful insight about the cognitive processes and psychological mechanisms that make conceptual change possible for science students in the 21st century. And that in turn, I believe, would help us to identify instructional strategies to summon those mechanisms into action in the modern science classroom.

The specific purpose of the research described here was to evaluate the validity and reliability of a new quantitative, closed-response instrument for assessing student conceptual change with respect to one notoriously counterintuitive and difficult science concept: the theory of evolution itself (e.g., Bishop & Anderson, 1990; Evans, 2008; Jensen & Finley, 1995; Nehm & Reilly, 2007; Shtulman, 2006; Sinatra, Brem, & Evans, 2008; Smith, 2009, 2010). That same theory – applied to human psychology – served as the theoretical foundation upon which this instrument was developed. It has two distinguishing design features:

1. The instrument is designed not only to gauge student mastery of the scientifically normative model of evolution, but also to elicit common misconceptions where they exist. In particular, it provides opportunities to endorse three deeply intuitive interpretations that are well known to compromise students' understanding of evolutionary theory: the projection of intentional agency, teleological directionality, and immutable essences onto biological phenomena.
2. Like most closed-response tests, one portion of the instrument obliges students to choose between scientific and unscientific statements. Unlike most closed-response tests, however, a second portion allows students to simultaneously endorse *both* scientifically normative and unscientific-yet-intuitive positions, without having to choose between them.

The first of these two design features was informed by a Darwinian perspective on the human mind itself: The three interpretive projections – intentional agency, teleological directionality, and immutable essences – may not only be *intuitive*, but also partly *innate*, built by natural selection to provide our hominin ancestors with pragmatic (if unscientific) ways of conceptually construing events in their physical, ecological, and social environments. The second design feature was also informed by a Darwinian perspective. If indeed those three intuitive projections are partly innate, then it is possible that they continue to inform a student's thinking even after she has undergone conceptual change. That is, pre-instructional intuitions may continue to *coexist* in the student's mind alongside the new scientific understanding. My experimental instrument was designed to permit both intuitive and scientific conceptions to surface side-by-side whenever they coexist for the student, while also searching for signs that the student can suppress the former in favor of the latter.

This instrument's unconventional structure was specifically tailored to support future research on the cognitive mechanisms – themselves plausibly products of biological evolution – that enable students to undertake conceptual change with regard to counterintuitive concepts, not only the theory of evolution but other scientific theories and models as well.

Research Problem and Questions

The problem tackled in this doctoral research was to develop, field-test, validate, and revise an instrument to support future testing of specific Darwinian hypotheses about the human capacity for conceptual change. The instrument was piloted with 340 members of the target audience – high school science students – from six different schools in a variety of communities, from rural to urban and from low SES to affluent, and representing a diversity of racial, ethnic, and cultural heritages. Their responses were analyzed with confirmatory factor analysis and other statistical methods for evidence of instrument validity and reliability. The study addressed three primary research questions:

1. Does the instrument provide valid and reliable estimates of scientific and unscientific conceptions regarding the theory of evolution, for the target audience?
2. Do individual test items and item choices, including distracters, consistently map onto the intended concepts – that is, the scientifically normative

conception *and* common intuitive preconceptions/misconceptions – held by each student?

3. What modifications are warranted to render the instrument more valid and reliable, and to align individual items with the intended concepts?

My hope is that this instrument, once validated and revised, will support a research program in which various classroom interventions are assessed for their ability to provoke conceptual change with respect to evolutionary theory. I hope that it will also prove useful to other researchers and professional science educators for gauging student understanding of evolution.

Significance and Justification

Few scientific models are as essential for scientific literacy as the theory of evolution. In biological science, it is the grand unifying framework, the interpretive lens through which biologists habitually view and explain biological phenomena (Dawkins, 1986; Dennett, 1995; Dobzhansky, 1973; Maynard Smith & Szathmáry, 1995; Mayr, 1991). The importance of integrating evolution into high school science classes as a central organizing theory has been stressed by the National Science Education Standards (National Research Council, 1998), the new Next Generation Science Standards (Achieve, Inc., 2013), and virtually all leading associations in science education research and practice (e.g., National Academy of Sciences, 1998; National Association of Biology Teachers, 2008; National Science Teachers Association, 2003). Without it, a sound grasp of biological phenomena and principles is not possible.

This doctoral study may contribute to the mission of evolution education in several ways. First, it may offer the science education research community another instrument – one with some distinguishing features not found in other instruments – for assessing student conceptual change with respect to evolutionary theory. Because it consists solely of closed-response items, and because it is easily administered online, with student responses instantly and anonymously transferred to the researcher, the instrument has potential to support large-scale studies.

Second, this study may offer classroom practitioners another diagnostic tool for gauging how well their own students grasp evolutionary theory. Moreover, the instrument's challenging concrete scenarios may also be useful for stimulating class dialogue and providing students with feedback to help them distinguish a scientific understanding of evolution from common misconceptions about it.

Third, in an age when local school districts are bound by learning standards and standardized assessments that have been instituted by state policymakers – with evolutionary theory usually a core component of secondary science standards – this doctoral study may offer a mechanism for local benchmark testing and data gathering that is easily administered district-wide via online testing, and easily scored. Similarly, for science education specialists in state level departments of education, it may offer a relatively simple means of sampling student conceptions of evolutionary theory statewide.

More broadly, as I describe in Chapters 2 and 3, the instrument's unconventional, tailor-made design has special potential to facilitate future research into the mechanisms of the human mind that make *all* scientific conceptual change possible. By extension, it has potential to promote the development of educational strategies and instructional interventions to cultivate conceptual change in the classroom.

Definitions

The term *conceptual change* has meant many different things to different researchers in different fields, from developmental psychologists to science educators to philosophers and historians of natural science. In Chapter 2 I will carefully delimit the particular species of conceptual change pertinent to my own research agenda and which my experimental instrument is designed to assess. Briefly, it is conceptual change that has three characteristics. First, it is induced by classroom instruction (as distinct from conceptual change that occurs naturally and spontaneously during the normal course of childhood development and ordinary life experience). Second, it is conceptual change that is highly challenging for students because it requires them to restructure, replace, or override an entrenched pre-instructional mental model of the natural world (as opposed to conceptual growth more generally, such as additive or assimilative learning of new concepts). Third, it involves a target concept that may be challenging precisely because it conflicts with natural intuitions that are themselves biologically evolved.

I also use the term *intuitive* in a restricted sense, to refer to ways of conceptualizing and construing the natural world that are (plausibly at least) biologically evolved, hence innate (e.g., Barkow, Cosmides, & Tooby, 1992; Carey & R. Gelman, 1991; Hirschfield & S. Gelman, 1994a; Pinker, 1997, 2002; Sperber, Premack, & Premack, 1995) My instrument is designed to elicit three such intuitive tendencies in particular, which I identify with the technical terms *teleological, intentional,* and *essentialist* (e.g., Atran, 1998; Baron-Cohen, 1995; Carey, 2009; Dennett, 1995; S. Gelman, 2003; Gopnik & Meltzoff, 1997; Inagaki & Hatano, 2002; Keil, 1989; Leslie, 1994; Premack & Premack, 1995; Tomasello, 1999; Wellman, 1990). The term teleological refers to a psychological tendency to project function or purposeful direction onto certain objects and phenomena, such as tools, biological adaptations, and sometimes even whole species and ecosystems (Atran, 1998; Dennett, 1995; Inagaki & Hatano, 2002; Keil, 1989, 1994, 1995). An intentional intuition is a personifying tendency to project willful agency not only onto other people, but also nonhuman animals and even purely physical phenomena such as weather events (Atran, 2002; Baron-Cohen, 1995; Boyer, 2001; Carey, 2009; Dennett, 1995; Gopnik & Meltzoff, 1997; Inagaki & Hatano, 2002; Leslie, 1994; Premack & Premack, 1995; Tomasello, 1999; Wellman, 1990). The term essentialist refers to a tendency to project a hidden, immutable inner "essence" onto plants and animals – both individual organisms and whole species – that defines them as a discrete "natural kind," governs their species-specific behavior, directs their growth and development, and preserves their identity throughout the life cycle (Atran, 1998; Carey, 2009; S. Gelman, 2003; Inagaki & Hatano, 2002; Keil, 1989, 1994, 1995). I define these terms more fully in Chapter 2, where I also carefully explain the sense in which these intuitions may be "innate."

I cannot yet provide strict operational definitions for these five conceptually defined constructs, because the very function of my instrument is to "operationalize" them – that is, to render them as measurable quantities – and then to seek evidence that the instrument measures them validly and reliably. The instrument generates two major scores for each test-taker: a scientific score and an intuitive score. The first, if assessed both before and after an instructional intervention, would serve as a measure of *conceptual change* with respect to the theory of evolution. The second would serve as a gauge of a student's *intuitive* understanding of evolution, in the sense defined above.

The three specific intuitive tendencies – *teleological*, *intentional*, and *essentialist* – are also to be gauged by this instrument, but indirectly so. These intuitions often act in concert, rather than in isolation, to spawn misconceptions about evolution. For example, all three intuitions are manifested in the well documented misconception that evolution is a linear, inexorable, and progressive development of entire species toward ever more sophisticated forms (Evans, 2008; Jensen & Finley, 1995, 1996; Nehm & Schonfeld, 2008; Poling & Evans, 2004; Samarapungavan & Wiers, 1997; Shtulman, 2006). My instrument therefore seeks to "operationalize" student intuitions by positing several new constructs, each encompassing a number of well documented misconceptions and each manifesting one or more of the three core intuitions. For example, the linear, progressive view of evolution belongs to what I dub the "transformationist" family of misconceptions. "Transformationism" is one of several umbrella concepts around which test items were developed and that were hypothesized as latent dimensions scored by the instrument. In Chapters 2 and 3 I will explain the connections between these latent dimensions that the instrument was designed to measure and the underlying teleological, intentional, and essentialist intuitions.

Assumptions, Limitations, and Delimitations

A pervasive and driving assumption in this study is that much of human psychology can be fruitfully interpreted through a Darwinian lens, even classroom learning and reasoning about abstract concepts. In particular, valuable insight about students' intuitive reasoning and the process of conceptual change may be gained by investigating them as the work of biologically evolved cognitive mechanisms, psychological adaptations that solved specific adaptive problems in ancestral environments. This position is also a deliberate delimitation, for it is the theoretical foundation upon which this entire doctoral study is based. I defend it carefully in Chapter 2, but it is far from a universal paradigm in contemporary learning theory and psychological science, and for that reason I list it here as both an assumption and a delimitation.

A more mundane assumption is that the teachers who administered the experimental instrument did so according to the protocol that I shared with them, and that the testing conditions were adequately consistent across the six schools. The protocol was simple and the testing "environment" was online, so there is warrant for accepting this assumption. Less warranted is the assumption that test-takers answered the questions with adequate effort, attention, and honesty. This assumption is almost certainly not warranted for every participant, but there were safeguards in place: Participation was voluntary and limited to students who had made the effort to secure parental consent, they were free to opt out at any point, and their responses were anonymous, such that they could answer genuinely and intuitively without concern that incorrect answers might affect their grades. On the other hand, the absence of rewards and penalties may have made it easier for some students to relax their concentration.

The study's gravest limitation is that it was impossible to draw a random, representative sample from the ideal target population – namely, all high school biology students in the United States. It is thus impossible to generalize results and conclusions cleanly to future audiences. If the instrument is used with other student populaces and school settings, its estimated validity and reliability must be treated with caution and the test results interpreted accordingly. Another limitation beyond my control was the students' cognitive readiness, reading level, English proficiency, and religious orientation (which can affect a person's affective disposition toward the theory of evolution). These may have compromised the instrument's ability to elicit a clear portrayal of their conceptual understanding of evolution. It was not possible to control for these potential modifying variables, but it should be mentioned that students limited by reading ability, cognitive readiness, and so on *are* members of the target population for which the instrument was developed.

As for deliberate delimitations, I already identified the Darwinian lens that I chose as the theoretical framework for developing my instrument. I also already circumscribed the particular species of conceptual change that my study targets and the type of "intuitive" preconceptions and misconceptions that it seeks to elicit. Although my research agenda aims to investigate scientific conceptual change more generally, I elected to use the theory of evolution as my "test case" instead of other science topics where conceptual change is difficult. And again, I selected high school biology students in the United States as my target audience, but also chose to restrict my study to a non-random sample of convenience from that audience.

Overview of Dissertation

In Chapter 2, through a review of literature in science education, cognitive science, evolutionary biology, and evolutionary psychology, I construct the theoretical framework that inspired and informed the development of my experimental instrument. From a Darwinian stance, I elaborate the two theoretical foundations that prescribed the two special design features of my instrument. The first design feature, again, is that it not only seeks to gauge students' mastery of the scientific theory of evolution, but also seeks to assess any intuitive yet unscientific mental model that they may hold. The theoretical foundation that served as a blueprint for this design feature is a research-based taxonomy of common preconceptions and misconceptions about evolution. The second design feature is a section of test items that permit students to freely endorse *both* scientific *and* intuitive positions, without having to choose between them. The theoretical foundation for this design feature is a hypothetical model of evolved human cognition, according to which pre-instructional intuitions may continue to *coexist* in the student's mind alongside the new scientific understanding, even after successful conceptual change.

In Chapter 3 I describe the process by which test items were composed for the new instrument, and then vetted by two university professors with expertise in evolutionary science. I describe the field-testing of the instrument with 340 high school students, and the analytical methods used to assess instrument validity and reliability, especially confirmatory factor analysis via structural equation modeling.

In Chapter 4 I display and discuss the results. I interpret the factor analysis output and other statistics as evidence both for and against instrument validity and reliability. I describe the process by which I gradually made post hoc modifications to the original structural equation model, attempting to improve its fit to the data, but always in theoretically defensible ways and within the confines of the original conceptual framework that had governed test design. One of these modified models yielded a superior fit to the data, and I discuss this revelation vis-à-vis the original conceptual framework. Finally, I critically reexamine individual test items whose validity and reliability were called into question by factor analysis, and I make recommendations for revising or replacing them.

In the final chapter I review the findings broadly and discuss them in light of the research purpose stated here in Chapter 1, the cognitive models and theoretical foundations put forth in Chapter 2, and potential significance for instructional planning and curriculum leadership. I underscore limitations of the study, especially with respect to generalizing conclusions to future audiences. I discuss how the instrument, once refined and cross-validated, might be useful not only to fellow researchers but also science educators – from classroom practitioners to district level instructional leaders to state level science education specialists – especially in the enduring (albeit shifting) climate of standardized testing. I also propose that the instrument could be useful not only for assessing student understanding of evolutionary theory, but cultivating that understanding as well, by using its challenging scenarios to stimulate classroom discussion and call attention to the intuitive appeal of its unscientific distracters. Finally, I suggest directions for future research to further investigate and improve the instrument's validity and reliability.

Chapter 2: Literature Review and Theoretical Framework

In the late 1960's, in a retired South Dakota gold mine beneath a mile of rock, Raymond Davis and John Bahcall filled a huge cylindrical tank with 100,000 gallons of chlorine-rich dry-cleaning fluid (Bahcall, 1969; Ferris, 1997). After letting the tank rest undisturbed for several weeks, they bubbled helium through it, then passed the helium through a series of filters – one chilled to -200°C – and a gas chromatograph. In this way they were able to isolate an extremely small amount of radioactive argon-37 that the helium had purged from the cleaning fluid. To quantify the argon, they sequestered it in a heavily shielded radiation counter for one year, allowing it to decay through 10 half-lives. The amount of radiation emitted during that year told them how much radioactive argon had been in the 100,000-gallon tank to begin with: only a few dozen atoms, as it turned out. Yet these argon atoms were *not* in the tank at the very start of the experiment. Rather, they were created in the weeks that followed when tiny, fast-moving cosmic particles penetrated the tank, collided with chlorine atoms in the cleaning fluid, and transmuted the chlorine into argon. What Davis and Bahcall had captured and counted with their giant tank, were solar neutrinos. These exceptionally small particles had been born in the belly of the sun, a product of nuclear fusion reactions. The renowned physicist Wolfgang Pauli predicted their existence four decades prior on *theoretical* grounds. Mathematical and quantum mechanical *models* suggested that they must pour forth from the sun in unimaginably prodigious quantities. Yet this was the first time any had ever been empirically detected – and not very many.

Davis and Bahcall's experiment ultimately proved Nobel-worthy, and its ingenuity and sophistication are awe-inspiring enough. But I have described it here because it is astonishing for a different reason: Evolutionarily speaking, the human brain should not – indeed could not – have evolved to detect neutrinos. Natural selection can only adapt organisms to tangible, statistically recurrent threats to their survival or reproductive success (Tooby & Cosmides, 1992). Our bodies and brains evolved to respond only to those features of the environment that bore upon the reproductive success of our ancestors, such as predators, parasites, calorie-rich carbon compounds, gravity, and solar radiation both in the visible spectrum and its invisible cancer-causing wavelengths.

Yet neutrinos are none of those things. They almost never interact with the "ordinary" matter that composes our bodies, brains, and Earthly environment, and so could not have influenced the course of our evolution. They are electrically neutral, have a vanishingly small mass, and travel at nearly the speed of light. This means they are virtually immune to the pull of gravitational and electromagnetic attractions that might otherwise cause them to "crash" into the protons, electrons, and neutrons of which our world, our flesh, and our DNA are made. Trillions pass through our bodies every second. We are transparent to neutrinos, and so is planet Earth. Davis and Bahcall did their experiment deep underground to shield out all other cosmic particles, like gamma rays, that might otherwise contaminate the chlorine solution, producing false positives. Neutrinos, by contrast, freely pass through rock and giant vats of cleaning fluid. Only on the rarest of rare occasions is one intercepted by an atomic nucleus, captured by an extremely short-range force (the “weak” force) that inhabits all atomic nuclei. If it happens to be the nucleus of a chlorine atom, the absorption of a neutrino will transform it into argon-37. In a 100,000-gallon tank containing a quadrillion quadrillion chlorine atoms, the daily shower of five billion trillion solar neutrinos will yield only one or two such collisions per day. Davis and Bahcall had done this math in advance – *based on underlying scientific theory* – and they knew that after a month or so, if neutrinos were real and the scientific models correct, their experimental setup should yield enough radioactive argon atoms (only a few dozen) to harvest and count.

Thus a rich scientific theory told the researchers what to look for, how to look for them, and where. And when at last they found the neutrinos right where they were looking for them, this lent tremendous empirical support to the theory that had spawned the predictions. In short, they had unearthed a deep and almost impossibly hidden truth about the natural universe. The history of science abounds with other examples of elusive truths brought to light *not* by sweeping exploration and discovery, but through a targeted, theory-driven search.

My own research ambitions are theory-driven, too, as I will show in this literature review. The overarching question that inspires them is this: How could *Homo sapiens* have evolved a brain able to uncover *and understand* the natural universe’s most hidden truths, even truths about things like neutrinos that never affected – neither directly nor indirectly – the reproductive fitness of our ancestors? The existence of neutrinos was predicted by theories of quantum mechanics, particle physics, and nuclear chemistry that are formidably *counterintuitive*. They posit concepts and principles that seem to fly in the face of the familiar physical laws governing our macroscopic world (Ferris, 1988/2003, 1997; Krauss, 2012). Those are the same macroscopic laws that determined who among our ancestors would survive and who would not, and so shaped their bodies and brains. Much of this macroscopic physics we are able to grasp quite *intuitively* – perhaps even instinctively (Carey, 2009; Carey & Spelke, 1994; Leslie, 1994, 1995; Pinker, 1997; Spelke, 1991; Spelke, 1994; Spelke, Phillips, & Woodward, 1995). Psychology experiments nearly as ingenious as Davis and Bahcall's show that even a two-month-old infant intuitively knows that one solid object ought not pass through another (e.g., Baillargeon, Spelke, & Wasserman, 1985), and nothing in evolution, it seems, could have prepared her newborn mind for violations of that intuition – like neutrinos. How then did natural selection fashion a brain that can suspend its own intuitions about physical phenomena in order to conceive such profoundly counterintuitive ideas as quantum indeterminacy, quantum entanglement, the wave-particle duality of light, the *ex nihilo* birth of new particles in the vacuum of empty space, and the relentless rain of neutrinos straight through our "solid" planet?

Clearly the answer lies largely in the fact that humans evolve not only *biologically*, but also *culturally*. Natural selection transformed our ancestors into a species that can transmit newly acquired knowledge and knowhow from one generation to the next (Richerson & Boyd, 2005, Sterelny, 2012; Tomasello, 1999, 2014; Tomasello, Kruger, & Ratner, 1993). The capacity for cultural transmission required the evolution of specialized mental machinery, including neural circuits in the brain that permit a person to adopt another's point of view and that generate imitative behavior – both rarities in the animal kingdom (Tomasello, 1999; Tomasello et al., 1993). Once such cognitive adaptations were in place, incremental improvements in tool technology and the social group's collective understanding of the natural environment could gradually accumulate. Modern science is both an extension and engine of such cultural evolution, made possible by brains built for sharing ideas across generations. On this view, contemporary quantum mechanics – counterintuitive though it is – is just another burgeoning technology, no different than the rise of Acheulean handaxes, agriculture, writing, and sailing ships in millennia past. We are able to unearth neutrinos and other deeply buried truths about the universe because we are *biologically* evolved to evolve *culturally*.

But I believe that this explanation is incomplete. By itself, I suggest, it cannot account for the biological evolution of a brain that can adopt culturally evolved ideas that are *counterintuitive*. It fails to address the enormity of the conceptual leap – the Kuhnian "paradigm shift" – that the quantum mechanics revolution represents (Carey, 2009; Kuhn, 1962/1970; Thagard, 1992). It leaves unanswered the question of how our minds could be cognitively adapted to abandon our most strongly held intuitions in order to apprehend and accept scientific concepts that are *not* intuitive, and sometimes deeply counterintuitive. *As a partial answer to that question, I offer this complementary hypothesis:*

> Paradoxically, what made modern science so spectacularly good at uncovering *truth* in the *natural* universe was the omnipresence of *deception* in the ancestral *social* environment. Our minds can suspend intuition, adopt counterintuitive truth claims, and uncover nature's most hidden truths, precisely because our ancestors were often pressured to detect untruths during their social interactions. Our brains evolved to entertain competing representations of reality – while spotting misrepresentations of reality – put forth by others in the social group about *social*

> affairs and *social* concerns. This capacity ultimately paved the way for scientists to craft and refine remarkably successful representations of *non-social* reality as well – that is, of *natural* phenomena – while weeding out poorer representations, *even those that are intuitively appealing.* In the ancestral social setting, “truth” was negotiable and slippery, and it was this that prepared human brains – hence human minds – to be open to the sometimes non-intuitive yet powerfully predictive models and explanatory theories of modern science.

This is the hypothesis that fuels my research ambitions and frames the present study. If it harbors some truth, then I believe it could have implications for science education and classroom practice. In this chapter I will develop this hypothesis broadly, while striving to connect it both to the present study and potential future research.

Theoretical Framework for Instrument Development: Two Models & Seven Theses

The new quantitative, closed-response instrument that was piloted in this doctoral study is designed to assess student understanding of one notoriously non-intuitive scientific model: the theory of evolution by natural selection. That same theory – applied to human psychology – served as the theoretical foundation upon which this instrument was developed. The instrument has two distinguishing design features:

1. The instrument is designed not only to gauge student mastery of the scientifically normative model of evolution, but also to elicit common misconceptions where they exist. In particular, it provides opportunities to endorse three deeply intuitive interpretations that are well known to compromise many students’ understanding of evolutionary theory – namely,

the projection of intentional agency, teleological directionality, and immutable essences onto biological phenomena.

2. Like most closed-response tests, one portion of this instrument obliges students to choose between scientific and unscientific statements. Unlike most closed-response tests, however, a second portion of this instrument allows students to simultaneously endorse *both* scientifically normative *and* unscientific-yet-intuitive positions, without having to choose between them.

The first of these two design features was informed by a Darwinian perspective on the human mind itself: The three interpretive projections – intentional agency, teleological directionality, and immutable essences – may not only be *intuitive*, but also partly *innate*, built by natural selection to provide our hominin ancestors with pragmatic (if unscientific) ways of conceptually construing events in their physical, ecological, and social environments. In other words, the human brain may be poorly evolved to grasp evolution itself.

The second of these two design features was also informed by a Darwinian perspective. If indeed those three intuitive projections are partly innate, then it is possible that they continue to inform a student's thinking even after she has undergone conceptual change. That is, pre-instructional intuitions may continue to *coexist* in the student's mind alongside the new scientific understanding. Probably the most pervasive metaphor in the educational literature is that student conceptual change requires either a *restructuring* or the *replacement* of an existing conceptual model. The evolutionary perspective adopted here, however, suggests a different metaphor: The challenge is not to restructure or replace one's intuitive pre-instructional conception, but to *suspend* and *suppress* it. My experimental instrument seeks to permit both intuitive and scientific conceptions to surface side-by-side whenever they coexist for the student, while also searching for signs that the student is suppressing the former in favor of the latter.

Two Models for Instrument Development

The theoretical blueprint for the first design feature above is captured in Table 1: a taxonomy of common pre-instructional, intuitive, and unscientific conceptions about the theory of evolution. The theoretical foundation for the second design feature is captured in Figure 1: a hypothetical model of the evolved, intuitive human mind. I will not elaborate these two models here, but develop and discuss them instead throughout this literature review.

In elaborating these two models, I will also lay the theoretical groundwork for a more general research agenda to empirically explore the *cognitive mechanisms* at work whenever students successfully undergo conceptual change. It is that agenda for which I developed my experimental instrument. Science educators have developed an astonishing variety of educational strategies and instructional interventions to help students undertake conceptual change with respect to specific science concepts. For 30 years, Duit (2009) has maintained a running bibliography devoted to the conceptual change literature. It now lists 8,400 entries and is almost 600 pages long, single-spaced. Yet researchers are far from consensus on any single, widely applicable model of the *cognitive processes and mechanisms* at work whenever successful conceptual change does occur. John Clement, long a leading scholar in the field, writes in the 2008 *International Handbook of Research on Conceptual Change:*

Table 1

Taxonomy of Common Pre-Instructional, Intuitive, and Scientifically Non-Normative Conceptions about the Theory of Evolution

Conceptual barrier / Non-normative conception / Intuitive bias or preconception	Causal stance / Mode of construal
Immutability (IMB): Natural kinds / essences are immutable.	
IMB.1 – Resistance to idea that species can change/evolve at all, from one "natural kind" into another (especially speciation and macroevolution). Essences are immutable. [1, 4, 5, 10, 11]	Essentialist
IMB.2 – Competition and "survival of the fittest" are interspecific interactions rather than intraspecific. Selection occurs between different natural kinds, not within a single kind. [9]	Essentialist
IMB.3 – Speciation and macroevolution occur through interbreeding of different species/natural kinds (blending essences). [8]	Essentialist
Transformationism (TFM): Evolution occurs across/between generations, driven by endogenous, progressive transformation of *all* members of a species rather than intraspecific variation and selection among them.	
TFM.1 – Difficulty grasping that intraspecific variation – that is, differences among members of a population – is what fuels natural selection. Belief that members of a natural kind are virtually identical or only superficially variant, such that differences do not affect survival or adaptation. [1, 2, 5, 6, 9, 11, 15]	Essentialist
TFM.2 – Difficulty grasping that evolution is a change in population membership and a shift in trait frequencies, not a transformation of the members themselves, nor the traits themselves, nor the species' essence. Tendency to view evolutionary changes as simultaneous adaptation of all members of a natural kind – transforming the species as a whole – rather than statistical adaptation of the population. Evolution is endogenously driven rather than a result of external selection – an ongoing metamorphosis of a natural kind's inner essence – such that offspring traits may differ directionally/progressively from their parents. [1, 2, 4, 5, 9, 13, 14, 16]	Essentialist + Teleological
TFM.3 – Evolution as linear instead of branching, and typically directional, progressive, quasi-developmental, and/or inherently driven in the direction of higher, more complex, more sophisticated forms – even perfection – perhaps as an intrinsic striving. Evolutionary adaptation occurs to prevent species extinction. Extinction is rare. [4, 7, 8, 9, 10, 11, 13]	Teleological + Essentialist (+ Intentional?)

(continued)

Table 1 (continued)

Taxonomy of Common Pre-Instructional, Intuitive, and Scientifically Non-Normative Conceptions about the Theory of Evolution

Conceptual barrier / Non-normative conception / Intuitive bias or preconception	Causal stance / Mode of construal
Within-Lifetime Adaptation (WLA): Evolutionary adaptation occurs within single generations/lifetimes via individual and typically heritable changes.	
WLA.1 – Adaptation is need-driven – a functional response to environmental conditions, changes, or duress – and/or directly caused by the environment itself. Beneficial mutations arise due to need, as an adaptive response, and/or are induced by environmental agents. Individuals "have to" adapt. An explanation of adaptive function or necessity for survival serves as a sufficient causal explanation of a trait's origin and existence. [1, 2, 3, 4, 5, 7, 8, 9, 10, 12]	Teleological
WLA.2 – Adaptation as behavior-based, effort-based, and/or a consequence of use/disuse coupled with the inheritance of acquired characteristics (quasi-Lamarckian). Learned behaviors are genetically heritable, and the fruits of repeated "practice" can be passed to offspring. Organisms can adapt, develop beneficial mutations, or evolve into new natural kinds through "wanting," "trying," and forward-looking, goal-directed behavior. [1, 2, 4, 5, 7, 8, 9, 10, 11, 12, 16]	Intentional (Teleological?)
WLA.3 – Conflation of long-term adaptation via natural selection with the daily behavioral and homeostatic "adaptation" of individual organisms to variable conditions. Confounding species- or population-level adaptation across generations with individual-level adaptation within lifetimes. [1, 2, 3, 4, 9, 13]	Essentialist + Teleological

Note. Citations are as follows: [1]D. Anderson et al., 2002, [2]Bishop & C. Anderson, 1990, [3]Brumby, 1984, [4]E. M. Evans, 2008, [5]P. Evans & D. Anderson, 2013, [6]S. Gelman, 2003, [7]Jensen & Finley, 1995, [8]Jensen & Finley, 1996, [9]Nehm & Schonfeld, 2008, [10]Poling & E. M. Evans, 2004, [11]Samarapungavan & Wiers, 1997, [12]Settlage, 1994, [13]Shtulman, 2006, [14]Shtulman & Calabi, 2013, [15]Shtulman & Schulz, 2008, [16]Sinatra et al., 2008.

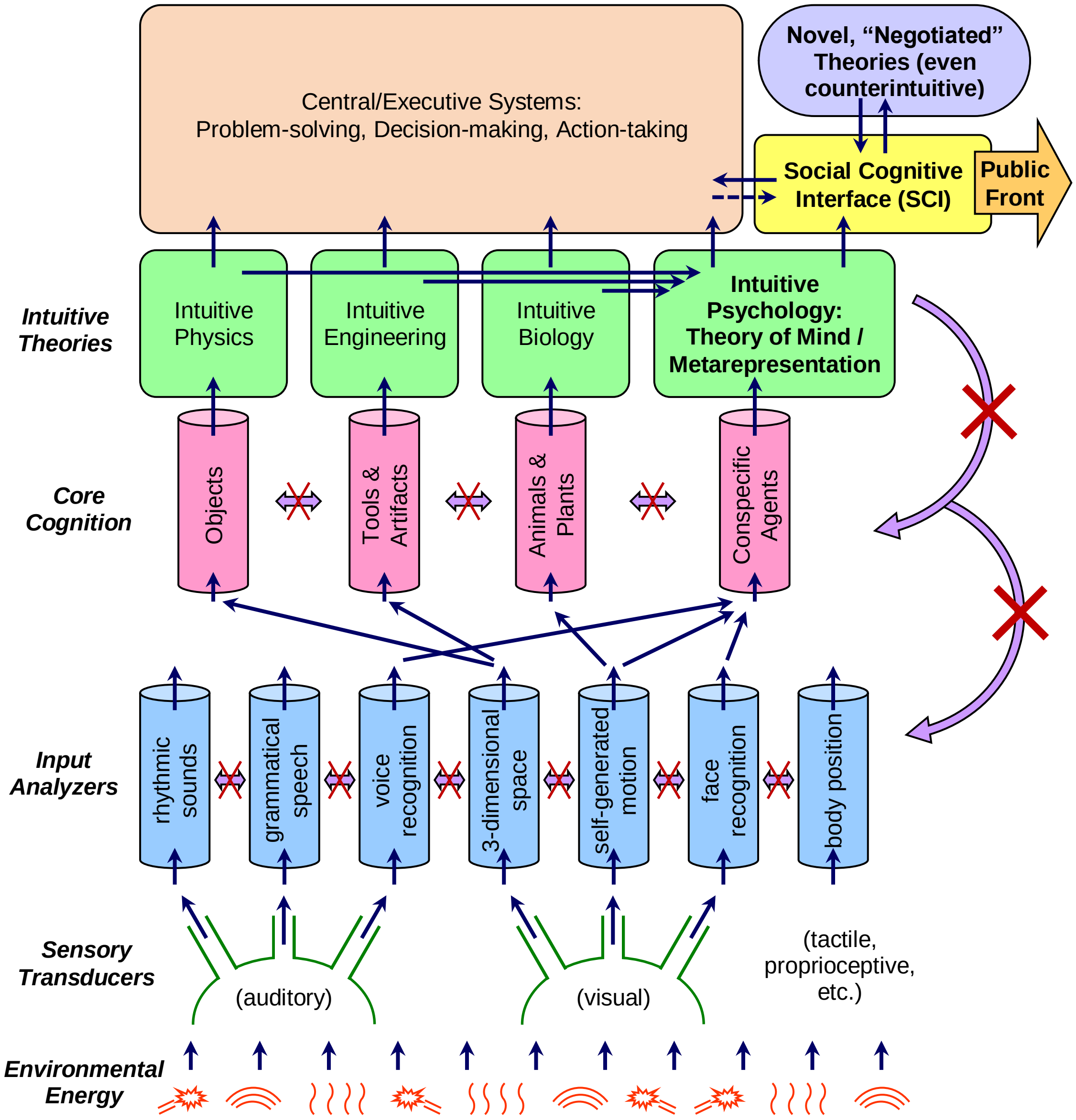


Figure 1. A hypothetical model of the evolved, intuitive mind. Information from the environment enters the cognitive system through sense organs at the bottom of the diagram. Distinct patterns of information are selectively organized into basic perceptual representations, which are then channeled upward into the appropriate conceptual systems. Four conceptual systems are shown, each dedicated to a different domain that was relevant to ancestral hominins and depicted here because they generate four forms of "intuitive science": physics, biology, psychology, and engineering. Their output is then piped to the executive and social systems, where for the first time mental representations from different functional domains may be integrated with one another and conceptual change may occur. The model is hybridized from the work of multiple scholars; see text for references. (Language centers are deliberately omitted from this model.)

> We still need to address an enormous gap that remains at the core of conceptual change theory: we do not have an adequate cognitive model of the basic conceptual change process….Most of the classical theory in science education…is either about *conditions* for change…or *factors* that make it easier or more difficult….What is missing is a fuller specification of *mechanisms* of change – causal descriptions of processes that produce conceptual change. (p. 417)

The theoretical framework that I will develop in this literature review aims to contribute to this quest for "an adequate cognitive model of the basic conceptual change process" – that is, a "causal description" of the cognitive "mechanisms" that incite conceptual change.

Here too, the guiding theory is that those cognitive mechanisms may themselves be products of evolution by natural selection. That is, the capacity for conceptual change may be a psychological *adaptation* – subserved by innate neural architecture – which evolved in early humans because it enabled them to solve certain adaptive problems in their physical, ecological, and/or social environments. Like tool use and spoken language, conceptual change is a distinctly human capacity. Tool use and language are species-specific *cognitive* and *behavioral adaptations* that evolved in our hominin ancestors through natural selection, acting upon *both body and brain* (Gamble, Gowlett, & Dunbar, 2014; Mithen, 1996; Pinker, 1994, 1997; Richerson & Boyd, 2005; Tomasello, 1999; Tooby & Cosmides, 1992.). Their adaptive benefits are plain. The cognitive capacity for conceptual change likewise must have evolved long ago through natural selection, yet its adaptive benefits are not so clear. Why did ancestral humans evolve an ability to abandon preexisting, intuitively compelling mental models of the world in favor of alternative, even counterintuitive ones? What ecological and/or social conditions prompted its evolution, and what adaptive problems did it solve? I believe that a consideration of such questions may yield useful insight about the cognitive processes and psychological mechanisms that make conceptual change possible for 21st century science students. And I believe that this in turn will suggest educational strategies and instructional interventions to summon them into action in the modern science classroom.

A Sequence of Seven Theses toward Instrument Development

In this chapter I will erect a Darwinian theoretical framework as the foundation for Table 1, Figure 1, and my experimental instrument. I will develop that framework step-by-step through the following sequence of 7 theses:

1. Evolution by natural selection is a process that inherently tends to produce *representations* – or “re-presentations” – of physical reality. Through evolution, certain recurrent patterns in the natural world come to be reflected or “re-presented” in the bodies and behaviors of living organisms.
2. Among those re-presentations, in organisms with complex nervous systems, are *mental* representations. In the hominin line, biological evolution has bestowed us with innate cognitive mechanisms for mentally representing, interpreting, construing, conceptualizing, and reasoning about the natural world. Many of these mechanisms are *domain-specific* and *content-* or *context-sensitive*: Each was “designed” by natural selection to make sense only of certain kinds of natural or social phenomena, and each is automatically mobilized in the mind by the appropriate context or content.

That is, certain contexts tend to arouse or activate particular mechanisms of the evolved mind, and thus tend to evoke certain *intuitive* interpretations or construals. Among these construals are a trio of intuitions that make it difficult for students to master the theory of evolution: the projection of intentional agency, teleological directionality, and immutable essences onto biological phenomena.

3. In many animal species, *learning*, too, is motivated and shaped by an innate cognitive architecture, such that each individual's behavior becomes calibrated to select features of the local environment in species-typical, fitness-promoting ways. This is by evolutionary design. In our own species, early childhood conceptual development in the domains of physics, biology, psychology, and engineering is largely *calibrational*: Through local experience, the child constructs an elaborate "intuitive theory" by fleshing out an innate "skeletal" framework already built into the mind. Calibrational learning is categorically distinct from *conceptual change*.
4. Once our hominin ancestors evolved cognitive machinery for acquiring knowledge and knowhow from others in the social group – that is, for cultural transmission – human populations could accumulate behavioral and conceptual changes non-genetically, yet these changes were still channeled and constrained by biologically evolved intuition. Eventually, however, cultural evolution produced the self-correcting empirical mechanisms of modern natural science, which have been able to generate and validate conceptual models that conflict with biologically evolved intuition.

Consequently, an especially challenging form of *conceptual change* – as opposed to calibrational learning – is necessary for students to master those models: They must restructure, replace, or at least override a pre-instructional theory that is so well entrenched, it may even continue to coexist in the student's mind alongside the newly acquired scientific understanding.

5. Natural selection might reasonably be expected to select *against* any mode of learning that violates or subverts innate intuitions. For this reason, the human ability to adopt counterintuitive concepts itself demands evolutionary explanation. One possibility is that *conceptual change* was in fact adaptive, evolving because it enabled each individual to construct an ever more accurate understanding of the natural world, thus conferring heightened cognitive and behavioral flexibility in the face of environmental variability. I call this the *individual constructivist hypothesis* of the evolutionary origin of conceptual change. In practice, many of the pedagogical strategies advocated by science education scholars implicitly assume such an evolutionary origin.
6. An ability to adopt ideas that violate or subvert innate intuitions may have been ancestrally adaptive *not* because it gave individuals the cognitive flexibility to rationally improve the predictive accuracy of their conceptual models (= the individual constructivist hypothesis), but instead because it enabled them to *become* cognitively flexible by actively assimilating the incremental improvements in conceptual models that had accumulated in their culture over previous generations. I call this the *cultural re-constructivist hypothesis* of the evolutionary origin of conceptual change. In practice, many

of the pedagogical strategies advocated by science education scholars implicitly assume such an evolutionary origin.

7. The cognitive mechanisms supporting scientific conceptual change may have originally evolved *not* for flexibly making sense of *natural* phenomena (= the individual constructivist hypothesis), *nor* for restructuring or replacing intuitive preconceptions in favor of *culturally* normative ones (= the cultural re-constructivist hypothesis), but instead for vetting and deciding among competing truth claims made by fellow members of one's social group about *strictly social* affairs. Only much later were those cognitive mechanisms recruited to *non-social* phenomena – that is, the objects of natural science. I call this the *social strategic hypothesis* of the evolutionary origin of conceptual change. Certain pedagogical strategies advocated by science education scholars accord with such an evolutionary origin.

As I will show, a corollary of this last thesis – the social strategic hypothesis – is that the psychological machinery that carries out conceptual change may function largely independently of the machinery that carries out calibrational learning and generates our core intuitions in the domains of physics, engineering, biology, and psychology. Consequently those core intuitions may continue to affect a student's thinking and expression of ideas even after she undergoes conceptual change. Again, my experimental instrument was deliberately designed to allow for that possibility, and potentially to reveal it.

Thesis #1 – The Meshing of Organic Life to World: The Evolution of "Representation" and Functional Fit

An adaptation is a reliably developing structure…which, because it meshes with the recurrent structure of the world, causes the solution to an adaptive problem.

John Tooby & Leda Cosmides (1992, p. 104)

To characterize conceptual change in evolutionary terms, I must first establish that the human brain/mind is "prewired" to construct distinctive, domain-specific *mental representations* of the world, and that these representations constitute potent "intuitions" that do not always accord with modern scientific models of reality. The human brain, like every other bodily organ, is a product of biological evolution through natural selection. Among its many functions, it evolved a constellation of psychological mechanisms that provided our hominin and pre-hominin ancestors with *pragmatically useful* ways of interpreting phenomena in their physical, ecological, and social environments (Barkow et al., 1992; Buss, 2004; Carey & R. Gelman, 1991; Hirschfield & S. Gelman, 1994a; Pinker, 1997, 2002; Sperber et al., 1995). I will show that although these mechanisms must be calibrated to local conditions through childhood experience, their operation is in a strong sense innate, generating many of the powerful, persistent constructions and construals with which we all make sense of experience. Our minds are predisposed to *mentally represent reality* by projecting a variety of stock interpretations onto the outside world. These projections are *domain-specific* and *content-* or *context-sensitive*: Each is "designed" (by natural selection) to make sense only of certain kinds of natural or social phenomena, and each is automatically evoked in the mind by the appropriate context or content. Though functionally handy and intuitively powerful, these projections are not necessarily scientific. It is amidst an ocean of such intuitions that the contemporary science student must strive to apprehend and adopt the often counterintuitive representations of reality put forth by modern science. Of course, conceptual change in the science classroom does not always represent a raw collision of scientific concepts with innate intuitions, but I will build a case that it often does.

The notion that many of our intuitions are innate is vulnerable to misinterpretation and so will demand careful articulation. In particular, people often equate "innate" with rigid reflex reactions, stereotyped stimulus-response behaviors, and inborn "instincts," as opposed to the products of "learning." Learning, it is thought, is the opposite of instinct, for it derives from empirical experience during childhood, and this produces the pre-instructional conceptions that students bring to the science classroom. The following passage from a seminal work on conceptual change exemplifies the empiricist stance of many science education researchers:

> Children develop ideas about natural phenomena before they are taught science in school….They have *experiences* of what happens when they drop, push, pull, or throw objects, and in this way they build up ideas and expectations relating to the way objects feel and move. Similarly, ideas about other aspects of the world around them develop *through experiences* with, for example, animals, plants, water, light and shadows, fires and toys….Many of the conceptions which children develop about natural phenomena *derive from their sensory experiences*….Children have ways of construing events and phenomena which are coherent and fit with their own *domains of experience* yet which may differ substantially from the scientific view. Studies also indicate that these notions

> may persist into adulthood despite formal teaching….Children's [initial] science conceptions are not idiosyncratic, nor are they in many cases heavily culturally dependent. *They are shaped by personal experience with phenomena.* (Driver et al., 1994, pp. 1-3; emphases added)

An implicit assumption here is that the many documented commonalities that exist among students' starting conceptions, even across diverse cultures, arise from universal *experiences* with similar natural phenomena in the same natural world.

Against this empiricist position, I will build a case that those commonalities also often stem from a universal architecture which natural selection built into the human mind. I will show that (etymology aside) "innate" does not strictly mean "inborn," for genes continue to organize our bodies, brains, and behaviors throughout the course of normal human development. I will show that the opposition of "instinct" to "learning" is a false dichotomy, for evolution has endowed us with "instincts for learning" (Marler, 1991). I will describe much of this "instinctive learning" as "calibrational": Through empirical experience, our innate expectations become attuned to – but not overwritten by – the particulars of the local environment. Analogously, a scientific instrument like a pH meter is designed to measure and "interpret" a specific feature of the environment, yet must still be periodically calibrated through exposure to solutions of known pH. Through such "experience," the instrument "learns" about the natural variation in its local setting. And this is by design. The capacity for calibration is "innate," deliberately built into the device by its engineers. Similarly, I will argue, local experience calibrates many of the innate learning mechanisms that evolution built into the human brain and mind – and this too is by (evolutionary) design. Consequently, we come to interpret experience in evolutionarily ordained ways, ways that gave our ancestors a flexible yet stable grip on a variable and sometimes unpredictable world. *Conceptual change*, I will argue, differs from such *calibrational learning*. It is not mere calibration. Calibration is intuitive learning; conceptual change is not.

The key point is that many of the pre-instructional concepts which a student brings to the classroom may have emerged during the course of his cognitive development through a complex, calibrational interaction between his innate psychological architecture and "personal experience with phenomena" (Driver et al., 1994; see quote above). If so, then a psychometric instrument seeking to assess conceptual change should be able to reveal the presence of innate intuitions where they exist or persist. *In this section, I develop the following thesis:*

> Evolution by natural selection is a process that inherently tends to produce *representations* – or "re-presentations" – of physical reality. Through evolution, certain recurrent patterns in the natural world come to be reflected or "re-presented" in the bodies and behaviors of living organisms.

This will set the stage for a discussion of the *mental representations* of physical reality that the human mind/brain evolved to construct, and for the role that "calibrational learning" plays in the development of those mental representations.

Four Forms of Biological Re-Presentation via Natural Selection

Mirror-like representations. "Representation" (re-presentation) is a natural outcome of evolution. It arises from the coupling of two general phenomena: recurrence and replication. Recurrence is simply the repetition in time or space of distinct patterns or events: the 24-hour spin of the earth on its axis, the crashing of cyclones into a coastline, the patchy distribution of oases in a desert, the presence of hidden predators at watering holes, the blossoming of fruit trees in spring, the migrations of fish in fall, the onset of illness after a spider bite, the efforts of males to woo females, the cries and coos of an infant in its mother's arms. Replication, by contrast, is the faithful (but not always perfect) *copying* of some preexisting pattern: the proliferation of crystals, the doubling of chromosomes, the transcribing of texts, the photocopying of sheet music, the retelling of a story, the mimicking of a behavior. Whenever a recurrent condition or event influences the rate at which a replicator replicates – and given the fact that copying errors inevitably creep in – evolution occurs (Darwin, 1859; Dawkins, 1976/1989; Dennett, 1995).

In the biological world the primary replicators are genes, discrete stretches of DNA that spell out the developmental recipe for building a living organism (Dawkins, 1986), including in some species a brain that regulates its behavior. Natural selection "adapts" that recipe to those statistically recurrent features of the environment that influence future reproductive fitness – that is, successful replication of genes (Tooby & Cosmides, 1992). It is a relentless "algorithmic" (Dennett, 1995) feedback cycle in which organic forms that fit well with the environment are automatically propagated at the expense of organic forms that fit less well. Incrementally, it fine-tunes the bodies, brains, and behaviors of organisms such that they increasingly "re-present" the fitness-relevant features of their surroundings.

An example: Beaches throughout the world are inhabited by ghost crabs (*Ocypode* spp.), so named because of their largely nocturnal habits and the cryptic coloration (camouflage) that gives them a phantom-like "transparency" against the sand. Each population's markings closely mimic – or "re-present" – the color and composition

Figure 2. Cryptic coloration in three representatives of genus *Ocypode* on three different beaches. Images: Dorothy Pugh, Chuck Elzinga, Ray Farm.

of the local beach, plainly an adaptation against visual predators on high (Figure 2). Individual crabs do not adjust their coloration in any way. If one were to transfer all the light-colored crabs from a white Bahaman beach to a volcanic island with black sand, their offspring would still develop into light-colored crabs. Nevertheless, there would almost surely be some slight, heritable differences in coloration, and if any crabs were lucky enough to survive, natural selection would favor the darkest (or least light) crabs in each generation. Having escaped predation, those crabs would reproduce, and as a result, genes for a darker carapace would be increasingly well represented in each successive generation. Moreover, if by chance a genetic mutation were to arise that made a crab darker still, that new gene would swiftly spread within the population. After many such cycles, a very dark phenotype would predominate. Over time, differential survival, reproduction, and gene replication would "mesh" the "design" of the ghost crab body with certain recurrent features of the environment (Darwin, 1859; Dawkins, 1986; Tooby & Cosmides, 1992).

Thus this "representational" meshing of body to background yields a kind of *mirror* of reality: The natural world produces a striking replica of itself in the body of a living organism. Later I will observe that *mental representations,* too, can sometimes be likened to mirror images or replicas of the outside world. (It is worth pointing out that, from the crab's perspective, the goal is not to represent but to *misrepresent* reality to its

Figure 3. Form follows function: The hand-in-glove fit of a hummingbird bill (*Calothorax lucifer*) to a flower (and vice versa); the oyster knife shape of an oystercatcher's bill (*Haematopus palliatus*), perfect for prying open bivalves; the elongated lower mandible of a black skimmer (*Rynchops niger*), for scooping up fish on the fly. Images: Charles W. Melton, Mia McPherson, and Joe Reynolds (left to right).

predators – that is, to deceive their nervous systems [Dawkins, 1976/1989, 1982; Trivers, 1985, 2002]. Later, the evolution of misrepresentation and deception will figure prominently in my "social strategic" hypothesis regarding the evolution of conceptual change, under Thesis #7.)

Functional fit representations. Far more often, natural selection "re-presents" reality not as a mirror image or replica, but as a *functional* fit between organism and environment. Bird bills are a nice example (Figure 3). Others are the parachute design of a dandelion seed, the cradling contour of a kangaroo pouch, the net-like nature of a spider web, and the lock-and-key fit of a digestive enzyme to the molecular structure of its substrate. There is a sense in which each of these forms – and indeed all adaptations – re-present something about the physical and biological world around them. Functionality harbors a kind of "truth" about the universe, a valid representation of the way the world really is. Nevertheless, the "truth" of it lies purely in its pragmatic usefulness. As I will show below, the same will be so of many of our mental representations: Although some may mirror the real world, more often they serve as handy, functional ways of interpreting reality, and thus may be suspect from a scientific point of view.

Ontogenetic and epigenetic representation. A third way that evolution can "mesh" organisms to their environments is through flexible *developmental* pathways. DNA is often described as a blueprint for building an organism, but that metaphor is misleading. Because development from fertilized egg to adulthood occurs through a sequential expression of genes, with different genes being "turned on and off" at different stages of development, DNA is not so much a static, iconic blueprint as a step-by-step "recipe" for building bodies, brains, and behaviors (Dawkins, 1986; Ridley, 2003). However, ontogeny (development) is never a linear, lockstep unfolding of a genetic code in an environmental vacuum; gene activation and deactivation depend crucially on context, and – *by evolutionary design* – genes and environment mutually condition one another (Fischer & Bidell, 1991; Futuyma, 1998; Gallistel, Brown, Carey, S. Gelman, & Keil, 1991; Ridley, 2003; Tooby & Cosmides, 1992). Evolution has often provided a developing organism with mechanisms for adjusting the ontogenetic sequence in response to environmental input – an interaction known as "epigenesis" (Bjorklund & Pellegrini , 2002; Fischer & Bidell, 1991; Gallistel et al., 1991; Geary 2007; Geary & Bjorklund, 2000). Epigenesis is *calibrational*: the "ontogenetic adaptation of phenotypes to the local ecology" (Geary & Bjorklund, 2000, p. 57). Thus vine tendrils and oyster shells conform to their substrates; a bear's fur coat thickens or thins according to local latitude, altitude, or season; and people who do honest work for a

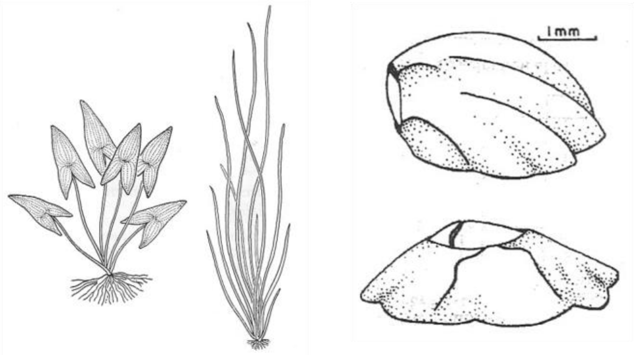


Figure 4. Dimorphism and phenotypic plasticity in arrowhead arum (*Sagittaria sagittifolia*) and acorn barnacles (*Chthamalus anispoma*). Sketches: Futuyma (1998) and Lively (1986).

living grow calluses on their hands. The formation of a new callus requires the activation and expression of genes, yet it is plainly a response to experience, *conforming* the body to a salient feature of the physical environment. Moreover, natural selection has sometimes supplied a species with more than one genetic recipe, hence ontogenetic sequence, that its development can follow – a phenomenon known as phenotypic plasticity. When the seed of an arum plant germinates in shallow water, it grows into the familiar "lilypad" form with arrowhead-shaped leaves and spongy, buoyant stems (Figure 4); but if the seed sprouts on dry land, it grows instead into a tall, wispy, grass-like morph (Futuyma, 1998). Larval acorn barnacles can develop either into a volcano-shaped "conical" morph or, if there are chemical traces of predatory snails in the local habitat, a lopsided "bent" morph that is better fortified against attacks, albeit at some cost in filter-feeding efficiency (Lively, 1986). Every arum plant and barnacle is born with DNA sequences for building *both* morphs, but an environmental cue throws the switch that steers it onto one ontogenetic trajectory rather than the other.

There are two important connections to human learning and the construction of mental representations. First, the existence of epigenesis and phenotypic plasticity precludes any rigid conceptual separation of genetic influences from environmental influences; evolution has utterly twined them together (Ridley, 2003). Later I will show that it is equally naïve to rigidly partition "learned" behavior from "instincts." Natural selection supplied us instead with "instincts to learn" (Marler, 1991), such that much of what we "learn" is shaped by our evolved genetic recipes just as surely as a barnacle morph, vine tendril, or callus (Gallistel et al., 1991; Tooby & Cosmides, 1992). Such learning is *calibrational*, adjusting and aligning the genetically organized neural/cognitive machinery of the mind to particulars of the local environment.

Second, changes in how a child mentally represents reality and construes phenomena need not reflect a simple imprinting of experience upon an infinitely malleable mind (Pinker, 2002). Instead, those changes could reflect normal ontogenetic development, a "maturational" emergence not unlike the bodily changes that happen during puberty. Or they may reflect an epigenetic emergence, triggered by cues in the environment, yet genetically organized all the same (Gallistel et al., 1991). In short, even though they change after birth, our mental representations may remain in a strong sense *innate*, the work of evolution by natural selection.

Behavioral representations. Finally, a fourth way exists for natural selection to "re-present" reality: by evolving a nervous system capable of conforming an organism's body to its physical surroundings. This is most vivid in certain species' ability to rapidly calibrate their coloration, shape, and texture to novel backgrounds (Figure 5). Whereas the ghost crab's camouflage develops over evolutionary time, the flounder, cuttlefish, and octopus carry it out in "real time." Like the cryptic coloration of ghost crabs, the chameleon-ship of flatfish and cephalopods *mirrors* something in the world. Here again, though, real-time calibration more often accomplishes a *functional fit* to the world.

Figure 5. Calibrational camouflage: A winter flounder (*Pseudopleuronectes americanus*) on a chessboard; a cuttlefish (*Sepia officinalis*) raises its tentacles and alters the texture of its skin to mimic a sprig of artificial algae; before and after photos of an octopus (*Octopus vulgaris*) changing its coloration, shape, and texture to match an algae-covered coral. Images: Field Museum of Natural History, Justine Allen, Roger Hanlon.

Striking examples include prehensile elephant trunks, opossum tails, and the nimble, manipulative hands of raccoons, spider monkeys, and humans.

The octopus's marvelous "meshing" ability does not constitute a single adaptation but a whole complex of coordinated adaptations. This includes a nervous system with dedicated neural mechanisms for processing visual input, for integrating that input to assess the landscape in three dimensions, and for reproducing select aspects of that landscape through several forms of motor control, such as pigment-changing chromatophores and contortions of skin and tentacles. Here we might suspect some kinships to human cognition. Do cephalopod brains house *mental representations* of their surroundings? Do they construct *mental models* of the world outside? Since they select only certain features to mimic, shall we grant them a capacity for *construing* their surroundings in particular ways, a sort of *interpretation* of the natural world?

Here too, I draw an analogy between cephalopod shape-shifting and human "calibrational" learning: Some of our mental machinery is designed to *conform to* certain variables in the natural environment, to *construct* mental representations of them, and to *construe* them in specific ways. Nevertheless, there may be limitations on the malleability of these mental representations – an important issue that I discuss next.

The Limits of Re-presentation: Calibration versus Conceptual Change

Although the flounder, cuttlefish, and octopus's calibrational re-presentations are impressive, limitations are also evident. The black and white "squares" on the flounder's flank, for example, are too irregular to support a proper game of chess. It is as if natural selection "expected" only a finite range of environmental variation and equipped the flatfish accordingly; nothing in its evolution prepared it for the straight lines and 90°

corners of a chessboard. Likewise, the cuttlefish's mimicry is far from convincing, for the artificial algae falls partly outside the range of forms that evolution "predicted" the animal would encounter, based on statistically recurrent features of past environments (see Cosmides & Tooby, 1994; Dawkins, 1976/1989; Tooby & Cosmides, 1992). It seems doubtful that even an octopus could realistically blend itself into a right-angled reef made of multicolored Lego blocks or a seafloor strewn with bowling balls and kitchen utensils (would it even try?). The problem in part is that Lego blocks, chessboards, and artificial algae are evolutionarily novel; they are very recent products of cultural, not biological, evolution.

In human cognition, we should expect similar constraints on the range of mental representations that the mind will readily construct. Concepts born of recent cultural evolution, such as those produced by natural science, may fall far afield of what natural selection "predicted" a person might have to learn (Geary, 2002, 2007). Mastering those evolutionarily novel concepts will not entail "calibrational" learning, but bona fide *conceptual change*. Like a flounder on a chessboard, a student may find it very difficult to form an adequately accurate mental representation of certain scientific concepts.

The cognitive anthropologist Dan Sperber (1994) makes a useful distinction in this regard. He distinguishes between the *proper domain*, the circumstances for which an adaptation originally evolved, and the more expansive *actual domain*, the circumstances which may summon it into action even though those are not the circumstances for which it evolved. Mottled sand flats belong to the proper domain of a flounder's capacity for camouflage; that is the context or "content" to which the fish's brain and body are designed to calibrate themselves. Chessboards belong to its actual domain; the fish attempts to calibrate, but its re-presentation is a mismatch. Elsewhere Sperber (1996) draws a kindred distinction: *dispositions* versus *susceptibilities*. Flounder are naturally “disposed” to mimic speckled seafloors, and they are (unnaturally) “susceptible” to mimicking chessboards as well.

Similarly, a scientific idea may fall outside the proper domain for which the interpretive mechanisms of the human mind evolved. But if the content happens to fall within their actual domain, then a science student may be “susceptible” to interpreting it in intuitive-yet-unscientific ways. In other words, she will *misinterpret* or *misrepresent* it. She will unwittingly attempt to calibrate when what is really wanted is conceptual change.

Summary, Thesis #1: The Evolution of Re-presentation

My purpose in this section was not to make mere analogies between animal adaptations and human cognition. Rather, I have shown that evolution is a process whose very nature is to shape living species and their body parts into “re-presentations” of certain features of the physical world. This should be just as true of the structure of the human brain and the mental representations that it generates. The representations that natural selection creates may be mirror-like “replicas” of the physical environment (ghost crab carapaces), but more often they constitute a “functional fit” to it (hummingbird bills). Natural selection can produce representations on several different timescales: the evolution of whole populations over multiple generations (ghost crabs, hummingbirds), the ontogenetic and epigenetic development of individual organisms during a single lifetime (barnacle morphs, vine tendrils, hand calluses), or “real time” behavioral transformations (octopus shape-shifting, prehensile elephant trunks). Epigenetic and behavioral re-presentation can be thought of as "calibrational" processes which flexibly attune bodies, brains, and behaviors to variable features of the natural world. However, they are not infinitely flexible: They adapt an organism only within a finite range of statistically recurrent environmental variability that was relevant to the reproductive fitness of its ancestors. When conditions or content fall outside that range, mismatches, misrepresentations, or misinterpretations are likely to occur.

Thesis #2 – The Meshing of Mind to World: Mental Re-presentations

There has been the evolution of a mesh between the principles of the mind and the regularities of the world such that our minds reflect many properties of the world....Our minds are always automatically applying a rich variety of frames to guide us through the world. Implicitly, these frames appear to us to be part of the world.

John Tooby & Leda Cosmides (1992, p. 72)

A central problem in Western philosophy – from Plato and Aristotle to Descartes and Locke to Kant, Hegel, and beyond – has been the relationship between reality-as-it-appears and reality-as-it-is, between things-in-the-mind and things-in-themselves (Bubner, 1981; Descombes, 1980; Plato, trans. 1974; Randall, 1926/1976; Smith 1923/1992; Solomon, 1983; see also Crotty, 1998; Pinker, 2002, 2007). Does the external world stamp its structure onto our minds, and if so, with how much fidelity? Do ideas accurately represent reality, or are they mere imperfect shadows of it? Or do ideas perhaps *precede* reality, such that in a sense we *construct* reality by projecting our ideas onto the world? Do our conceptual categories "carve nature at its joints," as Plato famously put it, or do they impose an artificial mental order upon it? As ideas change through human history, does "reality" change with it? Must we conclude, with Buddhist traditions and 20th century postmodernism, that in the end mental constructions (or social constructions) are all that exist, and that "reality" utterly reduces to our ideas about it? In the end, is what is "real" wholly relative to subjective perspective and/or sociocultural setting?

By introducing *natural* history into the mix, and the mechanism of natural selection, Darwin (1859) contributed mightily toward a solution to these persistent riddles (see Dennett, 1995). The reason that the human mind can and *must* in some sense be in touch with things-in-themselves is that the things-in-themselves *made a difference* as to which kinds of minds would survive and which ones would not:

> Natural selection operates through the testing of alternative designs through repeated encounters with evolutionarily recurrent situations....In our evolutionary history, design changes that enhanced their own propagation relative to alternative designs were selected for – that is, they caused their own successive spread until they became universal, species-typical features of our evolved architecture. (Cosmides & Tooby, 1994, pp. 86-87)

This "testing of alternative designs" holds not only for body parts, but brains and behaviors as well. It established Tooby and Cosmides' (1992) inevitable "mesh…between the principles of the mind and the regularities of the world" (p. 72). The universe came to *represent* (some sliver of) itself in the human brain. The mind reflects reality-as-it-is with some fidelity, because if it did not, reality would have killed it off long ago. Dire penalties await a mind that meshes poorly with a mean world. In this sense at least, reality precedes mind, and not the other way around.

This is a crucial point with which any theory about the nature of conceptual change must contend. It implies that evolution endowed us with fitness-enhancing "intuitions" about reality-out-there, and that it did so by weeding out inferior intuitions. Natural selection would swiftly punish any variants of brain/mind that generated less potent intuitions, or that disregarded the more potent intuitions inherited from our successfully surviving and reproducing forebears. *Yet that is what the phenomenon of conceptual change implies.* How did evolution fashion a mind able to erase, undo, or undermine the very intuitions that had served our ancestors so well? As I will show later under Theses #5-7, this question played a major role in developing the theoretical framework that inspired and informed the design of my experimental instrument.

In the preceding section I set the stage for a discussion of *mental representations* by showing that "re-presentation" more generally is a natural outcome of evolution. In this section I will plumb the cognitive science literature to show that evolution has indeed predisposed us to project an assortment of pragmatically useful – though not always scientifically accurate – interpretations onto the outside world. *I develop the following thesis:*

> In organisms with complex nervous systems, the re-presentations of the natural world that evolution creates include *mental* representations. In the hominin line, biological evolution has bestowed us with innate cognitive mechanisms for mentally representing, interpreting, construing, conceptualizing, and reasoning about external reality. Many of these mechanisms are *domain-specific* and *content*- or *context-sensitive*: Each was "designed" by natural selection to make sense only of certain kinds of natural or social phenomena, and each is

> automatically mobilized in the mind by the appropriate context or content. That is, certain contexts tend to arouse or activate particular mechanisms of the evolved mind, and thus tend to evoke certain *intuitive* interpretations or construals.

It is in the face of such innate intuitions that contemporary science students must undertake conceptual change, apprehending and adopting the often counterintuitive representations of reality generated by modern science. Among them are a trio of intuitions that make it difficult for students to master the theory of evolution: the projection of *intentional agency*, *teleological directionality*, and *immutable essences* onto biological phenomena (see Table 1). The experimental instrument piloted in this doctoral study was deliberately designed to evoke these three intuitions.

Four Plausible Domains of "Intuitive Science"

I hasten to repeat that although these intuitions are innate – belonging to a biologically evolved architecture of brain and mind – "innate" need not mean fully formed at birth. Like breasts and wisdom teeth, natural selection may have designed them to emerge ontogenetically via sequential gene expression during the course of normal post-natal human development. It is also possible that they must be epigenetically "calibrated" in accord with real world experience, through domain-specific, content-sensitive *learning*. In that case, much *learning itself* would be evolved and *instinctive*. I will argue under Thesis #3 that "calibrational learning" (as opposed to conceptual change) may be evolution's main mechanism for shaping individuals' intuitive theories about the natural world.

Under the current thesis, however, I will simply describe those intuitive theories. I will review the cognitive science literature on four plausible domains of intuition that may strongly inform students' pre-instructional understanding of natural phenomena: *intuitive physics, intuitive biology*, *intuitive psychology*, and *intuitive engineering*. Each appears to be driven by a powerful "mode of construal" (Keil, 1994) or "stance" (Dennett, 1989, 1995) that all people naturally adopt: the *mechanical* stance, *essentialist* stance, *intentional* stance, and *teleological* stance respectively. The last three are especially important to my research, because they are well known to compromise many students' mastery of the scientific theory of evolution (e.g., Evans, 2008; Shtulman, 2006; Sinatra et al., 2008; Smith, 2010). The instrument that I field-tested in this doctoral study was deliberately designed to elicit essentialist, intentional, and teleological misconceptions where they exist (see Table 1). These stances are among the pragmatically useful, interpretive "frames" that Tooby and Cosmides (1992) posit in the lead quote above: "Our minds are always automatically applying a rich variety of frames to guide us through the world. Implicitly, these frames appear to us to be part of the world" (p. 72).

I think it will be helpful to begin with a specific example.

An example of innate (but unscientific) intuition: Leslie's "launching event" studies. In a series of ingenious experiments, Leslie (1994) uncovered strong evidence that 6-month-old infants already construe contact causality in a manner similar to adults. He and his colleagues showed adults footage of "launching events," billiard-ball-like collisions in which a moving object slides into a stationary object, launching it into motion. In some cases, the second object took off immediately upon contact, while in other cases there was a half-second delay between collision and takeoff. Whenever the second object's reaction was immediate, adults perceived the event as *causal:* The first object is deemed to *cause* the second to fly away. Moreover, they intuitively projected distinct "mechanical roles" onto the two objects: One delivers, the other receives. The first does something *to* the second, while the second has something done to it. The incoming object seems to provide something – a push or force – that launches the second, and the human mind automatically construes this as the former *acting upon* the latter.

This way of causally construing a collision seems so obvious and commonsensical that it is easy to forget that it is not given by perceptual experience itself. The only thing that falls onto the observer's retina is the spatial convergence of the two objects, followed by their divergence. As the Scottish philosopher David Hume famously pointed out, we do not *perceive* causation in the world; rather, it is an interpretation, a construal, that our minds project onto phenomena (Smith 1923/1992). What seems to be a part of the world is in fact a construction of the mind. It is, however, a *pragmatically useful* construction, one that helps us successfully navigate the world of moving objects.

When the film was run backwards, adults perceived the objects' "roles" as reversed. Moreover, when they witnessed the same event with a half-second delay, the sense of causation and distinct mechanical roles vanished altogether. Leslie used these facts to investigate whether pre-linguistic infants construe collisions in a similar manner. He habituated some infants to an instantaneous collision and others to the delayed condition, showing the same footage over and over until each infant lost interest in the event and looked away. At that point he reversed the footage, thereby introducing a novel stimulus to recapture the infants' attention. In the *non-causal* condition (delayed launch) the only novelty was direction of motion, whereas in the *causal* condition (instantaneous), not only had the motion changed but the mechanical roles had reversed – a qualitative change, hence a higher degree of novelty. The infants exhibited a stronger renewal of interest (as measured by duration of recaptured gaze) in the causal scenario than the non-causal. Leslie interpreted this as evidence that a 6-month-old already *mentally represents* the events in a manner akin to adults, distinguishing between causal and non-causal events *and* projecting distinct mechanical roles onto the colliding objects.

This example is relevant to the issue of conceptual change in two respects. First, because the conception of distinct mechanical roles is available at such and early and pre-linguistic age, it suggests that certain commonsense ways of construing causation and interpreting natural phenomena may be innate, a handy psychological mechanism provided by natural selection. Says Leslie, "As a result of adaptive evolution, the infant is a specialized processor of information with an architecture that (in part) reflects properties of the world" (p. 119). Second, it shows that sometimes such construals, though useful, are contradicted by the models of natural science, which by comparison are conceptually "counterintuitive." Scientifically speaking, the intuitive projection of asymmetrical mechanical roles onto a launching event is erroneous, for Newton's 3rd law of motion dictates that in every collision, the two objects strike *each other* with equal and opposite force. When the incoming object strikes the resting object, the resting object *hits back* (and this stops, slows, or deflects the first object). Yet the Newtonian construal – *a product of cultural rather than biological evolution* – is neither natural nor intuitive. Adopting the normative scientific conception requires a physics student to undergo conceptual change, rejecting an intuitive and perhaps innate framing of natural phenomena for a counterintuitive one.

Intuitive physics and the “mechanical stance.” The manner in which adults construe “contact causality” during collisions might be fairly characterized as a component of the *intuitive physics* with which we all naturally negotiate the world of inanimate objects. Leslie’s (1994) launching experiments with infants suggest that a form of this intuitive physics may already be available soon after birth, hence innate. A host of other habituation experiments have revealed that within the first two months of life, children already have expectations about how the physical world should operate: Objects continue to exist even when out of sight, two objects cannot share the same space, moving objects should follow uninterrupted paths and retain their rigid shape during travel, collisions should alter trajectories, objects cannot influence each other’s trajectory at a distance, and so on (Baillargeon et al., 1985; Carey, 2009; Carey & Spelke, 1994; Spelke, 1991; Spelke, 1994; Spelke et al., 1995). Such principled expectations and constrained interpretations are almost certainly innate, for they are non-obvious and could not have derived from perceptual experience (Spelke, 1994), even if epigenetic interactions channel or calibrate their emergence (Fischer & Bidell, 1991; Baillargeon, Kovotsky, & Needham, 1995). In time, other elements of intuitive physics come on line. One is a “mechanical” stance (or “physical” stance; Dennett, 1995) which construes *causation* as a directional transmission of intrinsic forces from one body to another, strictly via direct contact (Leslie, 1994, 1995). Another seems to be a naïve belief that a moving object, say a pitcher’s fastball, is continually propelled by a sustained internal “impetus” rather than mere momentum (Pinker, 1997). Although this intuition functioned quite well in the day-to-day lives of our ancestors (e.g., during hominin baseball games), it is at odds with Newton's 1st Law of Motion and so must be corrected in the modern science classroom. Learning academic principles that are evolutionarily novel requires conceptual change in the face of countless generations of evolved cognitive inertia (Geary, 2007).

Intuitive biology and the "essentialist stance." Everyday life for our hominin and pre-hominin ancestors was like "a camping trip that lasts a lifetime" (Orians & Heerwagen, 1992, p. 556). As opportunistic foragers and hunters with a varied diet and nomadic niche, their interactions with other species were much more frequent, intimate, and salient to daily survival than in modern civilization. There was a strong selective pressure to evolve an "intuitive biology" to guide them in those interactions (Atran, 1995, 1998). Just as our intuitive physics imposes distinct "rules" for reading causation within that domain (impetus, contact, etc.), we seem to project domain-specific construals onto biological phenomena, too.

For example, for all people of all cultures, biological understanding is powerfully informed by an "essentialist" stance toward living organisms and species (Atran, 1995, 1998; S. Gelman, 2003). We readily project a hidden "essence" onto plants and animals – both individual organisms and whole species – intuitively regarding them as inhabited by an *intrinsic nature* which gives each organism its traits, prompts its species-specific behavior, directs its growth and development, and preserves its identity throughout the life cycle (Atran, 1995, 1998; S. Gelman, 2003; S. Gelman, Coley, & Gottfried, 1994; Keil, 1989, 1991, 1994; Wellman & S. Gelman, 1992). For our ancestors, essentialism may have been a cognitive adaptation that allowed individuals to respond to statistically recurrent features of the environment, conferring powers of prediction that permitted successful foraging, hunting, habitat selection, and avoidance of predators and toxic plants (Tooby & Cosmides, 1992). Still today, adults and children in all cultures from foraging peoples to modern civilizations continue to "essentialize" plants and animals at the species or genus level, and they use this intuitive projection to classify organisms into nested taxonomical hierarchies (Atran, 1995, 1998; Berlin, Breedlove, & Raven, 1973). This systematic clustering of organisms into "natural kinds" (Keil, 1989, 1991) according to their essences permits powerful inductive inferences; for instance, when two distant species are thought to share a particular trait, it gets generalized to all organisms that fall within that nested level (Atran, 1995, 1998; S. Gelman, 2003; Keil, 1989, 1991, 1994, 1995). To hominins on a lifelong camping trip, the benefits are plain: One need not risk repeated encounters or resort to trial-and-error learning to infer that *all* saber-toothed cats are likely to be hostile, *all* acacia pods are nutritious, *all* oleander flowers cause illness, *all* male gazelles can be tracked by their telltale territorial markings, and so on.

The essentializing tendency has crucial implications for students undertaking conceptual change with respect to the theory of evolution itself. Many well-crafted studies show that it reliably develops during early to late childhood and has fully emerged by the time students first tackle the theory of evolution in middle or high school (Evans, 2008; R. Gelman, 1990; S. Gelman, 2003; S. Gelman et al., 1994; S. Gelman & Kremer, 1991; S. Gelman & Wellman, 1991; Hatano & Inagaki, 1994; Hickling & S. Gelman, 1995; Inagaki & Hatano, 1993, 1996, 2002; Keil, 1989, 1994; Poling & Evans, 2002; Rosengren, S. Gelman, Kalish, & McCormick, 1991; Samarapungavan & Wiers, 1997; Wellman & S. Gelman, 1992). The essentialist stance may impede student mastery of

Darwinian theory by making it difficult for students to conceive that species can evolve at all, changing from one “natural kind” into another – i.e., deeming essences immutable (D. Anderson, Fisher, & Norman, 2002; Evans, 2008; Poling & Evans, 2004; Samarapungavan & Wiers, 1997). To sense how compelling this intuition can be, one need only consider another class of “natural kinds” that seem quite immutable and onto which we readily project essences: minerals, metals, and other natural substances (S. Gelman, 2003; Keil, 1989). It is no surprise that one obstacle which historically prevented evolutionary theory from gaining ground until the 19th century was the deep intuition that “reptiles could no more *turn into* birds than copper could turn into gold” (Dennett, 1995, p. 38; see also Mayr, 1991).

Even when the essentialist stance does not bar receptiveness to macroevolutionary change, it may inhibit a proper grasp of the microevolutionary mechanisms causing it. It may prompt the misconception that evolution occurs through interbreeding of different species, thereby blending their essences (Jensen & Finley, 1996). Or it may contribute to the common misconception that evolution is a simultaneous adaptation of *all* members of a natural kind, transforming the species and its essence *as a whole*, rather than a statistical shift in trait frequencies. In other words, the projection of a unifying essence may make it difficult to recognize that natural selection is fueled by *differences within* a species, and to grasp that evolution is a change in group *membership*, and not the members themselves nor their essences (D. Anderson et al., 2002; Bishop & C. Anderson, 1990; Mayr, 1991; Nehm & Schonfeld, 2008; Samarapungavan & Wiers, 1997; Shtulman, 2006; Shtulman & Schulz, 2008; Sinatra et al., 2008; Smith, 2010).

Table 1 documents more thoroughly the range of common misconceptions that may stem in part from students' tendency to essentialize living things. The taxonomy in Table 1 served as a key blueprint in the design of the experimental instrument piloted in this doctoral study.

Intuitive psychology and the "intentional stance." Another well-researched intuitive domain is folk psychology, the central cognitive component of which is the so-called "theory of mind" module (Baron-Cohen, 1995; Leslie, 1994; Premack & Premack, 1995; Wellman, 1990). Possessing a "theory of mind" means that one understands other minds *as* other minds. In its mature adult form, our intuitive psychology entails *mentally representing* the *mental representations held by others*, presumably in order to predict their actions. We have a strong cognitive tendency to project *intentional agency* onto other people. When we adopt the "intentional stance" (Dennett, 1989) toward an event, we construe it as originating in a fundamentally different way than the way we construe purely mechanical causation – namely from a willful, goal-directed purpose (Keil, 1994; Leslie, 1994, 1995; Wellman, 1990). Whereas the essentialist stance induces us to lump nonhuman animals into homogeneous "natural kinds," the intentional stance interprets each human as distinctly *individual*, driven by personal desires, goals, perceptions, and beliefs (Baron-Cohen, 1995; Leslie, 1994; Wellman, 1990). Like the essentialist stance but unlike the mechanical stance, causation here derives from *internal* sources – striving, aiming, willing – rather than external contact (Carey, 2009; Leslie, 1994). While the rules of mechanical causation forbid "action at a distance" – whether the divide is spatial or temporal – it constitutes the very marrow of intentional agency (Leslie, 1994). Across space, intentional agents perceive and respond to events without direct contact (ducking a high fastball, for instance). Across time, intentional behavior seeks to close a discrepancy between a present state of affairs and a desired future state, which involves making predictions and taking deliberate actions (Leslie, 1994). In short, the intentional stance that we adopt toward other people's behavior construes *causation* very differently than do the mechanical and essentialist stances within their own domains.

Robust evidence from studies of animals and autistic children (whose theory of mind is compromised) suggests that intuitive psychology may involve a suite of innate mechanisms built into our brains (Baron-Cohen, 1995; Leslie, 1994). Ingenious experiments with normally developing children reveal early signs of an emerging theory of mind during the first few months of life and its gradual flowering thereafter (Baron-Cohen, 1995; Bjorklund & Pellegrini, 2002; Carey, 2009; Gopnik & Meltzoff, 1997; Gopnik & Wellman, 1994; Leslie, 1994; Premack & Premack, 1995; Tomasello, 1999; Wellman, 1990). Toddlers are "mentalists" not "behaviorists," as seen in a pre-linguistic infant's imitative and pseudo-conversational interactions with adults, in "social referencing" (e.g., pointing), and in "joint attention" with others; they grasp that others are psychologically engaged with the same objects of the world as they are (Baron-Cohen, 1995; Gopnik & Wellman, 1994; Tomasello, 1999). By age 3 they are employing the language of mental states and beliefs, and by age 5 they can express their theory of mind explicitly (Baron-Cohen, 1995; Bjorklund & Pellegrini, 2002; Carey, 2009; Wellman, 1990).

Intuitive psychology constitutes a powerful, pervasive cognitive system which saturates our mental representations of reality. However, like the essentialist stance, it may hamper student conceptual change with respect to the theory of evolution (Table 1).

We readily project agency not only onto other people, but sometimes nonhuman phenomena as well: especially big-brained mammals like ourselves, but also “lower” animals, plants, machines, meteorological phenomena, and even craftily animated motions of dots on a computer screen (Baron-Cohen, 1995; Boyer, 2001; R. Gelman, Durgin, & Kaufman, 1995; Leslie, 1994, 1995; Premack, 1990; Premack & Premack, 1994). Like flounders on a checkerboard, we are “susceptible” (Sperber, 1996) to extending the intentional stance beyond its “proper domain” (Sperber, 1994), especially when faced with novel, culturally evolved concepts such as those of modern science. And this can lead to flounder-esque mismatches and mis-re-presentations. For example, this tendency may contribute to the common quasi-Lamarckian misconception that evolutionary adaptation is driven by animal behavior, effort, learning, and/or the repetitive use/disuse of body parts, with parents passing these within-lifetime “adaptations” and “acquired traits” to their offspring through biological inheritance (D. Anderson et al., 2002; Bishop & C. Anderson, 1990; Evans, 2008; Jensen & Finley, 1995, 1996; Mayr, 1991; Nehm & Schonfeld, 2008; Poling & Evans, 1994; Samarapungavan & Wiers, 1997; Settlage, 1994; Sinatra et al., 2008; Smith, 2010).

Intuitive engineering and the “teleological stance.” A fourth intuitive domain relevant to evolution education is that of tools and human-made artifacts. Far beyond any other species, we humans quickly, intuitively comprehend how to craft and employ all sorts of useful devices. Based on several million years of archaeological evidence, Mithen (1996) argues that hominin tool design reflects the evolution of an innate, specialized “technical intelligence,” able to work stone and fashion implements more sophisticated than anything general intelligence or trial-and-error learning could produce.

Also critical was the evolution of an exceptional cognitive faculty: the faithful imitation of others' goal-directed behaviors (Tomasello et al., 1993). This ability, which perhaps first evolved to support procedural learning such as tool use, is a species-specific psychological adaptation mostly beyond our closest relatives: "We say 'monkey see, monkey do,' and use ape as a verb, but in fact monkeys and even apes do not seem to be especially clever imitators compared to humans" (Richerson & Boyd, 2005, p. 109; see Tomasello et al., 1993). And it appears astonishingly early in life: Within an hour of birth newborns will mimic adult facial expressions (Gopnik & Meltzoff, 1997). Long before their first birthday, infants are captivated by contingencies between their own movements and the behavior of objects, and they are highly motivated to manipulate, explore, and eventually employ articles by hand (Gopnik & Meltzoff, 1997). Perhaps in this we are witnessing an ontogenetic and epigenetic emergence that manifests our deep history as toolmakers and tool users (Bjorklund & Pellegrini, 2002; Pinker, 1997).

The central intuitive construal in this domain is *functionality*. Hand a middle school biology student an odd, unfamiliar kitchen gadget, ask her what it is, and she will invariably speculate not about its material composition, its mechanical configuration, or its origins, but about its utility, its function, its purpose. Her automatic inference will be that it is *for* something. This reflects the *teleological* interpretation – or what Dennett (1995) dubs the "design stance" – that we readily adopt toward tools and artifacts, as well as the bodily and behavioral adaptations of plants and animals (Atran, 1995; Keil, 1989, 1994, 1995). In effect, we deem an artifact's function or "end" (Greek *telos*) as an adequate causal explanation of its actions and indeed its very existence; its purpose *precedes* its creation and its actions are a *consequence* of its function (Dennett, 1995).

As with the essentialist and intentional stances, clever experiments suggest that the seeds of the design stance emerge at a very young age, even infancy, and that it is domain-specific and content-sensitive: Preschoolers classify living things and natural substances – but not artifacts – according to perceived essences; as they grow older they increasingly identify *function* as the proper criterion for categorizing artifacts (Keil, 1989, 1995).

Here again, though – like flounders on a checkerboard – we readily extend the design stance beyond its "proper domain" (Sperber, 1994), and this may impede student conceptual change with respect to the theory of evolution (Table 1). While biology teachers can usefully recruit a student's teleological intuitions to make sense of bona fide adaptations such as body parts and instinctive behaviors, it is also a cognitive bias that can compromise correct Darwinian interpretations at other levels. For example, although we would never project the design stance onto liquid water by asking, "What's the purpose of water's transparency? What's the use of being clear?" (Keil, 1994) – these *are* questions that we can sensibly ask of a see-through jellyfish. However, students may take this too far by asking, "What's the purpose of the jellyfish itself?" – intuitively attempting to understand the animal's function within the ecosystem, or its *raison d'etre* relative to human needs. In general, the teleological stance may make students "susceptible" (Sperber, 1996) to viewing whole organisms, species, and even ecosystems as functional or purposeful (e.g., for humans' benefit, as part of the ecological or cosmological order, etc.). And when coupled with the intentional stance, it may make it easier for students to regard living organisms – with all their wonderful, functional adaptations – as the creative product of a purposeful, divine designer; the design stance makes it deeply counterintuitive to conceive such evident functionality as a product of mechanistic, mindless, foresight-free natural selection (Dennett, 1995; Mayr, 1991; Pinker, 2002; Poling & Evans, 2004).

Teleological intuitions also clearly foster the widespread misunderstanding that evolution is driven by individual or species “needs,” and that they evolve because they “have to,” as a functional response to environmental duress; many students believe that the origin of a trait is adequately explained simply by noting its adaptive functionality or necessity for survival (D. Anderson et al., 2002; Bishop & C. Anderson, 1990; Brumby, 1984; Evans, 2008; Jensen & Finley, 1995, 1996; Nehm & Schonfeld, 2008; Poling & Evans, 2004; Samarapungavan & Wiers, 1997). Together, teleological and essentialist projections may lead students to confound the long-term evolutionary adaptation of whole populations with an individual organism’s everyday “adapting” to changing conditions (Bishop & C. Anderson, 1990; Brumby, 1984; Evans, 2008; Jensen & Finley, 1995; Nehm & Schonfeld, 2008). The same combination may trip students into the common fallacy of “good-of-the-species” group selection (see Dawkins, 1976/1989; Williams, 1966) or the belief that throughout natural history species “had to” evolve to avoid extinction (Jensen & Finley, 1995; Nehm & Schonfeld, 2008; Poling & Evans, 2004; Samarapungavan & Wiers, 1997; Shtulman, 2006). Finally, the combination of all three intuitive stances – teleological, essentialist, and intentional – may contribute to the common misunderstanding of evolution as linear rather than branching, and as progressive, goal-directed, and/or quasi-developmental, driven intrinsically toward higher complexity and sophistication, perhaps by an internal striving (Evans, 2008; Jensen & Finley, 1996; Mayr, 1991; Nehm & Schonfeld, 2008; Shtulman, 2006).

Summary, Thesis #2: Intuitive Domain-Specific Re-presentations and Construals

In this section I reviewed literature from cognitive science and developmental psychology that bears out the Darwinian prediction that our brain and minds, no less than our bodies, were molded by natural selection to "mesh" with certain fitness-relevant, niche-specific features of the social and ecological environments of our hominin and pre-hominin ancestors. We are innately predisposed to infer certain meanings, mentally represent reality and causal relationships in certain ways, and conceptually carve up the world along certain boundaries. We *construe* particular phenomena in particular fashions so naturally, so unconsciously, that we do not readily realize that each is a *functional, interpretive* projection onto the phenomena, rather than a clean window onto them. Trouble arises – educationally – whenever science discovers that things are not, after all, exactly what they seem.

I dwelled especially on three ubiquitous "modes of construal" (Keil, 1994) that informed the design of the instrument which I field-tested in this doctoral study: the essentialist, intentional, and teleological stances. As I will explain in Chapter 3, the taxonomy in Table 1 served as the conceptual framework around which instrument items were developed.

Each stance appears to be a universal cognitive adaptation that predictably emerges during infancy or childhood. Some aspects of each may be available at birth, some may mature ontogenetically thereafter, and some may require epigenetic calibration via experience – not unlike barnacle morphs and hand calluses. Like octopus shape-shifting and prehensile elephant trunks, these interpretive lenses can flexibly conform to environmental features in "real time," albeit within a finite range. Yet they are *intuitive* precisely because natural selection "designed" them to emerge as they do. As in Leslie's (1994) launching event experiments, they are readily activated by relevant cues, allowing the mind to use small fragments of *perceptual* information to make elaborate *conceptual* inferences and interpretations (Tooby & Cosmides, 1992). "These modes of construal could be viewed as opportunistic, exploratory entities that are constantly trying to find resonances with aspects of real world structure" (Keil, 1994, p. 252). Together, they make comprehensible the behaviors of other people and animals that might otherwise be too complex to grasp using purely mechanical interpretations (Keil, 1994). But because they do not so much generate "true" replicas of reality (á la ghost crab carapaces) as provide a pragmatic and functional fit to select features of the social and ecological environment (á la hummingbird bills), they can both help and hinder learning in the modern biology classroom. On the one hand – like flounders on a checkerboard – students may put their intuitions to good use by mobilizing ancient cognitive adaptations in evolutionarily novel circumstances. For example, they may fruitfully invoke the teleological stance to understand how molecular enzymes function, or the essentialist stance to grasp DNA as a hidden, intrinsic, species-specific genetic code (Atran, 1995). On the other hand – again like flounders on a checkerboard – these same intuitions can lead to mismatches between ancestral ideas born of *biological* evolution and contemporary ideas born of *cultural* evolution, as when the essentialist, intentional, and teleological stances impede students' learning of the scientific theory of evolution. Hence the need for conceptual change.

Thesis #3 – Instincts to Learn: Calibrational Learning versus Conceptual Change

The common belief that "learning" is an alternative hypothesis to an evolutionary theory of adaptive function is a category error. Learning is a cognitive process. An adaptive function is not a cognitive process; it is a problem that is solved by a cognitive process. Learning is accomplished through psychological mechanisms…created through the evolutionary process….The issue is not whether a behavior is the result of natural selection "or" learning. The issue is, What kind of learning mechanisms would natural selection have produced?

Students often ask whether a behavior was caused by "instinct" or "learning." A better question would be "which instincts caused the learning?"

Leda Cosmides & John Tooby (1987, p. 292, and 1997, p. 11)

In the preceding section, I portrayed innate intuitions such as the essentialist, teleological, and intentional stances as interpretive projections that the mind's neural machinery automatically "cranks out" whenever mobilized by the correct cue or context. That was too simple. They are also "learning devices" (Carey, 2009). Part of their function is to facilitate rapid learning, serving as "footholds into acquiring much more elaborated belief systems in an extraordinary number of specialized domains" (Keil, 1994, p. 251). For example, the essentialist stance enables children to construct a mental catalog of local plants and animals grouped into nested hierarchies of "natural kinds" (Atran, 1998; Berlin, Breedlove, & Raven, 1973; Keil, 1989). Though the child's mind may come prewired to project essences, it must still be *calibrated* to the local wildlife. The child *learns* what evolution intended her to learn, in the face of experiences that evolution anticipated she would have. As described earlier, one way that natural selection “meshes” organisms with their surroundings – and in so doing “re-presents reality” – is by fashioning a nervous system able to conform to select features of the environment through “real time” behavior. Octopus shape-shifting and prehensile elephant trunks are vivid examples of such rapid calibration. Intuitive, domain-specific learning during childhood may be another.

Such learning, however, seems natural and intuitive compared to scientific conceptual change, which is often effortful and non-intuitive. In this section I will set the stage for more formal definitions of *conceptual change* (under Thesis #4) by contrasting it with what I call *calibrational learning*. *I develop the following thesis:*

> In many animal species, learning itself is motivated and shaped by an innate cognitive architecture, such that each individual’s behavior becomes calibrated to select features of the local environment in species-typical, fitness-promoting ways. This is by evolutionary design. In our own species, early childhood conceptual development in the domains of physics, biology, psychology, and engineering is largely calibrational: Through local experience, the child constructs an elaborate “intuitive theory” by fleshing out an innate “skeletal” framework already built into the mind.

As I will describe in Chapter 3, the instrument that was field-tested for this doctoral study was designed to assess the outcomes of both calibrational learning and conceptual change.

Instincts for Learning

The conventional separation of “instinctive” from “learned” behavior is a naïve and false dichotomy. Innate does not mean immutable. Animals are innately endowed with “instincts to learn” that foster, focus, and guide learning (Gallistel et al., 1991; Marler, 1991). For example, soon after fledging, young indigo buntings carry out a long migration by navigating by the stars. Since the flight path is new to them, something inborn must guide them. But it cannot be an innate map of the heavens, for such a map would quickly become outdated. Five millennia from now, an evolutionary eye blink, the “north star” Polaris will no longer be north, because Earth’s axis “precesses” over the centuries, wobbling like a top against the backdrop of the celestial sky. This perpetually alters which stars are most northerly, at a rate too fast for evolution to keep pace. How, then, do newborn buntings know which way to migrate? A clever experiment revealed the answer: When bunting hatchlings were raised in a planetarium, where researchers could manipulate the nightly rising and setting of the artificial stars, the hatchlings detected which constellation was making a tight rotation over the north pole (Emlen, 1975, as described in Carey, 2009, and Gallistel et al., 1991). If they *learned* that Polaris and the little dipper are north, then they *instinctively* migrated in that direction upon fledging. But if the researchers made some other constellation “north,” then the buntings’ navigational instincts calibrated to that constellation instead, and migrated accordingly. In effect, natural selection dealt with a shifty environment by crafting a cognitive mechanism that is flexible and shaped by experience, yet subservient to species-specific behavior. This is what I call *calibrational learning*.

Domain-general vs. domain-specific, species-typical learning. Skinnerian behaviorism searched for *general* laws of learning – such as classical conditioning – that *all* advanced animal species employ, regardless of content or context (Gallistel et al., 1991). Such learning is “domain-general.” In the wake of behaviorism, however, students of animal behavior have repeatedly found that learning rarely amounts to such equipotential association-making. Instead, an animal's attention and information-processing are constrained by innate cognitive architecture, such that, *for a given species*, it is easier to make some associations than others (Gallistel et al., 1991). Such learning is both species- and domain-specific. In effect, animals possess an evolved *preparedness* for learning that privileges certain content and biases them in the direction of particular "conclusions," in accord with the species' ecological niche.

There are, for instance, the famous experiments of Garcia and Koelling (1966, as described in Gallistel et al., 1991) which showed that rats, when punished with an immediate electrical shock, can readily learn to avoid *unflavored* water whenever a light flashes: an instance of domain-general learning. Yet when punished instead by a nausea-inducing poison whose effects are delayed by hours, they can only learn to avoid *flavored* water. Since plants in the ancestral wild would have advertised their toxicity with distinctive flavors, this *content-specific* learning looks like the work of an evolved calibrational mechanism built into the rat's mind.

Species-typical, domain-specific "learning instincts" extend even to innovation and creativity. For example, experiments show that although sparrow songs are plainly learned – that is, are *socially assimilated* – there is also an element of individual invention; nevertheless, both assimilation *and* invention are channeled by *innate* preferences (Marler, 1991). Song learning is a plastic process of imitating and calibrating to the local "dialect." It is flexible rather than fixed, and open to the introduction of novelty through the partitioning and recombination of existing songs. Yet the "rules" for doing so, as well as the degree of fidelity, differ from one species of sparrow to the next (Marler, 1991). Creativity itself is innately guided and constrained.

A human learning instinct: spoken vs. written language. The ethological literature abounds in other instances of “instincts for learning” (Marler, 1991), and we humans are no exception. Learning is not an alternative to biological explanations of behavior, but an *instance* of them (Cosmides & Tooby, 1987; see lead quote above). Evolution “expected” us to have certain environmental experiences and *prepared* us to learn from them (Gallistel et al., 1991; Geary, 2007). Language acquisition is a prime example. If a song sparrow’s species-typical instincts can channel social assimilation and even individual inventiveness, the same can be so for human language. Chomsky and his disciples produced a wealth of compelling evidence that all languages – beneath their peculiar vocabulary and syntax – are undergirded by a “universal grammar,” an abstract set of linguistic structures that reside innately in every human mind (Hirschfield & S. Gelman, 1994b; Pinker, 1994). As a result, any child can master the complexities of any language, and needs no instruction whatsoever to “learn” it. Moreover, Pinker (1994, 2007) argues that language acquisition involves mapping semantic relationships onto *preexisting* concepts and construals, such as those generated by a child’s intuitive physics, biology, psychology, and engineering (see earlier discussion). The skeletal conceptual and linguistic architecture is already there. A child’s “language instincts” must merely be calibrated to the local tongue, and this happens quite automatically and effortlessly so long as the child is within earshot of elders (Pinker, 1994; see also R. Gelman, 1991).

Written language, however, is another matter altogether. It is not calibrational. Reading and writing are very recent products of cultural evolution, and there has not been enough time for humans to evolve a special capacity for written language through biological evolution (Geary, 2007). Consequently, a child is unlikely to “discover” how to read and write on her own. These skills must be acquired through conscious, effortful, protracted practice and struggle under the guidance of an experienced adult (Geary, 2007). This, of course, is what schools are for.

Conceptual Calibration: The Development of Intuitive “Theories”

This distinction between spoken and written language acquisition has a parallel in human *conceptual* development. As discussed under Thesis #2, we naturally project mechanical contact causality, internal essences, intentional agency, and teleological function onto certain raw experiences. This in turn may fuel the formation of domain-specific “intuitive theories”: elaborate mental models that enable us to carve the world into categories, explain phenomena, and make useful cause-and-effect inferences and predictions. They govern the intuitive physics, biology, psychology, and engineering with which we navigate our ecological and social environments. Like language acquisition, developing these intuitive theories is arguably a case of calibrational learning: effortless, intuitive, and automatic (see Bjorklund & Pellegrini, 2002; Carey, 2009; Geary, 2007; Geary & Bjorklund, 2000; Keil, 1994). Bona fide conceptual change, on the other hand, is comparatively challenging, is not intuitive, and demands conscious effort. Whereas natural selection *prepared* the mind for calibrational learning, what seems to distinguish conceptual change – at least when it comes to science education – is how *unprepared* the mind is for it. Because evolution did not “expect” us to encounter the concepts, principles, and explanatory models of modern science, we are ill disposed to properly grasp them.

Scholars often characterize conceptual change as the replacement of a student's pre-instructional "theory" with a normative scientific theory, and the instrument piloted in this doctoral study was devised to detect elements of both. Below I review literature on the acquisition of intuitive theories as a special form of "instinctive learning."

"Skeletal principles." Cognitive scientists of a Darwinian bent have long debated the nature of our intuitive conceptual systems – sometimes hotly, but based upon compelling experiments and evidence from infants, children, and adults from diverse cultures (see Carey & R. Gelman, 1991; Hirschfield & S. Gelman, 1994a; Sperber et al., 1995; Vosniadou, 2008). Some scholars stress knowledge that is literally inborn – i.e., available at birth (e.g., Baron-Cohen, 1995; Spelke, 1991, 1994). Others emphasize the sequential emergence of mental models via ontogenetic maturation and/or epigenetic calibration (e.g., Baillargeon et al., 1995; Bjorklund & Pellegrini, 2002; Carey, 2009; Carey & Spelke, 1994; Gallistel et al., 1991; Geary, 2007; Geary & Bjorklund, 2000; R. Gelman, 1990, 1991; Keil, 1989; Leslie, 1994, 1995; Premack & Premack, 1994, 1995; Spelke et al., 1995; Wellman, 1990). Still others emphasize the malleability of these mental models, arguing that although we are born with innate ways of interpreting phenomena, evolution designed them to be flexibly modified in accord with experience (e.g., Gopnik & Meltzoff, 1997; Gopnik & Wellman, 1994; Karmiloff-Smith, 1991, 1992).

What seems utterly unlikely, however, is that our conceptual systems could arise *solely* from empirical experience and statistical associations, unguided by innate inclinations and the historical hand of natural selection. Pinker (2002) puts the problem like this:

> Nothing comes out of nothing, and the complexity of the brain has to come from somewhere. It cannot come from the environment alone, because the whole point of having a brain is to accomplish certain goals, and the environment has no idea what those goals are. A given environment can accommodate organisms that build dams, migrate by the stars, trill and twitter to impress the females, scent-mark trees, write sonnets, and so on. To one species, a snatch of human speech is a warning to flee; to another, it is an interesting new sound to incorporate into its own vocal repertoire; to a third, it is grist for grammatical analysis. Information in the world doesn't tell you what to do with it. (p. 75)

In a seminal paper, R. Gelman (1990) proffers the following solution to the issue Pinker raises:

> It is necessary to grant infants and/or young children domain-specific organizing structures that direct attention to the data that bear on the concepts and facts relevant to a particular cognitive domain. The thesis is that the mind brings domain-specific organizing principles to bear on the assimilation and structuring of facts and concepts, that learners can narrow the range of possible interpretations of the environment because they have implicit assumptions that guide their search for relevant data. (p. 79)

She dubs these implicit assumptions and organizing structures "skeletal principles": a bare skeleton upon which our conceptual universe gradually gets fleshed out via experience. When we enter this world as infants, we already have readymade frameworks built into the architecture of our brains/minds that focus our attention on particular patterns and organize the near infinity of sensory information. The answer to Pinker's question of how we can ever know "what to do with information in the world," she says, is that our minds constrain, or "narrow," the scope of inferences that we can draw from that information. Otherwise, we would never be able to construct meaningful interpretations in the first place.

Experiments show, for instance, that toddlers and even infants distinguish between animate and inanimate objects – that is, between those whose movement is self-initiated and those that move only when acted upon by external forces (R. Gelman & Massey, 1988; Inagaki & Hatano, 2002). Gelman (1990) contended that children are innately guided by an "innards principle" which, when triggered by certain cues – i.e., certain kinds of motion – causes them to attribute self-animated movement to something *internal,* a source or force *within* the object or organism. An automatic attentional bias narrows or constrains the interpretation. This constraint may be a "skeletal" precursor to the essentialist stance and the eventual development of such conceptual categories as "living" versus "nonliving" – core components of the child's emerging intuitive biology (see Carey 1985, 1995; S.Gelman, 2003; Inagaki & Hatano, 2002; Keil, 1989).

Similarly, a child's intuitive psychology and "theory of mind" are perhaps constructed around an innate skeletal framework that includes such well-documented, early-emerging biases as: a newborn's attraction to face-like patterns, an infant's ability to detect the direction of other people's gaze, and a pre-linguistic toddler's recognition that he and another person are simultaneously focusing attention on the same object (see Baron-Cohen, 1995; Gopnik & Meltzoff, 1997). Meltzoff famously showed that newborns as young as 42 minutes will mimic a hovering adult's tongue protrusions and other facial expressions (Gopnik & Meltzoff, 1997). The cognitive complexity of such a feat is astonishing: Not only must the infant recognize a similarity between herself and others, but also integrate sensory input from her eyes with feedback from proprioceptors (internal muscle position detectors), and orchestrate it all into a coordinated motor action (Gopnik & Meltzoff, 1997). This innate link between self and others may be another skeletal precursor upon which the child's intuitive psychology will ultimately be built, while the instinct to imitate may be a skeletal precursor to the imitative learning that lies at the heart of tool mastery and intuitive engineering (see Tomasello et al., 1993).

"The theory theory." According to R. Gelman (1990, 1991) and like-minded scholars, our innate cognitive constraints, biases, orientations, and skeletal principles serve as springboards for developing complex conceptual systems (see Carey, 2009; Gallistel et al., 1991, Inagaki & Hatano, 2002). These conceptual systems are coherent webs of interrelated mental representations, physically underwritten in the brain by a complex of neural associations (Bransford et al., 1999). Cognitive scientists often characterize them as "framework theories" (Wellman, 1990; Wellman & S. Gelman, 1992): holistic interpretive schemes – akin to scientific theories – each specific to a particular content domain such as rigid objects, biological beings, human behavior, or tools and artifacts. According to advocates of the "theory theory," a conceptual scheme is *theory-like* if it is (a) *interpretive* (supporting inferences about events in the world), (b) *explanatory* (grasping those events through cause-and-effect principles rather than mere description), (c) *predictive* (in the sense of principled predictions, as opposed to purely statistical, probabilistic projections), and (d) *ontological* (circumscribing and/or categorizing the kinds of things that belong within its content domain; Carey, 2009; S. Gelman et al., 1994; Gopnik & Meltzoff, 1997; Gopnik & Wellman, 1994; Hirschfield & S. Gelman, 1994b; Inagaki & Hatano, 2002; Keil, 1989; Wellman & S. Gelman, 1992; Wellman, 1990). Each intuitive theory allows the human mind to make useful inferences and predictions across a broad range of circumstances, albeit only within the appropriate context.

Our intuitive physics, biology, psychology, and engineering – discussed earlier – all meet these criteria as "theories." Each covers its own ontological domain, and each plies its own distinctive cause-and-effect construal of interpretation, explanation, and prediction: namely, contact causality, inner essences, intentional agency, and teleological function, respectively. Many scholars depict these theories as products of calibrational learning: elaborate conceptual cathedrals that children gradually *construct* around "skeletal" scaffolds already built into the evolved mind, sculpting them simultaneously in accord with empirical *experience* and *innate* cognitive constraints, biases, and "modes of construal" (see Bjorklund & Pellegrini, 2002; Carey, 2009; Geary, 2007; Geary & Bjorklund, 2000; Keil, 1994). Although the human mind may come pre-equipped to project functionality and purpose onto human artifacts, in the contemporary household this intuition must still be calibrated to whole collections of kitchen gadgets, garage tools, and electronic gizmos, all of which help a child construct his intuitive theory of "engineering" (see Keil, 1989). A child may flesh out her intuitive theory of biology by instinctively slotting local flora and fauna into a skeletal classification scheme according to perceived essences (Atran, 1998; Berlin, Breedlove, & Raven, 1973; Keil, 1989).

On this view, then, these "theories" are simultaneously grounded in lived experience and born of the interpretive filter of innate cognition. They would have been adaptive in ancestral environments, and so – *even though learned and constructed by each individual child* – these theories may nonetheless have been shaped by natural selection acting on whole populations. They may be products of epigenetic calibration and "instinctive learning," *no less evolved* than a barnacle morph, an elephant's trunk, or the shape-shifting acrobatics of an octopus on the run.

Summary, Thesis #3: Calibrational Learning

In this section I have showed that learning is not the opposite of instinct. Instead, evolution endowed us with "instincts to learn" (Marler, 1991). "Learning" to walk, to talk, to navigate by circumpolar constellations (if you are a bunting), to avoid certain flavors of water (if you are a rat), to mimic the local seafloor (if you are a flounder), to conform a dam to the contours of a stream (if you are a beaver), to operate a new tool (if you are a *Homo sapiens*), to anticipate the arc of a projectile, to forecast the seasonal fruiting of a particular species of tree, to predict the goal- and belief-driven behavior of other people – all of these are arguably epigenetic calibrations. Evolution "expected" us to have certain environmental experiences and *prepared* us to learn accordingly (Gallistel et al., 1991; Geary, 2002, 2007). It endowed us with what Geary (2007) calls "modular plasticity": natural selection crafted domain-specific mental machinery that is flexible within a finite, fitness-relevant range of variation, conforms to select local particulars, constructs useful mental representations of them, and construes them in a specific manner. Thus we come to interpret experience in evolutionarily ordained ways, ways that gave our ancestors a flexible yet stable grip on a dynamically shifting environment.

Among the products of such instinctive, calibrational learning are the “intuitive theories” of physics, biology, psychology, and engineering that enable us to generate the powerful, principled predictions, inferences, and cause-and-effect explanations that guide us as we navigate our social and ecological worlds. The instrument that was field-tested in this doctoral study was designed to bring to the surface certain elements of these intuitive theories – namely, those that can make the scientific theory of evolution counterintuitive for young people (see Table 1). For a student to abandon those products of calibrational learning in favor of the scientific model requires *conceptual change*. I will take up conceptual change under Thesis #4. But first, an interlude to schematically depict the model of the evolved, intuitive mind that I have developed so far.

Interlude: A Hypothetical Model of the Intuitive Mind

In the first half of this literature review, I showed how evolution may have shaped the human mind to develop – through *calibrational learning* – distinct, domain-specific, *intuitive theories* about the world. I dwelt especially on those intuitive theories that have potential to compromise a student’s grasp of modern scientific theories, the theory of evolution in particular. In the second half of this literature review, I will turn more explicitly to the issue of *conceptual change*, the cognitive process by which a student may replace an intuitive theory with a scientific one. First, however, I want to begin elaborating the schematic model of the evolved mind depicted in Figure 1, based on ideas developed so far in Theses #2 and #3 with respect to intuitive construals/projections, calibrational learning, and intuitive framework theories. Again, this schematic model was one of the theoretical foundations that inspired and informed the design of my experimental instrument.

Sense Organs and Perceptual Analyzers

The model (Figure 1) is hybridized from the work of multiple scholars in the fields of cognitive science, developmental psychology, and evolutionary psychology. It incarnates the "computational theory of mind" that came to dominate psychological science during the "cognitive revolution" of recent decades, according to which cognition is a form of *information-processing* carried out by *specialized computational mechanisms*, each subserved by dedicated neural structures (Fodor, 1983; Pinker, 1997; Thagard, 1992). On this view, the mind comprises an assortment of innate processors that translate and transform information from the environment into various *mental representations*. These processes and representations need not be conscious or attended by any sense of voluntary control, and most probably are not (Fodor, 1983; Kahneman, 2011; Pinker, 1997; Wilson, 2002).

The model should be read from the bottom up. The lower tiers derive from the influential work of Fodor (1983). Sense organs (green) harvest mechanical and electromagnetic energy from the environment (orange), and translate it into patterned neural impulses. This information is directly cabled (blue arrows) to the appropriate "input analyzers" (light blue; Fodor, 1983). These seek out specific patterns amidst the kaleidoscopic stream of sensory input, and when found, encode them into *perceptual* representations. Especially well-researched are input analyzers for visual processing. Watching a flock of pelicans in flight, or twisting a cereal box in one's hand, or merely turning one's head splashes a spectacle of shifting geometric forms onto the twodimensional surface of one's retina. Yet downstream of the eyes, innate visual processors assemble the erratic montage into a stable three-dimensional world of depth and durable objects (Pinker, 1997). Other input analyzers may listen for signatures of grammatical speech amidst ambient noise, detect face-like patterns, recognize individual voices, map the positions of our limbs and bodies in space, single out the self-animated movements of animals, and so on (Carey & Spelke, 1994; Fodor, 1983; Pinker 1994, 1997; R. Gelman, 1990).

Core Cognition and Intuitive Theories

These *perceptual* representations are then directly and selectively cabled to the appropriate *conceptual* systems. Following Carey (2009), I adopt the term "core cognition" for this tier of information-processing mechanisms (pink). Each is dedicated to a domain that was important to ancestral hominins. Many may exist, but my model posits four that are relevant to learning science: the domains of physical objects, animals/plants, other people, and tools/artifacts. As discussed earlier, these mechanisms interpret our perceptions, often projecting onto them a domain-specific form of causation: mechanical contact, hidden essences, teleological functions, or intentional agency. Whereas the operations and output of Fodor's perceptual analyzers have a "here and now" immediacy tied closely to the senses, core cognition is *conceptual* in that its mental representations play rich inferential roles (Carey, 2009). Again, none of this need be available to conscious awareness, reflection, or control.

The output of core cognition may now fuel the formation of domain-specific "intuitive theories" (green), elaborate mental models that enable one to carve the world into categories and make useful cause-and-effect explanations and predictions (Carey, 2009; Gopnik & Meltzoff, 1997; Wellman & S. Gelman, 1992). The four depicted here are a person's intuitive physics, biology, psychology, and engineering (see Geary, 2007; Pinker, 1997, 2002; Mithen, 1996). As discussed in the preceding section, these "theories" are grounded in lived experience – that is, they are *learned* – yet also born of the interpretive filter of innate core cognition, and would have been adaptive in ancestral environments. That is, they are produced by calibrational learning.

Vertical Flow vs. Horizontal Integration

A crucial feature of this model so far is that the flow of information has been strictly "vertical." It proceeds from sense organs to perceptual processors to core cognition to intuitive theories through neural pipelines that permit no lateral sharing of representations between cognitive mechanisms (purple two-headed arrows crossed out in red). That is, no "horizontal integration" occurs between domains (Fodor, 1983). Evolutionarily this may be the default state: Natural selection would not wire independent information-processing mechanisms together unless there were an adaptive benefit for doing so (Kurzban, 2010). For example, songbirds presumably have separate computational systems for nest-building and song-making – both of them innate yet requiring calibrational learning – and there is no reason to suspect that mental representations are shared between them (Fodor, 1983). Because cross-communication between nest-building and song-making mechanisms would have bestowed no adaptive advantage, the output of one never became input to the other.

Nevertheless, as I will later describe, virtually all scholars agree that conceptual change does require some degree of horizontal integration across domains. And since humans are plainly capable of conceptual change, an essential question is: What evolutionary pressure prompted the lateral integration of previously isolated sub-systems in our species? This is a key question that I will take up in the second half of this literature review, under the title "the riddle of horizontal integration." For now, my model only calls attention to the common view that horizontal integration is the work of the brain's "executive" system (tan), which receives representations from multiple domains and uses them to make decisions and selectively generate behavior. (Such a superordinate system may exist in songbirds, too, regulating the activation of each sub-system – e.g., governing when to build nests versus when to sing.) Later, however, I will hypothesize that there is a second candidate locus of horizontal integration, hence conceptual change. That hypothetical locus – the "social cognitive interface" – is intimately connected to the mind's mechanisms for strategically navigating interpersonal relations in the social domain. I will suggest that the adaptive value of horizontal integration lies in piecing together *other* people's intentions and interpretations, not one's own. That hypothesis partly inspired the development of the instrument that was field-tested in this doctoral study.

Furthermore, according to this model information moves from tier to tier not only vertically, but strictly "upward." Lower tier computations are automatic and not subject to top-down influence from the brain's "higher" systems (curved purple arrows, crossed out). In the case of *perceptual* mechanisms, Fodor (1983) illustrates this point with optical illusions (Figure 6): No amount of conscious effort will alter the unconscious operations of the visual input analyzer. Likewise, he writes, "you can't help hearing an utterance of a sentence (in a language you know) as an utterance of a sentence….You can't hear speech as noise even if you would prefer to" (pp. 52-53).

Figure 6. The Müller-Lyer illusion. Though the two horizontal shafts are equal in length, the automatic operation of the mind's visual analyzers misreads the top one as shorter. The illusion is immune to conscious effort to see the shafts accurately.

Operations at the *conceptual* tiers may be equally automatic. For example, Leslie's (1994) launching event experiments (described earlier) show that we automatically project unidirectional causality onto colliding objects, and this may be quite irrepressible, even for adults. So, too, may be the projection of teleological function onto tools, or hidden essences onto animal species. Argues Carey (2009): "Core cognition is elaborated during development [i.e., experientially and calibrationally]….[but] never overturned or lost….The representations that articulate core cognition…continue to operate throughout life" (pp. 68-69).

In short, core cognition and perhaps even intuitive theories might not be amenable to modification from above. *Even after conceptual change occurs, certain pre-instructional intuitions may continue to feed "upward" into the "higher" cognitive systems which generate the conceptual models that a student articulates to other people* (e.g., on science tests). In other words, conceptual change may require a student to *suppress* or *suspend* those intuitions, rather than restructure or supplant them. And even after conceptual change, two or more relatively independent conceptual models may *coexist* for the student, non-exclusively. The instrument that was field-tested in this study was specifically designed to (a) allow such coexisting models to surface, while also (b) eliciting signals that whenever a student is forced to choose between competing models, she will suppress certain models in favor of others.

Thesis #4 – Conceptual Change and the Divide between Biological and Cultural Evolution

Education is a technology that tries to make up for what the human mind is innately bad at.

Stephen Pinker (2002, p. 222)

So far I have made the case that *biological* evolution has provided us with pragmatically useful mechanisms for construing natural phenomena and constructing intuitive theories in the domains of biology, psychology, physics, and engineering. In this section I portray modern scientific models in these same domains as products of *cultural* evolution that do not always cleanly accord with our innate intuitions. I argue that for 21st century science students, mastering certain scientific models – including the theory of evolution – requires a distinct form of learning that goes beyond mere calibrational learning. It requires conceptual change. *I develop the following thesis:*

> Once our hominin ancestors evolved cognitive machinery for acquiring knowledge and knowhow from others in the social group – that is, for cultural transmission – human populations could accumulate behavioral and conceptual changes non-genetically, yet these changes were still channeled and constrained by biologically evolved intuition. Eventually, however, cultural evolution produced the self-correcting empirical mechanisms of modern natural science, which have been able to generate and validate conceptual models that conflict with biologically evolved intuition. Consequently, an especially challenging form

> of conceptual change is necessary for students to master those models: They must restructure, replace, or at least override a pre-instructional theory that is so well entrenched, it may even continue to coexist in the student's mind alongside the newly acquired scientific understanding.

A psychometric instrument that seeks to measure this species of conceptual change should be designed to allow both innate intuitions and scientific conceptions to surface side-by-side, in case they coexist for the student. The instrument should also contain items that oblige the test taker to suppress the former in favor of the latter. These were primary design specs for the instrument piloted in this doctoral study.

Natural Science as an Engine of Cultural Evolution

The biological evolution of cultural evolution.

A number of adaptations and capacities distinguish *Homo sapiens* from almost all other mammals. These include toolmaking, control of fire, cooking, symbolic language with complex grammar, cooperation and food sharing among non-kin, abstract conceptual reasoning, use of projectiles as weapons, a division of foraging labor between the sexes, high male parental investment in offspring, self-awareness, a theory of mind, opposable thumbs, bipedalism, and one adaptation that is especially relevant to this thesis: *cumulative culture*. The qualifier "cumulative" is essential. A few other animals arguably have "culture" in the sense of behaviors that are diffused socially rather than genetically. "Termite fishing" in chimpanzees and "bubble netting" in killer whales, for example, may cross the generations through cultural transmission. But only hominins have *cumulative* culture, by which new technologies and conceptual systems continually emerge and progress (Richerson & Boyd, 2005; Tomasello, 2014). Just as biological evolution fashioned such complex, exquisite adaptations as eyes and wings through the incremental accumulation of lucky mutations across many generations (Dawkins, 1996), so has "cultural evolution" produced such elegant "adaptations" as kayaks, boomerangs, and alphabetic writing through the gradual, cross-generational accumulation of small improvements (Richerson & Boyd, 2005).

Richerson and Boyd (2005) make a strong case that our capacity for cumulative culture is itself a species-specific adaptation, subserved by dedicated neural systems:

> Culture is as much a part of human biology as walking upright….Culture is taught by motivated human teachers, acquired by motivated human learners, and stored and manipulated in human brains. Culture is an evolving product of populations of human brains, brains that have been shaped by natural selection to learn and manage culture. (p. 7)

In short, we are *biologically evolved* to *evolve culturally* (see also Tomasello, 2014). Initially, what probably made cumulative culture possible was the evolution of an exceptional cognitive faculty: an ability to read others' intentions and to faithfully imitate their goal-directed behaviors (Tomasello et al., 1993). Later, of course, hominins evolved cognitive machinery supporting spoken language (Pinker, 1994), and this became a primary vehicle for cultural transmission. We also appear to have an innate disposition to emulate individuals of status, to adopt the practices of successful individuals, to conform our own beliefs to those of the majority, and to both embrace and enforce social norms (Richerson & Boyd, 2005; Tomasello, 2014).

Richerson and Boyd (2005) argue that our "cultural nature" evolved as part of a generalist, opportunistic niche on the African savannah which enabled hominins to exploit a wide diversity of ever-shifting habitats during the rapid climate oscillations of the Pleistocene epoch (see also Gamble et al., 2014; Mithen, 1996; Sterelny, 2012; Tomasello, 2014). Humans are evolved both for independent learning and for cultural learning, and the *combination* is a recipe for rapid cultural evolution: “When lots of imitation is mixed with a little bit of individual learning, populations can adapt in ways that outreach the abilities of any individual genius” (Richerson & Boyd, 2005, p. 13). Cumulative culture made human populations far more flexible than other social animals, able to adapt technologically and behaviorally to ephemeral environments and new niches.

Memetics and other models of cultural evolution. Invoking the term “evolution” for cultural change is no mere analogy; it is literal. Scholars have put forth various quasi-Darwinian models of cultural evolution via natural-selection-like processes. Each posits a mechanism for the *selective* spreading of information, ideas, technologies, and other slices of transmittable, transmutable culture. The most straightforward application of evolutionary theory to culture centers on Dawkins’ (1976/1989) construct of “memes” (see also Blackmore, 1999; Dennett, 1995, 2002). Derived from the Greek *mimesis* (imitation) and meant to echo the word “gene,” a meme is a nugget of information – an idea, a technology, a text, a custom, a fashion, a hummable tune – that (a) resides in human brains or is preserved in human artifacts such as books and tools, and (b) can be copied into other human brains or artifacts. Like genes, memes are replicators. Because the number of “brain vacancies” is finite, memes compete with one another for attention and adoption by human minds. Attractive and memorable memes spread, the rest spiral into extinction. Inevitably, copying fidelity is imperfect. Since heritable variation is thus coupled to differential replication, memes – like genes – evolve. This, say meme theorists, is cultural evolution.

Other scholars have criticized the meme model as too simple (e.g., Kirkpatrick, 2009; Sperber, 1996). Along with other “dual inheritance” models, it treats biological and cultural evolution as two separate vectors unfolding along largely independent trajectories, bound together only by the fact that the human brain evolved a general capacity for imitative learning, information replication, mutual perspective-taking, and teaching (see Blackmore, 1999; Dennett, 1995, 2002; Richerson & Boyd, 2005; Tomasello, 1999, 2014). What such models neglect, say the critics, is the richness of the human mind’s innate psychological architecture (Kirkpatrick, 2009; Sperber, 1996). That architecture evolved *biologically* before the advent of culture, and it is through this architecture that “cultural selection” occurs. Every meme, every truth claim, and every *conceptual representation* that enters the public sphere must pass again and again through the selective, interpretive filter of each individual’s innate psychology (Kirkpatrick, 2009; Sperber, 1996). Earlier, for example, I discussed several (among many) of the cognitive constraints, construals, and constructions that the biologically evolved mind imposes on experience and information. These surely influence which memes are memorable, which concepts are adoptable, and which directions cultural evolution will and will not take. Innate intuition channels and constrains cultural selection (Boyer, 2001; Sperber, 1996; Wilson, 1998).

Natural science: a divergence of cultural and biological evolution. The foregoing implies that even modern scientific concepts are inhabited by our innate intuitions about the natural world, since those intuitions historically favored the adoption of certain memes over others. Cultures everywhere, for example, embrace a "folk biology" that manifests such psychological tendencies as essentialism, teleology, and vitalism, and these same intuitions still take up subtle residence in the models and metaphors of natural science (Atran, 1998; S. Gelman, 2003; Inagaki & Hatano, 2002; Mayr, 1991). We readily transfer the intuitive notion of internal "essences" to such contemporary scientific constructs as "species" and "DNA," while the teleological stance persists in such constructs as "adaptation" and "ecological niche."

Nevertheless, the history of natural science is also a tale of successful *divergences* from folk theories. Science has driven the cultural evolution of many concepts that – against the backdrop of our biologically evolved psychology – are *highly novel* (Geary, 2007). The process, once again, looks Darwinian: Scientific concepts and theories are relentlessly, cyclically subjected to empirical scrutiny, and those of poor predictive or explanatory power are culled out (see Kuhn 1962/1970; Thagard, 1992; Toulmin, 1972). Only the fittest models and metaphors survive, with the result that they evolve to conform ever more closely to nature-in-itself. While scientific truth claims are always provisional and probabilistic, and while scientists are human beings whose interpretations are always to some extent subjective, value-laden, and historically, culturally contingent (Crotty, 1998; Willis, 2007), the scientific enterprise wields multiple mechanisms for systematically squeezing human subjectivity out of its interpretations and inferences (Thagard, 1992; Wilson, 1998). Among these mechanisms are the use of falsification to discriminate among competing hypotheses, replication both within and across studies, iterative peer review, and other forms of public discourse that permit criticism of methodology and conclusions. Every scientific truth claim must pass through an exacting trial-by-fire that is both empirical and social, such that the *collective* scientific enterprise is much more objective than individual scientists may be (Thagard, 1992; Wilson, 1998).

Modern natural science, then, is systematically self-correcting, smartly structured to remedy, neutralize, or minimize the effects of human biases – *including, it appears, biases born of our evolved psychology*. It persistently re-crafts its models and metaphors in a manner that meshes them ever more tightly with the material world – *even to the point of constructing concepts that we as individuals find non-intuitive, even counterintuitive*. It is a powerful engine of cultural evolution that both depends upon and spawns *conceptual change* (see Kuhn 1962/1970; Thagard, 1992; Toulmin, 1972).

Student Conceptual Change as Clash between Biological and Cultural Evolution

I argued under Thesis #1 that biological evolution is a process whose very nature is to sculpt "re-presentations" of material reality in the bodies, brains, and behaviors of living organisms. I argued under Theses #2 and #3 that among these evolved "re-presentations" are many of the mental, conceptual *representations of reality* with which we intuitively make sense of experience. And I have now argued that modern natural science is a form of *cultural* evolution that generates its own conceptual representations of reality, by systematically, experimentally, and empirically tightening the correspondence between those representations and the material universe.

Compared to the sluggish pace of biological evolution, however, the cultural evolution of modern scientific models and metaphors has been exceptionally rapid and recent. Consequently, as natural science grows ever more sophisticated, what we ask students to learn in the 21st century classroom becomes ever more "remote" from the ecological and social settings in which their brains evolved (Geary, 2002, 2007).

Inevitably, mismatches occur between scholarly conceptions and students' intuitive projections.

Thesis #4, then, is this: The contemporary science classroom is a crossroads of biological and cultural evolution where cognitive collisions inescapably occur. It is precisely this divide, I suggest, that can sometimes render a scientific concept "counterintuitive" for students. Crossing such a divide would require a special form of conceptual change. This is not to say that all conceptual change represents a raw collision of culturally evolved concepts with innate intuitions. In many cases, for example, the movement is from one culturally evolved conception to another culturally evolved conception. Indeed, the construct "conceptual change" has meant many different things to different researchers in different fields. In the following sub-sections, I delimit the particular species of conceptual change pertinent to my long-term research agenda, and which my experimental instrument was designed to assess.

Conceptual change as instruction-induced. First, my research focuses on conceptual change that is "instruction-induced" (Vosniadou, 2008b), occasioned by deliberate instructional interventions designed to transform a student's pre-instructional conception into a scientifically normative one. This distinguishes it from other forms of conceptual change – much studied in cognitive science and developmental psychology – that occur "spontaneously" during the normal course of childhood cognitive development, occasioned by everyday social interactions, cultural experiences, and/or real world encounters with natural phenomena (e.g., Carey, 1985; S. Gelman, 2003; S. Gelman et al., 1994; Gopnik & Meltzoff, 1997; Gopnik & Wellman, 1994; Hatano &

Inagaki, 1994; Inagaki & Hatano, 2002, 2008; Karmiloff-Smith, 1991, 1992; Keil, 1989, 1991, 1994, 1995; Keil & Newman, 2008; Leslie, 1994, 1995; Wellman, 1990).

An example of the latter: Carey (1985) has contended that young children do not yet possess an intuitive theory of biology, and thus understand animals' behavior and even our own bodily processes *anthropomorphically*. But by late childhood their mental model undergoes a dramatic transformation into a *mechanistic* understanding, whereby living organisms and the human body are conceived as "biological machines." This happens predictably, she argues, through ordinary experience and dialogue, even in the absence of formal schooling.

By contrast, my own research centers on conceptual change that is deliberately pursued in secondary and post-secondary science classrooms, especially the mastery of formal "explanatory models" like the theories of relativity, plate tectonics, evolution, etc., which arguably constitute the most important content goals in science education (Clement, 2008). Such conceptual change is effortful and often plagued by inertia, resistance, and relapse (Carey, 2009; Chi, 2008; Chinn & Brewer, 1993; Driver et al., 1994; Geary, 2007; Posner et al., 1982; Vosniadou, 2008). There is a sense, then, in which instruction-induced conceptual change disrupts a child's "normal" or "natural" cognitive/conceptual development. My position implies a clash between pre-instructional theories and scientifically normative ones, with classroom teachers acting as agents on behalf of cultural evolution over against biologically evolved intuition. As Pinker (2002) puts it, "Education is a technology that tries to make up for what the human mind is innately bad at" (p. 222).

Conceptual change as "high learning demand" and "strong restructuring." Second, my research centers on conceptual change of high "learning demand" (Leach & Scott, 2008; Mortimer & Scott, 2003), defined as the conceptual distance between a student's pre-instructional intuitions about a topic or phenomenon and the desired scientific conception. With some concepts the divide is wider than others, and these are more challenging to master. In the special case with which my research is concerned, learning demand is proportional to the degree that cultural evolution has diverged from biological evolution. My research slants toward the high end of the learning demand continuum: cases of conceptual change that are especially difficult for students.

This is to distinguish my focus from the many educational researchers who use the term "conceptual change" to encompass concept learning and cognitive growth quite generally, ranging from simpler additive and assimilative processes, to minor revising of existing mental models, to dramatic transformations in a person's fundamental working theories about the world's ontological categories and causal relationships (see the taxonomies of Chi, 2008; Clement, 2008; Thagard, 1992; Vosniadou et al., 2008). Other scholars, however, reserve the term only for the latter end of the spectrum. Among these scholars, a pervasive metaphor is that bona fide conceptual change requires the student to *"restructure"* an existing conceptual system, mental model, or framework theory (Carey, 1985; Chi, 1992, 2008; Clement, 2008; Keil, 1989; Keil & Newman, 2008; Sinatra & Pintrich, 2003; Southerland et al., 2007; Thagard, 1992; Vosniadou et al., 2008). In a seminal paper, Vosniadou and Brewer (1987) distinguish "weak restructuring" from "radical restructuring." The former means "enriching and elaborating existing theories" (p. 54), and is conceptual change of comparatively low learning demand. The latter involves high learning demand and entails major reformations of those theories. Carey (1985) draws a similar distinction between “weak” and “strong” restructuring.

While my research does focus on conceptual change at the “high learning demand” and “strong restructuring” end of the spectrum, there is a caveat, as I describe in the next section.

Conceptual change as (plausibly) a contest between coexisting mental models. Later in this chapter I will raise theoretical questions about the aptness of the prevalent metaphor of “restructuring.” I will suggest (on evolutionary grounds) that some difficult instances of conceptual change may embody a contest between two or more *independent* conceptual schemes which *coexist* within the same student’s mind, or at least between two or more relatively isolated cognitive systems. Such a contest between coexisting conceptions is especially likely to arise, I argue, where at least one of them is rooted in biologically evolved intuition. The hypothetical model of mind that I depicted in Figure 1 shows why: According to this model, information flows strictly “vertically” and “upward” within each intuitive domain. Perceptual representations are translated *unidirectionally* into conceptual representations, which then *unidirectionally* feed the formation of domain-specific intuitive theories. This is by evolutionary design. Natural selection would not have made these cognitive mechanisms open to input from “above” (nor “laterally”), unless there were some payoff in reproductive fitness. It is possible, then, that some of our biologically evolved intuitions endure inexorably into adulthood, impervious to conscious efforts to “restructure” them.

My research focuses on conceptual change that at least plausibly involves a contest between two coexisting conceptual systems: one the descendant of biological evolution, the other of cultural evolution. Conceptual change in such cases would not require the student to restructure a singular theory or mental model, but instead to override one theory in favor of another – that is, to *suppress* (not *supplant*) her pre-instructional conception in favor of the scientific one. As I will explain in Chapter 3, the experimental instrument that was field-tested in this doctoral study was specifically designed for such a possibility: Some test items permit a student's coexisting theories – both pre-instructional and scientific – to surface side-by-side. Others ask him to suppress the former in favor of the latter.

Summary, Thesis #4: Conceptual Change and Biological vs. Cultural Evolution

In this section I explained that we humans are *biologically* evolved to evolve *culturally* (see Richerson & Boyd, 2005). To a unique degree in the animal kingdom, our hominin ancestors evolved a cognitive capacity for cultural transmission, supported by such sophisticated competences as reading one another's intentions, imitating others' behaviors, and (later) communicating thoughts via spoken language. More specifically, our ancestors evolved an ability to acquire "cultural variants" (Richerson & Boyd, 2005) from others in the social group: different behaviors, practices, techniques, technologies, beliefs, ideas, norms, and so on. A given cultural variant may or may not spread and endure, depending on whether it finds comfortable lodging in the next generation of human minds. Its success depends largely upon the pre-cultural predispositions of the biologically evolved mind itself: Some cultural variants are innately, intuitively more appealing or memorable than others.

Over many generations, the iterative sifting of cultural variants accelerated cultural evolution to a breakneck pace (breakneck compared to biological evolution, that is). Hundreds of millennia later, culture at last evolved the modern enterprise of natural science, a special and speedy engine of cultural evolution in its own right. Through its own systematic, natural-selection-like self-corrections, science steadily tightened the predictive, explanatory grip that its models and metaphors have on the natural world. In the process, however, some of these models and metaphors evolved along conceptual trajectories that diverge strongly from *biologically* evolved ways of interpreting experience.

In this light, I made a case that learning and mastering scientific concepts and theories in the 21st century often entails a special form of conceptual change: one that is (a) induced by classroom instruction (as opposed to “natural” or “normal” conceptual change that occurs spontaneously in the course of life experiences); (b) highly challenging for students because it requires them to restructure, replace, or at least override an established pre-instructional mental model or framework theory (as opposed to more general conceptual growth, such as the additive or assimilative learning of new concepts); and (c) involves a target conception that is non-intuitive or counterintuitive precisely because it conflicts with biologically evolved intuition (as opposed to movement from one culturally evolved conception to another). I suggested, moreover, that even after a student undergoes this form of conceptual change, pre-instructional conceptions may continue to *coexist* in her mind alongside the new scientific understanding, and so must be actively suppressed even into adulthood. The experimental instrument that I developed for this doctoral study deliberately allows for this possibility: Some test items invite a student’s coexisting conceptions – some descended from biological evolution, others from cultural evolution – to surface side-by-side, while other test items ask her to suppress the former in favor of the latter.

Thesis #5 – The Evolutionary Origin of Conceptual Change: "Individual Constructivist" Hypotheses

It is not that children are little scientists but that scientists are little children.

Alison Gopnik & Andrew Meltzoff (1997, p. 32)

The genius creates good ideas because we all create good ideas; that is what our combinatorial, adapted minds are for.

Steven Pinker (1997, p. 362)

My first four theses collectively built a case that conceptual change sometimes embodies a clash between *biologically evolved* ways of interpreting experience and the *culturally evolved* theories of modern natural science. The next three theses now tackle the riddle of *how* biological evolution ever could have fashioned a mind capable of such conceptual change in the first place. Like tool use and spoken language, conceptual change is a distinctly human capacity. Tool use and language are cognitive/behavioral adaptations that evolved through natural selection, and their adaptive benefits are plain (Mithen, 1996; Pinker, 1994, 1997; Richerson & Boyd, 2005; Tomasello, 1999; Tooby & Cosmides, 1992). Our cognitive capacity for conceptual change likewise must have evolved long ago through natural selection, though its adaptive benefits are not so clear. How could biological evolution produce a mind/brain able to abandon its own innate intuitions to adopt counterintuitive ideas handed down by cultural evolution? Why did ancestral humans evolve an ability to forsake preexisting, compelling mental models of the world in favor of novel, less compelling ones? What ecological and/or social conditions prompted its evolution, and what adaptive problems did it solve?

I believe that a consideration of such questions may yield useful insight into the psychological mechanisms and cognitive processes which make conceptual change possible for 21st century science students, and that this in turn may suggest educational strategies and instructional interventions to summon them into action in the modern science classroom. In this section I put forth the first of three hypothetical origins of our capacity for conceptual change: an *individual constructivist* hypothesis. In subsequent sections I proffer a pair of alternatives: a *cultural re-constructivist* hypothesis and a *social strategic* hypothesis. Each posits a different adaptive problem, a different ancestral circumstance, as the stimulus which rewarded an ability to exchange intuitive ideas for counterintuitive ones, thus prompting the evolution of psychological mechanisms enabling conceptual change. I will further argue that in the modern science classroom, different pedagogical practices implicitly correspond to each of these hypothetical origins. That is, the way a teacher designs instruction for conceptual change implies an unspoken assumption about how the evolved human mind works, and how it is even able to undertake conceptual change in the first place. *Here I develop the following thesis:*

> Natural selection might reasonably be expected to select against any mode of learning that violates or subverts innate intuitions. For this reason, the human ability to adopt counterintuitive concepts itself demands evolutionary explanation. One possibility is that conceptual change was in fact adaptive, evolving because it enabled each individual to construct an ever more accurate understanding of the

> natural world, thus conferring heightened cognitive and behavioral flexibility in the face of environmental variability. In practice, many of the pedagogical strategies advocated by science education scholars implicitly assume such an evolutionary origin.

The experimental instrument that I developed for this doctoral study was not specifically motivated by this first hypothesis and its corresponding pedagogies, nor by the "cultural re-constructivist" hypothesis to be developed under Thesis #6. Nevertheless, these first two hypotheses form the essential backdrop against which I conceived the "social strategic" hypothesis of conceptual change (Thesis #7), which did inspire the design of my instrument.

The Cognitive Evolution of Conceptual Change: A Pair of Darwinian Riddles

Under Thesis #3 I showed that "innate" intuition need not preclude learning. Rather, like other animal species, we have "instincts to learn" (Marler, 1991), and within the intuitive domains of physics, biology, psychology, and engineering, learning is (arguably) "calibrational": We are prewired to construct "intuitive theories" in accord with certain conceptual expectations about the kinds of experiences that we will have during life, and this enables us to attune our mental models to particulars of the local environment. By contrast, conceptual change is a form of learning that (arguably) supersedes calibrational learning. From a Darwinian perspective, our capacity for conceptual change presents at least two theoretical puzzles, which I discuss below: an "evolvability" riddle and a "horizontal integration" riddle.

The riddle of evolvability. Under Thesis #1 I characterized natural selection as a process that "meshes" the bodies, brains, and behaviors of living organisms to statistically recurring, fitness-relevant features of the material world (Tooby & Cosmides, 1992). But evolution and its multimillion-year meshings did not expect us to encounter among our environmental experiences the very recent, very novel concepts, principles, and explanatory models of modern science (Geary 2002, 2007; see Thesis #4). Consequently we are ill disposed to properly grasp them. The problem is that our minds did not strictly evolve to discover scientifically verifiable truths about the world, but instead to make sense of the world by way of fast-and-frugal shorthand interpretations that in ancestral environments were "just good enough" to secure vital resources, negotiate threats to survival, and maximize reproductive success (Kahneman, 2001; Keil, 1994; Richerson & Boyd, 2005; Tooby & Cosmides, 1992). Innate intuitions evolved because they were functional; they solved an adaptive problem (Tooby & Cosmides, 1992). They provided our hominin ancestors with pragmatically useful ways of interpreting phenomena in their physical, ecological, and social environments.

And this presents a deep riddle. An essentialistic disposition toward toothy predators and venomous reptiles enabled our ancestors to make *accurate enough* predictions about animal behavior that might have been life-saving. A mechanistic (if non-Newtonian) stance toward projectile motion enabled hominin hunters to make *accurate enough* throws to put meat on the fire. A tendency to group flora into "natural kinds" enabled foragers to make *accurate enough* assumptions to avoid toxic vegetation. *Presumably these intuitive propensities long preceded the evolution of any capacity for conceptual change.* It is difficult to imagine an adaptive landscape that would have allowed the "erasure" of such useful psychological instincts in favor of a more general and sluggish learning process (Sperber, 1994; Tooby & Cosmides, 1992), or the evolution of a dedicated psychological mechanism whose function is to “transcend” them (Carey, 2009; see Atran, 1994, 1995; Kurzban, 2010; and Trivers, 2000 for kindred positions).

Of course, if natural selection could have seen Copernicus, Newton, Maxwell, Darwin, and Chomsky coming, it might have anticipated their arrival by building a capacity for conceptual change into our psychological architecture. But it could not. In the ancestral world, dismissing innate intuition and calibrational learning in favor of a less intuitive representation would likely have compromised one’s fitness-relevant predictions relative to one’s competitors (Kurzban, 2010): a less accurate projectile launch, perhaps, or a slower reaction to dangerous wildlife. An individual whose penchant for teleological interpretation was compromised would find herself at a disadvantage in wielding tools. Someone who suspended or suppressed his “theory of mind” would find it challenging to navigate the social environment, as we see today in autistic individuals who lack the intuitive psychology that most of us take for granted (Baron-Cohen, 1995). In short, natural selection should select *against* learning that violates or subverts our innate intuitions.

The question is one of “evolvability”: How *could* a capacity for conceptual change ever evolve? Or rather, since it plainly did, how *did* it? If the cognitive capacity for conceptual change is itself a psychological adaptation, underwritten by innate neural architecture, then what adaptive problems did it evolve to solve? At the very least, any model of conceptual change *as an adaptation* has to convincingly explain how the fitness payoffs could have compensated for the cost of eroding one’s fast, frugal, and otherwise *accurate enough* intuitions (Kurzban, 2010). What selective advantages could favor individuals who are able to discard, disregard, or suspend those patently useful intuitions?

The riddle of horizontal integration. A second Darwinian puzzle relates to the cognitive *processes* by which conceptual change occurs. Earlier I presented a hypothetical model of evolved human cognition (Figure 1), a model which strongly influenced the design of my experimental instrument. According to this model, incoming information flows "vertically" upward through perceptual and conceptual systems that are isolated from one another: There is no "horizontal" sharing of mental representations between different functional systems. Output from one intuitive domain does not become input to another intuitive domain. This accords with evolutionary theory. Natural selection would not cable together independent information-processing mechanisms unless there were an adaptive benefit for doing so:

> Because all the different modular systems in the brain are products of evolution, there is no sense in which connections among modules is the necessary or default state of affairs. Selection must act to link up systems in a way that enhances overall function…. Informational encapsulation – the *lack* of information flow across modules – is, oddly, the default. Evolution must act to connect modules, and it will only act to do so if the connection leads to better functioning. (Kurzban, 2010, p. 49)

Watch any animal go about its daily affairs: Its behaviors manifest a mosaic of separate, domain-specific systems. Thus in songbirds, nest-building and song-making are presumably governed by two separate systems (Fodor, 1983). While a superordinate system might exist which regulates their *activation* – dictating when to build, when to sing – that would not entail "horizontal" *integration* of the mental representations generated by each system.

Again, these vertical pathways can involve *calibrational* learning. Songbirds attune their songs to the local dialect (Marler, 1991), and nest-building improves with experience. Yet the learning itself may be isolated and domain-specific.

Bona fide *conceptual change*, however, appears to differ in this regard. Despite their many disputes and divergences, almost all conceptual change researchers in cognitive science, developmental psychology, and science education agree that conceptual change requires some form of "horizontal" integration across representational systems. Some stress the role of metaphor, making of analogies, or mapping of mental representations from one domain to the next (Carey & Spelke, 1994; Clement, 2008; Hofstadter & Sander, 2013; Pinker, 1997, 2007; Vosniadou & Brewer, 1987). Some stress the value of visual or physical models (Nersessian, 2008; Vosniadou & Brewer, 1987). Some stress the need for students to reassign natural phenomena from one interpretive, categorical, or ontological framework to another (Chi, 1992, 2008; Keil, 1994, 1995; Thagard, 1992). Karmiloff-Smith (1991, 1992) argues that the very essence of conceptual change is a steady tearing down of barriers between once insulated modular systems, through an iterative cognitive process she dubs "representational redescription." Atran (1994, 1995, 1998), even though he maintains that intuitive constraints and construals doggedly persist even in the thinking of professional scholars and scientists, contends that to cultivate new learning, modern education must import, recruit, and analogize from the folk domains. And while he rejects the idea that folk intuitions are organized into coherent, isolated framework "theories," diSessa (2008; Smith, diSessa, &

Roschelle, 1993) maintains that students undergo conceptual change by opportunistically combining and recombining conceptual elements from diverse intuitive domains. In a similar vein, Pinker (2007) argues that conceptual change is possible because our language and concepts constitute a "discrete combinatorial system" whose elements permit an infinity of novel permutations. Nersessian (2008) and Thagard (1992) likewise deem novel conceptual combinations a key mechanism of conceptual change. And Carey (2009) proposes that the most dramatic transformations occur through conceptual "bootstrapping," a back-and-forth dance of ideas from one domain to another.

The common thread here is that in order for conceptual change to occur, the conceptual representations generated within distinct domains must somehow, somewhere cross paths and interact. And this forces the question: What evolutionary benefit prompted the lateral integration of previously isolated cognitive sub-systems in our species? If modularization is the "default" state of evolving psychological architecture, and if inter-domain communication cannot evolve without positive selection pressure (Kurzban, 2010), then what adaptive advantages induced it among our ancestors? What adaptive problem did horizontal integration solve?

"Individual Constructivist" Models of the Origin of Conceptual Change

According to Gopnik and Meltzoff (1997), nothing could be more natural for *Homo sapiens* than scientific conceptual change. Not only do they champion the "theory theory" – the proposition that our intuitions are organized into coherent explanatory models analogous to *scientific* theories – they contend that children are literally born into the world with baseline theories already stitched into the synaptic structures of their brains. Although these baseline theories differ from adult theories, evolution has designed them to be “defeasible”: amenable to revision and even dramatic transformation. They were shaped by natural selection, say Gopnik and Meltzoff, as temporary conceptual footholds meant to be modified – á la scientific models – in accord with empirical evidence, including counterevidence and even *counterintuitive* evidence. Moreover, evolution bestowed children with an instinctive will to explore and experiment – that is, to test their quasi-scientific hypotheses – coupled with special cognitive machinery for reworking their extant theories whenever they encounter empirical shortcomings in them. As their personal experience with natural and social phenomena expands, children progress through a sequence of conceptual changes until they arrive at adult theories. On this account, the exploratory, experimental, and explanatory work undertaken by professional scientists is but a formalized, institutionalized extension of childhood theory-building: “It is not that children are little scientists but that scientists are little children” (p. 32).

Very explicitly, then, Gopnik and Meltzoff (1997) view conceptual change itself as an *adaptation* which natural selection crafted to permit *each individual* to improve the accuracy of the working theories with which she interprets the world around her: “Human beings are endowed by evolution with a wide variety of devices – some quite substantive and domain-specific, others much more general and multipurpose – that enable us to arrive at a roughly veridical view of the world” (p. 15). Many other models of learning in cognitive science, developmental psychology, and science education also correspond (at least implicitly) to such an evolutionary origin. They typically depict conceptual change as a self-evidently adaptive process by which the individual rationally reconciles her conceptions with empirical evidence regarding physical, biological, ecological, and psychological phenomena. The evolutionary benefits, presumably, were an ever-improving individual mastery of the natural world and greater behavioral flexibility in the face of environmental diversity and uncertainty. Individuals who excelled in this regard outcompeted others in their ability to secure resources and elude life-threatening perils, and so had greater reproductive fitness.

I will classify these collectively as “individual constructivist” models of conceptual change. Below I survey a handful of models belonging to this family, drawn from different corners of cognitive science, and discuss how they address the “evolvability” and “horizontal integration” riddles raised above. Thereafter I shift to science education and describe corresponding pedagogical strategies for cultivating conceptual change in the classroom.

Gopnik and Meltzoff: Theories all the way down. Gopnik and Meltzoff’s (1997) radical “theories all the way down” stance neatly sidesteps the evolvability riddle that I raised above: A capacity for conceptual change did not have to evolve against countercurrents of core cognition and calibrational learning, because the latter do not exist. Instead, “intuitive theories” that generate principled predictions are already available from birth, though imperfect and destined to be amended in the face of mounting counterevidence. Spelled out in their important 1997 synthesis *Words, Thoughts, and Theories,* Gopnik and Meltzoff anchor their argument in a wealth of empirical evidence, especially experiments with infants. Here they invoke Neurath’s famous metaphor of the boat that must be rebuilt plank by plank while still at sea. Innate theories are the original boats – crafted by natural selection – in which the child, armed with a handful of boat-building/theory-shaping tools, embarks into the sea of environmental experience. So long as he grows up with the sorts of experience that prevailed in ancestral environments, he will eventually construct the adult folk theories that evolution "intended" him to have – unless modern schooling pilots these cycles of conceptual change to a different shore.

Fodor: Central systems. In his influential *Modularity of Mind,* the cognitive scientist and philosopher Jerry Fodor (1983) paints a very different picture of conceptual cognition: An assortment of hardwired, specialized mechanisms – or "modules" – translate *perceptual* and *linguistic* input into a common "language of thought" (see the blue "input analyzers" in Figure 1 above). Their output is not yet *conceptual*, and the flow of information through them is isolated and strictly "vertical," but they pipe their representations into the mind's "central systems" where conceptual processes and horizontal integration at last begin (see Figure 1, yellow). A primary function of central systems is "fixation of belief": correcting discrepancies among perceptual inputs, something that requires global cross-referencing across different domains and against memories. Central systems also govern linguistic communication with other people, which likewise requires broad access to information from multiple domains and diverse memories. Presumably these two functions were ancestrally adaptive, and this explains the evolution of horizontal connectivity. They also presumably explain the evolution of conceptual change. Although Fodor does not use the language of conceptual change, he portrays "fixation of belief" and "scientific confirmation" as two sides of a single cognitive coin, made possible by "analogical reasoning" across domains:

> There have been frequent examples in the history of science where the structure of theories in a new subject area has been borrowed from…theories *in situ* in

> some quite different domain: what's known about the flow of water gets borrowed to model the flow of electricity; what's known about the structure of the solar system gets borrowed to model the structure of the atom; what's known about the behavior of the market gets borrowed to model the process of natural selection…."Analogical reasoning"…depends precisely upon the transfer of information among cognitive domains previously assumed to be mutually irrelevant. (pp. 107-108)

Carey: Conceptual bootstrapping. In her 2009 masterwork on conceptual change, *The Origin of Concepts,* Susan Carey similarly locates conceptual change in central systems that have broad "horizontal" access to mental representations from multiple domains. But in between these central systems and Fodor's perceptual/linguistic modules, she adds an intermediate tier of innate *conceptual* modules (the pink "core cognition" in Figure 1). Throughout life, she argues, these conceptual modules automatically feed powerful cause-and-effect intuitions, such as the mechanical, teleological, and intentional stances discussed earlier, into the central systems, where they influence day-to-day inference-making and fuel the childhood development of "intuitive theories." Under the right conditions, however, central systems can also radically *restructure* those intuitive theories into new understandings that truly "transcend" their precursors. These "qualitative" quantum leaps, says Carey, constitute genuine "discontinuities," for the concepts in the later theory are "incommensurable" with those in the original, and there can be no "translation" between them. Like Fodor she deems analogical transfer vital to such restructuring, but she goes further. The most challenging conceptual changes (e.g., in science classes) take place through a cognitive process so difficult, she dubs it conceptual “bootstrapping” – as in “pulling oneself up by one’s bootstraps”: an impossible task. First, the student learns how the new concepts are inter-defined amongst themselves, yet *without* understanding them – that is, *not* by mapping them onto his preexisting concepts. These memorized-but-meaningless (or misunderstood) symbols serve as “placeholders,” which must then be interpreted through “modeling processes” such as analogizing across domains, thought experiments, and inductive inference, *especially in light of empirical evidence*. Slowly and step-by-step the child may piece together an authentic grasp of the adult meaning inhabiting the placeholders, and if so, he will no longer be able to make sense of his old concepts.

What might be the ultimate evolutionary origin of this bootstrapping ability? Carey does not say, but here is a clue: They are carried out, she claims, by ordinary “problem-solving mechanisms that play a role in thought more generally” (p. 307), and by:

> Garden-variety learning processes: association, the mechanisms that support language acquisition….noticing analogies and making inductive and abductive leaps….[It depends] on integrating previously distinct representations….The bootstrapping process…maps onto each of its sources and thus serves to integrate them. (p. 328)

Like Fodor, then, Carey apparently accepts that evolution endowed humans with domain-general central systems that can access the domain-specific output of diverse sub-systems and freely, “horizontally” integrate them for consistency, linguistic communication, problem-solving, and so forth – all for coping with ordinary daily life.

Pinker: Metaphor and combinatorics. Like Carey and Fodor, the psycholinguist and evolutionary psychologist Steven Pinker ascribes conceptual change largely to the power of drawing analogies across different domains. His analysis, however, is much more deliberately Darwinian and grounded in subtle evolutionary reasoning. First, he explicitly associates metaphor and analogy with the evolution of a uniquely human trait: spoken language. In *The Language Instinct* – summoning evidence from genetics, morphology, paleontology, neuroanatomy, developmental psychology, and linguistics – he vigorously defends the thesis that humans are exquisitely evolved for speech, complicated grammar, and rapid language acquisition during childhood (Pinker, 1994). Our bodily organs (tongue, larynx, lungs) are specially adapted for speaking, while our brains house a suite of "language organs" of their own, for parsing and producing spoken propositions. Even grammatical frameworks are innate.

Then in *The Stuff of Thought*, Pinker (2007) argues that these linguistic adaptations were long preceded by more ancient cognitive adaptations: the evolution of innate, fundamental concepts such as space, time, substance, change, causation, events, forces, goals, intentions, internal essences, impulses, helping, possessing, preventing, and so on. These are the mental footings of our intuitive physics, psychology, biology, etc. Analyzing English and other modern languages as a "window into the machinery of thought" (p. 60), he shows how our linguistic minds routinely apply these core concepts "metaphorically to other domains, as when we count events as if they were objects or when we use space as a metaphor for time" (p. 26). And this ability "to combine [core concepts] into bigger assemblies and to extend them to new domains by metaphorical leaps goes a long way toward explaining what makes us so smart" (p. 24). Hence Pinker's answer to the riddle of conceptual change:

> Ever since Darwin and Wallace proposed the theory of evolution by natural selection, people have wondered how the human mind evolved the ability to reason about abstract domains such as physics, chess, or politics, which have no relevance to reproduction or survival….Conceptual metaphor points a way to solve the mystery….Imagine an evolutionary step that allowed the neural programs that carry out such reasoning to cut themselves loose from actual hunks of matter and work on *symbols* [emphasis added] that can stand for just about anything. The cognitive machinery that computes relations among things, places, and causes could then be co-opted for abstract ideas….[If] all metaphors are assembled out of biologically basic concepts, then we would have an explanation for the evolution of human intelligence. [It] would be a product of metaphor and combinatorics. Metaphor allows the mind to use a few basic ideas – substance, location, force, goal – to understand more abstract domains. Combinatorics allows a finite set of simple ideas to give rise to an infinite set of complex ones." (pp. 242-243)

Metaphor and combinatorics give us our capacity for drawing analogies – for discerning parallel *relationships* in otherwise dissimilar phenomena – which in turn has made modern science so spectacularly successful: "Our powers of analogy allow us to apply ancient neural structures to newfound subject matter, to discover hidden laws and systems in nature" (p. 276). Likewise, says Pinker, the use of analogy is critical for provoking conceptual change in the science classroom:

> Many scientific theories were first stated as analogies, and often are still best explained that way: gravity is like light, heat is like a fluid, evolution is like selective breeding, the atom is like a solar system, genes are like coded messages. (p. 254)
>
> Conceptual metaphors point to an obvious way in which people could learn to reason about new, abstract concepts. They would notice, or have pointed out to them, a parallel between a physical realm they already understand and a conceptual realm they don't yet understand. Analogies such as THE ATOM IS A SOLAR SYSTEM or AN ANTIBODY IS A LOCK FOR A KEY would be more than pedagogical devices; they would be the *mechanism that the mind uses to understand otherwise inaccessible concepts* [emphasis added]. (p. 241)
>
> The goal of education is to make up for the shortcomings in our instinctive ways of thinking about the physical and social world. And education is likely to succeed not by trying to implant abstract statements in empty minds but by taking the mental models that are our standard equipment, applying them to new subjects in selective analogies, and assembling them into new, more sophisticated combinations." (p. 439)

In sum, it was the evolution of spoken language as a "discrete combinatorial system," anchored in a more ancient constellation of instinctive concepts and construals, that paved the way both for scientific progress and for learning science. It permits limitless inventiveness in the form of fresh permutations of existing concepts. And precisely this, he says, is its *adaptive* value: "The genius creates good ideas because we all create good ideas; that is what our combinatorial, adapted minds are *for* [emphasis added]" (Pinker, 1997, p. 362). This is his solution to our twin Darwinian riddles of evolvability and horizontal integration. A fundamental, useful, and economical feature of symbolic language – the metaphorical transfer of core concepts/symbols to new material – was ramped up through combinatorics to weave theories and models far more sophisticated than anything natural selection "intended" (note the fruitful metaphor). Conceptual change evolved as part of our elaborately specialized linguistic nature, imparting cognitive creativity and flexibility that he deems self-evidently adaptive.

Karmiloff-Smith: Developmental de-modularization. In my reading, Fodor, Carey, and perhaps Pinker attribute conceptual change to the work of evolved central systems perched atop an assortment of innate information-processing modules, all forever isolated from one another (see Figure 1). In an essay aptly titled "Beyond Modularity," and later a book of the same name, Annette Karmiloff-Smith (1991, 1992) argues that this initially compartmentalized structure gradually breaks down during childhood cognitive development, and she believes her studies with young children have captured this de-compartmentalization in progress. Her fascinating theory is a developmental approach that attempts to answer the question: Whence cognitive flexibility? Citing Fodor, Carey, and others, she accepts that evolution bequeathed the human mind with many specialized modules, and with Pinker she denies that linguistic and conceptual development could ever get off the ground without them. But how, then, do we become such flexible thinkers? The answer, she proposes, is that as we progress from infancy to adulthood, mental representations previously confined to one module become available to *other* modules through an iterative process she calls "representational redescription":

> No one would hesitate to accept that the spider, the ant, the beaver, and the like use innately specified knowledge structures. So why not the human? But…then what is special about human cognition? My argument…is that although all species have knowledge *in* their cognitive systems, knowledge in the human mind subsequently becomes knowledge *to* other parts of the mind. (1991, p. 172)
>
> The representational redescription model attempts to account for the way in which children's representations become progressively more manipulable and flexible, for the emergence of conscious access to knowledge, and for children's theory building. It involves a cyclical process by which….implicit information *in* the mind subsequently becomes explicit knowledge *to* the mind, first within a domain and then sometimes across domains. (1992, pp. 17-18)

Modular multiplicity may indeed be the evolutionary default state (Kurzban, 2010), but in our species, says Karmiloff-Smith, concepts get repeatedly re-represented in different parts of the mind.

She documents this process empirically and experimentally in several distinct domains, including intuitive physics, intuitive psychology, language, number, and symbolic notation. Eventually, the blossoming of "horizontal" information flow permits conceptual change via "analogies, thought experiments and real experiments…possible only on the basis of prior representational redescription, which turns *implicit* information into *explicit* knowledge" (1992, p. 16). Conceptual change is uniquely human, part and parcel of our evolutionary niche:

> Far more than even its near cousin the chimpanzee…knowledge [in the human mind] becomes usable beyond the special-purpose goals for which it is normally

> used….Do chimpanzees, like children, play with *knowledge*, just as they play with physical objects and conspecifics?....In the child…representational redescription frequently follows behavioral mastery. The chimpanzee, by contrast, seems to be content to continuously repeat its successes; it does not go beyond behavioral mastery.…Human children spontaneously seek to understand their own cognition, and…this leads to the sort of representational manipulability that eventually allows them to become folk linguists, physicists, mathematicians, psychologists, and notators....Intra-domain and inter-domain representational relations are the hallmark of a flexible and creative cognitive system.…Let me go so far as to say that the process of redescription is, in Marler's terms, one of the human instincts for inventiveness. (1992, pp. 191-193)

This is a strong adaptationist position with respect to conceptual change. The fitness-promoting benefit, presumably, was that metacognition and the horizontal flow of information bestowed our ancestors with behavioral, technological, and cognitive flexibility.

Mithen: Evolutionary de-modularization. In The *Prehistory of the Mind*, Mithen (1996) tackles the riddles of conceptual change and horizontal integration from a different perspective: an anthropological one. Citing Carey, Fodor, Karmiloff-Smith, and other cognitive scientists, as well as the evolutionary psychologists Tooby and Cosmides, he is convinced by the evidence that the human mind houses innate computational mechanisms for reasoning in at least four domains: biology (which he calls "natural history intelligence"), psychology ("social intelligence"), language ("linguistic intelligence"), and physics and artifacts combined ("technical intelligence"). And yet, he observes, people do not tackle life's challenges in a compartmentalized way. Foraging peoples saturate the natural world with anthropomorphic projections (animism), identify human kin groups with plant and animal kinds (totemism), and imbue even the most functional tasks – such as crafting tools, garments, and shelter – with social significance (status, kinship, group membership, etc.). In modern society, children relentlessly anthropomorphize dolls and animals, and their cartoons are rife with violations of intuitive physics and psychology (flying people, talking objects, etc.): "This surreal world is understood effortlessly by young minds" (p. 52). Moreover, scientific, mathematical, and artistic genius go far beyond the sort of specialized, domain-specific reasoning that would have been adaptive for hominin ancestors. This is a puzzle:

> We are left with a paradox. The evolutionary psychologists make a very powerful argument that the mind should be like a Swiss army knife. It should be constituted by multiple, content-rich mental modules, each adapted to solve a specific problem faced by Pleistocene hunter-gatherers. One cannot fault the logic of their argument. I find it compelling. But as soon as we think about Cambridge dons, Australian Aborigines, or young children this idea seems almost absurd. For me it is the human passion for analogy and metaphor which provides the greatest challenge to Cosmides and Tooby's view of the mind. Simply by being able to invoke the analogy that the mind is like a Swiss army knife, Leda Cosmides appears to be falsifying the claim that is being made. How can we resolve this paradox? (p. 52)

How could evolution have drilled holes between the walls of our cognitive domains to let knowledge flow between them or to get replicated in different parts

> of the mind?...*This crossing-over between domains is after all exactly what…should not happen in evolution, since it can lead to all sorts of behavioral mistakes* [emphasis added]. (p. 64)

In effect, Mithen is raising our twin riddles of horizontal integration and the evolvability of conceptual change.

In search of an answer, he turns to ancestral artifacts. He argues that he can discern the step-by-step traces of horizontal integration in the archaeological remains left by our hominin ancestors: tools, artwork, cooking sites, lodging, etc. For most of that prehistory, thanks to their evolving "technical intelligence," hominins made sophisticated tools out of stone, yet ignored candidate materials from the *biological* domain such as bone, antler, and ivory. Moreover, although their innate "natural history intelligence" surely permitted shrewd and subtle analyses of prey, they could not creatively apply that knowledge to the design of hunting weapons specialized for each species, making only general-purpose tools instead. And though their "social intelligence" was equally sophisticated,

> It remains just as isolated from the thoughts about toolmaking and foraging as in the chimpanzee mind. There is no evidence that [early humans] used tools in social strategies….There is no imposition of social information on the tools. Similarly, there are no examples in the archaeological record of spatial structure in archaeological sites which might reflect a social use of space. Material culture was not used in social strategies. (p. 126)

Meanwhile, "linguistic intelligence" was evolving, but strictly for social functions – in effect, replacing the mutual grooming that other primates use to cement society together.

Language, he argues, was not yet being used to talk about toolmaking, hunting, plant-gathering, etc., and it had not yet "become transformed to have the general purpose functions which are familiar to us today – a means to communicate information regardless of the behavioral domain" (p. 161). In most respects, early humans became every bit as cognitively advanced as modern humans, with a Swiss army knife mind comprising impressive, complex modules for all four domains – except for one crucial difference: They lacked the horizontal connectivity, hence the "cognitive fluidity," that is the signature of the modern mind.

And then some time in the past 100,000 years, the floodgates opened between them. Archaeological artifacts reveal an increasingly free flow of thinking and ideas between psychological systems. Art emerged, such as cave paintings and carved figurines, typically coupling technical prowess either to the biological domain, the social domain, or both. Weapons were increasingly made from antler, bone, wood, plant parts, and other biological materials, and sometimes decorated with carvings of animals: a two-way crossing of technical and natural history intelligence. Beads, pendants, and other personal adornments appeared, probably to mark status and group affiliation – an intersection of the social and technical domains. All of this is plainly symbolic, so linguistic intelligence was now no doubt tapping into non-social domains.

What adaptive advantages does Mithen think prompted this new inter-wiring and cross-domain communication between intelligences? He suggests that hunting prowess may have been enhanced by an ability to make anthropomorphic projections onto prey behaviors (a habit seen in all modern hunter-gatherers), and to communicate these to other members of a hunting party. He also suggests that the use of bone and antler in weapons supported an increasingly varied repertoire of specialized hunting techniques for diverse target species, while the linkage of natural history intelligence to technical intelligence permitted innovations for trapping game, catching fish, processing food, and storing it. He suggests that one function of art was to store valuable information about the seasonal migrations of game, and perhaps even to teach or transmit it to others. And all the while, by expanding beyond its original social functions, linguistic intelligence supported valuable communications in the technical and natural history domains. But above all, like Karmiloff-Smith, he seems to hold that horizontal integration conferred a general "cognitive flexibility" that is self-evidently adaptive, especially in the face of rapid climatic and ecological fluctuations during the Pleistocene.

"Individual Constructivist" Pedagogical Strategies in Science Education

Although the cognitive models put forth by Fodor, Carey, Pinker, Karmiloff-Smith, Mithen, and Gopnik and Meltzoff differ in many respects, they appear to share this core position: The cognitive systems that support conceptual change evolved because they enable *each individual* to improve the accuracy of the working theories with which he makes sense of his natural and social environments. Although infants enter the world "pre-wired" with baseline expectations about what the natural and social world will be like and how to interpret their experiences, natural selection nonetheless designed these initial frameworks to be malleable and modified in accord with empirical evidence, counterevidence, and even *counterintuitive* evidence. Evolution bestowed our ancestors with dedicated cognitive machinery for reworking their preexisting mental models of reality. The adaptive benefits, presumably, were an ever-improving *individual* mastery of one’s physical, biological, ecological, and psychological surroundings, and greater behavioral flexibility in the face of environmental variability.

In the field of science education, this “individual constructivist” origin is implicitly presupposed in many of the instructional strategies that scholars advocate for helping students undertake conceptual change. In this section I provide a snapshot of such pedagogical practices.

Dissonance strategies. Many instructional methods strive to provoke conceptual change by confronting the student with empirical evidence, counterevidence, and firsthand experience – often via laboratory investigations, vivid teacher demonstrations, or real scientific data – that does not neatly accord with his pre-instructional conceptions. They are “dissonance strategies” (Clement, 2008) designed to arouse dissatisfaction that compels the student to rationally reevaluate his working mental models of reality and move toward the scientifically normative position (Chinn & Brewer, 1993; Clement, 2008; Posner et al., 1982; Strike & Posner, 1992; Vosniadou, 2008b). For example, a common intuition is that the more massive an object is, the more rapidly it will accelerate toward Earth during freefall. To counteract this misconception, a physics teacher might place both a feather and a penny into a vacuum chamber, evacuate all the air, and invert it, vividly demonstrating that in the absence of air resistance they will fall side-by-side. She might preface the exercise by asking each student to jot down a prediction, and afterwards ask them to rationally interpret the surprising outcome.

Southerland et al. (2007) dub this approach “Piagetian,” after Jean Piaget’s influential *empiricist* theory of cognitive development, which stressed that young children primarily construct new knowledge through individual empirical experience.

This constitutes the “classical approach” (Vosniadou, 2008) to conceptual change that prevailed for decades in the wake of seminal papers by Posner and Strike (Posner et al., 1982; Strike & Posner, 1992). They married Piagetian psychology to Thomas Kuhn’s (1962/1970) famous model of theory change during the history of natural science: Kuhn’s “normal science” and “paradigm shifts” correspond to Piaget’s “assimilation” and “accommodation” respectively. They argued that a science student will undergo conceptual change only if (a) empirical evidence or experience convinces her that her current working theory cannot adequately explain certain phenomena; and (b) only if some new (scientific) theory *can* plausibly explain those same phenomena; and crucially, (c) only if she finds that novel theory *intelligible*. (The kinship to Gopnik and Meltzoff’s model should be clear.) Frequently, they say, the third condition is the most formidable obstacle to conceptual change: The counterintuitive nature of many scientific constructs clashes with the student’s own powerful intuitions. Many researchers agree that passing through a phase of such “cognitive conflict” (Vosniadou, 2008) has a useful role to play in effecting conceptual change (Chinn & Brewer, 1993; Clement, 2008; Hatano & Inagaki, 2003; Posner et al., 1982; Southerland et al., 2007; Strike & Posner, 1992; Thagard, 1992; Vosniadou & Brewer, 1987).

“Dissonance” versus “constructivism.” Other educational scholars, though, have challenged the cognitive conflict strategy on the grounds that it flouts constructivist learning theory, which holds that a student can advance his understanding only by building upon prior conceptions, many of which *are* useful and valid even if others are misleading (diSessa, 2008; Smith et al. 1993). For this reason, many researchers also advocate the use of analogies, metaphor, modeling, thought experiments, mental simulations, and transfers from the familiar, all as strategies that fruitfully build upon a student's existing ideas and intuitions (e.g., Carey, 2009; Carey & Spelke, 1994; Clement, 2008; Nersessian, 2008; Pinker, 2007; Vosniadiou & Brewer, 1987; Vosniadou et al., 2008). To counter the freefall misconception, for example, a physics teacher might invite her students to logically contemplate a "thought experiment" like the one that helped Galileo himself reason his way to the correct scientific truth (Ferris 1988/2003): Imagine a pair of heavy steel spheres connected by a taut steel wire, dropped side-by-side from a tall tower and falling together as a single object. And imagine that during the descent the wire snaps, splitting the single object into two separate spheres of half the original weight. Will they now suddenly decelerate? If one sphere is heavier than the other, will it now begin to fall faster than the other, whereas before they fell together? The point of such exercises is to recruit the student's prior concepts and knowledge to compel him to abandon his flawed intuition in favor of a correct scientific principle.

Other pedagogies of this ilk are available. Nersessian (2008) advocates "model-based reasoning" – as opposed to linguistic, proposition-based logic – as the key to conceptual change, and she describes a toolkit of learning processes that employ analogical, visual, and spatial models or simulations. Clement (2008) has developed "bridging analogies," strategically sequenced analogies designed to move students incrementally toward the target concept. He also proposes a "model evolution" pedagogy that uses *both* cognitive conflict and constructivist analogies in tandem: The student advances from his pre-instructional conception to the final scientific model by stepping through a sequence of intermediate models, with each transition mediated by (a) a dissonance-inducing observation or thought experiment that challenges the previous model in the chain, plus (b) a familiar analogy that fosters construction of the next model in the chain. The literature is rich in other applications of a kindred spirit.

"Warm" versus "cold" conceptual change. Finally, some scholars have objected that the strategies above promote "a *cold*, or overly rational, model of conceptual change that focuses only on student cognition without considering the ways in which students' motivational beliefs about themselves as learners…can facilitate or hinder conceptual change" (Pintrich, Marx, & Boyle, 1993, p. 167; also Dole & Sinatra, 1998; Sinatra & Dole, 1998; Sinatra & Pintrich, 2003). A Darwinian perspective warrants such a position: The traditional partitioning of affective motivations from cognitive processes is flawed, for motivational mechanisms surely evolved to work hand-in-hand with information-processing, knowledge-seeking, and problem-solving mechanisms (Geary, 2002, 2007; Kaplan, 1992). Advocates of "warmer" pedagogies stress that teachers should encourage individual goal-setting, self-efficacy, personal interest, reflective valuation, metacognitive awareness, self-monitoring of progress, and conscious self-regulation (Bransford et al., 1999; Dole & Sinatra, 1998; Sinatra & Dole, 1998; Sinatra & Pintrich, 2003). Such practices complement the "colder," more rationalistic approaches to conceptual change, and thus also embrace the spirit of *individual constructivist* pedagogy.

Summary, Thesis #5: "Individual Constructivist" Models

In this section I showed that the human capacity for conceptual change is itself a cognitive ability that demands evolutionary explanation. I developed two Darwinian riddles: (1) an "evolvability" riddle – How *could* natural selection create psychological mechanisms for disregarding, discarding, or restructuring the more ancient and manifestly useful intuitions that natural selection had already built into the hominin brain? And (2) the riddle of "horizontal integration" – What evolutionary benefit prompted the lateral integration of previously isolated cognitive sub-systems in our species (which all researchers deem crucial for conceptual change)? I then surveyed a handful of answers to these riddles, proposed by prominent cognitive scientists with a Darwinian orientation. Though their explanations vary in many respects, they share an *individualist, empiricist,* and *constructivist* stance: Conceptual change is made possible by psychological mechanisms that evolved to permit each person – upon encountering new evidence and counterevidence during the course of real world experience – to expand and improve the precision of her working theories about the physical, ecological, and social environments. The apparent adaptive benefit was greater cognitive/behavioral flexibility for coping with new and changing habitats during the ever-fluctuating Pleistocene epoch. This flexibility was enhanced by a growing ability to integrate mental representations and concepts across once disconnected domains, and – especially with the advent of spoken language – by the evolution of an increasingly metaphorical mind, capable of mapping analogous, useful relationships from one conceptual scheme onto another.

Finally, I showed that such an "individual constructivist" origin is implicitly assumed in many of the pedagogical strategies that education researchers advocate for cultivating conceptual change in the modern science classroom. A prevailing practice is to confront students with evidence, counterevidence, and surprising experiences that challenge her pre-instructional intuitions, thereby compelling her to rationally rework her mental models of reality. Other instructional approaches employ analogies, metaphor, modeling, and thought experiments to help the student rationally construct a scientifically normative understanding of natural phenomena. Like the evolutionary models that tacitly inhabit them, these pedagogies are *individualist, empiricist,* and *constructivist,* and as such may be regarded as "Piagetian" approaches to provoking conceptual change. A wealth of research documents their efficacy, which by extension supports the underlying evolutionary hypothesis. Nevertheless, our cognitive capacity for conceptual change may also have special origins in the evolution of our *social* nature. It was this possibility – which I explore under the next two theses – which prompted the development of the experimental instrument that was tested in this doctoral study.

Thesis #6 – The Evolutionary Origin of Conceptual Change: "Cultural Re-Constructivist" Hypotheses

> *Thinking would seem to be a completely solitary activity. And so it is for other animal species. But for humans, thinking is like a jazz musician improvising a novel riff in the privacy of his own room. It is a solitary activity all right, but on an instrument made by others….after years of playing with and learning from other practitioners, in a musical genre with a rich history of legendary riffs....Human thinking is individual improvisation enmeshed in a sociocultural matrix. Individuals mediate their interactions with the world through the culture's artifacts and symbols from early in ontogeny (Vygotsky, 1978…), thus absorbing something of the wisdom of the entire cultural group and its history.*
>
> Michael Tomasello (2014, p. 1 and p. 142)

In this section I discuss a second candidate solution to the dual riddles of conceptual change: the "cultural re-constructivist" hypothesis. According to this explanation, the human mind became evolutionarily adapted for conceptual change in order to reap the full benefits of culturally accumulating technologies, knowledge, and knowhow. On this view, conceptual change evolved as one part of the psychological architecture that supports cultural transmission in our species. The ability to acquire culture from others is so central to our lives, yet seems so divorced from our more "biological" drives and behaviors, that it is easy to overlook the fact that it is a species-specific adaptation – that is, a product of biological evolution (Richerson & Boyd, 2005; Tomasello, 2014). It is made possible by a suite of evolved, species-specific cognitive mechanisms, such as the ability to adopt the perspective of others, read their intentions, and faithfully imitate their behaviors (Tomasello et al., 1993). As with "learning" versus "instincts," it is a bankrupt dichotomy to oppose "culture" to "nature" (or nature to nurture): We are *naturally cultural*; culturality was part of our ancestors' ecological niche (Richerson & Boyd, 2005; Tomasello et al., 1993; Tomasello, 1999, 2014; Tooby & Cosmides, 1992).

Just as a capacity for transmitting tool technology across generations bestowed hominins with an ecological flexibility that was both individually and collectively adaptive, so – according to this hypothesis – did a capacity for acquiring *novel mental models* via cultural transmission bestow hominins with a *cognitive* flexibility that was both individually and collectively adaptive (Richerson & Boyd, 2005; Tomasello, 1999, 2014). In this account, concepts are essentially "tools" for thinking and communicating, and as with handheld tools, there were Darwinian payoffs for those who could manufacture or master new ones, *even non-intuitive ones*. Those who excelled in this regard reaped the greatest benefits from accumulating knowledge and knowhow, and so enjoyed superior reproductive fitness. And as for the riddle of "horizontal integration," the cultural re-constructivist hypothesis might propose that this ability was adaptive because it enabled our ancestors to combine and recombine ideas from different domains in order to foster novel insights that enhanced tool technology, hunting strategies, and sociocultural practices (Mithen, 1996; for a complementary answer, see discussion below of Tomasello, 2014).

Here again, I will argue that such an evolutionary origin is tacitly presupposed in many of the pedagogical strategies now used to incite conceptual change in science education. If *individual constructivist* pedagogies may be dubbed "Piagetian," then *cultural re-constructivist* pedagogies might be dubbed "Vygotskyan" (Southerland et al., 2007), for they owe much to Lev Vygotsky's influential theory of childhood learning and development. Vygotsky (1978) metaphorically described culturally evolved concepts as "tools for talking and thinking," whose usage children "internalize" with the help of adult experts. Whatever a child's current level of development, Vygotsky observed, she can exercise novel concepts more proficiently with adult assistance than she ever could by herself. Adult teachers – as agents of cultural transmission – can steer and step a student along the road of internalization, challenging her at each moment just a bit beyond her current level of developmental readiness. The "just-a-bit-beyond" window within which this process occurs is Vygotsky's famous "zone of proximal development." Thus in an important sense, he contends, internalizing new concepts actually *precedes* the child's cognitive development. New conceptual tools are mastered, and development *follows*.

This is quite the inverse of Piaget, who held, for example, that a child can acquire abstract concepts only *after* she matures into the formal operational stage of cognitive development.

Cognitively, the progression is not a passive assimilation of cultural concepts but an *active reconstruction* of them. Dynamic dialogue between child and adult is central to the process, often augmented by discourse among peers, and supported by the use of external "signs": linguistic and non-linguistic *symbols* and *artifacts* (Vygotsky, 1978). Thus children are born into a system of spoken language, symbols, tools, and other artifacts, which "preorganize their worlds for them" (Tomasello, p. 1) and ultimately govern how they will come to conceptualize reality. Modern scientific ideas may be part of that system. In that case, scientific conceptual change amounts to mastery of a formerly non-intuitive *conceptual technology*.

In this section I will use Michael Tomasello's (1999, 2014) neo-Vygotskyan theory of human cognitive evolution to illustrate the cultural re-constructivist hypothesis of how our capacity for conceptual change may have evolved. And I will again survey corresponding pedagogical practices in science education. *I develop the following thesis:*

> An ability to adopt ideas that violate or subvert innate intuitions may have been ancestrally adaptive *not* because it gave individuals the cognitive flexibility to rationally improve the predictive accuracy of their conceptual models (= the individual constructivist hypothesis), but instead because it enabled them to *become* cognitively flexible by actively assimilating the incremental improvements in conceptual models that had accumulated in their culture over

previous generations. In practice, many of the pedagogical strategies advocated by science education scholars implicitly assume such an evolutionary origin. Along with the individual constructivist hypothesis, the cultural re-constructivist hypothesis forms the essential backdrop against which I conceived my "social strategic" hypothesis of conceptual change (Thesis #7), which in turn inspired and informed the design of the experimental instrument that I developed for this doctoral study.

A "Cultural Re-Constructivist" Model of the Origin of Conceptual Change: Tomasello's Natural History of Human Thinking

Under Thesis #5 I described how a half dozen evolution-minded cognitive scientists tackled the twin riddles of conceptual change: (1) What adaptive benefits made it "evolvable"? and (2) What adaptive benefits favored the "horizontal integration" of once isolated conceptual systems? Their answers, though diverse, all more-or-less represented an individual constructivist orientation. For Thesis #6 I will instead focus on only one scientist whose prolific work especially embodies a cultural re-constructivist (and Vygotskyan) orientation: the primatologist, developmental psychologist, and psycholinguist Michael Tomasello. His approach is thoroughly Darwinian. In a pair of books of astonishing scope and scholarship – *The Cultural Origins of Human Cognition* (1999) and *A Natural History of Human Thinking* (2014) – he spins a 7-million-year tale of cognitive evolution in the hominin line. To do this, he draws upon the anthropological record, his research on language acquisition during childhood, and his own and others' comparative studies of learning, cognition, and social interactions in apes versus human children. Although he never discusses conceptual change directly, and though there are other prominent Darwinian scholars who also arguably manifest a cultural re-constructivist stance – including Dennett (1995), Dunbar and colleagues (Gamble, Gowlett, & Dunbar, 2014), Richerson and Boyd (2005), and Sterelny (2014) – Tomasello offers the most compelling Vygotskyan account of how natural selection might have produced a mind ready to adopt ideas crafted by cultural evolution, even when they clash with biologically evolved intuition.

Tomasello's (1999, 2014) overall argument is that ever increasing brain size, intelligence, and cognitive flexibility in the hominin line – and even *individual reasoning itself* – was propelled mainly by their *social* evolution. More specifically, our cognitive evolution was driven by progressively more sophisticated *cooperation* in hominin societies. (Robin Dunbar and colleagues [Gamble et al., 2014] and Kim Sterelny [2012] stake out similar positions.) Tomasello presents a long phylogenetic sequence of how these changes might have taken place, and I will summarize his "natural history of human thinking" below, especially as it sheds light on possible evolutionary origins of conceptual change. To anticipate, he argues that *conceptual reasoning itself* first evolved for a specific social function: to "conceptualize situations for others…in cooperative and conventional communication" (2014, p. 135). Note the inversion: It is not just that we use language as a convenient vehicle for conveying our ideas; rather, we think conceptually and formulate ideas in the first place precisely for the purpose of being able to convey them linguistically to others. *We conceptualize in order to communicate*. Conversely, our minds are specifically adapted to "internalize" the conceptual communications of others within the sociocultural group. *And this (presumably) means that the mind is ready, by evolutionary design, to re-construct whatever culturally* *evolved conceptions have historically accumulated in a given culture, however non-intuitive they might be*.

Cultural learning. In a seminal paper, based on research with young children and chimpanzees, Tomasello and colleagues (1993) describe a suite of mechanisms peculiar to humans which enable "cultural learning": *imitation*, *instruction*, and *collaboration*. All three depend upon an especially well-developed ability to read the perspectives of other individuals, and this key adaptation has made humans, but not other apes, capable of the rapid, high fidelity transmission needed for cumulative cultural evolution (see Thesis #4). These three vectors of cultural learning emerge in children as roughly sequential stages – *imitative* then *instructed* then *collaborative* learning – a progression that depends upon the development of an ever more adult-like "theory of mind" (see Thesis #2 above).

Imitative learning, in which the child observes and then attempts to reproduce an adult's actions, requires an ability to recognize another person's *goals* and *intents*, and involves simple, unidirectional perspective-taking by the observer.

Instructed learning requires an ability not only to grasp another person's goals and intents, but also her *beliefs* and *thoughts*, which may differ from one's own. It is not a case of the child merely following an adult's spoken instructions, but instead involves alternating, reciprocal perspective-taking between learner and instructor. Through observation and dialogue, the child willfully adopts the instructor's perspective, comparing and contrasting the adult's conceptual representations against her own. On the adult's end, the instructor repeatedly reads the child's perspective, especially at critical decision points in the task at hand, and coaches him accordingly. Perspective-taking is thus alternating, coordinated, and mediated by language. In this way the child comes to *internalize the dialogue*, and this now permits him to “self-regulate” his own actions. Later, for instance, during difficult portions of a learned task, he may coach himself out loud, thus reenacting the adult’s instructions and perspective. Tomasello explicitly equates this process with Vygotsky’s “internalization” of culture within the child’s “zone of proximal development.” And evolutionarily, as I will show below, he deems the internalizing of dialogue a key move that ultimately gave our species a capacity for conceptual change.

Finally, the third vector of cultural learning – collaborative learning – involves a tight, mutual perspective-taking among two or more individuals collaborating on a shared task or problem. Here their roles are symmetrical: Neither party is the authority or expert, neither is the novice, and they “co-construct” new knowledge as they go. This kind of learning, too, represents a crucial development in Tomasello’s “natural history of human thinking” – which I turn to next.

Cognitive transition #1: From apes to early humans. To some extent here, says Tomasello (2014), ontogeny recapitulates phylogeny: The childhood developmental sequence just described manifests key moments in the evolution of human cognition. He proposes that during the last half million years or so, there were two major transitions in the direction of heightened *cooperation* in hominin societies. The first coincides roughly with the transition from apes to “early humans,” the second with the transition from early humans to modern humans. Together, these two “cooperative turns” produced our capacity for conceptual change.

Modern chimpanzees – taken by Tomasello to roughly represent the last common ancestor (LCA) between apes and humans 7 million years ago – possess innate "cognitive skills…for dealing with space, objects, tools, quantities, categories, social relationships, communication, and social learning" (1999, p. 7). That is, they have innate intuitions for negotiating both their physical and social environments. Moreover, recent experiments have demonstrated that chimps do grasp that fellow apes have goals, perceptions, and vantage points of their own, and thus "understand others as intentional agents" (2014, p. 20; see earlier discussion under Thesis #2 of the "intentional stance" [Dennett, 1989, 1995]). Tomasello believes that apes have an ability to run mental "simulations" about events in the world, conducting "thought experiments" about potential consequences of different actions (see Dennett, 1995, for a plausible sequence of how this ability might have evolved). At first these *internal simulations* probably concerned purely physical phenomena, as apes are quite adept at using tools and otherwise manipulating the physical world based on a *causal* understanding of it. But mental simulations were eventually recruited to the social sphere as well, based instead on an *intentional* understanding of the perceptions and desires of others. Here, too, the function of thought experiments is manipulation: As Tomasello's and others' research shows, chimps can skillfully influence others' actions by manipulating what they do and do not see, anticipating their responses, and communicating (sometimes deceptively) through arm and hand gestures. Tomasello (2014) argues that apes' interpersonal intuitions evolved primarily for *competitive* success within the social group. They developed a "Machiavellian intelligence" (citing Whiten & Byrne, 1988), whose adaptive function is to outwit and outcompete others for survival, sexual, and social resources (see also de Waal 1996, 2006; Gamble et al., 2014).

But in the hominin line, says Tomasello (2014), these competition-driven skills – mental simulations, the intentional stance, gestural communication, etc. – became transformed for much more *cooperative* and *collaborative* ends. His evidence for this comes partly from comparative studies of apes versus human children. Very young children – but not apes – adopt and pursue joint goals:

> From soon after their first birthdays, and continuing up to their third birthdays, they come to engage with others in collaborative activities that have a species-unique structure and that do not, in any obvious way, depend on cultural conventions or language. These young children coordinate a joint goal, commit themselves to that joint goal until all get their reward, expect others to be similarly committed to the joint goal, divide the common spoils of a collaboration equally, take leave when breaking a commitment, understand their own and the partner's role in the joint activity, and even help the partner in her role when necessary. (p. 41)

Phylogenetically, argues Tomasello, such evidently innate behaviors bespeak an ancestral transition from the ape stage of "individual intentionality" to the early human stage of "joint intentionality," as hominins began to form ad hoc dyadic or small group partnerships for collaborative foraging and hunting (presumably prompted by a shift in "feeding ecology").

Now, says Tomasello, social *cooperation* rather than competition began to drive our cognitive evolution. Most important for the question of conceptual change, hominins on a joint hunt or foraging expedition started to communicate ever more dynamically through pointing and pantomiming. For example, a partner spotting a snake in the underbrush might point and make a slithering hand motion, or while reading a gazelle's tracks might splay his fingers like antlers above his forehead and motion toward the prey's predicted flight path. They did not yet have spoken language, but as early humans became increasingly dependent on collaborative foraging, there was a strong selective pressure to become honest, helpful, lucid communicators. Crucially, says Tomasello, the ancestral ape's capacity for running mental simulations now turned to the task of formulating complex-yet-comprehensible gestures for others. Doing so required an evolving ability to imagine the communications from the partner's perspective. Herein lay the seeds of *internal dialogue* that would emerge during a second hominin transition and set the stage for conceptual change.

Cognitive transition #2: From early to modern humans. Tomasello's second "cooperative turn" was a transition from ad hoc, small group collaborations to cohesive societies marked by distinctive cultural conventions. Provoked by an intensifying demographic threat from *other* populations, each human social group became increasingly interdependent for foraging and even fighting with other groups. The more cohesive and internally cooperative a society, the better it fared in competition with other populations ("cultural group selection"):

> Modern humans became cultural beings by identifying with their specific cultural group and creating with groupmates various kinds of cultural conventions, norms, and institutions built not on personal but on cultural common ground….[They now] actively conform to the behavior and norms of the group, and even enforce

> conformity on others through teaching and social norm enforcement. (2014, p. 80)

Cognitively, this was a transition from the stage of “joint intentionality” to the stage of “collective intentionality”: Psychological mechanisms, motivations, and emotions that originally evolved for mutually profitable ad hoc partnerships now got “scaled up” for life in bona fide cultural systems. Cultural transmission via imitative, instructed, and collaborative learning (see description above) now became fully realized:

> Modern human culture…is fundamentally cooperative, as adults actively teach children, altruistically, and children actively conform to adults.…Teaching borrows its basic structure from cooperative communication in which we inform others of things helpfully, and conformity is imitation fortified by a desire to coordinate with the normative expectations of the group. Modern humans did not start from scratch but started from early human cooperation. (p. 82)

Most important for the issue of conceptual change was the way communication changed. Spontaneous pantomiming became conventionalized, such that all members of a social group came to use the same signals in the same situations. This freed their gestures to become shortened, stylized, and ultimately “arbitrary”: They no longer had to visually mimic their referents, as in the snake and slithering hand. Arbitrary signs had the advantage that they could signify much more abstract concepts than imagistic pantomiming, as well as complex combinations of concepts. Eventually they were supplanted by vocalizations (though how this may have happened, Tomasello does not say). Crucially, the advent of arbitrary, spoken signs now meant that an individual could no longer guess the meaning of signs – as she might with concrete imagistic gestures –

but instead had to learn the entire local symbolic system. *And this, claims Tomasello, not only meant that children inherited the local language, but also local ways of conceptualizing reality:*

> Children were now born into a group of people using a set of communicative conventions that their ancestors had previously found useful in coordinating their referential acts, and everyone was expected to acquire and use exactly these conventions. Individuals did not have to invent their own ways of conceptualizing things; they just had to learn those of others, which embodied, as it were, the entire collective intelligence of the entire cultural group over much historical time. Individuals thus “inherited” myriad ways of conceptualizing and perspectivizing the world for others. (2014, p. 96)

The special talents that early humans had evolved for collaborating with partners on joint tasks – namely, a highly developed intentional stance and an ability to assume one another’s perspective during dynamic communications – now permitted individuals to think in *genuinely novel*, perhaps even *non-intuitive*, ways:

> I would claim that the process of acquiring and using linguistic symbols fundamentally transforms the nature of human cognitive representation….Many researchers do not believe that acquiring a language has any great effect on the nature of cognitive representation because they view linguistic symbols as simply handy tags for already formulated concepts (e.g., Piaget, 1970)….The intersubjectivity of human linguistic symbols…means that linguistic symbols do not represent the world…directly…but rather are used by people to *induce others to construe certain perceptual/conceptual situations…in one way rather than in*

> *another* [emphasis added]….What I want to claim is that participation in these communicative exchanges is internalized by the child in something like the way Vygotsky envisioned it. Internalization is not a mystical process, as some envision it, but merely the normal process of imitative learning as it takes place in this special intersubjective situation: I learn to use the symbolic means that other persons have used to share attention with one another. In imitatively learning a linguistic symbol from other persons in this way, I internalize not only their communicative intention…but also the specific perspective they have taken….They give children *truly new ways of conceptualizing things* [emphasis added]. (1999, pp. 123-124 and 128-129)

In effect, through language acquisition the child comes to adopt the perspectives, hence the concepts, of cultural predecessors long past.

The birth of rational thinking and conceptual change. With Vygotsky, then, Tomasello regards concepts as "tools for talking and thinking" (Vygotsky, 1978), and language as a *conceptual technology* that is *culturally re-constructed* – or "internalized" – by each individual. Yet he goes even further. The mental simulations and thought experiments of our ape and early human ancestors, and the self-regulating speech seen in young apprentices practicing a new skill under the wing of an adult instructor, now become *imagined discourses* and *"inner dialogues"* spoken in the language and concepts of the local culture. *And this, literally, is the origin of conceptual reasoning itself:*

> When a communicator informs a recipient of something, she wants to be believed...But sometimes, there is not enough trust on the recipient's part…and so the communicator gives reasons for her informative statement. In reason-giving

> discourse of this kind, individuals are attempting to convince others….Humans evolved reasoning abilities not for getting at the truth but for convincing others of their views. The proposal that human reasoning, including individual reasoning, has a social-communicative origin is almost certainly correct. (2014, p. 110)

> The capstone of all of this – recognized by all modern thinkers who take a sociocultural view human thinking – is the internalization of these various interpersonal processes of making things explicit into individual rational thinking or reasoning. Making things explicit to facilitate the comprehension of the recipient leads the communicator to simulate…how his planned communicative act might be comprehended – perhaps in a kind of inner dialogue. Making things explicit to persuade someone leads the disputant to simulate ahead of time how a potential opponent might counter his argument, and so to make ready, in thought, an interconnected set of reasons and justifications – again perhaps, in a kind of inner dialogue. (2014, p. 112)

Herein lies Tomasello's apparent (but unspoken) answer to our twin riddles of conceptual change. We conceptualize in order to communicate. We reason to convey reasons. And since this all evolved to facilitate *cooperation*, evolution also made us receptive to the reasons, reasoning, and conceptual communications of others, no matter how non-intuitive and no matter which of life's domains they concern. Cooperative communication and linguistically mediated cultural transmission are what concepts themselves are *for*. Our cognitive capacity for developing new concepts – and for integrating them "horizontally" across domains – evolved as part of our "cultural nature." In this Vygotskyan model, the essence of conceptual change is individual re-construction of multiple perspectives, embodying a long history of culturally evolved ideas, explanations, analogies, metaphors, and entire conceptual systems – including even the loftiest abstractions and counterintuitive theories of natural science system.

“Cultural Re-Constructivist” Pedagogical Strategies in Science Education

Tomasello’s Darwinian, neo-Vygotskyan model of cultural learning via the triple vectors of imitative, instructed, and collaborative learning, and of conceptual change through external and internal dialogue, finds implicit expression in many instructional strategies endorsed within the science education community. I now outline some of these pedagogical practices.

Direct instruction coupled with independent practice. In cultural re-constructivist pedagogy, an adult teacher – as expert agent of scientific culture – must be the main vector for supplying students with new and non-intuitive conceptual “tools” for making sense of natural phenomena (Geary, 2002, 2007; Leach & Scott, 2008). Even so, this pedagogy should not be equated with “direct instruction” of the linear, unidirectional ilk where students are but passive recipients of new knowledge: The teacher tells, the professor professes, the expert explains. Students are not blank slates, and they are no more likely to have their intuitive conceptions transformed into counterintuitive scientific ones through such methods, as they are to master a difficult new tool, like a guitar, simply by having the process described to them by a seasoned veteran. While direct instruction plays an important role, students will master novel concepts only by actively employing them in teacher-specified tasks, using them to interpret concrete scenarios, or participating in teacher-led question-and-answer exchanges. During practice sessions, the teacher typically moves amongst them, scaffolding and coaching as needed in a back-and-forth dynamic. In this way they not only *construct* the new culturally normative understanding, but actively *reconstruct* it (Hatano & Inagaki, 2003; Leach & Scott, 2008; Sperber, 1994, 1996; Tomasello, 1999; Tooby & Cosmides, 1992; Vygotsky, 1978).

Thus, to return to our earlier example of acceleration during freefall, a physics teacher might begin by explaining freefall in terms of Newton's 2nd law of motion, $F = m a$. She might rearrange this equation to $a = F / m$ to highlight the independence of acceleration from mass: Double mass in the denominator, and you also double the force of gravity in the numerator (i.e., the object's weight), such that acceleration on the left remains constant. All objects therefore fall at the same rate, regardless of mass. The teacher will then oblige the students to practice with this new principle on their own, applying it to novel conceptual scenarios such as pendulum motion, and using it to solve various computational problems. During this time she will move among them, coaching as needed. Little by little, the novices master a new conceptual technology supplied by modern science, initially counterintuitive though it was.

Placeholders and scaffolding. In her role as expert agent of scientific culture, a teacher must often supply students with symbolic "placeholders" (Carey, 2009; Hatano & Inagaki, 2003), such as "F = m a" above. Typically these are tight linguistic definitions, formulations, expressions, and/or labels for the target concept – akin to Vygotsky's (1978) external "signs" – whose meaning the teacher grasps but which her students may not yet be ready to understand. Still novices, they do not yet own the new concept and cannot think with it habitually; they cannot yet wield the new thinking tool with the proficiency of an expert (Bransford et al., 1999). In the case of especially remote placeholders where more dramatic, difficult leaps are required – such as Carey's (2009) conceptual “bootstrapping” and Chi’s (1992, 2008) “categorical shifts” – neither definitions nor explanations alone will suffice to bring about conceptual change (Carey, 2009; Chi, 1992, 2008). Nevertheless, the placeholder serves as a beacon toward which a student can steer. Once he successfully undergoes conceptual change, the definitions at last become fully meaningful, as the placeholder crystallizes into an authentic scientific conception.

Some placeholders can function as intermediary “conceptual pegs” or “cognitive scaffolds” (Hatano & Inagaki, 2003), conceptually accessible waystations deliberately placed within a student’s “zone of proximal development” (Vygotsky, 1978) en route to the final target understanding. This belongs to the more general Vygotskyan strategy of “scaffolding” (Joyce, Weil, & Calhoun, 2009): In her capacity as scientific expert and vector of cultural transmission, the teacher steers her students in stepwise fashion through the conceptual change process, dynamically assessing and adjusting as needed along the way, and supplying them with a wide variety of instructional rungs to assist them in their climb toward an increasingly formal understanding of the concept at hand. Of course, these scaffolds may include “Piagetian” moves like analogical transfer and reasoning about empirical data. In practice, instruction for scientific conceptual change almost always involves a blend of individual constructivist and cultural re-constructivist pedagogies.

“Acquisition” versus “participation”: Inquiry, discourse, and argumentation. Many learning theorists have declared an allegiance to Vygotsky as a corrective to the excessively individualistic nature of the classical Piagetian approach (Southerland et al., 2007; Vosniadou, 2008). These scholars universally advocate *classroom discourse,* *group interactions*, *interpersonal relations*, and *social motivations* as essential to the conceptual change process (e.g., Hatano & Inagaki, 2003; Kelly & Green, 1998; Leach & Scott, 2008; Miyake, 2008; Mortimer & Scott, 2003). Yet even within the Vygotskyan camp, important differences exist. Leach and Scott (2008) distinguish "acquisition" models of conceptual change from "participation" models. The former emphasize the cultural transmission/acquisition of scientifically normative concepts, while the latter downplay this function and instead stress participation in social activities through which the learner becomes socialized into the scientific community and culture (e.g., Kelly & Green, 1998).

The distinction is well captured in the prominent practice of inquiry-driven instruction. Inquiry can be roughly defined as student participation in science-like activity that is centered on an investigative question and in which students draw conclusions based on empirical data, with students themselves often designing experiments and collecting the data (Southerland et al., 2007). Advocates of inquiry see it as a way to enter the "culture" of science through immersion in authentic scientific activity: The student learns scientific practices and processes, while developing proficiency in scientific thinking skills (Leach & Scott, 2008; Southerland et al., 2007). Some science education researchers further contend that inquiry can be useful for inducing *conceptual change* (e.g., Pintrich et al., 1993). Yet Leach and Scott (2008) dispute the efficacy of inquiry where conceptual change is the goal:

> If the aim of a sequence of teaching is to introduce aspects of scientific conceptual knowledge, then some form of clear and direct guidance by the teacher is essential. The scientific knowledge itself is authoritative in nature and some form

> of authoritative intervention by the teacher is therefore needed to introduce it to the social plane of the classroom. At the same time, of course, students should be given every opportunity to apply that knowledge as they talk and use it for themselves. From this point of view we believe that there are real limitations to using inquiry-based, participation driven approaches to teaching scientific conceptual knowledge. (p. 658)

Although *participating* in scientific inquiry can help students break into scientific culture and learn about the nature of science, then, it may not be effective in helping them *acquire* its more challenging, counterintuitive concepts. Once again, Leach and Scott would say, the teacher must channel those concepts, and students must practice independently with them.

Similarly, the past decade has seen a swelling emphasis in science education on student participation in scientific "discourse and argumentation":

> If science education is to help young people engage with the claims produced by science-in-the-making, science education must give access to these forms of argument through promoting appropriate classroom activities and their associated discursive practices. Such practices...are the means of socializing young people into the norms of scientific argument. (Driver, Newton, & Osborne, 2000, p. 288)

Advocates of argumentation stress evidence-based reasoning, public stance-taking, the pitting of competing explanations against one another and against empirical data, and above all the "social construction" of scientific knowledge through dialogue (Driver et al., 2000; Duschl, 2008; Kuhn, 1993, 2010; Nussbaum, 2008). The explicit goal is to "enculturate" students into the scientific community (Driver et al., 2000; Southerland et al., 2007), and the overwhelming focus has been on process and the “nature of science” rather than core content (Nussbaum, 2008). Yet it is one thing to grasp the *culture of science* – its norms, customs, methods, and rules of reasoning – and another to master its *culturally evolved concepts*, including the counterintuitive (Leach & Scott, 2008). Although Vygotskyan *participation* may suffice for the former, the latter may require Vygotskyan *acquisition* in the form of expert-to-novice transfer coupled with active student reconstruction.

On the other hand, perhaps there is a role for discourse and argumentation after all. Miyake (2008) cites evidence that “collaborative reflection” – a dialogue between two or more students tackling a shared task – can foster conceptual change. This is especially likely, she says, when students offer verbal explanations and make their ideas explicit; exchange metaphors, perspectives, and suggestions that provoke one another to rethink; test their ideas against each other and against observations; and adjust or repair them in response. Whereas expert-to-novice transfer plainly embodies Tomasello’s *instructed learning,* this pedagogy manifests his *collaborative learning* (Tomasello et al., 1993). Still other pedagogies scale these small-group exchanges up to the whole-class level (e.g., Hatano & Inagaki, 2003; Leach & Scott, 2008; Mortimer & Scott, 2003), with the teacher staging the classroom as a forum where students openly share and evolve their ideas. The teacher expects divergent views and patiently accepts them, but ultimately steers the students toward (re)construction of normative scientific concepts.

Nevertheless, I will suggest under Thesis #7 that some elements within these self-professed “Vygotskyan” instructional strategies correspond more closely to a rather different, perhaps more ancient, evolutionary origin of our capacity for conceptual change. I will hypothesize that discourse and argumentation, collaborative reflection, and whole-class dialogue – to the extent that they can foster scientific conceptual change – owe their effectiveness not solely to our *cooperative* and *cultural* nature, but crucially to our *competitive* and *"socially strategic"* nature as well. Such pedagogies, then, would be neither Piagetian nor strictly Vygotskyan, but partly an alternative to both.

Summary, Thesis #6: "Cultural Re-Constructivist" Models

In this section I showed that Tomasello's (1999, 2014) neo-Vygotskyan account of the evolution of human cognition constitutes a powerful, plausible alternative to more individualistic hypotheses about the origin of our capacity for conceptual change. According to his "natural history of human thinking," hominins inherited from their ape ancestors an ability to discern the desires, goals, and perspectives of fellow members of the social group (the "intentional stance"). They also inherited an ability to carry out "mental simulations" of events in the real world, wherein an individual imagines various outcomes of alternative actions both physical and social. In early humans, however, these abilities were redirected from competitive ends to cooperative ends. Individuals began to form ad hoc partnerships for foraging, hunting, and other tasks, and so their ability to run mental simulations and read each other's perspectives turned to the function of helpful, effective communication through gestural pointing and pantomiming. This paved the way for spoken language to evolve, and when it did, mental simulations morphed into imagined "inner dialogues." This marked the birth of conceptual reasoning itself: We conceptualize in order to be able to communicate well. Conversely, our minds became adapted to "internalize" the conceptual communications of others within the sociocultural group. Meanwhile, the ability to read one another's perspectives also launched “cultural learning” via imitation and instruction, and so set cumulative cultural evolution in motion. As a result, the human mind was made ready, by evolutionary design, to be receptive to and re-construct entire culturally evolved conceptual systems, no matter how non-intuitive. On this account, concepts are “tools” for talking and thinking, and conceptual change is tantamount to mastering a challenging new conceptual “technology.”

I then showed that such a “cultural re-constructivist” origin is implicitly assumed in many of the pedagogical strategies that education researchers advocate for cultivating conceptual change in the modern science classroom. A prevailing practice is direct instruction coupled with independent student practice: The teacher, in his capacity as scientific expert and agent of cultural transmission, channels the target concepts to students in the form of explanations, analogies, and linguistic or symbolic “placeholders.” Because these will not yet be fully meaningful to the students, he will oblige them to work with and think with the new conceptual tools in novel scenarios, while he coaches and scaffolds from the sidelines. Other explicitly Vygotskyan pedagogies engage students in small-group or whole-class discourse, thereby embracing the collaborative and dialogical core of Tomasello’s cultural re-constructivist model.

Under the next thesis, however, I will explore a third hypothetical origin of our capacity for conceptual change – one likewise rooted in our evolution as a social species, yet different in key respects. Its corresponding pedagogies are neither Piagetian nor strictly Vygotskyan, but partly an alternative to both. It was this last model of conceptual change – in explicit contrast to the first two – that inspired me to create my experimental instrument and informed its design.

Thesis #7 – Conceptual Change as Evolutionary Byproduct:

An Alternative "Social Strategic" Hypothesis

Social intelligence developed initially to cope with problems of inter-personal relationships…And it is, I believe, essentially the same intelligence which has created the systems of philosophical and scientific thought which have flowered in advanced civilisations in the last four thousand years.

Sir Nicholas Humphrey (1976, p. 312)

If deceit is fundamental to animal communication, then there must be strong selection to spot deception and this ought, in turn, to select for a degree of self-deception, rendering some facts and motives unconscious so as not to betray – by the subtle signs of self-knowledge – the deception being practiced. Thus, the conventional view that natural selection favors nervous systems which produce ever more accurate images of the world must be a very naïve view of mental evolution.

Robert Trivers (1976, p. xxvi)

In this section I discuss a third and final candidate solution to the dual riddles of conceptual change: a "social strategic" hypothesis. Unlike the previous two hypotheses, this one suggests that our capacity for conceptual change is not strictly speaking an evolutionary adaptation at all. Rather, it is an accidental "byproduct" (Williams, 1966) of psychological adaptations that evolved for quite a different function. That function was to strategically navigate the shifty social terrain of ancestral ape and hominin societies – playing what Nicholas Humphrey (1976) calls social "chess," engaging in what Franz de Waal (1996) calls primate "politics," and coping with what Robin Dunbar and colleagues (Gamble et al., 2014) call "the soap opera of daily life." For many millions of years, primate society has been a crucible of complex cooperative and competitive interactions, where individuals sometimes support each other, sometimes strive to outwit and outmaneuver each other, sometimes pursue joint goals, and sometimes manipulate and deceive each other (de Waal, 1996, 2005; Gamble et al., 2014; Humphrey, 1976, 1986; Sterelny, 2012; Tomasello, 2014).

Once hominins evolved spoken language, individuals could “represent reality” verbally to one another, and sometimes there would have been rewards for publicly portraying their own and others’ words, deeds, talents, and intentions in less than accurate ways. They may have evolved a tendency (whether conscious or unconscious) to slant truth claims in fitness-enhancing directions – that is, to *misrepresent* reality in the social sphere in a personally beneficial light (Kurzban, 2010; Kurzban & Aktipis, 2007; Sperber, 1996; Trivers, 2002). That in turn might have prompted the evolution of psychological counter-adaptations for *detecting* such misrepresentations (Trivers, 1985, 2000): a skeptical wariness, a thirst for corroborating evidence, and crucially, a special interest in counterclaims and alternative interpretations. Somewhat paradoxically, a critical eye toward the truth claims of others may go hand-in-hand with a receptive ear toward those same claims. It implies *openness to rival or alternative conceptions of reality*, including an ability to *suspend one’s own intuitions* long enough to weigh rival claims, scrutinize them against available evidence, and perhaps adjust one’s own interpretations accordingly. That – according to my social strategic hypothesis – bestowed our ancestors with psychological machinery capable of conceptual change.

However, in the ancestral environment the truth claims in question would have concerned strictly *social* affairs. It may have been only very recently, even after the rise of agriculture and civilization, that this psychological machinery was recruited to *non-social* phenomena in the intuitive domains of physics and ecology, thereby enabling individuals to adopt non-intuitive conceptions of the *natural* universe. Perhaps it was this latter-day transfer that made both science and science education possible. In that case, the capacity for scientific conceptual change is not an adaptation after all, but a side-effect – an evolutionary byproduct – of psychological adaptations that really evolved for negotiating interpersonal truth claims in the social sphere. If so, then in the modern science classroom, cues evoking a strategic social context might mobilize those cognitive mechanisms and so foster student conceptual change.

In this section I will develop this social strategic hypothesis, once again focusing especially on select scholars: Dan Sperber, Robert Kurzban, and Robert Trivers. And I will again survey corresponding pedagogical practices in science education. *I develop the following thesis:*

> The cognitive mechanisms supporting scientific conceptual change may have originally evolved *not* for flexibly making sense of *natural* phenomena (= the individual constructivist hypothesis), *nor* for restructuring or replacing intuitive preconceptions in favor of *culturally* normative ones (= the cultural re-constructivist hypothesis), but instead for vetting and deciding among competing truth claims made by fellow members of one's social group about *strictly social* affairs. Only much later were those cognitive mechanisms recruited to *non-social* phenomena – that is, the objects of natural science. Certain pedagogical strategies advocated by science education scholars accord with such an evolutionary origin.

Although the three hypotheses about the evolutionary origin of conceptual change need not be exclusive of one another, this third thesis is one that I favor, and the instrument developed for this doctoral study was specifically designed to support research to test it. A corollary of this hypothesis, as I will describe later, is that the psychological machinery that carries out conceptual change may function largely independently of the machinery that generates our core intuitions in the domains of physics, engineering, biology, and psychology. Consequently those core intuitions may continue to affect a student's thinking and expression of ideas even after she undergoes conceptual change. In that case, the mind's social strategic mechanisms would neither *replace* nor *restructure* her pre-instructional conception so much as *suspend* and *suppress* it. For that reason, as I will describe in Chapter 3, one section of the instrument piloted in this study was designed to permit a student to endorse intuitive and scientific concepts side-by-side, while another section was designed to oblige her to suppress one in favor of the other.

A Social Strategic Answer to the Riddle of Horizontal Integration

The "riddle of horizontal integration" posed earlier was this: What evolutionary pressure prompted previously isolated conceptual sub-systems to start communicating with one another – something that all researchers deem essential for conceptual change? The "social strategic" hypothesis might answer that riddle as follows: The mind's conceptual sub-systems originally began communicating with one another *not* because it was adaptive for individuals to integrate their *own* mental representations across conceptual domains, but because it was adaptive to integrate those held by *other* people. An integrated picture of others' knowledge, goals, and ideas enables one to anticipate their intentions and actions in any arena where *competitive* and *cooperative* interactions might affect one's own fitness. For example, even before language evolved, success on collaborative hunting expeditions might have been enhanced by an ability to grasp a colleague's intuitions about the behavior of prey: a marriage of intuitive psychology and intuitive biology. Likewise, in a confrontation where an aggressor reaches for a stone or cutting tool, it might have advanced one's fitness to be able to read the mechanics of projectile motion (intuitive physics) and the functional employment of tools (intuitive engineering) from the *other's* vantage (intuitive psychology). In such situations, the fitness advantage goes to individuals who can best form an *integrated* grasp of other people's conceptual representations. Perhaps this was the selective advantage that first prompted "horizontal" sharing of information between cognitive domains.

Sperber's Module of Metarepresentation. In his 1996 collection of essays, *Explaining Culture: A Naturalistic Approach*, the cognitive anthropologist Dan Sperber develops such a position, especially in connection with conceptual reasoning. Sperber has been sharply critical of Vygotskyan models of human thinking. Even so, Sperber's own view of human cognitive evolution overlaps substantially with Tomasello's (2014) neo-Vygotskyan model, albeit with a couple twists. Like Tomasello, Sperber attributes our cognitive flexibility and talent for advanced conceptual reasoning largely to our *social* nature. But whereas Tomasello sees *cooperative* processes like collaborative foraging, communicating honestly and helpfully, deliberately sharing tool technology, and conforming to cultural norms as the key evolutionary waystations en route to conceptual agility, Sperber also emphasizes our *competitive* and *strategic* nature. Even for our closest ape relatives – chimpanzees and bonobos – social interactions comprise a dazzlingly dynamic interplay of competition, cooperation, coalition-forming, and kinship relations; “primate politics” run deep in our phylogenetic lineage (de Waal 1996, 2005; Gamble et al., 2014). Consequently, much more than in non-social species, an ability to predict others’ actions – *in a variety of domains* – would have been adaptive for our ancestors.

Thus for Sperber (1996), like Tomasello (1999, 2014), a key innovation in our cognitive evolution was the power to read others’ perspectives and beliefs (the “intentional stance” and a “theory of mind”; see Thesis #2). This, he says, gave us a capacity for “metarepresentation,” an ability to form mental representations *of* mental representations *as* mental representations:

> Humans have the ability to form mental representations of mental representations….The metarepresentational module is a special conceptual module…a second-order one, so to speak. Whereas other conceptual modules process concepts and representations of things, typically of things perceived, the metarepresentational module processes concepts of concepts and representations of representations. (p. 60)

Here he is positing the evolution in *Homo sapiens* of a new and very special cognitive sub-system: a module of metarepresentation (MMR). Higher on the cognitive hierarchy, the MMR receives as input the mental representations output by all other domain-specific conceptual systems, such as intuitive physics, engineering, biology, and psychology (see Figure 1). And this is his solution to the riddle of horizontal integration. Where Fodor (1983) and Carey (2009) locate horizontal integration in domain-general, content-neutral “central systems,” and where Karmiloff-Smith (1991, 1992) and Mithen (1996) attribute it to de-compartmentalization or de-modularization of the human mind, Sperber (1996) instead contends that horizontal integration occurred through the *addition of yet another specialized cognitive system*. He doubts it could be otherwise:

> Loosening the domain of a module will bring about not greater flexibility, but greater slack in the organism's response to the problem. To the extent that evolution goes toward improving a species' biological endowments, then, we should generally expect improvements in the manner in which existing modules perform their task, [or] emergence of new modules to handle other problems, but not demodularization. (p. 45)

His reasoning echoes our riddle of "evolvability": Since natural selection has no "foresight" and can act only on individual differences in the here and now, how could it ever let speedy, specialized, time-tested responses be supplanted by slower, less specialized controls? If anything, says Sperber, we expect complexity and specialization to increase over time, such that more and more cognitive mechanisms accumulate – hence the latter day evolution of a module of metarepresentation.

MMR's function: Strategic social influence. The MMR's original adaptive function, says Sperber, was to help hominins strategically navigate the social sphere. The primary benefit lies in the ability to understand, anticipate, and influence *others'* mental representations:

> The ability to understand…behavior, not as mere bodily movements, but in terms of underlying mental states, is an essential adaptation for organisms that must *cooperate and compete with one another in a great variety of ways* [emphasis added]. (p. 60)

> Once you have mental states in your ontology, and the ability to attribute mental states to others, there is but a short step, or no step at all, to your having desires about these mental states – desiring that she should believe this, desiring that he should desire that – and to forming intentions to alter the mental states of others. Human communication is…a way to satisfy metarepresentational desires…A communicator, by means of her communicative behavior, is deliberately and overtly helping her addressees to infer the content of the mental representations she wants him to adopt. (pp. 60-61)

Sometimes such “metapresentational desires” coincide between two parties, as in collaborative undertakings (Tomasello, 2014), but sometimes they do not. Especially once coupled to spoken language, metarepresentation led our ancestors to reciprocally strive to influence each other’s desires, intentions, beliefs, and concepts, for both cooperative and competitive ends. Given this dynamic, it also would have been strategically adaptive for the MMR to include *one’s own* desires, intentions, beliefs, and concepts in the metarepresentational mix. And since this mix derives from different corners of life, the MMR receives and re-represents input from multiple intuitive domains (see Figure 1).

Horizontal integration, according to this hypothesis, is thus accomplished by a special cognitive mechanism with a specific end: strategic success in an increasingly dynamic and dialogical social environment. A similar mechanism with a similar function will play a central role in tackling the other riddle of conceptual change. I turn to that next.

A Social Strategic Answer to the Riddle of Evolvability

Our other Darwinian riddle was this: How could natural selection ever produce psychological mechanisms for disregarding, discarding, or restructuring the more ancient and patently useful intuitions that natural selection had already built into the hominin brain? Although he does not address conceptual change specifically, the evolutionary psychologist Rob Kurzban echoes that riddle in his 2010 book *Why Everyone (Else) is a Hypocrite*. As a general rule for any animal species, he argues, "true knowledge" – or at least pragmatically accurate knowledge – is generally adaptive, while misrepresentations or poor representations of reality are maladaptive. Natural selection should favor cognitive mechanisms able to develop a veridical understanding of the natural environment, for such an understanding will yield useful predictions that promote survival and successful acquisition of resources. For our hominin ancestors, that selective pressure presumably produced our intuitive physics, engineering, biology, and psychology. These intuitions evolved because they worked. Though not strictly "scientific," they would have been more pragmatically predictive than any alternative concepts that might erupt into an individual's head (whatever they might be and however that might happen). In a dangerous world, natural selection should have swiftly weeded out any tendency to adopt less accurate, or at least less practical, mental models of the world. Although modern science did ultimately produce more accurate models of reality – that is, more pragmatically predictive ones – it did so only through many generations of conceptual toil, occasioned by a certain "leisurely" liberation from the pressing exigencies of survival. Our hominin ancestors, however, were on a lifelong, never-ending "camping trip," with wilderness always all around (Orians & Heerwagen, 1992), and their evolved intuitions served them well in this wilderness. Given the potential costs of abandoning intuition, then, what fitness payoffs could have allowed natural selection to cobble together the cognitive machinery that supports conceptual change?

Kurzban (2010) also points the way to a possible solution to this "evolvability" riddle. There is one environment, he observes, where natural selection *would* sometimes favor the adoption and affirmation of inaccurate beliefs about reality: *the social environment* (see also Trivers 2000). Although pragmatically "true" mental representations are always adaptive in the *natural ecological* environment, they can sometimes be maladaptive in *public social* interactions. When our ancestors on their lifelong camping trip circled around the campfire at the hominin campground, there may have been rewards for *publicly* representing reality in less than accurate ways. Before the advent of writing, audio-recording, videotaping, and mass media, "truths" about past social events (who said and did what), about personal character and private motivations (both one's own and others'), and about semantic intents (the meanings behind words and actions) all would have been highly "negotiable" (Kurzban, 2010; Kurzban & Aktipis, 2007; see also Trivers, 1985). And in small societies where individual reproductive success was significantly influenced by personal reputation, competition for status, and the formation of cooperative alliances with kin and non-kin, those ambiguous "truths" would have been vigorously negotiated (see Baron-Cohen, 1995; Haidt, 2007, 2012; Humphrey, 1976, 1986; Kurzban, 2010; Kurzban & Aktipis, 2007; Sperber 1994, 1996; Trivers, 1985, 2000). This in turn would have prompted the evolution of psychological mechanisms for judging the merit of others' potentially inflated or distorted truth claims, which in turn would have prompted the evolution of mechanisms for crafting a consistent and credible worldview and public image (Kurzban, 2010; Kurzban & Aktipis, 2007; see also Trivers, 1971, 1972, 1974, 1985, 2000).

In short, cognitive machinery evolved for making decisions about *which* representations of reality – both one's own and others' – to adopt and *publicly* affirm. Herein, I hypothesize, lie the seeds of conceptual change.

Kurzban's social cognitive interface (SCI). "Ignorance is at its most useful," says Kurzban (2010, p. 85), "when it is most public." In contrast to ecological domains, there are occasions in social settings when *lacking* veridical knowledge can enhance one's fitness. For example, under the watchful eyes of others, "people's reputations suffer to the extent that they are perceived not to discharge a duty" (p. 80). Yet duties are also costly to discharge. The calories, time, and attention might be better spent on activities more immediately pertinent to personal survival and reproduction. An individual's own best interests may be compromised by third-party monitoring, expectation, proscription, and moral judgment within the social group (DeScioli & Kurzban, 2009). However, we will be judged only according to what we are aware of, or rather, what others are aware we are aware of (Kurzban, 2010; Trivers, 2000). We may escape judgment when we can plausibly plea ignorance.

Kurzban and Aktipis (2007) thus propose that the conscious "self" comprises a suite of modules that *by design* are deprived of access to information held in other, non-conscious modules. They dub this "self" the *social cognitive interface* (SCI) and posit among its primary functions the presentation of a public front that advances one's strategic social interests (see Figure 1):

> A modular view of the mind implies that there is no unitary "self" and that the mind consists of a set of informationally encapsulated systems, many of which have functions associated with navigating an inherently ambiguous and competitive social world…. There are a set of cognitive mechanisms…designed for strategic manipulation of others' representations of one's traits, abilities, and prospects. Although constrained by plausibility, these mechanisms are not necessarily designed to maximize accuracy or to maintain consistency with other encapsulated representational systems….We refer to this potentially large but integrated collection of subsystems as the social cognitive interface (SCI). (Kurzban & Aktipis, 2007, p. 131)

A key to the SCI's success is its *ignorance* of certain truths compartmentalized elsewhere in the mind, including cognitive mechanisms that are monitoring one's state and steering one's behavior *unconsciously* behind the scenes. This enables the SCI to represent oneself as more healthy, trustworthy, skilled, competent, good-intentioned, altruistic, and valuable as an ally than may be warranted by the facts:

> Because many facts about the world are not objectively knowable, they are subject to negotiation and persuasion….We hypothesize that this is a primary function of the SCI: to maintain a store of representations of negotiable facts that can be used for persuasive purposes in one's social world. (Kurzban & Aktipis, 2007, pp. 134-135)

The SCI's function, then, is public relations. It is not the executive in charge of making decisions, but a "press secretary" or "spin doctor" whose job is to (a) communicate with others, and (b) frame one's talents, status, desires, beliefs, and past and present deeds in the most positive possible – *yet still plausible* – light. “The modules that you experience as ‘you’ can be thought of [as] a mouthpiece for the organization. ‘You’ are the Machiavellian spin doctor” (Kurzban, 2010, p. 60; see Haidt, 2012, for an almost identical depiction of post hoc fabrications and rationalizations made during moral reasoning; other scholars have also identified the “conscious self” with our strategic social nature, most famously Humphrey, 1983, 1986).

Trivers: Deception and self-deception. The influential evolutionary biologist Robert Trivers (1985, 2000) has proposed a very similar model under the rubric of *self-deception*. For decades he has written about the ubiquity of deception and deception-detection in the animal kingdom, from camouflage and mimicry in predator-prey interactions, to threats and bluffs between males of the same species, to the subtle manipulations and misrepresentations that occur between prospective mates, potential allies, and offspring and their parents – especially in our own species (Trivers, 1971, 1972, 1974, 1985, 2000, 2002). These contexts, he proposes, led to the evolution in humans of self-deception, defined as “active misrepresentation of reality to the conscious self” (2000, p. 114). Much like Kurzban, he asks how self-deception could ever be “evolvable,” and then answers in kind:

> What evolutionary forces *favor* mechanisms of self-deception?….In trying to deal effectively with a complex, changing world, where is the benefit in misrepresenting reality to oneself? Only in interactions with other organisms, especially con-specifics, would several benefits arise. Because deception is easily selected between individuals, it may also generate self-deception, the better to hide ongoing deception from detection by others. In this view, the conscious

> mind is, in part, *a social front* [emphasis added], maintained to deceive others. (2000, pp. 114-115)

This implies a complex, compartmentalized mind:

> Of course, it must be advantageous for the truth to be registered somewhere, so that mechanisms of self-deception are expected to reside side-by-side with mechanisms for the correct apprehension of reality. The mind must be structured in very complex fashion, repeatedly *split into public and private portions* [emphasis added], with complicated interactions between the subsections. (1985, p. 416)

This structure prevents the conscious mind from inadvertently revealing concealed facts and hidden motivations, "rendering [them] unconscious so as not to betray – by the subtle signs of self-knowledge – the deception being practiced" (Trivers, 1976, p. xxvi). Its self-serving manifestations within the strategic social sphere are the same as those yielded by Kurzban's SCI: invoking plausible deniability and "fictitious narratives of intention," "self-promotion, self-exaggeration," "creation of a public persona as an altruist," "biased memory, biased computation, changing from active to passive voice when changing from positive to negative outcomes, and so on" (2000, pp. 117-118). Such phenomena are richly documented in the psychological literature (e.g., Ariely, 2008, 2012; Gazzaniga, 2011; Kahneman, 2011; Tavris & Aronson, 2007; Wilson, 2002).

Kurzban and Trivers' theories share a feature that was crucial to the design of the experimental instrument that I developed for this doctoral study: A single mind may simultaneously hold more than one representation of the same reality. In the next section I will propose that it was the evolution of the SCI that paved the way for scientific conceptual change. Indeed, I will suggest that it may even be the SCI that *carries out* conceptual change, grappling with and eventually adopting non-intuitive science concepts, and constructing a normative scientific representation of them. If so, then intuitive representations generated “elsewhere in the mind” may remain unchanged and continue to voice themselves naturally but unwittingly through the SCI as “mouthpiece for the organization” – e.g., on tests of scientific understanding. My instrument was designed to listen for *both* voices – intuitive and scientific – at the same time.

An arms race: The evolution of skepticism, an open mind, and a will to consistency. Kurzban and Trivers’ split between the private and public mind poses a strategic dilemma: “While some modules are guiding what we say, other modules might be guiding action, leading to potential inconsistencies” (Kurzban, 2010, p. 67). Although “there is no particular reason to believe that the mind is designed…to maintain consistency among its various representational systems” (Kurzban & Aktipis, 2007, p. 134), there is a *social pressure* to maintain at least the *appearance* of consistency. The pressure to present a coherent public front with respect to one’s words, deeds, observable abilities, and truth claims about oneself and others, I propose, set the stage for a classic Darwinian “arms race” among our ancestors.

Many – perhaps most – of the macro-adaptations that we see throughout the living world were driven into their present form by co-evolutionary arms races (Dawkins, 1982; Ridley, 1993). For example, cryptic coloration (camouflage) can be understood as a strategy to deceive or manipulate other animals’ nervous systems (Dawkins 1976/1989, 1982). As nocturnal ghost crabs, color-changing flounder, and shape-shifting octopi evolved ever better camouflaging, their predators (and prey) evolved keener powers of visual discrimination to combat the deception. This in turn generated selective pressure toward even subtler deception by crabs, flounder, and octopi, which compelled predators to evolve sharper eyesight still, and so on. The ultimate result: Camouflage of astonishing exactness and eyes of astonishing acuity.

Arms races occur not only between species, but within them as well, and in highly social species they become especially intense (Cronin, 1991; Dawkins 1976/1989, 1982; Krebs & Davies, 1993; Ridley, 1993). For a social species, behavior is never merely an individual adaptation to the prevailing physical and ecological environment, for the shifty dynamics of the social environment also pressure individuals to adapt *to each other*. Human psychology has surely been shaped by evolutionary arms races of deception, detection, manipulation, and misrepresentation among parents and their offspring (Trivers, 1972), prospective mates (Trivers, 1974), prospective allies (Trivers, 1971), and entire social groups (Trivers, 2000). Indeed, what probably drove our ancestors to evolve such large brains and high intelligence was the ever intensifying cognitive demand of navigating the ever escalating complexities of competition, cooperation, and collective coordination in hominin populations of ever increasing size (Gamble, Gowlett, & Dunbar, 2014; Geary, 2007; Humphrey, 1976; Sterelny, 2012; Tomasello, 2014).

I suggest that once “truth” became “negotiable” within the social sphere (Kurzban, 2010), it made inevitable a particular psychological arms race that paved the way for scientific conceptual change. This arms race was amplified by the evolution of linguistic communication. Says Trivers (1985):

> In human evolution, processes of deception and self-deception were greatly heightened by the advent of language. Language permits individuals to make

> statements about events distant in time and space, and these are least amenable to contradiction. Thus, language permits verbal deception of many kinds. Since contradictory information is not available at the moment a deception is being practiced, there may be heightened attention to signs of conscious intent to deceive, and this will favor mechanisms of self-deception….Individuals readily create entire belief systems with self-serving biases, and the more skillfully these self-serving components are hidden from both the self and others, the more difficult it will be to counter them. (p. 416)

My hypothesis is that the evolution of our capacity for conceptual change originated precisely here. Given the inevitable visibility of personal inconsistencies in the public sphere, we should expect suspicion and skepticism to evolve as counter-adaptations. In other words, our minds evolved "spin" detector mechanisms. Importantly for my hypothesis, we might also expect a concomitant demand for empirical evidence to corroborate others' truth claims. Then in turn, as a counter-counter-adaptation amidst a sea of skepticism, we should expect the SCI to further evolve a propensity for spinning credible, coherent narratives. Importantly for my hypothesis, we might also expect a concomitant discomfort whenever one is at risk of *appearing* inconsistent in the public eye, and a reluctance to display any incongruities between words, deeds, and visible facts. Surprisingly, says Kurzban, the tangle of twisted truths in the social sphere might have fostered a genuine *will to consistency* – if not rigid accuracy – in one's public professions about what one believes, what one does, and how one represents reality: "'Self-consistency' need not, in itself, necessarily be a deep, fundamental motive. Instead, people might be motivated to appear consistent, *which in turn leads them to* *actually be consistent* [emphasis added]" (Kurzban & Aktipis, 2007, p. 140). At the same time, however, one's narratives should (unconsciously) bend in order to reap the potential benefits of exaggerating, slanting, or twisting the "truth" of one's positions – though not too much and always within the bounds of plausibility. The ideal would be to render a credible, not-too-inconsistent social front in the eyes of others that permits one to win resources, mates, and allies (Kurzban, 2010; Kurzban & Aktipis, 2007).

Somewhat counterintuitively, we should predict such a spiraling arms race to generate psychological adaptations not only for "skepticism" but also "open-mindedness" about the representations of reality put forth by others. In a dynamic social setting in which alliances shift, words and actions are remembered, and reputations bear upon reproductive success, decisions do have to be made. At some point in the deliberations, one must adopt and *publicly affirm* a position of one's own. The decision-making may require careful, conscious weighing of discrepant representations of reality against one another. And so in a sense, all these discrepant representations become *options for adoption* by the SCI. They amount to competing truth claims – again, strictly in the social sphere where "truth" is ambiguous and negotiable – from which the individual SCI (spin detectors and all) must at some point select. Somewhat paradoxically, I suggest, a critical eye toward the truth claims of others goes hand-in-hand with a receptiveness toward those competing representations. Skepticism implies *openness to alternate conceptions of reality*.

The social arena thus comprises an endless dynamic of claim-staking and stance-taking where truth is not so much discovered as strategically negotiated. One outcome of the associated evolutionary arms race may have been a capacity to *suspend intuition and* *judgment*, consciously weigh competing representations of reality against each other, scrutinize them against available evidence, arrive perhaps at unexpected and *even non-intuitive conclusions*, adjust one's interpretive and explanatory positions accordingly, and craft a visibly coherent, consistent portrayal of "truth" in the public eye. Plainly this would not be "calibrational," but something much more loose and flexible. It would be conceptual change.

From Social to Scientific: Conceptual Change As Evolutionary Byproduct

Social strategic solutions to the twin riddles of horizontal integration and evolvability leave one key question unanswered: If the cognitive machinery that carries out conceptual change – Sperber's MMR and Kurzban's SCI – evolved to navigate the *social* arena, to negotiate truth claims about *social* affairs, and to formulate consistent representations of *social* realities, then how did it come to mediate conceptual change with respect to *non-social* phenomena in the physical and ecological domains ruled by core intuitions, folk theories, and calibrational learning? That shift was unlikely to happen during the lifelong camping trips of our foraging ancestors. There would have been neither occasion nor inclination for making inflated, distorted, or counterintuitive truth claims about biological, ecological, meteorological, technological, or physical matters – namely, the stuff of modern natural science. I suggest that it may have been only very recently, even after the post-Pleistocene rise of agriculture and civilization, that this psychological machinery got recruited for making sense of the natural universe.

A byproduct hypothesis. Here, then, is the last piece of the social strategic hypothesis of conceptual change. In the modern world, including the modern science classroom, something happens that would have been exceptionally rare in the ancestral world: People find themselves confronted with *socially expressed,* intuitively discrepant claims about *non-social* realities. Our strategic social systems get mobilized – unusually – to grapple with our intuitive theories and calibrational learning. Skepticism, receptivity, and the will to consistency (or semblance of consistency) come to bear upon matters that are ordinarily immune to such scrutinies. Through conscious deliberation, the SCI is compelled to render decisions about which representations of reality to adopt both publicly and personally, even about things that in the ancestral environment never would have played into one's strategic social concerns. Meanwhile, the module of metarepresentation (MMR) – as gateway to the SCI (see Figure 1) – acts as a locus of communication between intuitive domains, a conceptual crossroads where processes that facilitate conceptual change can occur: analogies, metaphor, modeling, cross-domain mapping, categorical transfers, ontological reassignments, representational redescription, novel conceptual combinations, and/or conceptual bootstrapping (Atran, 1994, 1995, 1998; Carey, 2009; Carey & Spelke, 1994; Chi, 1992, 2008; Clement, 2008; diSessa, 2008; Karmiloff-Smith, 1991, 1992; Keil, 1994, 1995; Nersessian, 2008; Pinker, 1997, 2007; Smith et al., 1993; Thagard, 1992; Vosniadou & Brewer, 1987).

On this view, the cognitive machinery that supports scientific conceptual change did not evolve to supplant, supersede, or restructure our intuitions about the natural world. It evolved in spite of them. In Sperber's (1994) language, the objects, phenomena, and concepts studied in natural science do not belong to the SCI's *proper domain*. Their proper domains are intuitive physics, biology, psychology, and engineering. Yet after many generations of cultural evolution, they have become part of the SCI's *actual domain*. Analogously, a mottled seafloor falls within the proper domain of a flounder's camouflaging mechanisms, but a checkerboard does not. A checkerboard falls only within its actual domain, arousing an adaptive response in a setting that evolution never anticipated.

Likewise, certain classroom settings and experiences might arouse the SCI in the service of educational ends that evolution never intended. If the social strategic hypothesis is valid, then a science teacher who wants to cultivate conceptual change should design instruction to mobilize the mind's social strategic mechanisms. He should perhaps craft the classroom as a public forum where students discuss and deliberate rival interpretations of observed phenomena. Or he should perhaps stage a fictional debate, or weave in other cues that evoke a strategic social setting.

Conceptual change on this view is not an evolutionary adaptation but an evolutionary byproduct. The "individual constructivist" and "cultural re-constructivist" hypotheses explain conceptual change as an *adaptation* for enhancing cognitive/behavioral flexibility and predictive/inferential accuracy in the natural world: We evolved *for* conceptual change. By contrast, the social strategic hypothesis explains conceptual change as an incidental *byproduct* – a phylogenetic spinoff or side-effect – of psychological adaptations that really evolved to solve some other adaptive problem (Buss 2004; Tooby & Cosmides, 1992; Williams, 1966). Similarly, flounder did not evolve to mimic checkerboards, but they can. Elephant trunks are not an adaptation for lifting circus performers onto their heads; this is just a spin-off of trunk technology. Humans can use their nimble hands (and nimble neural machinery) to play guitars, operate cigarette lighters, and throw screwballs, even though their hands are really an adaptation for clutching branches and wielding stone tools. That a surfer can hang ten is a happy accident bestowed by our bipedalism and an exquisitely evolved inner ear: “Our hunter-gatherer ancestors were not tunneling through curls in the primordial soup. The fact that we can surf and skateboard are mere by-products of adaptations designed for balancing while walking on two legs” (Cosmides & Tooby, 1997).

In every example just given, both human and non-human, the byproduct behavior emerges against the backdrop of cultural history, not natural history. So it is, I hypothesize, for conceptual change. Biological evolution built the necessary mental machinery, but only cultural evolution was able to summon it into action – and perhaps not until quite recently. I discuss this final chapter of the social strategic story next.

Coming of age in the Milky Way: The rise of Western science. We were *social* long before we were *cultural*. The cultural re-constructivist model of conceptual change is compelling, but it may be incomplete. It treats novel concepts simply as new “tools for talking and thinking” (Vygotsky, 1978), such that learning a difficult scientific theory amounts to mastering a challenging new conceptual “technology.” This characterization finds Darwinian heft in Tomasello’s (1999, 2014) neo-Vygotskyan natural history of human cognition, which credits our capacity for conceptual change to the same evolutionary breakthrough that originally made cultural transmission of tool technology possible: a theory of mind. An ability to read and identify with the intentions of others is what made it possible to faithfully imitate their behaviors (Tomasello et al., 1993).

But we were social first, *just* social and not cultural. Although sophisticated sociality is rare in the animal kingdom, it runs deep in the primate line, going back tens of millions of years and millions of generations, long before the advent of culture (Gamble et al., 2014). A theory of mind did not first evolve for imitative learning and cultural transmission per se, but for navigating a complex social (and pre-cultural) landscape, increasingly complicated by deceptive and manipulative strategies in addition to cooperative ones (Tomasello, 2014). We see rudiments of the intentional stance and a theory of mind in chimpanzees and bonobos, who are socially sophisticated and engage in canny "primate politics," yet are poor at imitative learning, lack symbolic language, and do not accumulate culture (de Waal 1996, 2005; Richerson & Boyd, 2005; Tomasello, 1999, 2014).

The social strategic hypothesis proposes that the most challenging conceptual changes entail more than a straightforward cultural transfer of "technology." At a certain level, after all, learning to wield a new tool is quite intuitive. The teleological stance and our ready grasp of "functionality" make it so. Mastering a tool such as a bicycle or bow and arrow may be difficult, but the process is basically calibrational: Nothing about it runs counter to deep intuitions like the essentialist and intentional stances, or even the teleological stance itself. The same is not true of certain scientific concepts, such as the theory of evolution. According to the social strategic hypothesis, mastering those requires tapping a more ancient suite of mechanisms and motivations.

Here, then, is a hypothetical history of scientific conceptual change: Metarepresentation and a theory of mind first evolved for strategic social functions (Sperber, 1996). Much later, spoken language also evolved, initially only for communication within social domains, later becoming wired into the technical and natural history intelligences (Mithen, 1999; also see Pinker, 1994). Once symbolic language entered the strategic fray, "truth" in the social sphere became negotiable, prompting the evolution of the social cognitive interface (SCI; Kurzban & Aktipis, 2007), including spin detection and the skeptical scrutiny of rival truth claims, as well as an open-mindedness toward them and a public striving for consistency. Along the way, the MMR and a theory of mind also led to (a) horizontal integrations within the mind across both social and non-social domains (Sperber, 1996), and (b) a blossoming capacity for cumulative culture (Ricerhson & Boyd, 2005; Tomasello et al., 1993). Cumulative cultural evolution made human *populations* far more flexible, able now to adapt technologically and behaviorally to ephemeral environments and exploit new niches (Richerson & Boyd, 2005). Meanwhile, horizontal integration made human *individuals* more cognitively flexible (Karmiloff-Smith, 1991, 1992; Mithen, 1996). Finally, only fifty to a hundred thousand years ago, the combination of a flexible, horizontally integrated mind and the availability of increasingly diverse cultural forms paved the way for great leaps of creativity, artistic imagination, and technological innovation (Gamble et al., 2014; Mithen, 1996). Cultural evolution swiftly accelerated.

In all these ancient societies, however, the intuitive “folk theories” that make sense of physical and ecological phenomena rarely clashed with truth claims being negotiated in the social sphere. *There was still no conceptual change with respect to the natural universe.* Perhaps it was not until cultural evolution began to build a technological world with writing, careful measurements, and information-gathering instruments – especially after Aristotle – that *empirical observations* of natural phenomena began to clash conspicuously with *publicly espoused* models of reality (Kuhn, 1962/1970; Thagard, 1992). Perhaps it was only as civilization began to accumulate a written history of its own representations of reality that physical and ecological phenomena stepped – for the first time in our species' long history – onto a *social, dialogical* stage where the SCI's skeptical-yet-open-minded machinery is on full alert and representational inconsistencies are little tolerated. Perhaps only then did competing and increasingly counterintuitive representations of *natural* reality become the *unnatural* stuff of *social* negotiation. Perhaps only then, as a byproduct of the evolution of our social strategic nature, did *Homo sapiens* at last undertake bona fide conceptual change.

Thus was Western, post-Aristotelian natural science born, as *empirical scrutiny* was married to *social scrutiny* in an ever-revolving wheel of self-correction (Thagard, 1992). And as this wheel churned over the centuries, the public record steadily accumulated concepts, theories, and models that, although predictively powerful, are not always intuitively easy to understand. The result: the ever-widening mismatches between our stone-age psychology and the memes of modern civilization that make contemporary schooling necessary, and that present science education with its toughest challenge.

With that in mind, I close this hypothetical history with a quote from one of my favorite books, Timothy Ferris's *Coming of Age in the Milky Way* (1988):

> The truth is beautiful, but the beautiful is not necessarily true: However aesthetically pleasing it may have been for the Sumerians to imagine that the stars and planets swim back from west to east each day via a subterranean river beneath a flat earth, such a conception was quite useless when it came to determining when Mars would go into retrograde or the moon would occult Jupiter. Consequently the idea slowly took hold that an adequate model of the universe not only should be internally consistent, like a song or a poem, but should also

> make accurate predictions that could be tested against the data of observation. The ascendancy of this thesis marked the beginning of the end of our cosmological childhood. Like other rites of passage into adulthood, however, the effort to construct an accurate model of the universe was a bittersweet endeavor that called for hard work and uncertainty and deferred gratification. (pp. 23-25)

In that last sentence, he could just as well be describing a student struggling through conceptual change in a modern science classroom.

“Social Strategic” Pedagogical Approaches in Science Education

If the (un)natural history of natural science just portrayed is valid – that is, if science developed its powerful-though-counterintuitive models of reality by ushering *non-social* phenomena into psychological systems whose real function is to negotiate truth claims about *social* phenomena – then conceptual change in the science classroom would require similar processes. If it is not in our nature to adopt novel, non-intuitive conceptions of the physical and ecological world, then we would want to find ways to import them into the social sphere where students’ concepts may be more amenable to modification. In particular, students would not adopt counterintuitive truth claims unless and until their social strategic modules were aroused. That might mean asking them to write, draw, speak, discuss, and otherwise *publicly* represent their conceptions during the learning process. It might mean engaging them in public discourse or debate about natural phenomena and scientific principles. It might mean obliging them to take stances, stake claims, make arguments, weigh rival representations against one another, justify their positions, and ultimately reevaluate their intuitions by way of open dialogue with teacher and peers. And it would mean nudging them toward public consistency and coherence.

Such instructional strategies do appear frequently in the conceptual change literature. I survey a few of these below.

The HEI method: Cycles of social and empirical scrutiny. Many Japanese science classrooms facilitate conceptual change through a method that meshes well with the hypothetical history of natural science that I depicted above. Called Hypothesis-Experiment-Instruction (HEI), the pedagogy actively engages students in alternating cycles of *empirical scrutiny* and *social scrutiny* (Hatano & Inagaki, 2003; Miyake, 2008). A typical sequence:

(1) Students confront a problem or puzzling phenomenon centered on a particular scientific concept, along with several alternative solutions or explanations in the form of testable predictions. One alternative is scientifically correct, the rest intuitively appealing distracters intended to evoke common misconceptions.

(2) Each student endorses one of the solutions, privately at first, followed by a vote in which results are publicly tallied.

(3) Students discuss, explain, and defend their positions in a public forum.

(4) Students revote, with the option to change their original positions.

(5) Students test their predictions/positions against empirical observations or evidence, through one of several means: actual experiment, teacher demonstration, graphing of real scientific data, reading a passage, etc.

(6) Students discuss and interpret results vis-à-vis their original votes.

(7) Teacher provides formal instruction about the concept at hand.

The HEI sequence includes a “Vygotskyan” tactic from *cultural re-constructivist* pedagogy: formal instruction by the teacher in step #7 (which presumably would include a round of student practice with the new concept, aided by teacher coaching). Also embedded in the sequence is a “Piagetian” tactic from *individual constructivist* pedagogy. The empirical experience in step #5 is a “dissonance strategy” (Clement, 2008) designed to provoke dissatisfaction with pre-instructional intuitions and compel the student to rationally reevaluate them. But here Hatano and Inagaki (2003) issue a caution:

> This strategy must be applied with great care, because students are *not always open-minded* [emphasis added]. For example, students tend to interpret new observations or anomalous data in biased ways so that they can be harmonious with their prior knowledge….A good [practice] is to present a phenomenon that disconfirms students’ predictions based on their misconception *after they state their predictions in public* [emphasis added], because the students cannot ignore the cognitive incongruity…in such a social situation. (p. 414)

The key to success with Piagetian and Vygotskyan tactics, then, is to place them within a broader public forum, just as the *social strategic* hypothesis dictates.

Vicarious mobilization of the SCI. Many other science education researchers also advocate the taking and sharing of stances both verbally and in writing, often in a public forum, often as a contest between explicitly juxtaposed scientific and unscientific ideas, and often in the face of empirical evidence (e.g., Bransford et al., 1999; Chinn & Brewer, 1993; Chinn & Buckland, 2012; Clement, 2008; Dole & Sinatra, 1998; Driver et al., 1994; Driver, Newton, & Osborne, 2000; Hynd, 2003; Kelly & Green, 1998; Leach &

Scott, 2008; Miyake, 2008; Mortimer & Scott, 2003; Nussbaum, 2008; Smith et al., 1993; Vosniadou et al., 2008; Vosniadou & Brewer, 1987; Zhou, 2010). Evolutionarily speaking, however, if the social strategic hypothesis is correct, then the real key to such practices is not merely that they engage students in rational discourse and debate about science and nature, but that they unwittingly evoke a *strategic contest* between *socially motivated* truth claims, and that they oblige students to *publicly* decide which representations of reality to adopt. If so, then alternative instructional tactics might be available. For example, a teacher might involve students *vicariously* in a strategic social contest by embedding scientific concepts in a fanciful scenario where fictional characters are faced with a pressing social dilemma. The characters would voice competing solutions to the dilemma, some representing a sound scientific understanding, others intuitively appealing but unscientific. These characters might be represented as having socially expedient motives, such as an award or financial gain. A teacher might also stage an imaginary time-travel debate between, say, Galileo and Aristotle, or Darwin and Lamarck, or Einstein and Newton, or Chomsky and Skinner. And these fictional or historical characters might make rival predictions about the outcome of an actual experiment or demonstration that the class will soon conduct (as in HEI).

Discourse and argumentation reconsidered. I already discussed the literature on scientific "discourse and argumentation" under Thesis #6 in connection with *cultural re-constructivist* pedagogy. But there I suggested that although advocates of discourse and argumentation profess allegiance to Vygotsky, this burgeoning pedagogical movement might better be classed as *social strategic*. In argumentation pedagogy, students are obliged to weigh competing claims, inhabit different perspectives, evaluate arguments and counterarguments, view it all with an open-minded yet healthily skeptical eye, and make decisions about what positions to embrace (Driver et al., 2000; Duschl, 2008; Kuhn, 1992, 1993, 2010; Nussbaum, 2008). They also make claims and cases of their own, and are pressed to defend their positions both theoretically and with empirical evidence, often committing to their positions in writing and/or a public forum (Driver et al., 2000; Duschl, 2008; Kuhn, 1992, 1993, 2010; Nussbaum, 2008). In class dialogue, the teacher plays the role of moderator, conductor, and sometimes *provocateur* (Driver et al., 2000). All of this seems tailor-made to mobilize the SCI in the service of science education.

Echoing Sir Nicholas Humphrey's (1976) famous "social intelligence hypothesis" (see lead quote above), Deanna Kuhn (1992) contends that social, dialogical argument is the very essence of intelligence itself, a position captured in her expressions "thinking as argument" (1992) and "scientific thinking as argument" (1993). Kuhn, the leading scholar in argumentation pedagogy, maintains that rational thought is tantamount to an "interiorized dialogue….with an imagined interlocutor who the [individual] attempts to convince" (2010, pp. 818-819). This accords well with the social strategic hypothesis, according to which conceptual change is mediated by cognitive mechanisms designed to "negotiate truth" in social encounters.

Argumentation for conceptual change? There is a crucial disconnect, however, between conceptual change and argumentation pedagogy. Most of the research on argumentation has focused on scientific reasoning and epistemology *unmoored from specific content and concepts,* and rarely as a mechanism for effecting conceptual change (Nussbaum, 2008). Content and concepts merely provide the occasion for cultivating expertise in argumentation, not the other way around. Indeed, Kuhn (1992, 1993, 2010) has long urged that "argumentive" reasoning skills be taught as ends in themselves, divorced from content objectives. Above all, the aim is to help students grasp the "nature of science" and socialize them into the scientific community.

Such was the position of Driver, Newton, and Osborne in a seminal 2000 paper that helped thrust argumentation to the fore of science education research. But in the following passage they at last began to weave content learning – including conceptual change – into the discourse, argumentation, and science "enculturation" process:

> Because science involves a process of social construction of knowledge…the terms, models, and ways of seeing the world agreed upon by scientists are human products....Giving learners access to these "ways of seeing"…means inducting learners into the *particular ways of representing the world* used by scientists and socializing them into adopting the *conceptual tools* of that culture. Through this process learners are introduced to a new language…that enables them to portray the world in new ways – a world inhabited by new entities such as genes, chromosomes, electric fields, atoms, and ions. This process of enculturation into science comes about in a very similar way to the way a foreign language is learned – *through its use*. Students need opportunities…to *practice using* the ideas themselves….It is our view that *conceptual change* is dependent on the opportunity to socially construct, and reconstruct, one's own personal knowledge *through a process of dialogic argument*. (p. 298; emphases added)

The orientation here is plainly Vygotskyan, and many of the metaphors locate it within *cultural re-constructivist* pedagogy: "conceptual tools," "practice using," "through its use," "socially reconstruct." Yet the final nod to "dialogic argument" supports the hypothesis that conceptual change might also require tapping the mind's more ancient *social strategic* mechanisms.

Research on argumentation for conceptual change. Since that 2000 paper, a few researchers have finally begun using experimental or quasi-experimental study designs to explore argumentation pedagogy as a candidate vehicle for content and concept learning, some targeting conceptual change with respect to difficult concepts. Although the research so far is thin (Nussbaum, 2008; Zhou, 2010), and most experiments are not designed to tease apart "social strategic" effects from "individual constructivist" or "cultural re-constructivist" effects, there is evidence that argumentation can foster students' grasp of challenging science concepts. For example, compared to more conventional instruction, students exhibit greater conceptual change regarding difficult physics principles when computer simulations are used to stimulate "cognitive conflict" – much in the vein of individualistic, empiricist Piagetian approaches – while embedded within a social argumentation framework that includes competitive prediction-making and stance-taking among peers (Zhou, 2010). And when students are explicitly taught argumentive reasoning skills and obliged to use them during the learning process, they master advanced genetics principles better than those who learn the material instead through text reading, teacher explanation, and individual practice problems – all elements of cultural re-constructive pedagogy (Zohar and Nemet, 2002).

A promising young approach that deliberately weds argumentation to content learning and conceptual change is "model-based instruction" (Chinn & Buckland, 2012):

> Our model-driven instructional methods center on the dialogic discourse and epistemic practices of coordinating explanatory modeling with evidence….The focus…is on the development of students' abilities to construct and revise models, to coordinate models and model revisions with evidence, and to engage in effective written and oral argumentation in support of this coordination. (p. 219)

Model-based instruction can yield significant gains in conceptual mastery of a variety of topics, including notoriously challenging ones (Chinn & Buckland, 2012; Lombardi, Sinatra, & Nussbaum, 2013).

Finally, a recent and robust pair of carefully controlled experimental studies evaluated the efficacy of argumentation in cultivating conceptual change with respect to the theory of evolution. In their first experiment, Asterhan and Schwarz (2007) deliberately partitioned the effects of "argumentative" interaction between students from the effects of mere "collaborative" interaction. In the second experiment they compared the effects of "dialogical" argumentation between students to mere "monological" sharing of interpretations. The authentically dialogical and argumentative interactions yielded superior, more durable conceptual change. Moreover, students who engaged in more dynamic and dialectical argumentation showed greater and more enduring conceptual gains than those who engaged in shallower one-sided arguments.

Researchers in this area typically attribute such conceptual gains to the "active thinking" (Zohar & Nemet, 2002) or "superior processing of information" (Asterhan & Schwarz, 2007) that dynamic argumentation demands. But it is possible that what makes the difference, at least in part, is the arousal of the mind's social strategic mechanisms.

The dark side of cognitive dissonance. One final caution about social strategic pedagogy merits mention. HEI, discourse and argumentation, model-based instruction, and kindred methods all risk provoking certain undesirable consequences of arousing "cognitive dissonance." This is a strong feeling of discomfort that occurs when one is confronted with evidence or arguments that contradict one's beliefs and expectations (Tavris & Aronson, 2007). The prominent psychological theory of cognitive dissonance is often taken to imply that dissonance normally motivates an individual to reevaluate and modify his conceptions in accord with evidence, thereby reducing the discomfort of dissonance. In reality, though, the theory predicts that the opposite will often occur: The individual may dismiss the evidence as flawed or fraudulent, or he may (mis)construe it in a selective manner that fits it to his prior conception, or he may simply hold fast to his initial position despite the evidence (Chinn & Brewer, 1993; Tavris & Aronson, 2007). These responses may be especially likely to occur when a person has *publicly* espoused his stance in a *social* setting, especially in a *competitive* climate; indeed, they may be part and parcel of our strategic social psychology (Kurzban, 2010; Trivers, 2000). A number of conceptual change researchers warn classroom practitioners of these potential responses and stress the importance of cultivating classroom norms and climate in which students respect one another, are willing to risk public answers even when unsure, are receptive to others' ideas and open to change, and adhere to a collective goal of conceptual comprehension for everyone – *especially* where dialogue and debate are part of the conceptual change process (Chinn & Brewer, 1993; Clement, 2008; Hatano & Inagaki, 2003; Hynd 1998, 2003).

Summary, Thesis #7: "Social Strategic" Models

The social strategic hypothesis that I developed in this section postulates a historical irony: It was the presence of *deception* in the ancestral *social* environment that ultimately led to the spectacular success of modern science at unearthing even the most elusive and counterintuitive *truths* about the *natural* universe. Evolutionarily, what made the human mind capable of developing natural science's often non-intuitive *representations of reality* was the pressure for our hominin predecessors to detect *misrepresentations* of reality amongst their peers. Our minds can suspend intuition, adopt counterintuitive truth claims, and craft evolutionarily novel explanations of natural phenomena precisely because our ancestors were often pressured to detect untruths during their social interactions.

Specifically, I made a case that neural communications first opened up between previously isolated conceptual sub-systems because it enabled individuals to form an integrated grasp of *others'* behavior within the social group. In particular, the wedding of our intuitive psychology and a theory of mind to the intuitive domains of physics, engineering, biology, etc., made it possible to anticipate others' desires, intentions, and actions during all sorts of cooperative and competitive interactions. Sperber (1996) proposed that this wedding was the work of a newly evolved "module of metarepresentation" (MMR), which permits an individual to construct second-order mental representations of the mental representations held by others and oneself, and thus to grasp them *as* mental representations. By effectively dislocating mental and public *representations* of reality from *reality itself*, the MMR enabled individuals to see those representations as malleable and manipulable. And so, says Sperber, especially with the advent of spoken language, the MMR became a tool for strategically influencing the desires, beliefs, and concepts held by others in the social group.

Among the tactics that likely now developed for influencing others' ideas was strategic *misrepresentation* of reality. Here I turned to evolutionary hypotheses proposed by Kurzban (2010; Kurzban & Aktipis, 2007) and Trivers (1985, 2000), who argue on theoretical grounds and from empirical evidence that the human mind evolved an often *unconscious* tendency to publicly distort the truth about past events, words, deeds, status, talents, and future intentions – whether one's own or others'. Kurzban specifically posits the evolution of a special constellation of cognitive mechanisms, perhaps including Sperber's MMR, whose function is to communicate clearly with others while presenting a public front that "represents reality" in personally beneficial, fitness-enhancing ways. He dubs this the "social cognitive interface" (SCI) and equates it with the consciously perceived "self." This is an eminently *social* self, with a penchant for strategic reasoning, post hoc rationalization, and imagining future dialogues with others. A key feature of the SCI – important in the development of the instrument piloted in this doctoral study – is that by evolutionary design, it remains largely unconscious of certain knowledge, motives, and *intuitions* that are guiding the individual's behavior and that continue to surreptitiously influence the SCI's communications.

Finally, I made a case that the plurality of distorted truth claims in ancestral social settings would have launched a rapidly escalating evolutionary "arms race" of deception and deception-detection, and that this arms race would have fashioned psychological adaptations for cautious *skepticism* toward others' truth claims, a thirst for *corroborating evidence*, an interest in *alternative interpretations* of the same incidents and affairs, and an *openness* toward the rival representations of reality put forth by others. Following Kurzban, I proposed that there would have been a selective pressure for individuals to *publicly* adopt and express positions that were both internally consistent and consistent with available evidence. Arriving at those positions might require an ability to *suspend intuition* while consciously weighing rival interpretations and possibly revising one's own interpretations in unexpected, even non-intuitive directions. That ability – although originally designed for negotiating truth about strictly *social* concerns – was eventually recruited to *non-social, natural* phenomena as well. And that, I proposed, ultimately paved the way for modern natural science to construct – through self-correcting cycles of *empirical scrutiny* coupled to *social scrutiny* – its remarkably successful models and theories, while abandoning poorer conceptions of reality, however intuitively appealing they might be. Our capacity for scientific conceptual change, on this account, is a recent *evolutionary byproduct* of ancient psychological adaptations for coping with misrepresentations of reality during strategic social discourse.

Thus when biologically evolved intuition collides with counterintuitive, culturally evolved truth claims in a social setting, such as a science classroom, there may be resistance, but also the potential for conceptual change. The key may be to supply cues that evoke a cooperative/competitive social context in order to arouse Sperber's MMR, Kurzban's SCI, and the mind's social strategic machinery. This need not be incompatible with the use of *individual constructivist* and *cultural re-constructivist* pedagogies. I surveyed a number of instructional strategies that embed Piagetian and/or Vygotskyan tactics within a broader public forum where students weigh scientific and unscientific explanations against one another, take stances, stake claims, make arguments, justify their positions, and ultimately reevaluate their intuitions through open dialogue with teacher and peers.

Research to evaluate the efficacy of such classroom strategies should include experiments designed to partition the effects of social strategic tactics from the effects of Piagetian or Vygotskyan ones. Such experimental designs would require instructional interventions that can be manipulated to control for Piagetian versus Vygotskyan versus social strategic influences. They would also require an instrument designed to detect signals that during conceptual change, a student is suspending and suppressing intuitive concepts, rather than replacing or restructuring them. That was a primary design feature of the instrument developed for this doctoral study.

Closure: A Hypothetical Model of the Intuitive Mind Revisited

As stated at the start of this literature review and repeated throughout, the new quantitative, closed-response instrument that was piloted in this doctoral study was designed with two distinguishing design features:

1. It is designed not only to gauge student mastery of the scientifically normative model of evolution, but also to elicit common misconceptions where they exist (Table 1).
2. It is designed to:
   (a) permit both scientific and intuitive conceptual models to surface *simultaneously*, side-by-side, whenever two or more relatively independent models *coexist* in the same student's mind, and

(b) elicit evidence that whenever a student does demonstrate conceptual change, she does so by suppressing her still active intuitive conceptions in favor of the scientific one.

The second design feature was implemented by dividing the test into two sections, each with a different format. Like most closed-response tests, one section of this instrument obliges students to choose between scientific and unscientific statements. Unlike most closed-response tests, however, the other section of this instrument allows students to simultaneously endorse *both* scientifically normative and unscientific-yet-intuitive positions, without having to choose between them. I will describe this structure more thoroughly under Methods in Chapter 3.

I now close this literature review with a return to the schematic model of mind depicted in Figure 1, which along with Table 1 was the main theoretical framework that inspired and informed the development of my experimental instrument. During this chapter, I gradually elaborated the rationale behind Figure 1 (and Table 1) through a sequence of seven theses. In this final section I show how this model of mind – especially in concert with the social strategic hypothesis – prescribed the design features just mentioned.

Upward Flow of Representations: Perceptual to Conceptual to Central Systems

As I described in the “Interlude” between Theses #3 and #4, the model should be read bottom up. Perceptual “input analyzers” (blue) seek out particular patterns amidst the kaleidoscopic stream of input from the sense organs (orange), and when found, encode them into *perceptual* representations. These perceptual representations are then selectively cabled to the appropriate *conceptual* systems (pink), each dedicated to a domain that was important to ancestral hominins, such as inanimate objects, animals/plants, other people, and tools/artifacts. These conceptual mechanisms interpret our perceptions, often projecting onto them a domain-specific form of causation, such as mechanical contact, hidden essences, teleological functions, or intentional agency. Through calibrational learning, this conceptual output may now guide the construction of domain-specific "intuitive theories" (green), such as our intuitive physics, biology, psychology, and engineering. Finally, our intuitive representations are piped to "higher" cognitive systems, where at last they can be "horizontally" integrated and where conceptual change presumably occurs.

This hypothetical model actually posits two loci where horizontal sharing and conceptual change might happen. The first is the mind's *central/executive* system (tan), which receives representations from multiple domains and uses them in solving problems, making decisions, and selectively generating behavior (e.g., Fodor, 1983; Carey, 2009). This is where the *individual constructivist* hypothesis – with its emphasis on flexible reasoning in the face of environmental uncertainty – might locate horizontal integration and conceptual change. Presumably the adaptive benefit would lie in improved prioritizing, problem-solving, and long-range planning. The *cultural re-constructivist* hypothesis might likewise locate horizontal sharing in some "central" crossroads of the mind, where – for our ancestors – ideas from different domains crossed paths, combined and recombined, and fostered novel insights that enhanced tool technology, hunting strategies, and sociocultural practices.

An Alternative Locus of Conceptual Change: The Social System (MMR and SCI)

The second locus where horizontal integration and conceptual change might occur is the mind's social system, responsible for interacting and communicating with other people – that is, with other minds. This system comprises Kurzban's *social cognitive interface* (SCI, yellow; Kurzban & Aktipis, 2007) and Sperber's (1996) *module of metarepresentation* (MMR, green), which is closely connected to our "theory of mind" and intuitive psychology. This is where the *social strategic hypothesis* would locate horizontal integration and conceptual change. Metarepresentation with respect to strategic social concerns is the very thing that would have set in motion an arms race producing cognitive mechanisms for "spin" detection, skepticism, open-mindedness, and vetting others' truth claims: "An organism endowed with a metarepresentational module can represent concepts and beliefs qua concepts and beliefs, *evaluate them critically, and accept or reject them* [emphasis added]" (Sperber, 1996, p. 60). And it is the SCI that would field those claims, warily scrutinizing them against available evidence and against intuitions upwelling from below. At the same time, however, the SCI should be designed to sometimes suppress those intuitions, if only partially or provisionally, in order to construct and communicate a credible and consistent worldview (at least in the public eye). Closely coupled to the MMR, it could avail itself of "horizontally" integrated mental representations from multiple domains. It would thus be able to suspend intuition and entertain novel truth claims not only about social affairs but also non-social phenomena in the physical and ecological environment. That might enable it to construct and store novel, "socially negotiated" theories about the natural world (Figure 1, purple). These newly adopted, newly (re)constructed theories might even include culturally evolved science concepts that are *counterintuitive,* having been unmoored from rigid intuition by passage through a cognitive system whose function is to navigate situations where “truth” is ambiguous and negotiable. And that would be conceptual change.

Coexisting Conceptions: Conceptual Change as Suppression, Not Restructuring

According to the social strategic hypothesis, then, conceptual change is the work of the SCI, a comparatively isolated sub-system near the “top” of the cognitive hierarchy. By evolutionary design, it is insulated from, or kept “ignorant” of, certain information elsewhere in the mind, especially the executive systems that are governing behavior behind the scenes (Kurzban, 2010) – hence the dashed arrow between them in Figure 1. Its relationship to our core intuitions is likewise largely unconscious. For example, the mechanical, intentional, teleological, and essentialist stances are evoked reflexively and imperceptibly. There is no reason to suppose that the SCI (or the executive system) exerts any “downward” influence that alters the routine operation of the “lower” perceptual and conceptual systems – hence the crossed out curved purple arrows in Figure 1. Those systems would continue to generate the natural, normal conceptual output for which natural selection designed them. Thus even after the SCI accomplishes a scientific conceptual change, certain pre-instructional intuitions would continue to quietly feed “upward” into the “higher” cognitive systems that author the conceptual models that a student articulates to other people, such as a science teacher on a science test.

In that case, although the SCI might be able to *override* a student’s intuitive theories, it would not be able to *restructure* or *replace* them. Two relatively independent conceptual models would *coexist* in the same brain: one intuitive, the other “negotiated” by the SCI. The continued coexistence of pre- and post-instructional conceptions, even after conceptual change, is a possibility acknowledged by some leading researchers in the field (e.g., Chi, 1992; Clement, 2008; Inagaki & Hatano, 2002), and it is supported by recent empirical evidence (Shtulman & Valcarcel, 2012). Moreover, this possibility is implicit in Vosniadou's (1994; Vosniadou et al., 2008) oft-cited observation that students, when asked to explain difficult scientific concepts, often construct either a "mixed model" or a "synthetic model" that combines or hybridizes an intuitive conception with the scientific one. And more generally, a robust body of psychological research suggests that conscious, effortful learning is carried out by systems that exist *alongside* our fast and frugal intuitive systems (Kahneman, 2011).

Once again, the instrument that I field-tested for this doctoral study was specifically designed to allow coexisting conceptual models to surface simultaneously, while also eliciting evidence that whenever a student does demonstrate conceptual change, she may be doing so by effortfully suppressing her still active intuitive conceptions in favor of the scientific one.

Preview of Chapter 3

In Chapter 3 I describe how this model of the evolved human mind (Figure 1) inspired and informed the development of the experimental instrument that was piloted for this doctoral study. As I have shown here, the evolutionary origin of our capacity for conceptual change may have real pedagogical implications for the 21st century science classroom, whether it is best explained by a Piagetian "individual constructivist" hypothesis, a Vygotskyan "cultural re-constructivist" hypothesis, or my own "social strategic” hypothesis – or perhaps more likely, some combination of two or all three of these hypotheses. I believe these hypotheses and their corresponding instructional practices are worth putting to the test in classroom intervention studies. Such studies require an appropriately designed, field-tested, and validated instrument to assess student conceptual change. The next chapter describes in detail the special design features of this instrument, the method by which it was field-tested with 340 high school students, and the statistical methods by which it was analyzed for validity and reliability.

Chapter 3: Research Methods

This study piloted and attempted to validate a new quantitative instrument for assessing student understanding of the theory of evolution. It was field-tested with 340 members of the target audience of secondary school students. The instrument was designed not only to gauge student mastery of the scientifically normative model of evolution, but also to elicit common misconceptions where they exist. In particular, as explained in my Chapter 2 literature review, it provides opportunities to endorse three deeply intuitive interpretations that are well known to compromise many students' understanding of evolutionary theory: the projection of intentional agency, teleological directionality, and immutable essences onto biological phenomena. As I argued in Chapter 2, these three interpretive tendencies are not only *intuitive,* but may also be partly *innate,* fashioned by natural selection to provide our hominin ancestors with pragmatic – though unscientific – ways of conceptually construing events in their physical, ecological, and social environments. In short, the human brain may be poorly evolved to grasp evolution itself.

From this Darwinian perspective I further reasoned in Chapter 2 that contrary to the prevailing view, conceptual change regarding evolution might *not* entail *restructuring* a single pre-instructional mental model into a scientific model, nor *replacing* the former with the latter. Instead, it might entail a competition between two or more mental models which *coexist* in the same mind: an innate, enduring intuitive model and the novel, newly acquired scientific one. Even after conceptual change, both models may continue to coexist for the student, such that conceptual change may be more a habitual *suppression* of intuitions than a supplanting or transformation of them. My experimental instrument was designed to allow for that possibility. Whereas most closed-response tests oblige students to choose between scientific and unscientific propositions, a portion of this one permits them to endorse both simultaneously, such that scientifically normative and unscientific-yet-intuitive conceptions can both surface where they coexist.

This new instrument was developed, piloted, and evaluated in several steps. First, based on the theoretical reasoning elaborated in Chapter 2, I established a list of design features that a quantitative, closed-response instrument should possess for assessing student understanding of evolutionary theory, and a framework of scientific concepts and common misconceptions that such an instrument should target. Second, I surveyed seven existing instruments and borrowed heavily from them in composing new candidate test items for a new instrument, custom-designed to support my own future research agenda. Third, content experts fluent in evolutionary theory appraised these items for scientific validity, conceptual soundness, clarity, and relevance to the guiding framework, and test items were revised in accord with their recommendations. Fourth, a large sample drawn from the target population of secondary biology students took the test under classroom conditions. Fifth, I used confirmatory factor analysis and other statistical methods assess how well student test responses aligned with the intended framework of scientific concepts and common misconceptions. Finally, I used the results to identify items that should be removed, replaced, or revised to improve instrument validity and reliability.

In this chapter I detail this six-step process.

Instrument Design and Development

The instrument – a test of evolutionary understanding – was designed in accord with the cognitive model of conceptual change elaborated in my Chapter 2 literature review. It is intended for secondary biology students, and my long-term research goal is to use it in a series of classroom intervention studies with large samples drawn from diverse student audiences across a wide range of school settings. Those interventions will be implemented by classroom teachers, who will test students before and after each intervention. I tailored my instrument to support such a research agenda, with the goal of satisfying seven specific requirements.

Seven Requisite Instrument Features

For my purposes, an instrument assessing conceptual change with respect to the theory of evolution should bear seven features:

1. It should be age-appropriate for the target audience (high school students).
2. It should employ closed-response items that provide quantitative data, rather than rely on open-response items that harvest qualitative data. The latter are potentially more time-consuming for students to take and more labor-intensive to code and analyze. As a practical matter, my long-term research agenda requires an instrument that teachers can easily administer to their classes and that generates data in a form that can be readily transferred to the researcher, scored, and analyzed in large quantities to test multiple hypotheses.
3. It should address the full suite of common misconceptions with respect to evolution, as documented in the science education literature and captured in Table 1 (see Chapter 2 for discussion).

4. It should provide opportunities for each student to endorse both scientifically normative and unscientific-yet-intuitive concepts, permitting both to surface if they coexist (see Chapter 2 for the theoretical rationale behind this design specification). Although open-response questions or open-ended interviews are ideal for eliciting such dualities (Nehm & Schonfeld, 2008), they are disallowed by the goal to build an instrument comprising only closed-response items. Some items should therefore be closed-response yet *not* “forced choice” in the sense that the student is compelled to choose between scientific and intuitive ideas, but is freely allowed to endorse both.
5. It should contain other items that do ask the student to choose scientific concepts over intuitive ones. That is, in addition to “non-forced choice” items it should include some “forced choice” items obliging the student to identify scientific propositions amidst a field of intuitively appealing distracters. If, as the prevailing view holds, conceptual change entails the *restructuring* or *replacement* of a single mental model, then forced-choice items are necessary to reveal whether the desired restructuring or replacement has taken place. Likewise, if conceptual change instead requires a student to *override* or *suppress* his still extant intuitive conceptions – as I hypothesized in Chapter 2 – then this too will best be elicited by forced-choice items. In sum, the instrument should include both forced and non-forced choice items.
6. It should contain multiple items per conceptual category. Not only does such “internal replication” heighten test reliability (Gall, Gall, & Borg, 2007), but also invites variance into the data which can then be partitioned into pools and

subjected to statistical analysis. For example, the presence of dissectible variance permits factor analysis and cluster analysis (Bryant & Yarnold, 1995; Gall et al., 2007; Hair & Black, 2000), which might be used to reveal whether or not students' scientific and/or alternative conceptions tend to cluster into coherent and principled "theoretical orientations" (Shtulman, 2006; Shtulman & Schulz, 2008; Shtulman & Calabi, 2012). Part of my research agenda is to use my instrument to investigate the long enduring "theory theory" issue in the conceptual change literature (e.g., Carey, 2009; Carey & R. Gelman, 1991; diSessa, 2008; Gopnik & Meltzoff, 1997; Hirschfield & S. Gelman, 1994; Smith et al., 1993; Sperber et al., 1995; Vosniadou, 1994; Vosniadou et al., 2008; Wellman & S. Gelman, 1992; see Chapter 2 for discussion).

7. It should be evaluated through a strong validity and reliability study.

Table 2

Presence or Absence of Key Features in Seven Instruments for Assessing Student Understanding of Evolutionary Theory

Feature	CINS	CINS-MS/HS	B&A/ORI/ACORNS	J&F	Shtulman
Age-appropriate	–	+	+ / –	–	–
Quantitative / closed response	+	+	–	+ / –	+ / –
Most misconceptions allowed to surface	+	+	+	+	+
Non-forced choice items	–	–	+ / –	+ / –	+ / –
Forced choice items	+	+	–	+	+
Multiple items per conceptual category	–	–	+	+ / –	+
Evaluated for validity and reliability	+	+	+	–	+ / –

Note. CINS = Conceptual Inventory for Natural Selection (Anderson et al., 2002); CINS-MS/HS = CINS adapted for middle and high school students (Evans & Anderson, 2013); B&A/ORI/ACORNS = three closely related open-response instruments: Bishop and Anderson (1990), Nehm and Schonfeld's (2008) Open Response Instrument, and Nehm et al.'s (2012) Assessing COntextual Reasoning about Natural Selection; J&F = Jensen and Finley's (1995, 1996) instrument; Shtulman = Shtulman's (2006) instrument.

Development of a New Instrument

In search of a fitting instrument for my own research, I reviewed seven existing instruments for assessing student understanding of evolutionary theory. Despite their abundant strengths, none fully satisfied my seven required/desired criteria, as documented in Table 2. I therefore decided to develop a new instrument, custom-designed to match my research agenda and its underlying theory. Rather than build a test entirely from scratch, however, I borrowed many items from these existing instruments and adapted them for my own instrument.

Scientifically normative framework: Darwin's theory of evolution by natural selection. Table 3 outlines 11 essential scientific concepts which are targeted in my new instrument. Like many other researchers who have developed instruments in evolution education (e.g., Anderson et al., 2002; Nehm & Reilly, 2007; Shtulman, 2006), I fashioned my framework largely after the work of eminent biologist and historian Ernst Mayr. Mayr (1991) identifies 3 facts and 5 inferences that collectively guided Charles Darwin to his model of evolution via natural selection (Table 3). I also added three other core components of modern evolutionary theory: speciation, random mutation, and the random (i.e., non-directional) variation of offspring genotypes and phenotypes from their parents (Freeman & Herron, 2001; Futuyma, 1998; Mayr, 1991; Ridley, 2004).

Framework of intuitive preconceptions/misconceptions. In Chapter 2 I developed a taxonomy of common intuitive, scientifically non-normative preconceptions and misconceptions known to compromise many students' understanding of the theory of evolution. Table 1 presented that taxonomy. As explained in my literature review, these conceptions seem to arise from three deeply intuitive "modes of construal" (Keil, 1994):

Table 3

Essential Elements of the Theory of Evolution Targeted in the New Instrument

Components of Darwin's explanatory model of natural selection (Mayr, 1991)
SCI.1 – Fact #1: Potential for exponential increase in population size
SCI.2 – Fact #2: Steady-state stability of populations (fixed size; finite carrying capacity)
SCI.3 – Fact #3: Resource limitation (finite resources/niche spaces)
SCI.4 – Inference #1: Struggle for existence among individuals (intraspecific competition)
SCI.5 – Fact #4: Uniqueness of the individual (phenotypic variation)
SCI.6 – Fact #5: Heritability of much individual variation (genetic variation)
SCI.7 – Inference #2: Differential survival (natural selection, whenever heritable variation bears upon fitness; hence differential reproduction)
SCI.8 – Inference #3: Evolution over multiple generations (descent with modification; changing frequency of genes/traits; shift in mean genotype/phenotype)
Other elements of contemporary evolutionary theory (Freeman & Herron, 2001; Futuyma, 1998; Mayr, 1991; Ridley, 2004)
SCI.9 – Offspring variation relative to parental phenotype is random (non-directional)
SCI.10 – Random introduction of genetic variation (mutations)
SCI.11 – Allopatric speciation (evolutionary divergence of species in isolated habitats)

the projection of intentional agency, teleological directionality, and immutable essences onto biological phenomena (Evans, 2008; Shtulman, 2006; Sinatra et al., 2008; Smith, 2010). Individually or in combination, these three construals apparently conspire to yield distinct, recurring misunderstandings, not only for students but also adults, including scientists and scholars throughout history (Mayr, 1991).

Along with the scientific concepts in Table 3, the taxonomy of intuitions in Table 1 served as the framework for developing items for my new instrument. Each test item – or each alternative in the case of multiple choice questions – targets one of the following conceptual families: (1) a correct scientific understanding (SCI); (2) an intuitive belief that species are immutable (IMB); (3) an intuitive belief that evolution is the progressive transformation (TFM) of a species as a whole, rather than a statistical shift in trait frequencies and population membership; and (4) an intuitive belief that evolution occurs through within-lifetime adaptation (WLA) as individuals adapt to their environments during their own lifetimes and pass those changes to their offspring. The test items seek to sample across several sub-categories within these four families, as shown in Table 1 and Table 3: SCI.1 through SCI.11, IMB.1 through IMB.3, TFM.1 through TFM.3, and WLA.1 through WLA.3.

Instrument structure and sequence. I initially composed a large pool of age-appropriate, closed-response items: 17 “forced choice” items (multiple choice) and 85 “non-forced choice” (yes/no). Some were original, but most were derived from or inspired by source items published in existing instruments. To make the test simpler, more accessible, and faster for students, I built these test items around a single concrete scenario: microevolution and speciation of ghost crab populations on isolated islands. From this pool I selected a final set of 10 forced choice (FC) and 24 non-forced choice (NFC) items, with multiple items targeting each conceptual category. Table 4 displays the relative representation of scientific concepts (SCI.1-11) and the three sub-categories of misconceptions (IMB.1-3; TFM.1-3; WLA.1-3) in the final instrument. Appendix A presents the instrument itself.

Each FC item obliges the test-taker to select a scientifically normative statement amidst a field of two or more intuitively appealing distracters. Each NFC item is an isolated statement that embodies *either* a scientifically normative concept *or* an intuitively attractive but unscientific one, and the student is asked to judge whether it could be part of a biologist’s explanation of the phenomenon at hand. The student is thus obliged to make yes/no decisions, but these are “non-forced choice” in the sense that she is not strictly compelled to select scientific statements against unscientific ones; she can

Table 4

Representation of Scientific Concepts and Common Misconceptions in the Final Instrument

Number of opportunities to endorse					
Scientific	NFC	FC	Misconception	NFC	FC
SCI.1	1	0	Immutability		
SCI.2	1	0	IMB.1	1	3
SCI.3	1	2	IMB.2	2	3
SCI.4	3	2	IMB.3	1	1
SCI.5	5	7	Total (IMB)	4	7
SCI.6	2	4	Transformationism		
SCI.7	5	6	TFM.1	1	3
SCI.8	2	3	TFM.2	1	1
SCI.9	0	4	TFM.3	2	3
SCI.10	1	4	Total (TFM)	4	7
SCI.11	0	1	Within-lifetime adaptation		
Total (SCI)	21	33	WLA.1	2	5
			WLA.2	1	5
			WLA.3	1	2
			Total (WLA)	4	12
Item Totals					
All Misconceptions (IMB + TFM + WLA)				12	26
Scientific Concepts (SCI)				8[a]	10[a]
Nonsense Items (NON)				4	3

Note. For definitions of each scientific concept (SCI.1-11), see Table 3. For definitions of each misconception (IMB.1-3, TFM.1-3, and WLA.1-3), see Table 1.

[a] The total number of test items representing scientific concepts is less than the itemized sums for SCI.1-11 above, because some test items harbor more than one scientific concept.

endorse both. This format contrasts sharply with the multiple choice and Likert-style items employed in existing instruments (e.g., Anderson et al., 2002; Evans & Anderson, 2013; Jensen & Finley, 1995, 1996; Shtulman, 2006). Not only do those instruments oblige students to choose between scientific and unscientific conceptions, but also often force them to choose between more than one misconception that may be intuitively compelling. By contrast, the NFC items here permit students to embrace multiple models of evolution. The NFC section is designed to inventory any and all conceptions – whether scientific, intuitive, or both – that a student may hold (or generate on the spot; Vosniadou, 1994; Vosniadou et al., 2008). This feature follows from the hypothesis presented in Chapter 2 that even after a student undergoes conceptual change, she may continue to harbor her old intuitive mental model of evolution alongside her new scientific understanding.

The instrument was administered to participants online using Qualtrics software, which transferred data automatically and anonymously to the researcher. Qualtrics presented the 24 NFC items *one at a time* to diminish any sense of having to choose among competing statements, and in *random sequence* to dilute any order effects. In contrast, the FC section aims to heighten the sense of competition and choice between competing claims, and thus appeared only after students had completed all 24 NFC items.

NFC items have the following prompt: “How do you think a scientist – like a professional biologist who studies crabs – would explain this change? Please indicate whether each statement could be a part of the biologist’s explanation, or if it would not be part of her explanation.” Similarly, FC items have this prompt: “Please choose the answer that a biologist would select.” Such wording was adopted because the goal is not to assess students' personal acceptance or non-acceptance of evolution, but her conceptual grasp of scientific theory (Nehm, Beggrow, Opfer, and Ha, 2012). Asking students to adopt a biologist's vantage encourages them to hold personal beliefs aside.

The instrument also contains a set of "nonsense" statements/choices, which are illogical, neither scientific nor intuitive, and plainly false. These were included to encourage students to reject at least some of the non-forced choice statements. Also, they constitute a quality control mechanism: Any participant who accepts a significant fraction of these items would have to be suspected of poor effort, motivation, attention, English proficiency, or cognitive readiness, which might warrant omission of his/her data.

The final step of instrument development was revision and replacement of items per recommendations made by two content experts. I describe this vetting process later as part of instrument validation.

Scoring. Each test-taker receives four primary sub-scores, two for the NFC section and two for the FC section. The "NFC Scientific" score is the number of scientific statements endorsed (out of 8). The "NFC Intuitive" score is the number of unscientific statements endorsed (out of 12). The latter score can be subdivided for each specific family of misconceptions: immutability (IMB), transformationism (TFM), and within-lifetime adaptation (WLA). (Endorsements of "nonsense" statements are not scored.)

The "FC Scientific" score is the number of items (out of 10) on which the correct scientific statement was chosen over intuitive distracters. The "FC Intuitive" score is the number of items (out of 10) on which an intuitive distracter was chosen. Here again, incorrect responses can be sub-tallied as IMB, TFM, and WLA.

Further, the two scientific sub-scores can be combined into a single "Overall Scientific" score, equal to the total number of scientifically correct endorsements (out of 18 opportunities). The unscientific scores can be combined into a single "Overall Intuitive" score (out of 22 opportunities).

Data Collection and Analysis

Participants and Sample Size

As a practical necessity, I used a non-random sample of convenience. Striving for a demographically and academically diverse group of participants, I recruited a pool of secondary science teachers from twenty public high schools who were willing to administer the instrument to their students. All of these teachers were professional acquaintances of mine or my university adviser, or else colleagues of an acquaintance. I then followed each school district's formal application procedure for conducting research, tailoring a letter of parental consent and student assent for each. I had previously secured approval from The College of William & Mary Institutional Review Board to conduct this research with human subjects. Six of the twenty districts granted permission to pursue the study.

Seven science teachers at six schools from six different school districts invited their students to participate and secured the required parent, guardian, and student signatures. A total of 350 students ultimately took the test, and I was able to keep and use the responses of all but 10 of these, for an overall sample size of 340. Table 5 summarizes demographics and other characteristics of the schools and participating students.

Table 5

Summary of Participating School and Student Demographics and Other Characteristics

School			Participants									
Setting	Size	% Free or Reduced Lunch	Number of Participants	Mean Age	Mean Grade	% American Indian or Alaska Native	% Asian or Pacific Islander	% Black or African American	% Hispanic or Latino	% White	% Other	% Studied Evolution
Suburban	813	<1	118	15.9	10.7	3.4	27.6	2.6	4.3	81.9	1.7	100
Rural	1761	28	72	15.7	9.9	1.4	2.8	4.2	4.2	86.1	2.8	97.2
Suburban	1646	31	71	15.5	9.9	2.8	9.9	22.5	16.9	67.6	7.0	97.2
Urban	1537	41	47	16.8	10.5	13.0	4.3	58.7	4.3	26.1	19.6	71.1
Suburban	1936	24	24	16.6	10.5	17.4	13.0	17.4	17.4	65.2	26.1	87.5
Rural	1033	10	8	15.8	9.9	0.0	0.0	0.0	14.3	100.0	0.0	100.0
Overall		25	340	15.9	10.3	5.1	13.7	15.8	8.1	71.3	7.2	94.0

One school was a small school in an affluent west coast suburb (118 participants, 35% of the sample), and the other five were in the mid-Atlantic region: a large urban school with a substantial fraction of economically disadvantaged students (47 participants, 14% of the sample); two large suburban schools in middle to upper middle class communities (24 and 71 participants respectively, 28%); and two mid-sized to large rural schools with local populaces ranging widely in socioeconomic status (8 and 72 participants respectively, 24%). Participants ranged from age 14 to 18, 46% male and 54% female. Five percent identified themselves as American Indian or Alaska Native, 14% as Asian or Pacific Islander, 16% as Black or African American, 8% as Hispanic or Latino, 71% as White, and 7% as "Other," with 12 students identifying with two or more racial/ethnic groups.

Ten percent of participants were in the 9th grade, 58% in 10th grade, 20% in 11th, and 11% in 12th. Although I had hoped for a balanced mix of those who had and had not yet studied evolution in a biology class, the vast majority of participants (94%) already had.

Participant anonymity was protected by administering the instrument online using Qualtrics software; neither student names nor any other information that could be used to identify specific individuals were collected.

Instrument Validation

This study addresses three main research questions:

1. Does the instrument provide valid and reliable estimates of scientific and unscientific conceptions regarding the theory of evolution, for the target audience?

2. Do individual test items and item choices, including distracters, consistently map onto the intended concepts – that is, the scientifically normative conception *and* common intuitive preconceptions/misconceptions – held by each student?
3. What modifications are warranted to render the instrument more valid and reliable, and to align individual items with the intended concepts?

In educational testing, "test validity" refers to the accuracy and thoroughness with which an instrument measures the concepts, constructs, outcomes, or competencies that its users intend; specifically, it is the extent to which the interpretations and conclusions that a researcher draws from it are supported by prior evidence (Bryant, 2000; Gall et al., 2007). Following Gall, Gall, and Borg (2007), I originally proposed to gather three classes of such evidence: (1) *content-related evidence*, provided by content experts (professional scientists); (2) *evidence from internal structure*, generated by administering the test to members of the target audience and then subjecting the data to confirmatory factor analysis; and (3) *evidence from relationship to other variables*, specifically "convergent evidence," generated by triangulating test results against students' scores on two existing instruments that measure the same concepts: the Conceptual Inventory of Natural Selection (CINS; Evans & Anderson, 2013) and the Open Response Instrument (ORI, Nehm & Schonfeld, 2008). Unfortunately, logistical hurdles prevented my teacher volunteers from administering the CINS and ORI to their students as originally proposed.

Content-related evidence provided by content experts. On any test of conceptual understanding, the test items are designed to sample students' knowledge and comprehension within the conceptual domain of interest. *Content-related evidence* bears upon the question of how representative that sample is (Bryant, 2000; Gall et al., 2007). Items need not sample exhaustively from the conceptual domain, but they should provide a balanced, relevant, and representative sample of it. Whether they do so is typically judged by content experts in the field (Bryant, 2000; Gall et al., 2007). Two university professors – one a veteran instructor and researcher in evolutionary biology, the other a veteran instructor and researcher in evolutionary psychology – scrutinized my experimental instrument using a customized content validation form. The form asked these experts to identify which of the 11 science concepts (SCI.1-11; Table 3) was being targeted by each item and statement meant to represent the normative model of evolution by natural selection. They also rated the item’s scientific validity and its relevance to the target concept. They likewise judged each item and distracter meant to violate the normative scientific model, identifying the specific misconception being targeted (IMB.1-3, TFM.1-3, or WLA.1-3; Table 1) and rating its relevance. They were asked to recommend ways to make items more valid or relevant, and to comment on whether the items provided an adequately representative sample of the scientific and intuitive conceptual domains.

The two experts evaluated the instrument independently on paper at first, and then later through mutual dialogue with one another and the researcher. Items were modified in accord with their ratings and recommendations, and resubmitted for evaluation. Some items were stricken and replaced with new ones. After several cycles of feedback, discussion, and revision, consensus was reached and the finalized instrument (Appendix A) was submitted for testing with the target audience. The experts deemed this final version adequately representative of the two conceptual domains (scientific and intuitive), and this constitutes evidence of the instrument's content validity.

Evidence from internal structure. A second kind of evidence that can bear upon test validity is the emergence of inter-correlations among variables (test items) when the test is taken by members of the target audience (Gall et al., 2007). If participants reply to related questions in the same direction – that is, if they respond similarly to items meant to map onto the same concept or construct – then this suggests that the instrument aligns as intended with its underlying conceptual framework (Blunch, 2013; Bryant, 2000; Gall et al., 2007). Put differently, it boosts confidence that the instrument is measuring what it is supposed to measure, and that conclusions drawn therefrom are valid.

I used factor analysis and structural equation modeling to assess how well my conceptual frameworks (Tables 1 and 3) fit the responses of the 340 participants. Because my instrument was built from an a priori framework, I employed *confirmatory factor analysis* (CFA) rather than exploratory factor analysis or principle components analysis to assess dimensionality in the data (Blunch, 2013; Bryant & Yarnold, 1995). CFA has the additional benefit that it provides measures of instrument *reliability* (as distinct from validity), loosely defined as the consistency of responses that a test would yield when administered to the same audience in similar conditions on different occasions. And unlike conventional measures such as Cronbach's alpha which only estimate reliability for the instrument as a whole, CFA estimates reliability for each individual test item (Blunch, 2013).

In the following sections I document the process by which I evaluated the validity and reliability of my experimental instrument using CFA and other statistical methods.

Data Analysis

Culling, coding of variables, and imputation of missing values. Once the instrument had been administered to all participants, I imported the raw data from Qualtrics into Microsoft Excel, where I organized it by school and checked for anomalies that might disqualify certain participants' responses. Of the 350 test takers, I eliminated 10 who did not finish the test. I also flagged 58 participants who had endorsed 3 or more of the 7 "nonsense" items as candidates for omission, since that might be a sign of poor effort, motivation, attention, English proficiency, or cognitive readiness. However, after discussion with my doctoral committee, I decided not to disqualify these participants, because other factors – such as confusing wording or simply misreading a question – might account for such mistakes. Moreover, students of limited English proficiency or cognitive readiness *are* part of the target population for which the instrument was designed, and their responses therefore bear upon the instrument's validity.

In this instrument, all variables (test items) are categorical – some dichotomous (yes/no), some nominal (A, B, C, D, or E). Because CFA requires data on a continuous or ordinal scale (Blunch, 2013; Bryant & Yarnold, 1995), I coded each student response as 0 or 1, thereby rendering items as binary variables. For the 20 NFC items, I coded each "yes" as 1 and each "no" as 0. For the 10 FC items, I coded *each* choice (A, B, C, D, and E) as either 0 or 1, rendering them as 37 *separate* binary variables.

Across the dataset, missing values were rare (43 out of 6800 responses), isolated, and scattered with no apparent pattern, hence probably neither systematic nor veiling significant information that could bias analysis and interpretation. However, missing values prevent the statistical software (SPSS AMOS package) from providing modification indices, which can be helpful in deciding whether to strike certain test items or how to modify the CFA model. I therefore used model-based regression imputation to replace the 43 missing values prior to each CFA run (Arbuckle, 2014).

Correlation matrices, Cronbach's alpha, and other statistics. For every possible bivariate combination of items, I calculated the phi correlation coefficient (ϕ_2), the statistic of choice for "true dichotomous" variables (ϕ_2 ranges from -1 to +1 and is equivalent to Pearson's r; Gall et al., 2007; Zar, 1997). These were entered into a correlation matrix. I also calculated the correlation between each item and the mean of all the other items within that same conceptual domain (SCI, IMB, TFM, or WLA). These are point-biserial coefficients (r_{pbis}), the bivariate statistic of choice when one variable is "true dichotomous" and the other continuous (r_{pbis} ranges from -1 to +1 and is equivalent to Pearson's r; Gall et al., 2007). By gauging the degree of inter-correlation within each conceptual cluster (SCI, IMB, TFM, or WLA), these phi and point-biserial coefficients serve as indicators of the validity of each hypothesized latent factor, while also flagging individual items for possible revision, removal, or replacement (Blunch, 2013). I also calculated the frequency (%) with which students endorsed each NFC statement and FC choice. This serves as an indicator of within-item variance, also helpful in identifying items for possible revision or replacement (Blunch, 2013).

Cronbach's alpha (α, 0 to 1 scale) was calculated for each latent dimension (that is, each sub-score: SCI, IMB, TFM, and WLA) as an estimation of internal consistency, hence *reliability* for that portion of the instrument (it is not appropriate to calculate α for the entire instrument when multiple dimensions are being measured; Field, 2009; Gall et al., 2007). Cronbach's alpha also serves indirectly as an indicator of construct validity, as low alpha is a symptom of poor inter-correlation within that conceptual cluster (Blunch, 2013). (Technically the statistic here was Kuder-Richardson 20, a relative of Cronbach's alpha that SPSS automatically calculates whenever the data is binary.)

Finally, overall scores and sub-scores were calculated for each participant, and averaged across the entire sample (n = 340).

CFA: Models and sequence. I conducted CFA on the NFC (non-forced choice) and FC (forced choice) sections separately, in effect treating them as two separate instruments, because the difference in format could in theory elicit different latent dimensions. In particular, the forced-choice structure of the FC section is designed to yield a sharp inverse correlation between scientific and unscientific conceptions, whereas the NFC section deliberately allows for the possibility that these conceptions coexist in the student's understanding, which thus ought to dampen that correlation. In a sense, then, the two sections are designed to mobilize two different cognitive phenomena (if they exist; see Chapter 2 for theoretical discussion). The NFC section seeks to gauge the degree to which a student *harbors* both scientific and intuitive understandings of evolution. The FC section seeks to gauge the degree to which a student *suppresses* the latter in favor of the former. These two phenomena are unlikely to reveal themselves identically even if they are tapping into the same underlying mental models. The CFA model that best fits the NFC data may therefore differ from the CFA model that best fits the FC data – their shared conceptual framework notwithstanding – and for this reason I analyzed the two sections in isolation from one another.

I began with the 20 NFC items (excluding the 4 "nonsense" items), assessing the fit of two "base models" to the data. The first base model is represented schematically in Figure 7, using the conventional notation for structural equation modeling (Blunch, 2013; Bryant & Yarnold, 1995; Klem, 2000). Large ovals signify the four hypothesized latent factors (SCI, IMB, TFM, WLA) around which the instrument was developed. Rectangles represent the actual test items that were intended to map or "load" onto each latent factor, hence the actual data drawn from 340 participants. Unidirectional arrows signify that the latent factor is hypothesized to influence or drive how students respond to each item; that is, variance in the latent factor causes or "explains" variance in each test item. Ovals on the left are error terms, comprising any variance not explained by the underlying latent factor, including both random measurement error and systematic variance stemming from sources outside the latent factor. Later schematics will contain curved two-head arrows signifying hypothesized correlations between latent factors. No correlations are posited in this initial base model.

This first "base model" is the simplest model that can be derived from the original conceptual framework that drove instrument development. It is the fundamental model for which instrument validity and reliability is sought. When subjected to confirmatory factor analysis, each of the four hypothesized latent factors was fit to its corresponding data in isolation from the other three constructs. No correlations or higher order relationships are posited between latent factors. The goal was simply to assess how well each test item maps onto its intended construct.

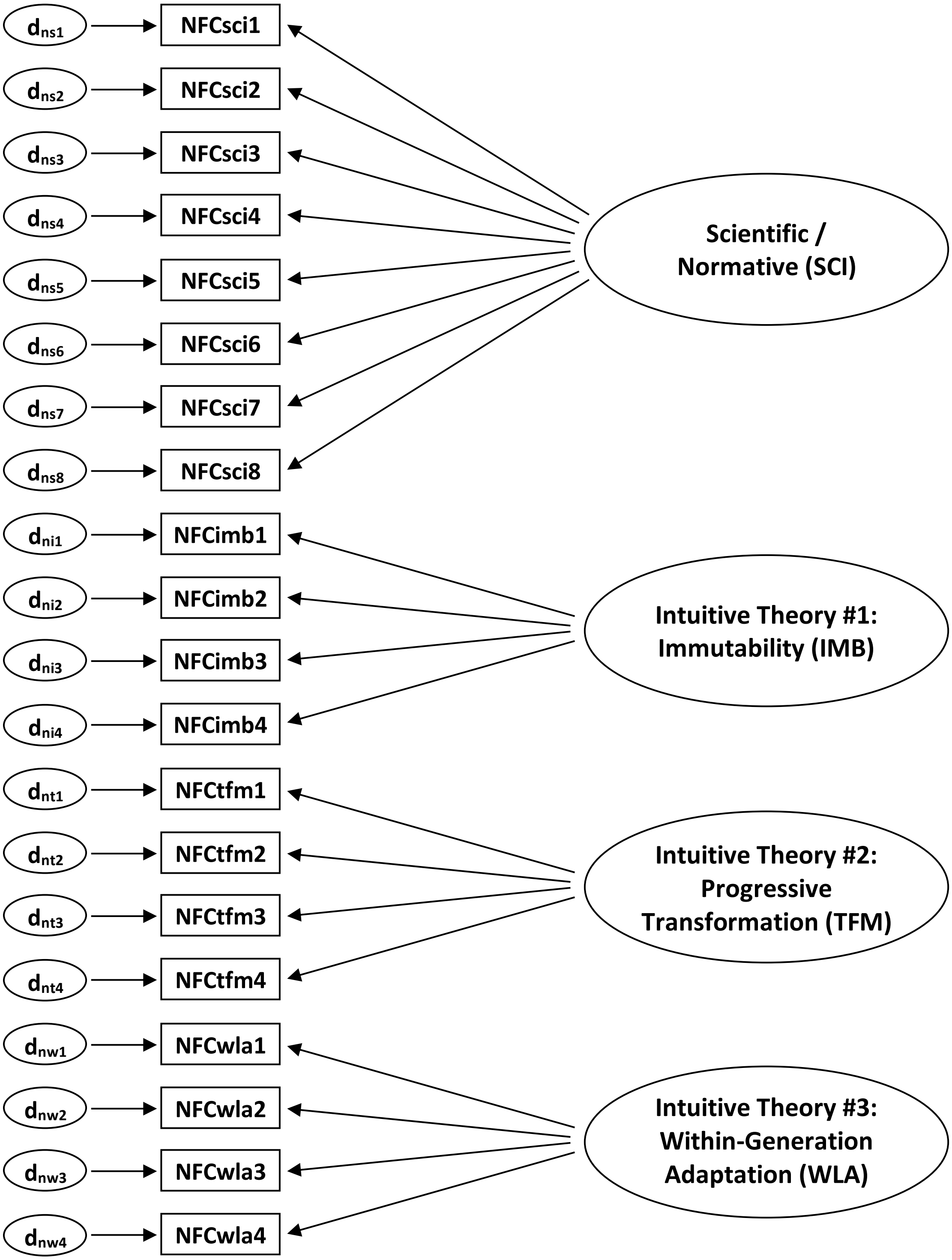


Figure 7. Base Model #1 for confirmatory factor analysis via structural equation modeling. Large ovals are the hypothesized latent factors. Rectangles are the 20 measured variables (test items) that are intended load onto the latent factors. Small ovals are error terms.

I interpreted the loadings, fit indices, and other statistics – including the phi correlations, point-biserial correlations, and Cronbach's alpha – to judge the validity and reliability of the instrument as a whole and of each individual test item. I also used these statistics to identify items that should be removed, replaced, or revised.

I then repeated the process for the second base model. I will present schematics for that model and every other model tested in Chapter 4, where I will augment them with the factor loadings and fit statistics emerging from each CFA run. After assessing the two base models, I analyzed the loadings, diagnostics, and other statistics for clues about other relationships that might exist between the latent factors and variables. I then modified the models in ways that were warranted not only empirically, but also *theoretically* (Blunch, 2013; Bryant & Yarnold, 1995; Klem, 2000; Thompson, 2000), and subjected these revised models to CFA. I will save the rationales for testing each model until Chapter 4. Finally, as many scholars urge (e.g., Anderson, 2008; Thompson, 2000), I compared the fit of these "rival" models and interpreted them theoretically.

I then repeated the analysis on data from the FC section of the instrument.

CFA: Estimation methods and assumptions. I initially ran each CFA model in SPSS AMOS using Maximum Likelihood Estimation (MLE). A strong assumption of MLE, however, is multivariate normality, which requires that all variables (test items) conform to a normal distribution and that all bivariate pairs regress linearly onto one another (Bryant & Yarnold, 1995; Blunch, 2013). Because binary data obeys a binomial distribution, it is inherently non-normal, as tests for multivariate normality on my data confirmed.

I therefore attempted CFA with Asymptotically Distribution-Free Estimation (ADF), which makse no distributional assumptions (Arbuckle, 2014; Blunch, 2013). Unfortunately, ADF generally requires an enormous sample size (n > 2000) and AMOS was unable to find a solution on my small dataset (n = 340). I also attempted CFA using Bayesian Estimation, with my variables recoded as ordinal-categorical, another method that does not assume multivariate normality (Arbuckle, 2014; Blunch, 2013). Unfortunately, in all cases AMOS failed to converge on a solution.

I then conducted CFA on several models using Unweighted Least Squares Estimation (ULS), which unlike MLE makes no distributional assumptions (Blunch, 2013). ULS is less desirable than other methods in that it cannot estimate statistical significance for loadings and correlations, nor provide fit indices and modification indices. Nevertheless, I used it as a check against output under MLE. The parameter estimates under ULS were very similar to those under MLE, rarely varying by more than .05 and never more than .10 on a standardized scale of 0 to 1. I interpret this as evidence that MLE – despite the violation of multivariate normality – provided acceptable solutions for my models and data, as its output is nearly identical to a method that does not assume multivariate normality. In Chapter 4, therefore, I will report MLE results with cautious confidence, rather than ULS estimates, since MLE can provides fit indices, p-values, and modification indices.

Finally, in an effort to minimize the ill effects of non-normality, I used bootstrapping (1000 re-samplings) after each MLE run to push parameter estimates as close as possible to their ideal values (Blunch, 2013). All tables and statistics in Chapter

4 report the bootstrapped estimates. In each run I also asked AMOS for various fit indices and modification indices, which I will describe and interpret in Chapter 4.

Internal and External Validity

Because the participant samples here were not drawn randomly, estimates of validity and reliability will generalize only to the (unknown and unknowable) population represented by the sample, and not cleanly to future audiences. Instrument validity and reliability are not inherent properties of the test, but specific to the sample and whatever “accessible population” it represents (Gall et al., 2007). External validity – i.e., generalizability – is therefore severely compromised in this study.

Internal validity refers to a study’s internal structure and whether there is adequate researcher control over unwanted variation within and between treatment groups (Gall et al., 2007). Since this study did not employ comparison groups, it is immune to most internal validity threats. One important threat, however, is variation in testing conditions across the six schools, including possible inconsistencies in instrument administration by the classroom teachers. To guard against this threat, all teacher volunteers were given the same careful guidelines. Students took the test on school computers, never at home or elsewhere, so the “online environment” was presumably consistent across sites – with one major exception: One of the two rural schools was unable to provide computers for testing, and its 72 participants took the test on paper instead. Because the online version randomizes the sequence of NFC items, these students were given 10 different paper versions, each with a different, randomized order of NFC items. No student names or other identifying information were recorded. Interpretation of the results, including any attempts to generalize findings to future student audiences, may be compromised by this unwanted variation in test administration.

These threats to internal and external validity demand that judicious constraints be placed on the interpretation and communication of results. I will address this in Chapters 4 and 5.

Preview of Chapter 4

The next chapter reports scores and descriptive statistics from the 340 high school students who took my experimental test, drawn from six different schools serving diverse communities from rural to urban and from economically disadvantaged to affluent. I interpret analytical statistics that bear upon instrument validity and reliability, including the results of confirmatory factor analysis on the two hypothetical base models described above. I then describe a series of post hoc modifications to these two conceptual models, made in light of empirical CFA output coupled with theoretical reasoning. One of these modified models yielded a superior fit to the data, and I discuss this revelation vis-à-vis the original conceptual framework that guided test design. I summarize evidence both in support of instrument validity and against it, at least for whatever audience my nonrandom sample might represent. Finally, I critically examine individual test items of questionable validity and reliability, and propose recommendations for removing, revising, or replacing them.

Chapter 4: Results and Interpretation

In this chapter I present the results of field-testing my experimental instrument with 340 high school science students and analyzing the data for evidence of instrument validity and reliability via confirmatory factor analysis and other methods. I describe post hoc modifications to the instrument's original conceptual model, made in light of empirical CFA output coupled with theoretical reasoning. Also based on statistical results, I vet individual test items of concern and offer recommendations for revising or replacing them. Once again, the primary research questions are:

1. Does the instrument provide valid and reliable estimates of scientific and unscientific conceptions regarding the theory of evolution, for the target audience?
2. Do individual test items and item choices, including distracters, consistently map onto the intended concepts – that is, the scientifically normative conception *and* common intuitive preconceptions/misconceptions – held by each student?
3. What modifications are warranted to render the instrument more valid and reliable, and to align individual items with the intended concepts?

The instrument (Appendix A) is divided into two distinct sections:

1. "Non-forced choice" (NFC): twenty isolated statements, each representing either a scientifically normative conception of the theory of evolution or an

intuitively appealing but unscientific conception, with the student asked either to endorse each statement as scientifically valid, or reject it. In this section the student is not obliged to choose directly between scientific statements juxtaposed to unscientific ones; he may freely endorse both.

2. "Forced choice" (FC): ten multiple choice questions in which the student is obliged to select a scientific statement amidst a field of intuitively appealing distracters.

The scientific statements are hypothesized to embody a single scientific dimension (SCI), while each unscientific statement/distracter is hypothesized to represent one of three intuitive dimensions: a position that species are immutable (IMB), a position that evolution transforms whole species as a unit (transformationism, TFM), or a position that evolution occurs through individual organisms adapting to local circumstances within their lifetimes (within-lifetime adaptation, WLA). Each test taker receives four primary scores: NFC scientific, NFC intuitive, FC scientific, and FC intuitive. There are also sub-scores for each intuitive dimension (IMB, TFM, WLA).

I analyze the NFC results first, followed by the FC results.

Non-Forced Choice Section (NFC): Descriptive and Diagnostic Statistics

Table 6 shows the frequency with which participants endorsed each NFC statement, as well as their mean scores on each scale and sub-scale. Most of the students (94%) had already studied evolution in a biology class, and on average, participants correctly endorsed 81% of scientific statements. When presented with an unscientific but intuitively appealing statement, students rejected that statement only about half the time on average, and this was true for all three intuitive families (IMB, TFM, and WLA).

Table 6

Non-Forced Choice Section (NFC): Endorsement Frequency for Each Test Item, Mean Participant Scores (Scientific vs. Intuitive), and Mean Sub-Scores (Immutability, Transformationism, and Within-Lifetime Adaptation)

Scientific Statements		Intuitive Statements					
Item	%	Item	%	Item	%	Item	%
SCI		IMB		TFM		WLA	
NFCsci1	76.5	NFCimb1	38.5	NFCtfm1	43.8	NFCwla1	63.8
NFCsci2	81.8	NFCimb2	52.6	NFCtfm2	42.6	NFCwla2	40.0
NFCsci3	76.2	NFCimb3	59.7	NFCtfm3	81.5	NFCwla3	30.3
NFCsci4	84.1	NFCimb4	48.8	NFCtfm4	55.3	NFCwla4	64.4
NFCsci5	80.3	IMB sub-score		TFM sub-score		WLA sub-score	
NFCsci6	80.3	Mean	49.9	Mean	55.8	Mean	49.6
NFCsci7	85.6	SD	31.8	SD	26.6	SD	29.6
NFCsci8	82.4	SE	1.72	SE	1.44	SE	1.60
Scientific score, total		Intuitive score, total					
Mean	80.9	Mean	51.8				
SD	19.6	SD	20.2				
SE	1.07	SE	1.09				

In what follows, I present additional descriptive and diagnostic statistics, especially with the aim of assessing the validity of individual test items and the latent dimensions to which they are hypothesized to belong. Tables 6, 7, 8, and 9 contain useful data for identifying individual test items that may need to be revised, replaced, or removed from the instrument.

Endorsement Frequencies

The frequency of endorsement (Table 6) serves as a proxy for each item's variance. Items near the extremes – i.e., endorsed by a vast majority or small minority of participants – may be less useful for discriminating between test-takers (Blunch, 2013). All eight SCI items have a somewhat unfavorably high frequency (low variance), endorsed by 76-87% of participants. Again, however, 94% of them had already studied evolution. The instrument is intended for use in before-after intervention studies, and

Table 7

Cronbach's Alpha for Each Sub-Scale and Effect on Alpha of Deleting Items

Scientific			Intuitive		
Sub-scale / Item	Cronbach's α	α if item deleted	Sub-scale / Item	Cronbach's α	α if item deleted
Scientific	.573		Immutability	.537	
NFCsci1		.568	NFCimb1		.437
NFCsci2		.539	NFCimb2		.353
NFCsci3		.568	NFCimb3		.518
NFCsci4		.563	NFCimb4		.534
NFCsci5		.550	Transformationist	.279	
NFCsci6		.508	NFCtfm1		.266
NFCsci7		.514	NFCtfm2		.222
NFCsci8		.506	NFCtfm3		.235
			NFCtfm4		.172
			Within-lifetime	.467	
			NFCwla1		.312
			NFCwla2		.424
			NFCwla3		.474
			NFCwla4		.361

presumably more favorable variances would have emerged had the instrument been piloted with a higher representation of those who had not yet studied evolution. The intuitive items were all endorsed by a moderate number of participants (30-70%), with one exception: TFM3 (82%).

Cronbach's Alpha

Table 7 gives Cronbach's alpha for each sub-scale, a measure of "internal consistency" as an index of instrument reliability. Although none of the alphas are good (α > .7 is desirable; Field, 2009), TFM shows an especially worrisome lack of internal consistency. No single item appears markedly culpable, however, as alphas cannot be raised by removing any item (except WLA3, and only negligibly: from .467 to .474).

Table 8

Each Individual Item's Correlation with the Mean Score of All Other Items Belonging to the Same Sub-Scale/Conceptual Domain (Point-Biserial Correlations, r_{pbis})

SCI		IMB		TFM		WLA	
Item	r_{pbis}	Item	r_{pbis}	Item	r_{pbis}	Item	r_{pbis}
NFCsci1	$.20^{***}$	NFCimb1	$.36^{***}$	NFCtfm1	$.11^{*}$	NFCwla1	$.35^{***}$
NFCsci2	$.28^{***}$	NFCimb2	$.44^{***}$	NFCtfm2	$.15^{**}$	NFCwla2	$.24^{***}$
NFCsci3	$.21^{***}$	NFCimb3	$.25^{***}$	NFCtfm3	$.13^{*}$	NFCwla3	$.18^{***}$
NFCsci4	$.20^{***}$	NFCimb4	$.24^{***}$	NFCtfm4	$.18^{***}$	NFCwla4	$.30^{***}$
NFCsci5	$.24^{***}$						
NFCsci6	$.37^{***}$						
NFCsci7	$.36^{***}$						
NFCsci8	$.38^{***}$						

$^{*}p < .05$. $^{**}p < .01$. $^{***}p < .001$.

Within-scale Correlations

Table 8 presents another useful statistic for identifying items of concern: the strength of correlation between each individual item and all other items belonging to the same sub-scale and conceptual domain. For example, the 340 participant responses to test item SCI1 have a .20 correlation with the 340 mean scores of items SCI2 through SCI8. Stronger correlations are desirable, as this suggests that the hypothesized latent factor is coherent and perhaps unidimensional, hence valid (Blunch, 2013). The TFM inter-correlations are weakest, suggesting that the TFM items do not form a very tight conceptual cluster.

Matrix of Inter-item Correlations

In a valid instrument with valid constructs, every test item will correlate positively with every other test item within each hypothesized latent factor (Blunch, 2013). Table 9 shows that this is so for all four hypothesized factors (SCI, IMB, TFM, and WLA). All eight scientific items correlate positively with one another, forming an inter-correlative cluster, although many of the correlations are weak. The same is true

Table 9

Matrix of Correlations between Non-Forced Choice Items (Phi Correlations, ϕ_2)

Test Item	SCI								IMB				TFM				WLA			
	NFC sci1	NFC sci2	NFC sci3	NFC sci4	NFC sci5	NFC sci6	NFC sci7	NFC sci8	NFC imb1	NFC imb2	NFC imb3	NFC imb4	NFC tfm1	NFC tfm2	NFC tfm3	NFC tfm4	NFC wla1	NFC wla2	NFC wla3	NFC wla4
SCI																				
NFCsci1	1	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
NFCsci2	.10^{ns}	1	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
NFCsci3	.06^{ns}	.15**	1	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
NFCsci4	.08^{ns}	.05^{ns}	.00^{ns}	1	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
NFCsci5	.13*	.11*	.07^{ns}	.08^{ns}	1	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
NFCsci6	.10^{ns}	.17**	.22***	.18**	.17**	1	-	-	-	-	-	-	-	-	-	-	-	-	-	-
NFCsci7	.17**	.19***	.08^{ns}	.19***	.15**	.24***	1	-	-	-	-	-	-	-	-	-	-	-	-	-
NFCsci8	.09^{ns}	.24***	.14*	.18***	.18**	.25***	.28***	1	-	-	-	-	-	-	-	-	-	-	-	-
IMB																				
NFCimb1	-.01^{ns}	-.09^{ns}	-.12*	-.05^{ns}	-.08^{ns}	-.20***	-.15**	-.21***	1	-	-	-	-	-	-	-	-	-	-	-
NFCimb2	-.05^{ns}	-.05^{ns}	.02^{ns}	-.08^{ns}	.00^{ns}	-.08^{ns}	-.14*	-.15**	.34***	1	-	-	-	-	-	-	-	-	-	-
NFCimb3	.12*	-.06^{ns}	.03^{ns}	.01^{ns}	.06^{ns}	.09^{ns}	-.05^{ns}	-.01^{ns}	.17**	.31***	1	-	-	-	-	-	-	-	-	-
NFCimb4	-.02^{ns}	.04^{ns}	-.06^{ns}	-.04^{ns}	-.10^{ns}	.00^{ns}	-.08^{ns}	-.17	.22***	.23***	.07^{ns}	1	-	-	-	-	-	-	-	-
TFM																				
NFCtfm1	-.19***	.02^{ns}	-.13*	-.11^{ns}	-.10^{ns}	-.08^{ns}	-.12*	-.15**	.16**	.11*	.05^{ns}	.08^{ns}	1	-	-	-	-	-	-	-
NFCtfm2	-.09^{ns}	.03^{ns}	-.03^{ns}	.0^{ns}	-.06^{ns}	-.12*	-.04^{ns}	-.07^{ns}	.03^{ns}	.08^{ns}	-.03^{ns}	.09^{ns}	.14*	1	-	-	-	-	-	-
NFCtfm3	.09^{ns}	.06^{ns}	.10^{ns}	.17**	-.10^{ns}	.08^{ns}	.09^{ns}	.03^{ns}	-.03^{ns}	-.07^{ns}	.05^{ns}	.03^{ns}	.02^{ns}	.02^{ns}	1	-	-	-	-	-
NFCtfm4	.08^{ns}	.02^{ns}	.07^{ns}	.12*	.07^{ns}	.10^{ns}	.02^{ns}	.07^{ns}	.03^{ns}	-.01^{ns}	.00^{ns}	-.03^{ns}	.05^{ns}	.10^{ns}	.21***	1	-	-	-	-
WLA																				
NFCwla1	-.09^{ns}	.03^{ns}	-.06^{ns}	-.01^{ns}	-.09^{ns}	-.03^{ns}	-.11*	-.12*	.03^{ns}	.09^{ns}	-.07^{ns}	.13*	.21***	.23***	.11*	.25***	1	-	-	-
NFCwla2	-.07^{ns}	-.10^{ns}	-.13*	-.05^{ns}	-.14**	-.13*	-.11*	-.11*	.12*	.11*	.06^{ns}	.14**	.15**	.02^{ns}	-.08^{ns}	-.07^{ns}	.23***	1	-	-
NFCwla3	-.02^{ns}	-.08^{ns}	-.10^{ns}	-.16**	-.01^{ns}	-.13*	-.03^{ns}	-.08^{ns}	.18***	.04^{ns}	.07^{ns}	.19***	.11*	.18**	-.03^{ns}	.08^{ns}	.10^{ns}	.15**	1	-
NFCwla4	-.04^{ns}	.04^{ns}	.01^{ns}	.00^{ns}	-.04^{ns}	.04^{ns}	-.06^{ns}	-.09^{ns}	.08^{ns}	.04^{ns}	.02^{ns}	.17**	.19***	.25***	.03^{ns}	.20***	.35***	.12*	.13*	1

^{ns} not significant. $^{*}p < .05$. $^{**}p < .01$. $^{***}p < .001$.

of all three intuitive domains, but once again the TFM inter-correlations appear especially weak. Correlations between scientific and non-scientific items are generally weak, sometimes positive and sometimes negative. Most WLA items correlate positively with IMB and TFM items, suggesting that the intuitive items may form a larger, higher-order inter-correlative cluster. However, TFM correlations with IMB items are weak and often negative. Overall, TFM items seem the least patterned, correlating poorly with one another and erratically with the other three domains.

Non-Forced Choice Section (NFC): Model Analysis and Modification

In this section I present the results of confirmatory factor analysis on the instrument's non-forced choice section, starting with a pair of "base models" derived from the theoretical blueprint that guided test design (Table 1). I then describe a sequence of post hoc modifications to these base models, made in light of empirical CFA output coupled with theoretical reasoning. The model that ultimately best fit the data fused the transformationist latent factor (TFM) and the within-lifetime adaptationist factor (WLA) into a single "Lamarckian" factor.

Confirmatory Factor Analysis: NFC Base Model #1

The first "base model" tested (Figure 8) is the fundamental model that can be derived from the original conceptual framework that guided instrument development. At this point no correlations or higher order relationships are posited between latent factors. The goal is simply to assess how well each test item maps onto its intended latent factor in isolation from the other three latent factors.

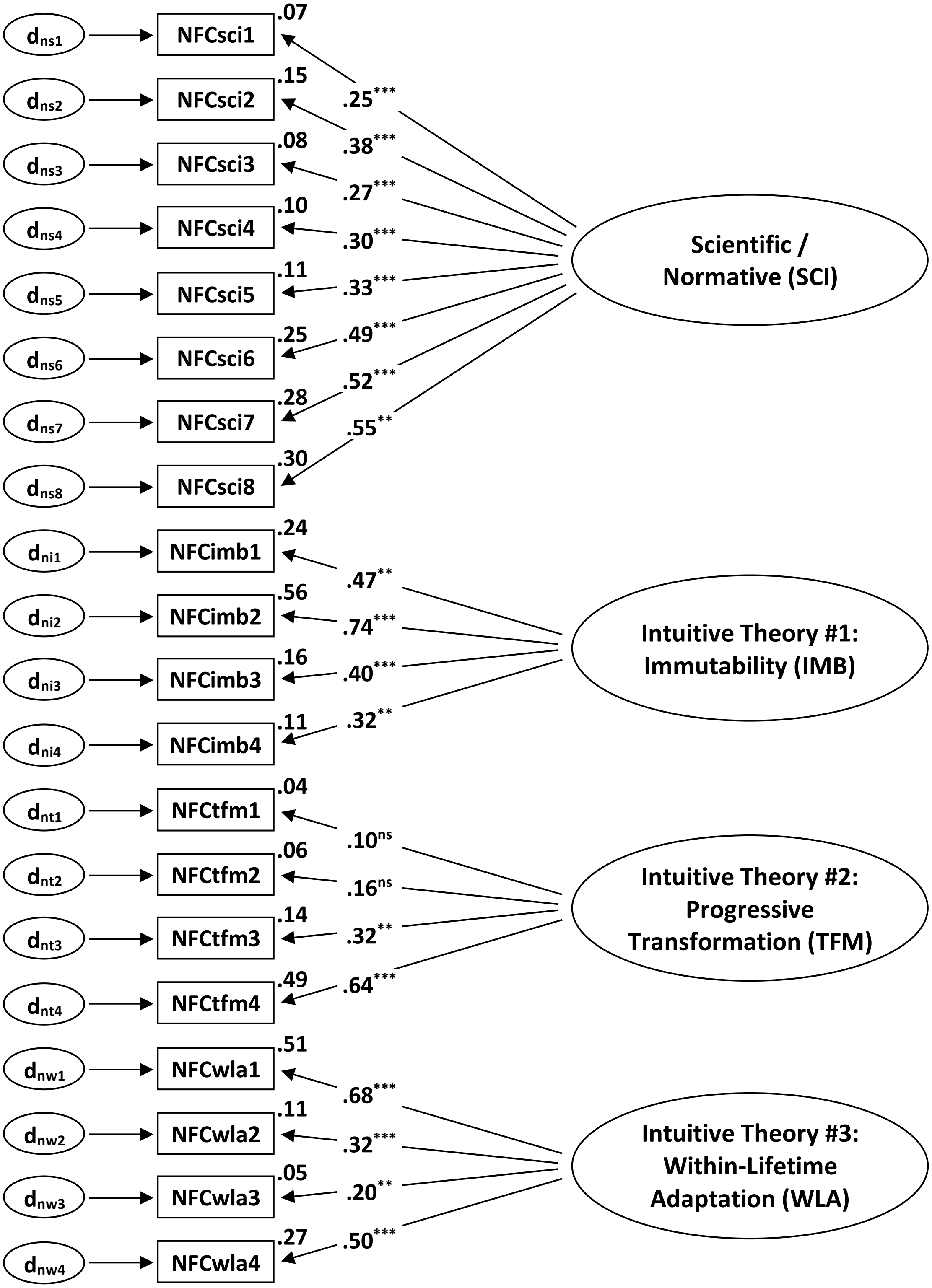


Figure 8. Results of confirmatory factor analysis on NFC Base Model #1. The number overlapping each arrow is the item's loading onto its latent factor (a standardized regression coefficient). The number on each item's upper right corner is the Squared Multiple Correlation, equivalent to R^2. Significance: *p < .05, **p < .01, ***p < .001, ns = not significant.

Loadings and goodness-of-fit. Figure 8 displays the loading (λ) of each test item onto its hypothesized latent factor (standardized regression weights, 0 to 1 scale, displayed over each arrow). It also gives the corresponding squared multiple correlation (equivalent to R^2, 0 to 1 scale, upper right corner of each variable), which estimates the fraction of variance in the test item that is "explained" by variance in the latent factor. These statistics are also listed in Table 10. Table 11 gives seven different fit indices as gauges of how well the model fits the data. Because no single fit index optimally reflects the consonance of a model with data, I have reported multiple indices (Blunch, 2013; Klem, 2000). Some indices suggest a satisfactory fit (χ^2/DF, PGFI), others not quite satisfactory, though close (CMIN/DF, PCFI, RMSEA; see note below Table 11 for details on interpreting these fit indices).

Many of the loadings are weak – though usually statistically significant – with latent factors explaining less than 20% of the variance in 12 of 20 test items. Once again, it is the TFM items that map least well onto the hypothesized latent factor. The TFM1 and TFM2 loadings, alone among 20 test items, are statistically nonsignificant. Also highly suspect are SCI1, SCI3, SCI4, SCI5, IMB4, and WLA3.

Modification indices. Besides loadings and fit indices, I asked SPSS AMOS to provide modification indices for this Base Model. Modification indices identify changes to the model that, if implemented, would improve its fit to the data (specifically χ^2). Most of the proposed modifications were either meaningless, such as positing a causal arrow (loading) between one test item and another, or else theoretically dubious, such as positing a correlation (two-headed arrow) between an error term and a different test item or latent factor (Blunch, 2013; Bryant & Yarnold, 1995; Klem, 2000; Thompson, 2000).

Table 10

Factor Loadings and Squared Multiple Correlations (R^2) for Five CFA Models

	Base Model #1		Base Model #2		Model #3		Model #4		Model #5	
Test Item	λ	R^2	λ	R^2	λ	R^2	λ	R^2	λ	R^2
Scientific										
SCI1	$.25^{***}$	.07	$.25^{***}$	.07	$.26^{***}$	.07	$.26^{***}$	.07	$.25^{***}$	.07
SCI2	$.38^{***}$	.15	$.38^{***}$	.15	$.36^{***}$	.13	$.36^{***}$	.15	$.36^{***}$	.14
SCI3	$.27^{***}$	.08	$.27^{**}$	.08	$.28^{**}$	.08	$.28^{**}$	.08	$.27^{**}$	.08
SCI4	$.30^{***}$	.10	$.30^{***}$	.10	$.30^{***}$	.10	$.30^{***}$	.10	$.30^{***}$	.10
SCI5	$.33^{***}$	.11	$.32^{***}$	.11	$.33^{***}$	.11	$.33^{***}$	.11	$.32^{***}$	.11
SCI6	$.49^{***}$	.25	$.49^{***}$	.25	$.48^{***}$	.23	$.48^{***}$	.25	$.47^{***}$	.23
SCI7	$.52^{***}$	.28	$.52^{***}$	.28	$.52^{***}$	.28	$.52^{***}$	.28	$.53^{***}$	.28
SCI8	$.55^{**}$	.30	$.54^{***}$	.30	$.57^{***}$	.33	$.57^{***}$	.30	$.57^{***}$	.33
Immutability										
IMB1	$.47^{**}$	.24	$.30^{**}$	.11	$.40^{**}$	.18	$.40^{**}$	.24	$.57^{***}$	.33
IMB2	$.74^{***}$	.56	$.29^{*}$	.10	$.36^{**}$	.15	$.36^{**}$	.56	$.65^{***}$	.43
IMB3	$.40^{***}$	.16	$.12^{ns}$	.03	$.15^{ns}$	.04	$.15^{ns}$	.16	$.34^{***}$	.13
IMB4	$.32^{**}$	.11	$.35^{**}$	.13	$.38^{***}$	.15	$.38^{***}$	.11	$.37^{***}$	.14
Transformationism										
TFM1	$.10^{ns}$	.04	$.35^{***}$	.13	$.37^{***}$	.14	$.33^{***}$	.11	$.36^{***}$	.14
TFM2	$.16^{ns}$	.06	$.35^{***}$	.13	$.31^{***}$	.11	$.38^{***}$	.15	$.38^{***}$	.15
TFM3	$.32^{**}$	.14	$.07^{ns}$	.01	$.00^{ns}$	.01	$.13^{**}$	.02	$.09^{ns}$	.02
TFM4	$.64^{***}$	.49	$.24^{*}$	.07	$.14^{ns}$	.03	$.34^{***}$	.12	$.29^{**}$	.09
Within-Lifetime Adaptation										
WLA1	$.68^{***}$	.51	$.52^{***}$	.29	$.44^{**}$	.21	$.65^{***}$	.43	$.63^{***}$	.40
WLA2	$.32^{***}$	.11	$.29^{***}$	.09	$.33^{***}$	.12	$.24^{***}$	.07	$.29^{***}$	.09
WLA3	$.20^{**}$	.05	$.31^{***}$	.11	$.34^{***}$	.12	$.24^{**}$	.06	$.27^{***}$	.08
WLA4	$.50^{***}$	.27	$.51^{***}$	.27	$.42^{***}$	.19	$.56^{***}$	.32	$.54^{***}$	.30

*p < .05, **p < .01, ***p < .001, ^{ns} not significant

Table 11

Fit Indices for Five CFA Models

Fit Index	Base Model #1	Base Model #2	Model #3	Model #4	Model #5
χ^2	196	194	193	193	191
χ^2 / DF	1.15	1.14	1.14	1.14	1.14
CMIN/DF	2.07	2.10	1.98	1.72	1.58
PCFI	.579	.569	.604	.682	.714
RMSEA	.056 (p = .11)	.057 (p = .08)	.054 (p = .23)	.046 (p = .76)	.041 (p = .94)
PGFI	.731	.729	.728	.743	.737
AIC_C	625	631	614	566	557

Note. χ^2 = Bollen-Stine Chi-square goodness of fit index after bootstrapping. All χ^2 values here are large and highly significant (p < .001), which signifies poor fit. However, the large sample sizes needed for CFA inevitably inflate χ^2, and for that reason Chi-square is not an ideal fit index for structural equation modeling (Blunch, 2013; Bryant & Yarnold, 1995; Klem, 2000). χ^2/DF divides Chi-square by the degrees of freedom for a more appropriate measure, ideally < 3 (Klem, 2000). CMIN/DF near 1.0 indicates good fit, and can go quite high (Blunch, 2013). PCFI and PGFI = parsimony adjusted comparative fit and goodness of fit indices; values > .60 indicate satisfactory fit (Blunch, 2013). RMSEA = root mean square error of approximation; values < .05 with high p-values indicate good fit (Blunch, 2013). By itself, AIC_C – the consistent Akaike information criterion – is a relative fit measure; it tells nothing about an isolated model's fit, but is the most appropriate statistic for comparing the fit of one or more non-nested models (Anderson, 2008). A lower AIC_C indicates improved fit. PGFI can also be used to compare models, but other indices are not suitable for this.

*p < .05, **p < .01, ***p < .001, ^{ns} not significant

Even so, these proposals could be clues to which items and factors behave least predictably within the given system. TFM and WLA appeared far more often in the indices than SCI or IMB items.

Other modification indices were more theoretically tenable, proposing that a test item be loaded onto a *different* latent factor, or that a correlation be posited between two latent factors. TFM1, which hardly loads at all onto its own latent factor, would load much better onto IMB or WLA, while TFM2 and TFM4 would load better onto WLA. Conversely, WLA1 and WLA4 would load better onto TFM than their own latent factor. Most striking, the single modification that would, by far, most improve the model's fit would be a correlation term (two-headed arrow) placed between the TFM and WLA latent factors – that is between an intuitive theory of progressive transformationism and an intuitive theory of within-lifetime adaptation.

Collectively, then, the modification indices suggest that the model's fit might especially improve if the TFM and WLA components of the model were "set free" (Blunch, 2013) to associate with items and factors outside their hypothesized domains, perhaps especially with one another. I explore such modifications in later models.

Confirmatory Factor Analysis: NFC Base Model #2

The second "base model" (Figure 9) does permit TFM items and WLA items to interact with one another and other items in the system. This base model reflects the prevalent view among conceptual change scholars that a student's pre-instructional conceptions about evolution (or any other scientific model) all stem from a singular intuitive (mis)understanding of the phenomenon. The schematic in Figure 9 ascribes all student misconceptions to a single "intuitive theory" (INTV), comprising all intuitions driven by the essentialist, teleological, and intentional stances (see Chapter 2 and Table 1). A negative correlation is not (yet) postulated between a scientific understanding and an intuitive one, since the NFC section of my instrument deliberately allows for the possibility that even after conceptual change, an intuitive understanding may continue to coexist for a student alongside her new scientific understanding; they need not be inversely related (see Chapter 2).

The SCI loadings are identical to Base Model #1, as this portion of the model was not altered. The side-by-side comparisons in Tables 10 and 11 show that the loadings of the 12 intuitive test items are perhaps slightly weaker than they were in Model #1, while the fit indices are virtually identical (PGFI and AIC_C are the most suitable indices for

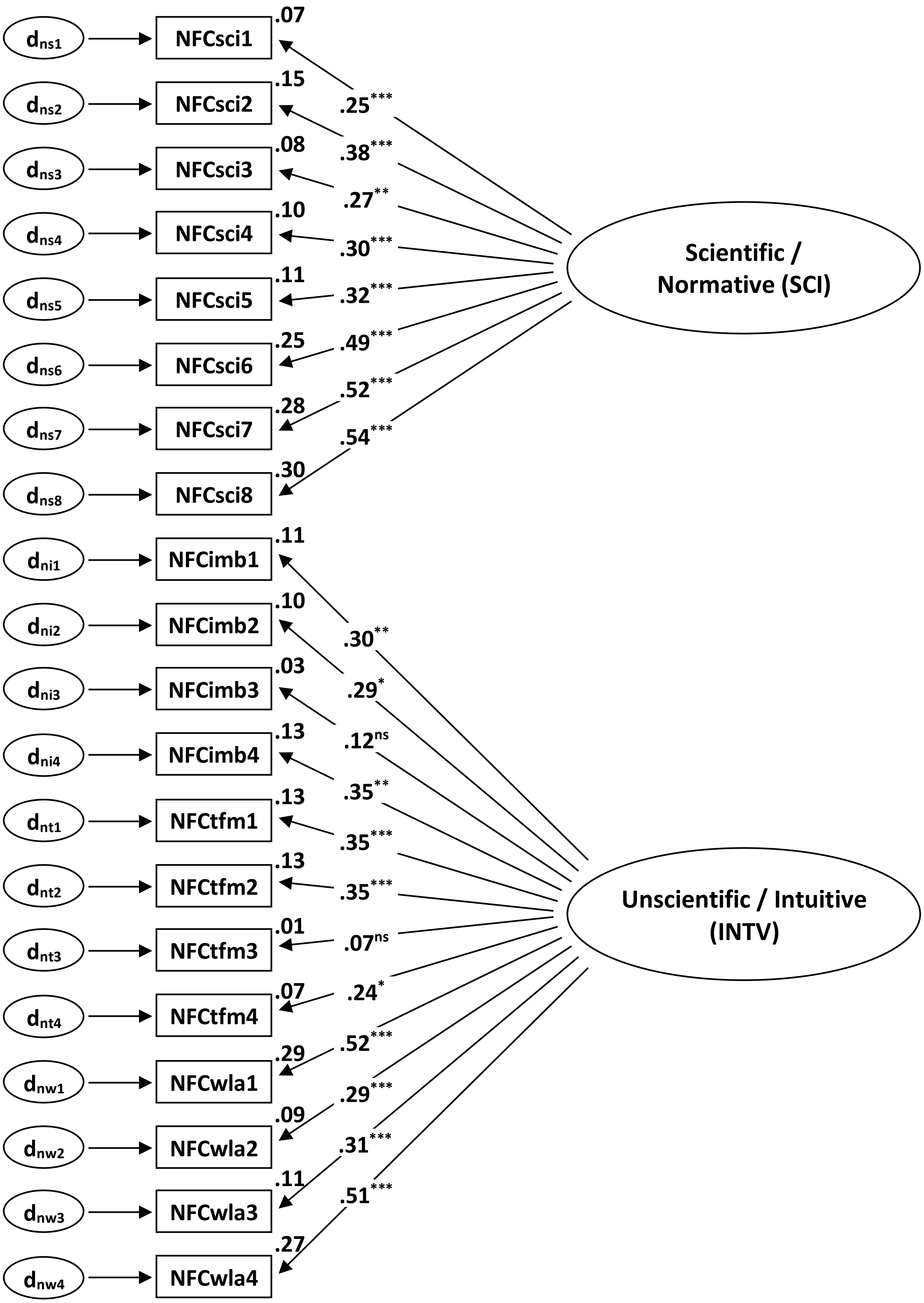


Figure 9. Results of confirmatory factor analysis on NFC Base Model #2. The number overlapping each arrow is the item's loading onto its latent factor (standardized regression coefficient). The number on its upper right corner is the Squared Multiple Correlation, equivalent to R^2. Significance: *p < .05, **p < .01, ***p < .001, ns = not significant.

comparing two models; Anderson, 2008). The TFM loadings remain most questionable, although TFM1 and TFM2 have switched places with TFM3 and TFM4 as the weakest and least significant. Most striking, the loadings of three of the four IMB items – IMB1, 2, and 3 – are dramatically diminished in Model #2. It is possible that this change stems from IMB's new interaction with TFM and WLA items, since both a transformationist stance and within-lifetime adaptationist stance would accept that species are mutable, whereas an immutability stance denies it. I explore this possibility in later models.

In the modification indices, IMB items now appear far more frequently than TFM and WLA items, whereas in Model #1 it was the converse. Once again most of the modifications proposed are either meaningless or theoretically dubious. However, there is one tenable proposal that would substantially improve model fit: the positing of a correlation (two-headed arrow) between the SCI and INTV latent factors – that is, between a scientific understanding and an intuitive one. That proposal was put to the test in the next model.

Confirmatory Factor Analysis: Modifications to the Base Models

NFC Model #3: A correlation between SCI and INTV. Base Model #2's modification indices recommended the insertion of a correlation term between the SCI and INTV latent factors. If such a correlation were negative, it would accord with the prevailing view that conceptual change amounts to a *restructuring* or *replacement* of a singular mental model: The student either transforms his pre-instructional conception into a scientific one, or else abandons the former in favor of the latter. After conceptual change, the old intuitive mental model would not – as hypothesized in Chapter 2 – continue to coexist for the student alongside his new scientific understanding.

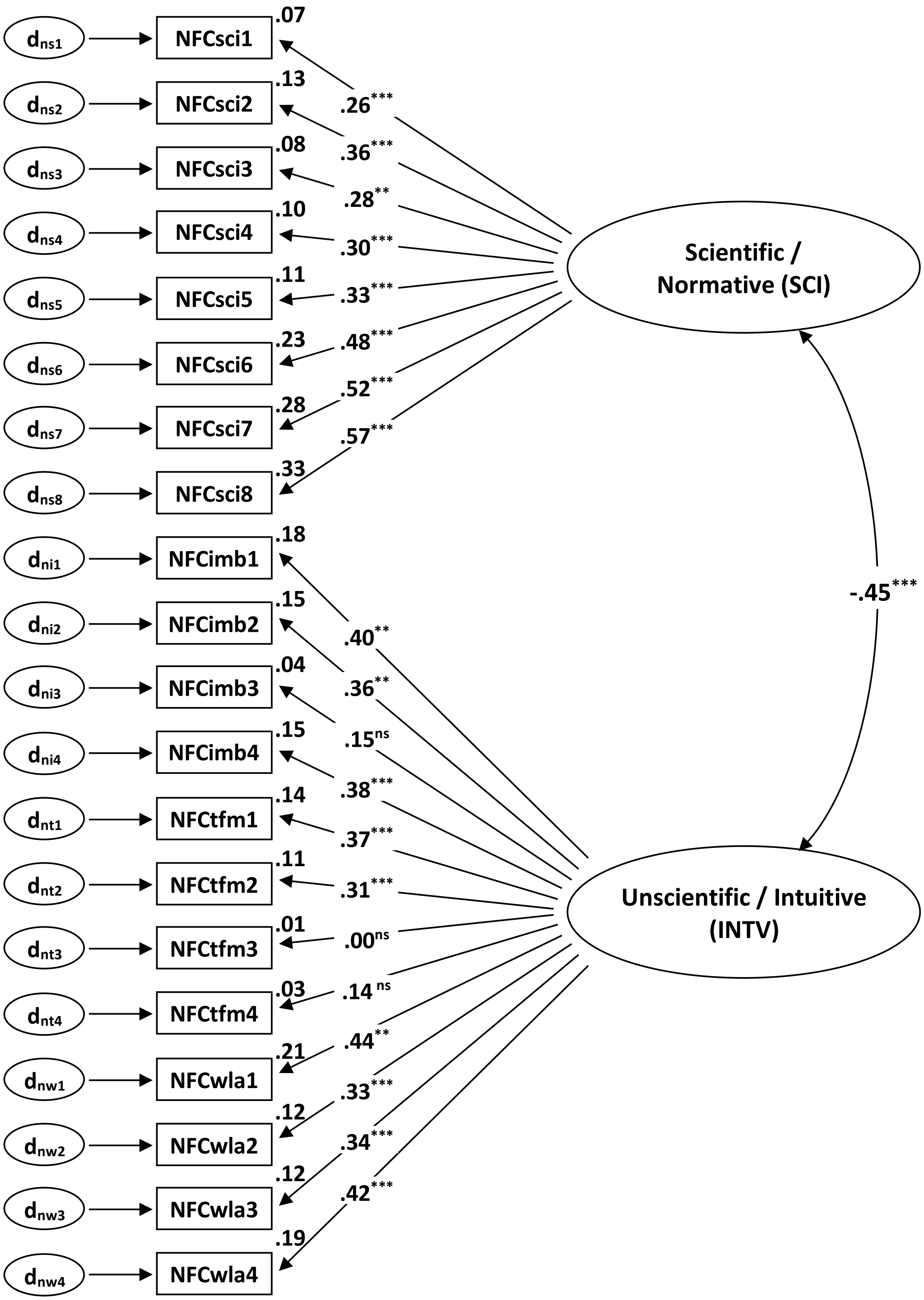


Figure 10. Results of confirmatory factor analysis on NFC Model #3. The number overlapping each arrow is the item's loading onto its latent factor. The number on its upper right corner is the Squared Multiple Correlation, equivalent to R^2. A correlation of -.45 exists between SCI and INTV. Significance: *p < .05, **p < .01, ***p < .001, ns = not significant.

There is, then, both theoretical and empirical justification (i.e., Model #2's modification indices) for adding this correlation term (Blunt, 2013; Bryant & Yarnold, 1995; Klem, 2000; Thompson, 2000). It appears in Figure 10 as a curved two-headed arrow joining the SCI and INTV ovals. The loadings were essentially unchanged (Table 10), but a statistically highly significant, *negative* correlation of -.45 emerged between SCI and INTV. This change also modestly improved the model's fit (e.g., AIC_C dropped from 631 to 614; Table 11).

Modifications to Base Model #1: Other correlations. On the same theoretical grounds – namely, that an inverse relation may exist between scientific and intuitive conceptions – I attempted a pair of changes to Model #1 and its four standalone factors (SCI, IMB, TFM, and WLA). First, I posited (negative) correlations between SCI and each of the three intuitive terms. AMOS was unable to resolve this model, even after various manipulations to render it soluble, and it was abandoned. Next, I posited the existence of a superordinate "global" factor (again dubbed INTV), with IMB, TFM, and WLA nested as sub-factors beneath it. This model, too, failed to converge and was abandoned. I then pursued another theoretically tenable lead that emerged under the Base Models: the possibility that a transformationist stance (TFM) and/or within-lifetime adaptationist stance (WLA) might correlate *negatively* with an immutability stance (IMB), since the former accept the mutability of species while the latter denies it. This model, too, failed to converge and was abandoned.

NFC Models #4 and #5: A new Lamarckian factor. Finally, my Base Model analyses prompted me to pursue one more modification: the merging of TFM and WLA into a single latent factor. Not only was this move empirically supported – e.g., Base

Model #1's modification indices promised that correlating TFM and WLA would substantially improve fit – but also theoretically defensible. Historically, transformationist notions have often gone hand-in-hand with notions of behavior-based adaptation (both are teleological), most famously in Jean Baptiste Lamarck's pre-Darwinian model of evolution (Mayr, 1991; I discuss Lamarck's important theory in Chapter 5). Model #4 (Figure 11) therefore now fuses TFM and WLA into a single "Lamarckian" factor, dubbed LMK.

The loadings are comparable to those in Model #1 (Table 10), but the fit is considerably closer (Table 11), as AIC_C has dropped from 625 to 566 and PGFI has increased from .731 to .743. The .682 PCFI is quite satisfactory (> .60 is considered good fit), while RMSEA has now crossed the conventional good-fit cutoff of .05. Moreover, RMSEA's now high p-value of .76 is also very favorable (as with χ^2, the higher p, the better the fit).

Finally, the modification indices recommend adding correlation terms between all 3 latent factors. This is consistent with empirically and theoretically tenable modifications already explored in earlier models. These new correlations were predicted in Model #5 (Figure 12): (1) an inverse relation between scientific and intuitive factors, and (2) an inverse relation between the immutability factor (IMB) and the new Lamarckian factor (LMK = TFM + WLA), which accepts the mutability of species.

Statistically significant correlations did indeed emerge between all three latent factors. The correlations between SCI and both IMB and LMK are negative, as expected (-.37 and -.27 respectively). The correlation between IMB and LMK, however,

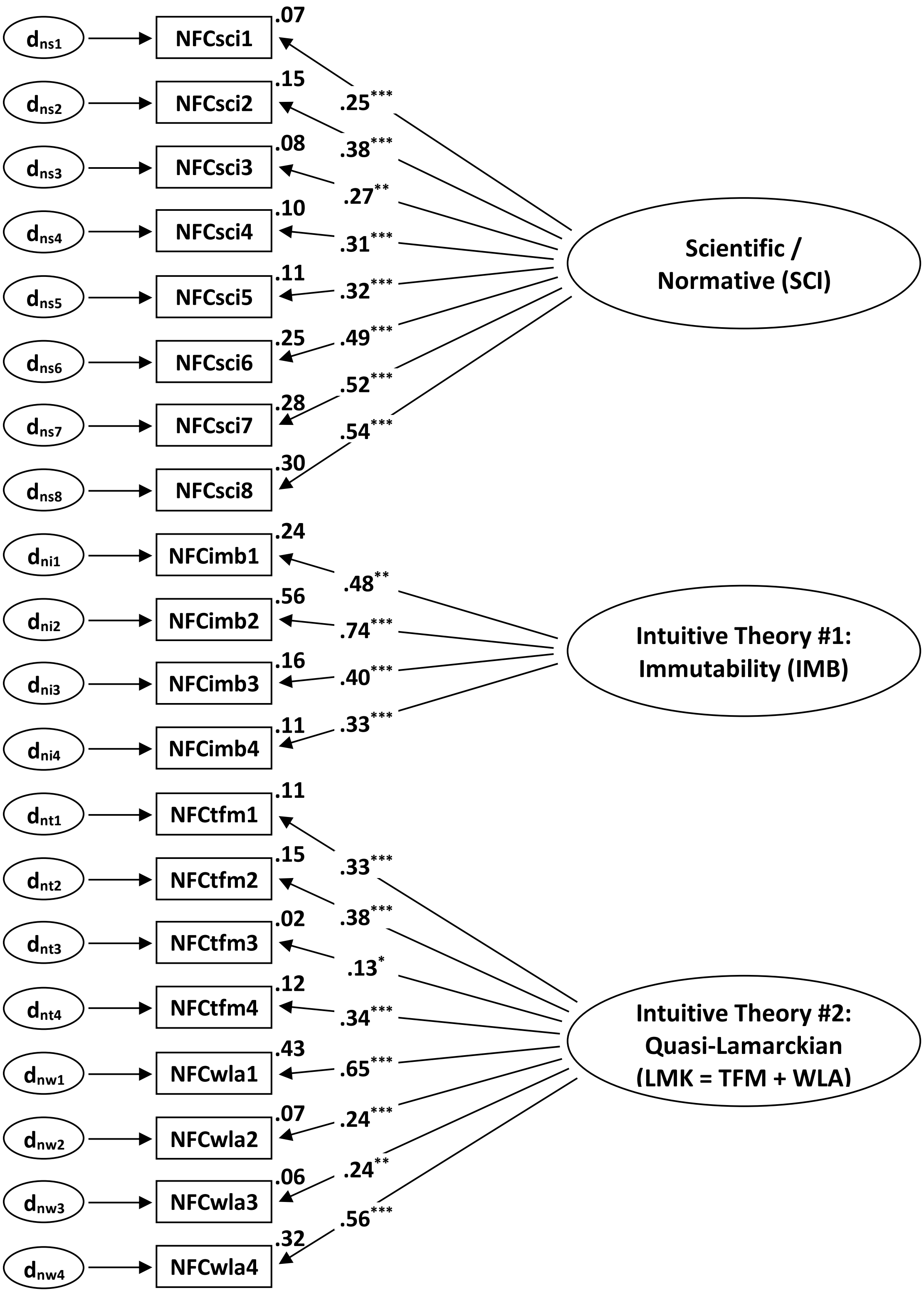


Figure 11. Results of confirmatory factor analysis on NFC Model #4. The number overlapping each arrow is the item's loading onto its latent factor (standardized regression coefficient). The number on its upper right corner is the Squared Multiple Correlation, equivalent to R^2. Significance: *p < .05, **p < .01, ***p < .001, ns = not significant.

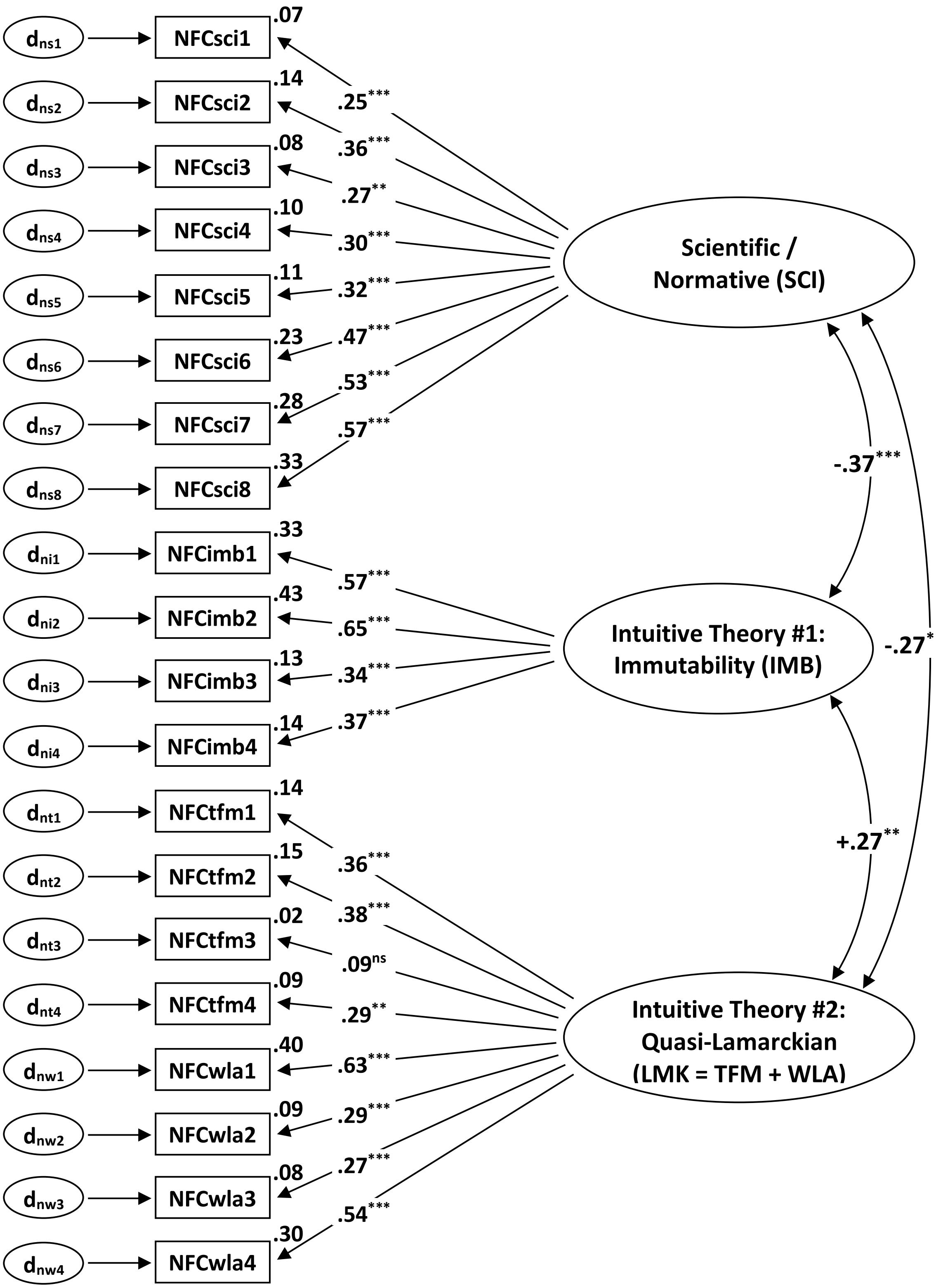


Figure 12. Results of confirmatory factor analysis on NFC Model #5. The number overlapping each arrow is the item’s loading onto its latent factor (standardized regression coefficient). The number on its upper right corner is the Squared Multiple Correlation, equivalent to R^2. Significance: *p < .05, **p < .01, ***p < .001, ns = not significant.

is positive (+.27) – not what was expected but consistent with the underlying theory that both immutability and transformationist intuitions are driven by an essentialist stance toward species as “natural kinds” (see Chapter 2 and Table 1). Model #5 has the best fit indices of any models tested (Table 11): high PCFI, fairly low CMIN/DF, low RMSEA with a very high (hence good) p-value of .94, and the lowest AIC_C of the 5 models.

As for individual items, the validity of all four TFM items remain highly questionable. Items WLA2, WLA3, IMB3, and IMB4 also warrant reevaluation, as do SCI1 through SCI5.

Summary and Conclusions: CFA Analysis and Model Evolution (NFC Section)

Table 12 presents a descriptive comparison of the five CFA models tested (for details see Figures 8-12 and Tables 10 and 11).

Sequence and strategy: Empirical evidence coupled to theoretical reasoning. Statisticians who have expertise in structural equation modeling, confirmatory factor analysis, and comparative model testing (e.g., Anderson, 2008; Arbuckle, 2014; Blunch, 2013; Bryant & Yarnold, 1995; Klem, 2000; Thompson, 2000) urge a cautious approach to making post hoc modifications to one’s CFA models: In deciding how to modify and retest one’s models, one should not go haphazardly on whatever empirical clues emerge in the CFA output (e.g., modification indices), for that risks capitalizing on perceived patterns that were really produced by random chance. Rather, one should only pursue modifications that are theoretically or at least pragmatically tenable as well – that is, plausible in the real world, especially given the conceptual framework within which one is operating. It was in this spirit that I progressed step-by-step from Model #1 to #5.

Table 12

Descriptive Comparison of the Five CFA Models Tested (NFC Section)

	Model #1	Model #2	Model #3	Model #4	Model #5
Hypothesized latent factors	Scientific (SCI) vs. three intuitive stances: Immutability (IMB), Transformationism (TFM), and Within-Lifetime Adaptation (WLA)	Scientific (SCI) vs. Unscientific (INTV)	Scientific (SCI) vs. Unscientific (INTV)	Scientific vs. two intuitive stances: Immutability (IMB) and Quasi-Lamarckism (LKM)	Scientific vs. two intuitive stances: Immutability (IMB) and Quasi-Lamarckism (LMK)
Description	This is the base model, positing four standalone factors – SCI, IMB, TFM, and WLA – around which clusters of test items were developed. Correlations are not posited between latent factors, because the NFC section was designed to allow for the possibility that a student may simultaneously hold both a scientific conception and one or more intuitive conceptions.	This model posits only two latent factors: a scientific understanding (SCI) versus unscientific (INTV, comprising all IMB, TFM, and WLA test items). This model conforms to the common view that a student holds only one singular mental model of the phenomenon in question: either scientific or unscientific – or else a transitional hybrid of both. No correlations posited.	Same as Model #2, except that an inverse correlation is predicted between the two latent factors. This accords with the prevailing view that conceptual change entails replacing an intuitive understanding with a scientific one, or else restructuring the former into the latter.	Same as Model #1, except that progressive whole-species transformation (TFM) has been fused with individual within-lifetime adaptation (WLA) into a single quasi-Lamarckian factor (LMK), echoing Jean Baptiste Lamarck's famous (but discredited) pre-Darwinian theory of evolution. No correlations posited.	Same as Model #4, but with a negative correlation predicted between a scientific understanding and each of the two intuitive mental models: IMB and LMK. A correlation is also allowed between IMB and LMK.
Correlations emerging between factors	(none posited)	(none posited)	Strong inverse correlation between SCI and INTV	(none posited)	Inverse correlation between SCI and both IMB and LMK. Positive correlation between IMB and LMK.
Goodness of fit	Marginal	Marginal	Improved	Good	Strong

I began by assessing the two NFC "base models" (Figures 8 and 9) that logically stem from the conceptual framework that guided the design of my instrument. The NFC items appear to the test-taker as isolated statements that she can either endorse or reject. The order is randomized, and she can freely endorse *both* scientific and unscientific but intuitively appealing propositions. This is in accord with the hypothesis developed in Chapter 2 that scientific and intuitive conceptions may continue to coexist for a student even after undergoing conceptual change. In both base models, therefore, the scientific (SCI) and intuitive factors (IMB, TFM, WLA, INTV) were held separate from one another. There was no expectation that they would correlate inversely.

The fit of the two base models to the data was marginal (Table 11). A majority of test items loaded weakly onto their intended latent factors and the various fit indices judged them borderline satisfactory at best. Neither model was superior to the other. When I analyzed the statistics and diagnostics for signs of other relationships that might exist among the latent factors and test items, three theoretically plausible possibilities emerged:

1. A correlation (presumably negative) between scientific and intuitive items/latent factors. Not only is such an inverse relationship theoretically plausible; it is in fact the prevailing view among conceptual change researchers, who conceive conceptual change as either the replacement or a restructuring of a student's pre-instructional mental model: After genuine conceptual change, the old intuitive model does not continue to coexist alongside the correct scientific one.

2. A correlation between immutability items/factors (IMB) and transformationist (TFM) and/or within-lifetime adaptationist (WLA) items/factors. That correlation might be positive, with all three factors belonging to the same overall intuitive model of evolution. However, the second base model had already cast doubt on that possibility, as IMB loadings had dramatically diminished when hypothesized to share a common global factor (INTV) with TFM and WLA items. Alternatively, the correlation might plausibly be negative, since the transformationist and within-lifetime adaptationist stances do accept that species are mutable, whereas an immutability stance denies it.
3. A correlation (presumably positive) between transformationist (TFM) and within-lifetime adaptationist (WLA) items/factors. Such a relation seems theoretically plausible not only because both stances might belong to the same overall intuitive model of evolution, but also because both are teleological and historically have often gone hand-in-hand, most famously in Lamarck's theory of evolution.

I then ran a series of new models that incorporated each of these candidate relationships into the base models. The fit of Base Model #2 was modestly improved by allowing the SCI and INTV factors to correlate (see Model #3 in Figure 10 and Table 11). The correlation that emerged was negative as expected, and substantial (-.45). A much greater improvement in fit, however, was achieved by incorporating all three candidate relationships into Model #5: (a) correlations between scientific and intuitive factors, (b) the fusion of TFM and WLA into a single quasi-Lamarckian factor (LMK), and (c) a correlation between IMB and the newly fused TFM and WLA factors (Figure

12 and Table 11). An inverse correlation emerged between the scientific factor and both intuitive factors, as expected, while a positive correlation emerged between the immutability factor and the new quasi-Lamarckian one.

Conclusion. Although none of these changes did much to improve the loading of test items onto their latent factors, Model #5 may be the theoretical framework for which the NFC section of my pilot instrument is "most valid." That tentative finding is gratifying because a quasi-Lamarckian mental model was one of the a priori models that I set forth in my original research proposal. In fact, all of the models tested were close approximations of a priori models that I had proposed. Yet here my decision to test each model was based not only on a priori theory, but also a posteriori results that were emerging throughout the analytical process: empirical evidence coupled to theoretical reasoning. That Model #5 emerged from amidst a field of principled models may be the strongest evidence in support of its validity:

> The most persuasive case that a model has been correctly specified is created when a researcher finds a differentially better fit of a given model as against the fit of numerous other defensible, thoughtfully formulated, rival plausible models; therefore, multiple models should be evaluated in an SEM project. (Thompson, 2000, p. 269; also see Anderson, 2008, and Blunch, 2013)

Conversely, the fact that Model #5 is theoretically and empirically defensible, and fits the experimental data reasonably well, speaks to the *instrument's* validity as well. Of course, the poor loading of many individual test items onto their latent factors signifies a lack of validity in the instrument. I will turn to that issue shortly. But the reasonably close fit to the data of a parsimonious and theoretically sound conceptual framework – a framework that moreover is consistent with the instrument's main purpose of identifying both scientific conceptions and common intuitive misconceptions – at least serves as some small evidence of the instrument's validity (Blunch, 2013; Gall et al., 2007).

Cautions. Even so, it is important to point out that I tested only a dozen or so of the "infinitely many models [that] can fit any given dataset…[and] the fit of a single tested model may always be an artifact of having tested too few models" (Thompson, 2000, pp. 277-278; also see Anderson, 2008, and Blunch, 2013). My decision to use CFA rather than exploratory methods such as EFA and PCA was a principled one: I had built the instrument from an a priori theory and conceptual framework, and so did not deem it appropriate to adopt an exploratory search for dimensionality in the data (Anderson, 2008; Blunch, 2013; Bryant & Yarnold, 1995). The post hoc modifications that led to Model #5 were all made within the confines of the antecedent theory. A consequence, however, is that I may be leaving better models, including conceptually sound ones, unexplored. I can only conclude that the favorable fit of Model #5 is "not inconsistent with" the proposition that my instrument exhibits a degree of construct validity. It provides supportive "evidence from internal structure" (Gall et al., 2007) that the instrument as a whole measures what was intended, and that future researchers could use it to draw valid conclusions about whether a student holds a correct scientific understanding of the theory of evolution and/or certain common misunderstandings of it.

The next step in instrument development should be a cross-validation study: field-testing with a new audience and comparison of results – including a reevaluation of Model #5 and the others (Blunch, 2013; Klem, 2000; Thompson, 2000). But first, some test items will have to be revised or replaced. I turn to the question of validity and reliability for individual test items next.

Non-Forced Choice Section (NFC): Validity and Revision of Individual Test Items

An important function of instrument validation is to judge the validity and reliability of individual test items, and to identify items that should be removed, replaced, or revised. The “classical” approach does this by examining inter-item correlations, Cronbach’s alpha for each sub-scale within the instrument, and other statistics that describe relationships solely among the measured variables – that is, the test items (Blunch, 2013). But the classical approach has a key limitation: neither positive correlations between items nor high Cronbach’s alpha guarantee that a sub-scale is unidimensional (Blunch, 2013; Field, 2009).

Factor analysis thus attempts to discern dimensionality in the data in one of two ways: either by positing a priori latent factors (CFA) or by extracting them a posteriori (EFA, PCA; Bryant & Yarnold, 1995; Blunch, 2013). By evaluating latent factors in addition to measured variables, CFA strengthens one’s conclusions about an instrument’s construct validity – that is, the degree to which it measures the constructs that it is supposed to measure (Blunch, 2013). Like Cronbach’s alpha, CFA provides insight into the instrument’s internal consistency, hence reliability. Unlike Cronbach’s alpha, however, CFA also provides a reliability estimate for each individual test item (Blunch, 2013). In a structural equation model, an item’s reliability is “the proportion of measured variance that can be traced back to [a latent factor]” (Blunch, 2013, p. 32). Mathematically, it is the coefficient of determination (R^2) when a measured variable is regressed onto a latent variable, which SPSS AMOS reports as the squared multiple correlation (see Table 10; Blunch, 2013). In short, test item reliability is the consistency with which students' latent concepts are mirrored in responses to that item.

In this section I use both the classical approach and confirmatory factor analysis to vet individual test items in my experimental instrument, and to recommend changes to them.

Trouble with TFM

Again and again during data analysis, signs appeared that called into question the validity of the transformationist (TFM) sub-scale. Either the test items align poorly with the hypothesized latent factor, or the factor itself is not unidimensional. Or it simply does not exist. (Or possibly there were not enough test items to sample it adequately.) Even when fused with the WLA factor to create a new Lamarckian construct (LMK) – thereby improving the model's fit to the data – the four TFM test items performed no better. Neither the originally hypothesized TFM factor nor the newly hypothesized LMK factor does much to explain variance in these four test items. Among the warning signs were the following:

1. The matrix of correlations (Table 9) shows that the four TFM variables correlate weakly with one another. Strong inter-correlation (and also correlations of similar magnitude) is desirable (Blunch, 2013).
2. The weak correlations between each TFM item and the mean of the other TFM items (Table 8) also show that they do not form a very inter-correlative cluster (Blunch, 2013).
3. Table 9 shows that TFM items correlate erratically – sometimes positively, sometimes negatively – with items belonging to the other three factors

(especially SCI and IMB; the correlations with WLA items were mostly positive). This is evidence that the hypothesized TFM factor may not be unidimensional and/or that the wording of TFM test items inadvertently evokes other conceptual dimensions.

4. Item TFM3 was endorsed by 82% of test takers even though the vast majority had already studied the theory of evolution (Table 6). This percentage might go even higher with students who have not yet studied evolution. An item with greater variance would be more useful for discriminating between test-takers.
5. Cronbach's alpha for the TFM sub-scale is unacceptably low ($\alpha = .279$), and substantially poorer than the other three sub-scales (Table 7). The problem appears to reside in all four items, as α cannot be improved by deleting any single item.
6. TFM loadings in every model tested were poor and sometimes statistically nonsignificant (Tables 10 and 11). Moreover, there were usually two items in each run that hardly loaded at all, yet in some models it was TFM1 and TFM2 that had the anemic loadings, while in others it was TFM3 and TFM4 – a sign of sub-scale instability.
7. The squared multiple correlations (R^2) for each TFM test item (Tables 10 and 11), which is a measure of item reliability (see above), were usually quite low.
8. TFM items were abundantly flagged in each model's modification indices. These indices sometimes suggested that TFM items would map better onto WLA or IMB.

I critically reexamine the wording of these four test items below.

Items TFM 3 and TFM4. In the CFA runs, TFM3 was by far the most problematic item, followed by TFM4. Item TFM3 reads: *Evolution is the process by which nature improves itself over time, and in this case it will push the crab population toward a perfect fit with the environment.* Only 18% of participants were able to resist endorsing this statement. Item TFM4 reads: *The species of crab became darker so that it would not go extinct*. Only 44% of participants were able to resist endorsing this statement. The correlation between these two items was .21. While that may not seem like an especially strong association, it is actually stronger than 94% of the other inter-item correlations in Table 9, and 77% of within-factor correlations. And in fact, the two items are conceptually similar: Both reflect a strong *teleological* and *progressivist* (and Lamarckian) perspective. TFM4 is also subtly *essentialistic* (it is the species as a whole that becomes darker), but TFM3 less so.

At the same time, these two items barely correlate at all with the other two TFM items, TFM1 and TFM2 (ϕ_2 = .02, .02, .05, and .10, all nonsignificant). Indeed, they correlate more strongly with almost all the *scientific* statements (SCI) than they do with TFM1 and 2. (That may not be a grave concern, however, since it is not necessarily the purpose of my instrument's NFC section to find negative correlations between scientific and intuitive conceptions, given the hypothesis that they could coexist for a student even after conceptual change.)

In sum, both TFM3 and 4 are strongly teleological and progressivist, correlate fairly well with one another, and proved irresistible to a majority of test takers. It may be that they belong to a separate dimension (latent factor) of their own. A reexamination of items TFM1 and TFM2 will reinforce that possibility.

Items TFM1 and TFM2. Item TFM1 reads: *Now that the population has adapted to the newly blackened beach, all the crabs are equally dark and well camouflaged*. Item TFM2 reads: *Although the crabs in each generation are basically the same color, their offspring are born a bit darker*. Unlike TFM3 and 4, a majority of participants were able to resist endorsing these two statements: 54% and 57% respectively. Though the correlation between them was only .14, they do share conceptual traits that distinguish them from TFM3 and 4. First, although both reflect a view that evolutionary change occurs at the whole-species level – a transformationist stance – they are *not* especially progressivist (TFM1 not at all and TFM2 only subtly so). Similarly, both plainly embody an *essentialist* stance (all crabs are equally dark), but are *not* teleological. By contrast, items TFM3 and 4 express a very teleological and progressivist perspective, yet are not very essentialistic.

This contrast again suggests that the TFM sub-scale does not correspond to a unidimensional latent factor, but perhaps two conceptually distinct dimensions. Also supporting this interpretation is the fact that during the running of Models #1-5, it was sometimes TFM1 and 2 that loaded most poorly, while on other runs it was TFM3 and 4: The two pairs traveled together (see Tables 10 and 11). It seems likely that this duality prevented AMOS from settling on more satisfying solutions to the system of equations. (Also, during attempts to fit the models through hyper-iterative Bayesian estimation, unwanted oscillations emerged around TFM variables, suggesting that this duality was making it impossible to converge on a solution.)

Upon recognizing this, I made a post hoc attempt to explore the effects of splitting these pairs into separate factors, and also of removing one pair or the other. Unfortunately, with only two measured variables per latent factor, the models were “unidentified” (Blunch, 2013) and AMOS could not find a solution. I also explored the effect of lumping TFM1 and 2 in with the four IMB items, since they too are strongly essentialistic, while lumping TFM3 and 4 with the four WLA items, since they too are teleological (albeit at the level of individual organisms rather than whole species). This model did converge cleanly, but its fit was little better than Models #1-3, and not nearly as good as Models #4 and #5.

Conclusions and recommendations. All four TFM items definitely embody common misconceptions about evolution that students express, as documented in Table 1 and Chapter 2, and as affirmed by the two content experts – veteran teachers of evolutionary science – who evaluated the content validity of these items. It may be that they should be revised or replaced by items more representative of the intended mental model. However, I would first want to explore the possibility that they really represent two distinct and valid latent factors: a “progressivist teleological” orientation versus an “essentialist whole-species transformationist” (though non-progressivist) orientation. That would entail composing additional test items for each new factor, followed by another cycle of field-testing.

Other Test Items Reexamined

Among the 20 NFC test items, TFM items were plainly the most pathological, but nine others were also persistently unreliable ($R^2 < .20$) across all models tested: IMB3 and 4, WLA2 and 3, and SCI1 through 5. They often correlated feebly with fellow items in their own domain and erratically with items in other domains. During each CFA run, they tended to load weakly onto their intended latent factor and appeared often among the modification indices. I critically reexamine them item-by-item below.

Items IMB3 and IMB4. In a field of test items (other than TFM) that almost always loaded onto their latent factors with high statistical significance ($p < .01$), IMB3 was the only one whose loadings were sometimes statistically nonsignificant. It reads:

> *The dark species must have been on the island all along, but before the eruption, the light species dominated the white sand beaches. But as the beach grew darker, most of the light crabs were eaten by predatory birds. Now the dark species came to dominate the beach.*

Although this was not the only item that contained more than one sentence, it is complex and may have stretched a student's working memory. Some test takers may have misunderstood it or failed to analyze it with care. Moreover, by itself, the second sentence does validly express a central Darwinian tenet, while the other two echo Darwinian language; it is only the subtle inclusion of the "species" concept that would cause a biologist to reject this proposition. Indeed, over 60% of participants endorsed it as scientific. Though it has content validity, the wording may be making it map inconsistently onto the intended latent factor, compromising its construct validity. This item should be replaced.

Item IMB4 reads: *The light crabs mated with a species of dark crab, creating a hybrid species that was better camouflaged against the newly darkened beach*. This statement is patently unscientific and any biologist would instantly reject it, but here again it is the somewhat subtle "species" concept that makes it so. Perhaps if worded in a way that more emphatically draws out the species element, it would more reliably evoke intuitions of species immutability and thus map more consistently onto its latent factor. For example: “The *original species* of light crabs mated with a *separate* species of dark crabs, creating a *new* hybrid species that was better camouflaged against the newly darkened beach.”

Items WLA2 and WLA3. WLA2 reads: *Exposure to the volcanic dust caused helpful mutations in the crabs’ DNA*. A majority of students (60%) correctly rejected this statement, and it correlates negatively with all 8 scientific (SCI) statements. It also correlates moderately well with the other 3 WLA items. However, it also correlates positively with all 4 IMB statements, and rather strongly with two of them, and this is likely the reason that it does not load cleanly onto its own latent factor. The reason for the correlation with IMB is unclear, as there is no apparent overlap in content or wording. But it is notable that this item attributes phylogenetic change to “mutation” induced by an *external* agent (volcanic dust). Perhaps that makes it more acceptable to students with an immutabilist orientation who believe that organisms’ *inner* essences do not change. Furthermore, of the 4 WLA items this one is the least Lamarckian: It is only moderately teleological, does not ascribe evolutionary change to animal behavior, and does not evoke the intentional stance. It does capture a common student misconception, but it should either be replaced or perhaps assigned to some new latent factor, even IMB itself.

Item WLA3 reads: *Some crabs learned that eating certain foods would make their shells darker, and passed this learning to their offspring*. Seven out of ten students correctly rejected WLA3, and responses correlate negatively with all 8 scientific items. Yet despite its consistent inverse association with the SCI domain, WLA3 was one of the least reliable test items across all models tested, with R^2 ranging from .05 to .12. A look at the correlation matrix makes clear why: Like WLA2 above, it correlates almost as well with IMB items as with other WLA items. And like WLA2, it may somewhat sidestep an essentialist resistance to species mutability, for it does not attribute the crabs' gradual adaptation to a change in inner "essence," but instead to learning and a sort of "cultural" transmission – something that students with an immutabilist orientation might find quite acceptable. Like WLA2, then, WLA3 should be replaced or perhaps even reassigned to the IMB factor itself.

Upon recognizing that items WLA2 and WLA3 might actually jibe with an immutabilist perspective, I made a post hoc attempt to explore the effects of reassigning items WLA2 and WLA3 onto the IMB latent factor. I performed this reassignment with Models #4 and #5. The item-by-item loadings did not shift much, but the fit indices were almost identical to the fine fit indices previously yielded by Models #4 and #5. Future field-testing should explore this unanticipated result, and WLA2 and WLA3 should be rewritten to wed them more explicitly either to a quasi-Lamarckian factor or to the immutability factor.

Items SCI1 through SCI5. Five of the eight SCI items did not perform as reliably and validly as hoped:

> SCI1: *Even before the volcano began spreading black dust on the beaches, some crabs were already slightly darker than others.*
>
> SCI2: *In each generation some crabs are darker and some are lighter, but the average darkness of the population is changing from one generation to the next.*

SCI3: *Coloration varies quite a bit from crab to crab, and those whose markings best match the beach are able to reproduce most successfully*.

SCI4: *As generations go by, genes for lighter colors get weeded out of the population*.

SCI5: *Once in a while, a chance mutation gives an offspring a lighter or darker shade*.

Each of these seems a straightforward statement of one or more core tenets of Darwinian theory (Table 3), and all were endorsed by 75-85% of participants (almost all of whom had already studied the theory of evolution). They also all correlate positively with one another, if not always strongly, and their content validity was affirmed by two evolutionary scientists. Given this apparent consistency, it is puzzling that their loadings are so weak and that they do not support a stronger claim to the SCI dimension's construct validity. In SCI3, the phrase "quite a bit" is perhaps relative and misleading, and probably should be stricken. Otherwise, however, I do not have changes to recommend, nor do I think these essential items should be replaced. It may be that their limited reliability and questionable validity will have to be accepted, and test results interpreted with due caution.

Conclusions and recommendations. The two poorly performing IMB items should be reworded or replaced to make the main thrust clearer to students, whether by shortening the item (IMB3) and/or underscoring the pivotal "species" concept (IMB3 and 4). The two poorly performing WLA items may need to be replaced by items more representative of the intended latent factor. However, they do express a pair of common misconceptions about evolution, and there are clues that both may inadvertently circumvent an essentialist resistance to species mutability. I would therefore want to explore the option of reassigning them, with some rewording, to the IMB factor. The five poorly performing SCI items express essential Darwinian concepts and should not be abandoned, although I do not have specific recommendations for revising them. Their questionable reliability and validity may simply have to be accepted.

Summary of Instrument Validation: Non-Forced Choice Section (NFC)

In educational testing, "instrument validity" refers to the accuracy and thoroughness with which an instrument measures the concepts, constructs, outcomes, or competencies that its users intend; specifically, it is the extent to which the interpretations and conclusions that a researcher draws from it are supported by prior evidence (Bryant, 2000; Gall et al., 2007). I have presented four bodies of such evidence for the NFC section of my experimental instrument:

1. Content-related evidence: Two content experts (professional scientists) vetted test items and helped revise them for the final version of the instrument. Content validity was deemed acceptable.
2. Evidence from internal structure: Test responses from 340 members of the target audience were subjected to confirmatory factor analysis. After several post hoc changes were made to the original CFA model – always within the confines of the original theoretical framework – a decent fit to the data emerged. This bolsters confidence in the instrument's construct validity.
3. Evidence from internal structure: Reliability estimates (Cronbach's alpha) for all four sub-scales were unacceptably low. This is evidence *against* instrument validity, since low reliability compromises the predictability of

instrument performance from one administration to the next, and thus undermines confidence in conclusions drawn from the test results (Gall et al., 2007).

4. Evidence from internal structure: During CFA, many individual test items mapped weakly onto their intended/hypothesized latent factors. This is evidence *against* the instrument's construct validity. Recommendations were offered for revising or replacing these items.

In sum, certain modifications are in order, but there is cause for optimism that the instrument's NFC section can be rendered into a valid and reliable instrument for assessing students' understanding of evolutionary theory as well as their adherence to common preconceptions/misconceptions about evolution. I hasten to repeat that this tentative conclusion is based on data from a non-representative sample of the target population. After the recommended revisions are made, the instrument will have to be cross-validated with new audiences.

Forced Choice Section (FC)

The second section of my experimental instrument comprises ten "forced choice" (FC) questions that oblige the test-taker to select a scientific statement amidst a field of intuitively appealing distracters. I have analyzed it separately from the NFC section, because the two sections were designed for two different functions: The FC section is designed to yield a sharp inverse correlation between scientific and unscientific conceptions, whereas the NFC section deliberately allows for the possibility that these conceptions may coexist in the student's understanding, which thus ought to dampen that correlation. In a sense, then, the two sections are designed to mobilize two different cognitive phenomena (if they exist). The NFC section seeks to gauge the degree to which a student *harbors* both scientific and intuitive understandings of evolution. The FC section seeks to gauge the degree to which a student *suppresses* the latter in favor of the former.

Here again, I present descriptive and diagnostic statistics, the results of confirmatory analysis, and an examination of individual items for possible revision or replacement.

Descriptive and Diagnostic Statistics (FC Section)

Each FC question contains a scientifically correct choice (SCI) and one or more distracters that are intended to map onto one of three hypothesized intuitive factors (IMB, TFM, WLA). Table 13 shows that on average, students selected the scientific choice about half the time (49%) and an intuitive distracter the other half (48%). On all ten questions, the scientific statement was selected more often than any of the distracters, from 34% to 68% of the time. On average, students selected an immutability (IMB) statement on 15% of the 7 opportunities to do so, a transformationist (TFM) statement on 27% of the 8 opportunities, and a within-lifetime adaptationist (WLA) statement on 37% of the 12 opportunities.

Cronbach's alpha (α) for the scientific sub-scale is an acceptable .775 (Table 14), and reliability cannot be appreciably improved by removing any single question, an indication of good reliability (Field, 2009). For all three intuitive sub-scales Cronbach's alpha is unacceptably low, and several items appear especially problematic: TFM2, TFM5, IMB6, and WLA9. It may not be realistic, however, to expect high reliability for

Table 13

Forced Choice Section (FC): Endorsement Frequency for Each Scientific Choice and Intuitive Distracter, Mean Participant Scores (Scientific vs. Intuitive), and Mean Sub-Scores (Immutability, Transformationism, and Within-Lifetime Adaptation)

	Scientific		Intuitive					
Test Item	Correct choice	%	IMB distracter	%	TFM distracter	%	WLA distracter	%
1	FCsci1	67.6	FCimb1	2.7	FCtfm1	19.6		
					FCtfm2	10.1		
2	FCsci2	66.1			FCtfm3	21.5		
					FCtfm4	5.0		
3	FCsci3	42.8	FCimb2	12.1			FCwla1	19.5
							FCwla2	20.9
							FCwla3	4.7
4	FCsci4	66.7			FCtfm5	6.9		
					FCtfm6	20.4		
5	FCsci5	46.3	FCimb3	8.3			FCwla4	26.4
							FCwla5	19.0
6	FCsci6	33.5	FCimb4	17.7	FCtfm7	19.2	FCwla6	29.6
7	FCsci7	41.0					FCwla7	24.4
							FCwla8	25.3
							FCwla9	9.3
8	FCsci8	44.3	FCimb5	24.4			FCwla10	14.9
			FCimb6	16.4				
9	FCsci9	46.3					FCwla11	43.0
10	FCsci10	36.2	FCimb7	15.9	FCtfm8	32.1	FCwla12	15.9
Sub-scores			IMB		TFM		WLA	
			Mean	16.1	Mean	26.7	Mean	35.7
			SD	16.0	SD	22.1	SD	24.5
			SE	0.87	SE	1.20	SE	1.33
Total scores	Scientific		Intuitive					
	Mean	48.5	Mean	48.0				
	SD	27.8	SD	25.7				
	SE	1.51	SE	1.39				

Table 14

Cronbach's Alpha (α) for Each Sub-Scale and Effect on Alpha of Deleting Items

Scientific			Intuitive					
Sub-scale / Item	α	α if item deleted	Sub-scale / Item	α	α if item deleted	Sub-scale / Item	α	α if item deleted
SCI	.775		TFM	.169		WLA	.400	
FCsci1		.770	FCtfm1		.102	FCwla1		.381
FCsci2		.766	FCtfm2		.201	FCwla2		.362
FCsci3		.738	FCtfm3		.132	FCwla3		.404
FCsci4		.766	FCtfm4		.173	FCwla4		.361
FCsci5		.741	FCtfm5		.183	FCwla5		.378
FCsci6		.764	FCtfm6		.152	FCwla6		.373
FCsci7		.740	FCtfm7		.099	FCwla7		.389
FCsci8		.780	FCtfm8		.149	FCwla8		.381
FCsci9		.743				FCwla9		.437
FCsci10		.749	IMB	.266		FCwla10		.403
			FCimb1		.272	FCwla11		.301
			FCimb2		.180	FCwla12		.356
			FCimb3		.172			
			FCimb4		.118			
			FCimb5		.274			
			FCimb6		.364			
			FCimb7		.234			

Table 15

Matrix of Correlations for FC Scientific Choices (Phi Correlations, ϕ_2)

Test Item	FCsci1	FCsci2	FCsci3	FCsci4	FCsci5	FCsci6	FCsci7	FCsci8	FCSci9	FCSci10
FCsci1	1	-	-	-	-	-	-	-	-	-
FCsci2	.25***	1	-	-	-	-	-	-	-	-
FCsci3	.26***	.29***	1	-	-	-	-	-	-	-
FCsci4	.15**	.18***	.24***	1	-	-	-	-	-	-
FCsci5	.15**	.23***	.46***	.24***	1	-	-	-	-	-
FCsci6	.15**	.15**	.31***	.10^{ns}	.35***	1	-	-	-	-
FCsci7	.20***	.24***	.46***	.31***	.43***	.26***	1	-	-	-
FCsci8	.13*	.12*	.15**	.16**	.17**	.09^{ns}	.13*	1	-	-
FCsci9	.23***	.29***	.36***	.27***	.36***	.30***	.40***	.21***	1	
FCsci10	.15**	.18**	.38***	.19***	.42***	.30***	.42***	.19***	.31***	1

^{ns} not significant. *p < .05. **p < .01. ***p < .001.

distracters, because each multiple choice question contains several intuitive distracters, often with more than one belonging to the same sub-scale. Consequently, intuitively appealing statements are in direct competition with one another, even within the same sub-scale. The frequency with which participants select *any* intuitive answer will, of course, just be the inverse of the frequency with which they select scientific answers; thus the *overall* intuitive scale will be just as reliable as the scientific scale.

Table 15 shows the correlations between all scientific items. They are all positive, as they must be (Blunch, 2013), and almost all statistically significant. Item SCI8 correlates somewhat weakly with the other nine items and so may warrant revision or replacement. The rest appear to form a tight inter-correlative cluster.

Correlation matrices are not shown for the three intuitive domains, but their inter-correlations are often weak, statistically nonsignificant, and sometimes negative. Here again, this is partly an inevitable consequence of the fact that most multiple choice questions contain two or even three distracters belonging to the *same* dimension (IMB, TFM, or WLA). Items within the same sub-scale are thus often in direct competition with each other, such that their inter-correlations are bound to be weak and sometimes negative.

Confirmatory Factor Analysis (FC Section)

Figure 13 shows the fundamental model for CFA on student responses to the ten FC questions. Although there are only ten questions, each contains up to five choices, each meant to map onto one of the four hypothesized dimensions. There are therefore 37 measured variables in the model. The scientific latent factor (SCI) is hypothesized to correlate negatively with each of three intuitive factors, because the FC section of the test, in contrast to the NFC section, is designed to assess the degree to which students can suppress their intuitions in favor of scientific theory. Or – in the prevailing language among conceptual change scholars – the FC section aims to assess whether students have successfully restructured or replaced their pre-instructional theory with a scientific one.

Unfortunately, when I attempted to fit this model to the data, SPSS AMOS was unable to converge on a solution. This failure was almost certainly due to the systematic interdependence of the measured variables: Within any given multiple choice question, the variables are in direct competition and so bear a linear relationship with one another. (As soon as a student selects an answer, the values of the other answers are automatically set to zero.) Linear relationships between variables will make a system insoluble via CFA (Arbuckle, 2014; Blunch, 2013). Other models were tried, including one that posited correlations between all the error terms within each multiple choice question, but without success.

Consequently, the only feasible analysis was to run CFA on the ten SCI items in isolation from the rest, since these items are not systematically interdependent (each multiple choice question has only one SCI answer). That model and its loadings are shown in Figure 14, and the loadings and fit indices are listed in Tables 16 and 17 respectively. The model fits well to the data. Despite the large sample size, chi-square is statistically nonsignificant (which is desirable). The fit indices all indicate an acceptable fit. Most of the loadings are respectable, with the latent factor explaining more than a third of the variance in five of ten items, though once again item SCI8 is suspect and SCI1 also merits reexamination.

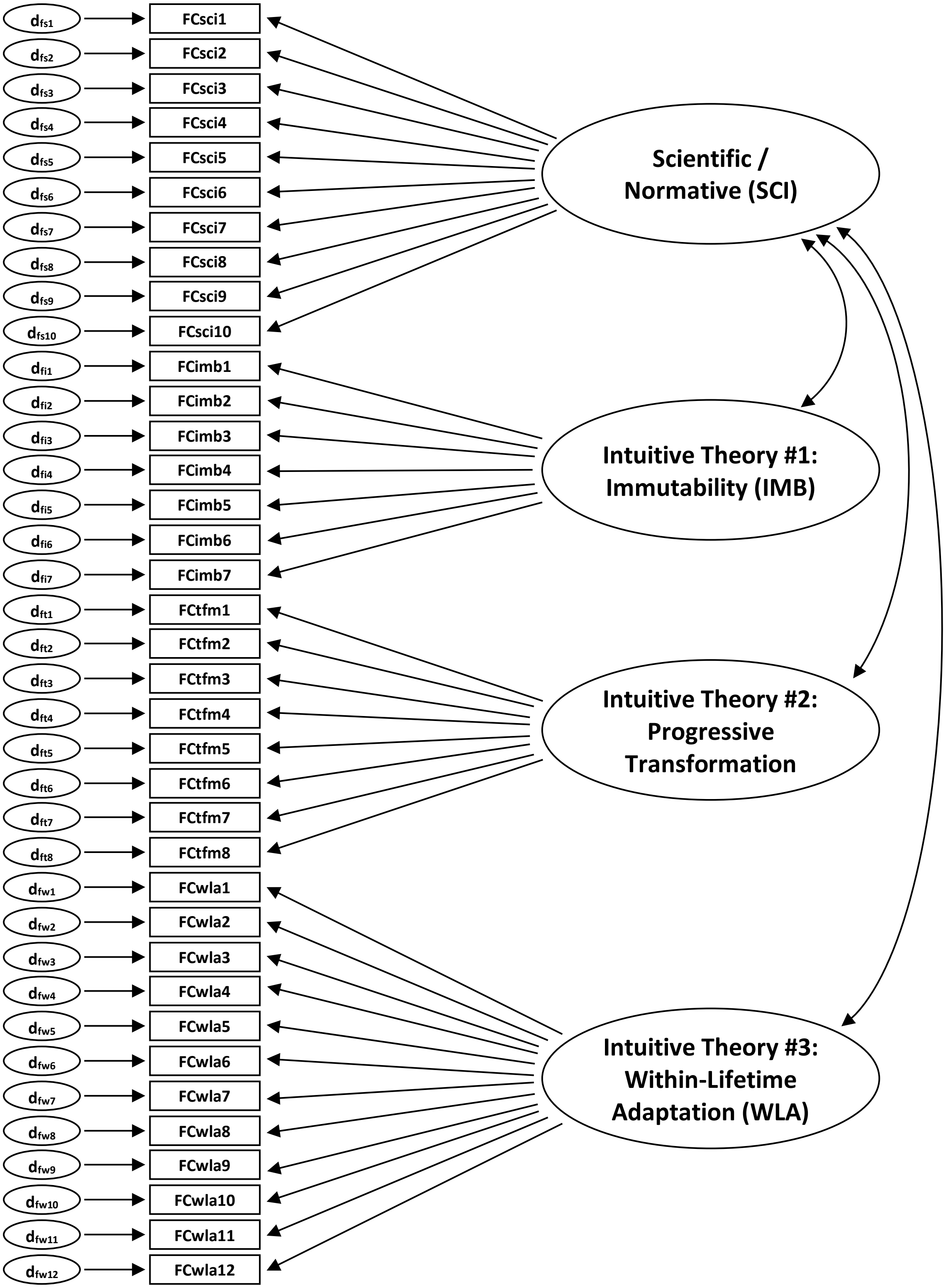


Figure 13. Base Model for confirmatory factor analysis on FC data.

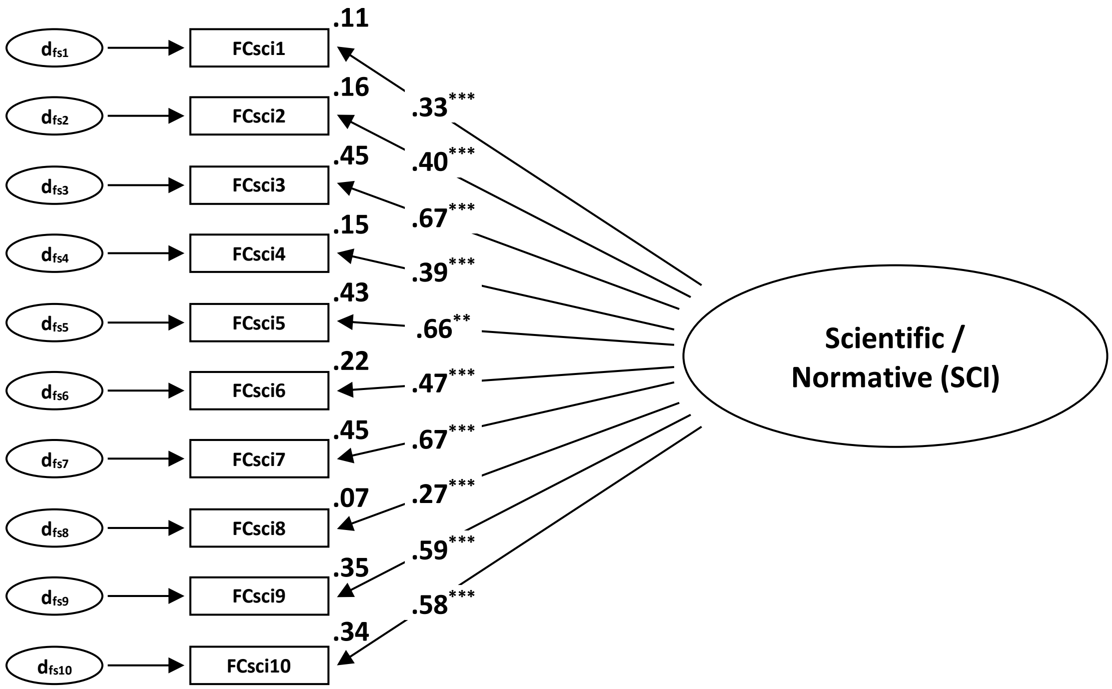


Figure 14. Results of confirmatory factor analysis on FC SCI items only. The number overlapping each arrow is the item's loading onto its latent factor (standardized regression coefficient). The number on its upper right corner is the Squared Multiple Correlation, equivalent to R^2. Significance: *p < .05, **p < .01, ***p < .001, ns = not significant.

Table 16

Factor Loadings and Squared Multiple Correlations (R^2) for Forced Choice Scientific Items

Test Item	λ	R^2
FCsci1	$.33^{***}$	.11
FCsci2	$.40^{***}$	.16
FCsci3	$.67^{***}$	.45
FCsci4	$.39^{***}$	.15
FCsci5	$.66^{***}$	.43
FCsci6	$.47^{***}$	.22
FCsci7	$.67^{***}$	.45
FCsci8	$.27^{***}$	.07
FCsci9	$.59^{***}$	.35
FCsci10	$.58^{***}$	.34

*p < .05, **p < .01, ***p < .001, ^{ns} not significant

Table 17

Fit Indices for Forced Choice Scientific Items

χ^2	χ^2 / DF	CMIN/DF	PCFI	RMSEA
40 (p = .256)	1.14	1.15	.631	.021 (p = .98)

Note. χ^2 = Bollen-Stine Chi-square goodness of fit index after bootstrapping. The nonsignificant p-value signifies acceptable fit. χ^2/DF, which divides Chi-square by the degrees of freedom is acceptably low (ideally < 3; Klem, 2000). CMIN/DF near 1.0 indicates good fit (Blunch, 2013). PCFI = parsimony adjusted comparative fit, with values > .60 indicate satisfactory fit (Blunch, 2013). RMSEA = root mean square error of approximation, with values < .05 and high p-values indicating good fit (Blunch, 2013).

*p < .05, **p < .01, ***p < .001, ^{ns} not significant

Individual Test Items (FC Section)

The dilemma of item interdependence compromises interpretation of the loadings, inter-correlations, and Cronbach’s alpha for all the intuitive distracters. For that reason I will not attempt to identify distracters for revision. Instead I will focus on the ten scientific choices.

Item SCI8 appears to be the one especially problematic test item: It loads poorly onto its latent factor, correlates comparatively weakly with the other nine SCI items, and is the only SCI item which if removed would improve Cronbach’s alpha (though negligibly so). It reads:

> *Suppose a scientist now conducts an experiment: She transplants 50 sandy island crabs to the rocky island. She predicts that “natural selection” will occur – sometimes called “survival of the fittest.” What does she mean?*
>
> A. *Because the species is not fit for this harsh new habitat and cannot change, the species will be weeded out.*
>
> B. *Crabs that keep themselves physically fit will have a chance to survive in this harsh new habitat. Less fit crabs will perish.*

C. *Some crabs will outlive others because they happen to have traits that are better suited for the new habitat.*

D. *Because the native rocky island species is already well adapted for this habitat, the native crabs will quickly outcompete the new crabs.*

The purpose of this question is to reveal students' understanding of the expressions "natural selection" and "survival of the fittest," both of which are frequently misunderstood. Choice C is the correct answer. Choice B, chosen by 15% of participants, reflects the common confusion of Darwinian fitness with "physical fitness" (a quasi-Lamarckian belief). Choices A and D, chosen by 24% and 16% of participants respectively, reflect an essentialist intuition that Darwinian fitness refers to how well entire species "fit" into their environments and that natural selection occurs at the whole-species level. This intuition regards "survival of the fittest" as a contest *between* species (interspecific competition) rather than *within* species (intraspecific competition).

Forty-four percent of students did answer this question correctly, a success rate almost as high as the section average (49%). Apparently many students who have a decent scientific understanding of evolution nonetheless missed this question, while some who lack a sound understanding of evolution got it right. Table 18 shows that of all ten FC questions, this one does bear the weakest correlation with students' overall scientific scores, both for the FC section and for the test as a whole (FC and NFC scores

Table 18

Correlation (r) of Each FC Scientific Item with Students' Overall Scientific Scores

Scientific Score	FCsci1	FCsci2	FCsci3	FCsci4	FCsci5	FCsci6	FCsci7	FCsci8	FCsci9	FCsci10
FC SCI Score	.46***	.51***	.69***	.50***	.67***	.53***	.67***	.42***	.65***	.62***
Overall SCI Score (NFC + FC)	.24***	.29***	.25***	.27***	.27***	.17**	.28***	.13*	.25***	.21***

$^{*}p < .05$. $^{**}p < .01$. $^{***}p < .001$.

combined). This question therefore does not appear to discriminate well between students who do and do not have a sound understanding of evolution. In fact, one of the content experts who evaluated the instrument observed that this item, unlike all the rest, focuses on a semantic distinction (i.e., the correct connotation of "survival of the fittest") rather than students' core conceptual understanding and intuitions. The question should probably be removed from the instrument.

The other item that loaded especially poorly onto its latent factor is SCI1. The question is visual (Figure 15) and challenging, demanding complex visual and conceptual analysis on the part of the test taker. Its difficulty and complexity may account for its unreliable mapping onto the latent factor, and should perhaps be discarded. However, it does correlate fairly well with students' overall scientific score (NFC and FC combined; Table 18). Thus, although challenging, it may validly discriminate students with a strong grasp of evolution from those without it. I recommend retaining the question and reassessing it after a cross-validation study.

Suppose that after 250 years of volcanic activity, the once white beach has become medium gray, as shown in the background above. If a biologist took a random sample of 10 crabs from the population, which of the above would she expect her sample to be like? Choose the best answer: A, B, C, or D.

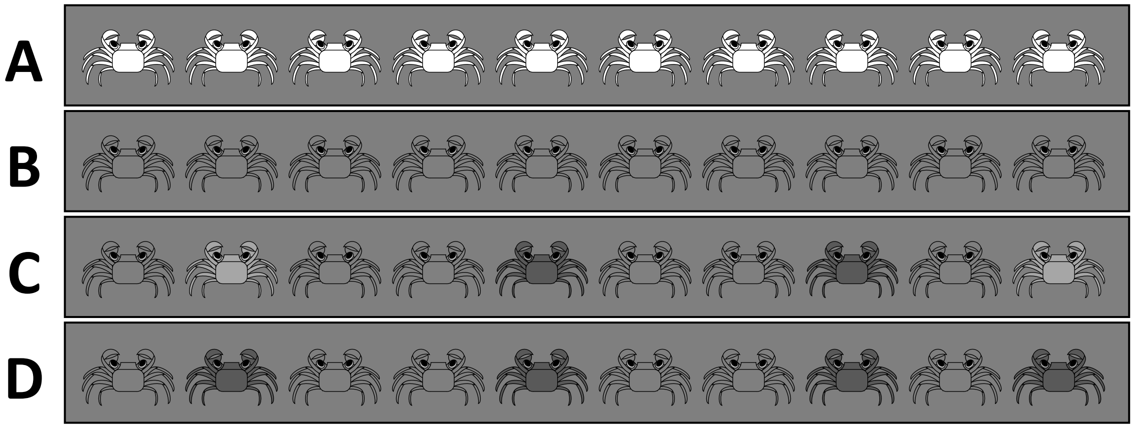


Figure 15. Forced choice question #1. Choice C is the best scientific answer.

Summary of Instrument Validation: Forced Choice Section (FC)

Whether conceptual change requires a student to *restructure* a pre-instructional mental model into a scientific model, or *replace* the former with the latter, or *suppress* the former in favor of the latter, the FC section of my experimental instrument appeared to perform well as a tool for assessing such conceptual change. There is evidence that the ten multiple choice questions can validly and reliably distinguish students who hold a correct scientific understanding from those who do not:

1. Content-related evidence: Two content experts (professional scientists) vetted test items and helped revise them for the final version of the instrument. Content validity was deemed acceptable.
2. Evidence from internal structure: Confirmatory factor analysis on the 10 SCI items produced a strong fit to the data. This bolsters confidence in the instrument's construct validity.
3. Evidence from internal structure: The reliability estimate for the scientific sub-scale was satisfactory (Cronbach's $\alpha = .775$). This, too, bolsters confidence in test validity, since high reliability is a necessary prerequisite for validity (Gall et al., 2007).
4. Evidence from internal structure: During CFA, most of the SCI items mapped satisfactorily onto their latent factor. This, too, is evidence of construct validity. Validity will be further strengthened by the removal of one problematic item (SCI8).

However, while the FC section appears valid as a gauge of scientific understanding, it *cannot* be used validly to assess student adherence to the three hypothesized intuitive dimensions (IMB, TFM, WLA). The systematic interdependence between intuitive distracters compromises the interpretation of the response frequencies, scores, loadings, inter-correlations, and Cronbach's alpha within the intuitive domain. The IMB, TFM, and WLA sub-scales are not valid and should not be used.

Preview of Chapter 5

In the final chapter I review these findings more broadly and discuss them in light of the research purpose stated in Chapter 1, the cognitive models and theoretical foundations put forth in Chapter 2, and potential significance for instructional planning and curriculum leadership. I underscore limitations of the study, especially offering cautions about generalizing conclusions to future audiences. I discuss how the instrument, once refined and cross-validated, might be useful not only to fellow researchers but also science educators – from classroom practitioners to district level instructional leaders to state level science education specialists – especially in the enduring (albeit shifting) climate of standardized testing. I consider possible implications for evolution education of the quasi-Lamarckian factor's emergence as a candidate driver of student intuitions about the process of evolution, over against Darwinian natural selection. I also propose that the instrument could be useful not only for assessing student understanding of evolutionary theory, but cultivating that understanding as well, by using its challenging scenarios to stimulate classroom discussion and call attention to the intuitive appeal of its unscientific distracters. Finally, I suggest directions for future research to further investigate and improve the instrument's validity and reliability.

Chapter 5: Discussion

The research described in the preceding four chapters confronted an educational problem – the challenge of conceptually mastering the Darwinian paradigm – within a Darwinian paradigm. The motivating premise was that the human mind may be poorly evolved to understand the theory of evolution itself. Beginning in early childhood, and apparently across all cultures from small foraging societies to large industrialized nations, young people develop certain distinct "intuitive theories" to govern their interactions with their physical, biological, ecological, and social environments. Though constructed during the course of real world experience, these theories are sculpted around a handful of powerful, pervasive ways of construing cause-and-effect relationships that may well be partly innate. Among these different species of causation are mechanical forcing, helping and hindering, permitting and preventing, personal goals and intentions, teleological function and purpose, and inner essences and intrinsic nature (e.g., Atran, 1998; Baron-Cohen, 1995; Carey, 2009; S. Gelman, 2003; Gopnik & Meltzoff, 1997; Inagaki & Hatano, 2002; Keil, 1989; Leslie, 1994; Pinker, 2007; Premack & Premack, 1995; Spelke, 1994; Tomasello, 1999; Wellman, 1990). Each of these causal construals would have been adaptive for our hominin (and pre-hominin) ancestors during their interactions with each other, with wildlife, and with physical phenomena; thus they may be psychological adaptations built into the innate architecture of the human mind (Tooby & Cosmides, 1992). Although some innate intuitions may be active at birth, others may emerge developmentally or facultatively only later in life, no less “innate” than the advent of wisdom teeth, hand calluses, sun tans, sexual attraction, and parental devotion; they are innate precisely to the degree that their predictable, distinctive emergence reflects “adaptive design” via natural selection in generations past (e.g., Dawkins, 1986; Gallistel et al., 1991; Ridley, 2003; Tooby & Cosmides, 1992; Williams, 1966).

Three of these core intuitions – the tendency to project immutable essences, teleological directionality, and intentional agency onto biological phenomena – are well known to compromise many students’ grasp of evolutionary theory (e.g., Evans, 2008; Shtulman, 2006; Sinatra et al., 2008; Smith, 2010). By the time a student first studies evolution in middle or high school, her “intuitive biology” may be so well entrenched (Atran, 1998; S. Gelman, 2003; Inagaki & Hatano, 2002; Keil, 1989) that she finds the scientific model counterintuitive and struggles to master it. Or perhaps she finds the scientific model comprehensible enough, but at the same time finds it difficult to suppress her intuitive biology, such that the ideas she expresses in class and on tests are colored by essentialist, intentional, and teleological misconstruals. Perhaps what materializes during a research interview or an open response test are “mixed” or “synthetic” models (Vosniadou et al., 2008) which constitute either a fusion or an interweaving of intuitive and scientific conceptions – in effect, an on-the-spot hybridization of pre- and post-instructional understandings (Chinn & Brewer, 1993; Poling & Evans, 2004; Shtulman, 2006; Shtulman & Calabi, 2012; Strike & Posner, 1992; Vosniadou, 1994; Vosniadou et al., 2008). Students might be especially prone to hybridize in this manner whenever their core intuitions are in fact innate, since innate intuitions may at some level be unshakeable, destined to inform thinking throughout life (e.g., Atran, 1998; Carey, 2009).

Even after conceptual change, then, the interpretations that a student communicates to teachers about the biology and behavior of living organisms may in part reflect older intuitive concepts that continue to coexist alongside her new Darwinian ones. In that case, rather than "restructure" or "replace" her pre-instructional understanding – as scholars often characterize conceptual change – she may need to learn to override and suppress it. Conceptual change may be more a habitual *suspension* of intuitive concepts than a transformation, supplanting, or superseding of them.

If so, then as I will argue in this chapter, there are important implications for science education in the domains of classroom pedagogy, assessment of student learning, conceptual change research, instructional leadership, and curricular standards and policy.

Research Purpose and Chapter Overview

The purpose of this study was to develop, pilot, and evaluate a new instrument for assessing student conceptual change with respect to the theory of evolution. The instrument employs a novel, unconventional format derived from the foregoing considerations about conceptual change and the evolution of human intuition.

The Instrument: A Novel Format

The instrument was designed not only to gauge student mastery of the scientifically normative model of evolution, but also to elicit essentialist, intentional, and teleological misconceptions where they exist. Half the instrument consists of typical multiple choice items where students are obliged to identify a scientifically normative statement amidst a field of intuitive appealing distracters. By itself, this conventional "forced choice" structure corresponds to the prevailing view of conceptual change as a restructuring or replacement of a singular pre-instructional mental model: The student's interpretations are either scientific or they are not. Or else they are in transition from unscientific to scientific.

The instrument's other half, by contrast, consists of isolated statements – some scientifically normative, some intuitively appealing – such that the student is not strictly obliged to choose scientific concepts over unscientific ones, but may freely endorse both. This "non-forced choice" structure corresponds to the alternative view that even after a student learns the theory of evolution, her pre-instructional mental model may continue to coexist and compete with the newly acquired Darwinian model. In effect, this part of the instrument allows the student – in a closed response format – to express a "mixed" or "synthetic" interpretation of evolution, hybridizing her pre- and post-instructional understandings (Vosniadou et al., 2008).

In combination, the "non-forced choice" and "forced choice" sections allow for the possibility that conceptual change entails active suppression or suspension of a pre-instructional mental model, rather than restructuring or replacing it: One section permits intuitive and scientific conceptions to surface side-by-side if and when they coexist for the student, while the other asks her to suppress the former in favor of the latter.

Theoretical and Pedagogical Context: The "Social Strategic" Hypothesis

The instrument's unconventional design was motivated by the alternative view of conceptual change as an act of suspension and suppression rather than replacement or restructuring. And this alternative view originally emerged from the "social strategic" hypothesis that I elaborated in Chapter 2 to explain the evolutionary origin of the human capacity for scientific conceptual change ("Thesis #7"), along with the corresponding model of the intuitive human mind that I postulated there (Figure 1). According to this hypothesis, we 21st century *Homo sapiens* are able to adopt the counterintuitive truth claims of modern science because in ancestral societies, our hominin forebears were frequently confronted by competing and potentially deceptive or distorted truth claims put forth by others in the social group. Consequently, the human mind evolved an ability to *suspend intuition* while maintaining both a receptiveness and a healthy skepticism toward rival propositions being proclaimed in the public sphere – propositions, that is, about strictly *social* affairs. Only much later in human history, on this view, was this same mental machinery recruited to make sense of *non-social, natural* phenomena as well. Thus did selective pressures in our ancestors' strategic social environment prepare human brains – hence human minds – to be open to the sometimes non-intuitive yet powerfully predictive models and explanatory theories of modern science.

If this hypothesis harbors some truth, then as I argued in Chapter 2, there would be critical implications for pedagogy in the contemporary science classroom. To cultivate conceptual change, teachers would want to engage students in dynamic dialogue in a public forum that evokes the strategic social sphere of our ancestors. They would want students to confront rival explanations of natural phenomena – namely, scientific explanations versus intuitively appealing but unscientific ones. And they would want to oblige students to make decisions and take positions of their own in a publicly visible manner. The goal of all this would be to mobilize the "strategic social" machinery of the evolved mind to enable students to suspend and suppress their natural intuitions about natural phenomena, at least long enough to embrace, internalize, and accurately communicate the scientifically normative understandings of those phenomena.

Chapter Overview

The social strategic hypothesis and its corresponding pedagogy form the backdrop for some of the educational policy, planning, and leadership implications that I discuss in this chapter, as does the more general idea – empirically supported in this study – that students' pre-instructional intuitions can persist even after conceptual change. I begin by summarizing results and conclusions from an evaluation of the pilot instrument's validity and reliability, as well as recommendations for modifying future versions. I then underscore the study's inherent limitations, urging caution in generalizing conclusions about the instrument's validity and reliability to other student audiences. With these cautions in mind, I suggest directions for future research and further instrument validation studies.

I then proffer several functions that my instrument, once refined and cross-validated, might ultimately serve: (1) It may support research into the cognitive mechanisms that permit students to undertake scientific conceptual change, especially in classroom intervention studies designed to evaluate pedagogies born of the social strategic hypothesis. (2) It may support diagnostic testing and data gathering by professional educators in the enduring (albeit shifting) climate of standardized testing, from classroom practitioners to instructional leaders at the school district and state levels. And (3), it might actually serve as a valuable instructional tool in its own right, providing teachers and students with vivid, challenging scenarios around which to stage classroom dialogue, debate, and decision-making about scientific versus intuitively appealing "truth claims." In other words, the instrument might be useful for instructional planning in the very spirit of the social strategic hypothesis that originally inspired its design.

Finally, I also discuss implications for science education policy, curricular standards, and assessment. In light of evidence from this study that students' intuitions can persist even after successfully learning the theory of evolution, I propose that learning standards should not only articulate the scientific concepts that students are to master, but should also explicitly identify common misconceptions that they may need to "unlearn." I especially discuss the potential significance of the quasi-Lamarckian ideas that surfaced in student responses and that may often be at work in student preconceptions/misconceptions about how evolution occurs. In the same vein, I suggest that science assessments like state end-of-course biology tests should include items to gauge students' adherence to common misconceptions, and should sometimes do so in a way that permits those misconceptions to surface side-by-side with correct scientific conceptions. I propose that my instrument's unconventional "non-forced choice" section shows one potential strategy that might be feasible for statewide assessments.

Summary and Conclusions

Two conceptual frameworks served as blueprints for constructing my experimental instrument – one to represent the modern scientific theory of evolution (Table 3), the other to represent common misconceptions born of the essentialist, teleological, and intentional stances (Table 1). Prior to field-testing, a pair of professional experts in evolutionary science evaluated and helped revise the test items in light of these two blueprints, ultimately affirming the instrument's content validity. More than three hundred high school students then took the test. Their responses to the "forced choice" and "non-forced choice" items were separately analyzed for evidence of construct validity and reliability, both with respect to the scientifically correct framework and with respect to the framework of common misconceptions.

Validity with Respect to the Scientifically Normative Framework

Confirmatory factor analysis and other statistical indicators provided compelling evidence that the conventional multiple choice questions in the instrument's "forced choice" section had validly distinguished students who held a correct scientific understanding from those who did not. While some modifications are in order, the forced choice section of this pilot instrument already exhibits real promise as a valid and reliable gauge of conceptual change, whether that requires a student to restructure his pre-instructional mental model into a scientific one, to replace the former with the latter, or to suppress the former in favor of the latter.

There was also evidence – though less compelling – that the "non-forced choice" section also yielded moderately valid and reliable estimates of participants' scientific understanding of evolution. All eight scientific statements were correctly endorsed by a majority of test takers, almost all of whom had already studied evolution in a science course. These eight test items aligned as intended with one another, though some items did so inconsistently. Revising some of those items may improve construct validity and reliability.

Validity with Respect to the Framework of Common Misconceptions

Scientific statements in both the forced and non-forced choice sections were written as components of a single dimension – namely, the scientifically normative theory of evolution by natural selection. By contrast, each unscientific statement in the non-forced choice section and each unscientific distracter in the multiple choice section was written to represent one of three dimensions, a distinct family of common misconceptions. Some items were meant to capture an “immutabilist” position that species and their inner essences are immutable (a strong essentialist stance). Other items were written to reflect a “transformationist” and often “progressivist” position that evolution transforms whole species as a unit, usually in the direction of increasing sophistication and complexity (a blend of teleological and essentialist intuitions, and perhaps the intentional stance as well). Still others were written to capture the common misconception that evolution occurs when individual organisms adapt to local circumstances during their lifetimes and then pass these changes to offspring, a “within-lifetime adaptationist” explanation of evolution (a blend of teleological and intentional intuitions).

In my analysis of the forced choice data, it was not possible to partition these three intuitive families from one another, due to systematic interdependence among the intuitive distracters (since each multiple choice question contained more than one distracter). That is, the forced choice items cannot be used to draw valid conclusions about a student’s specific adherence to an immutabilist position, nor to a transformationist position, nor to a within-lifetime adaptationist position. They did, however, appear to provide a valid and reliable assessment of participants’ *general* adherence to intuitive/unscientific misconceptions.

In my analysis of the non-forced choice data, on the other hand, it was possible to evaluate the reliability and construct validity of the immutabilist, transformationist, and within-lifetime adaptationist sub-scales independently of one another. Unfortunately, the instrument did not appear to yield very valid and reliable assessments of any of them.

Especially suspect was the transformationist/progressivist construct, which did not appear to exist at all as a distinct category of misconceptions. In Chapter 4 I reevaluated the most problematic items and offered suggestions for revising or replacing them, but it may be that my tripartite classification of common misconceptions is simply invalid.

Even so, the data did *not* support the possibility that a student's intuitions belong instead to a single general, unidimensional mental model embodying all three categories. In fact, there were clues that construct validity might be improved by abandoning the immutabilist/transformationist/adaptationist typology and reclassifying, reanalyzing, and/or rewriting items according to an essentialist/teleological/intentional typology instead. Although many of the test items, as written, embody two or even all three core intuitions, they often express one more strongly than the others. For example, certain transformationist items are more teleological, while others are more essentialist. Items would have to be substantially reworked, but my analysis suggested that the non-forced choice section might be made more valid and illuminating by grouping the most teleological items together, the most essentialist items together, and so on.

The Emergence of a Quasi-Lamarckian Orientation

On the other hand, my analysis also suggested that construct validity could be just as well improved simply by merging the transformationist and within-lifetime sub-scales into a single hybrid sub-scale – that is, a single higher order dimension. Indeed, I had already hypothesized this possibility prior to data collection, on theoretical and historical grounds: For centuries, transformationist/progressivist notions have often gone hand-in-hand with notions of behavior-based adaptation (both are teleological), most famously in Jean Baptiste Lamarck's pre-Darwinian model of evolution (Mayr, 1991). For that reason in Chapter 4 I dubbed this new hypothetical hybrid dimension "quasi-Lamarckian." I will discuss this finding at length later in the chapter.

Foundations for a Valid Instrument

In sum, certain modifications are in order – both to the underlying conceptual framework and to individual test items – but there is cause for optimism that this test can be rendered into a valid and reliable instrument for assessing students' understanding of evolutionary theory as well as their adherence to common preconceptions/misconceptions about evolution.

Limitations

All the results and conclusions just summarized are tentative and must be received with care, not only because this is a pilot study, a work in progress, but also because they are surely compromised by important threats to internal and external validity. I discuss these threats below.

Internal Validity

"Internal validity" refers to a study's internal structure and the degree to which the researcher can control extraneous and potentially confounding variables (Gall et al., 2007). It is the degree to which cause-and-effect conclusions are warranted. In general, internal validity is threatened whenever a researcher is in danger of attributing observed outcomes to the study's independent variable or treatments, when in fact the outcomes are at least in part the result of factors or conditions external to the study. Of particular concern in many studies is unwanted variation between treatment groups, especially where it is impossible to assign participants and/or treatments randomly. But because my study did not employ comparison groups, it is immune to such threats. Other threats have to do with unanticipated events or unwanted changes that occur over the duration of the study, especially when the data are drawn from pre- versus post-testing. However, many of these threats do not apply to my study, since data collection took place in a single brief testing session under teacher supervision.

Nevertheless, even in a short, single-cohort study there are certain sources of unwanted change and within-cohort variation that can erode confidence in conclusions and interpretation of outcomes. Most significant in this study are differences across the six schools in testing conditions and/or instrument administration by teachers – variants that may affect students' responses in ways that have nothing to do with their understanding of the theory of evolution. Also significant is the likelihood that the test itself altered some students' initial disposition toward it, perhaps inducing fatigue, haste, or confusion about the meaning of questions, and that this in turn changed their test responses in ways that do not really reflect their conception of evolution.

In this section I review the twelve primary threats to internal validity in educational studies (Gall et al., 2007), discussing how each may bear upon my own study and describing measures I took to lessen their impact.

History and maturation. Given the brief duration that each participant was involved in the study (less than half an hour), it is unlikely that their responses were anomalously affected by cognitive growth or unusual events during data collection. The testing environment was supervised by a professional teacher, and none reported any incidents that could have compromised results.

Testing effects and regression toward the mean. These two threats pertain mainly to studies in which an instrument is administered more than once to the same cohort, and so do not apply to my study. There was no opportunity for students to improve their scores through a "practice effect," nor did this study selectively target students of high or low science proficiency, and in any case there was no post-testing where chance error could bias the scores of such students.

Instrumentation/instrument change. This threat, too, bears mainly upon protracted studies during which the mechanism of data collection could change over time, as when observers develop a bias, a shift in perspective, or a change in skill, effort, or attention. In my study data was objectively transferred from participants to database in a single brief testing session, and so was not vulnerable to this threat. However, *participants'* perspectives or dispositions toward the mechanism of data collection could have changed during the study, thereby introducing an unwanted variable – a threat that I discuss below.

Differential selection and selection-maturation interaction. These two threats amount to a "selection bias" in which membership differs in important ways between two or more treatment groups, either because the participants are dissimilar from the start or become so by maturing at different rates. Since my study did not employ comparison groups, it is immune to this threat. (Differences between participants at the six schools are a threat to *external* validity, as I discuss later.)

Nevertheless, this may be the best place to discuss a kindred limitation that undermines confidence in the cause-and-effect conclusions drawn from my study. Selection bias sometimes goes under the heading "participant characteristics," the idea being that key causal characteristics of participants may differ across groups. In this study, there were almost certainly participant characteristics other than their actual conception of evolution that influenced their test responses. Probably among these peripheral causal factors were the students' cognitive readiness, reading level, English proficiency, and religious orientation (which can affect a person's affective disposition toward the theory of evolution). These may have compromised the instrument's ability to elicit a clear portrayal of their conceptual understanding of evolution. It was not possible to control for these potential modifying variables, but it should be mentioned that students limited by reading ability, cognitive readiness, and so on *are* members of the target population for which the instrument was developed.

Also likely playing into student responses were such peripheral factors as effort, attention, and honesty. These, too, were beyond my control, but there were safeguards in place: Participation was voluntary and limited to students who had made the effort to secure parental consent, they were free to opt out at any point, and their responses were anonymous, such that they could answer genuinely and intuitively without concern that incorrect answers might affect their grades. On the other hand, the absence of rewards and penalties may have made it easier for some students to relax their concentration.

As a further safeguard, I incorporated seven "nonsense" choices into my instrument, statements that I deemed illogical, neither scientific nor intuitive, and plainly false. These were included in part as a quality control mechanism: Any participant accepting a significant fraction of these items might reasonably be suspected of weak effort, motivation, attention, reading level, English proficiency, or cognitive readiness, and so might warrant omission from the dataset. However, after discussion with my doctoral committee, I decided not to disqualify any participants on these grounds, since other factors – such as confusing wording or simply misreading a question – might account for such mistakes.

Experimental treatment diffusion, compensatory rivalry, and resentful demoralization. These three threats to internal validity also pertain primarily to studies with two or more treatment groups, where aspects of an intervention might “leak” from the experimental group to the control group because implementers find it attractive, or where knowledge of being relegated to the control group might lead participants to deliberately bias their performance either positively or negatively. Such threats do not apply to my study. However, rivalry and demoralization are sometimes listed under the heading “participant attitudes,” something that *is* a concern in my study and which I discuss below under the threats of “location” and “experimental mortality.”

Implementation and location. These two threats refer to potential inconsistencies in the execution or environment of a study, an unwanted variable that can compromise cause-and-effect interpretation of results. This is an obvious threat in my study, where seven different teachers in six different school settings administered the instrument. To guard against it, all teachers were given the same careful guidelines for administering the test, and the protocol was simple. Students took the test on school computers, never at home or elsewhere, so the “online environment” was presumably consistent across sites – with one major exception: One of the schools was unable to provide computers for testing, and its participants took a paper test instead. It is possible that the paper format affected student performance. In particular, unlike the online format, the paper format permitted students to refer back to previous questions in answering each new question, and to change those earlier answers. If they did so, it would especially undermine the non-forced choice section's purpose of eliciting raw intuitions and dampening the sense of competition between items. More so than with the online format, the paper test had potential to *teach* students about evolution, leading them to respond to questions in ways that do not reflect the conception(s) of evolution that they had when they began testing.

A related concern is that the timing of test administration differed between the schools. For four of the six school districts, permission to conduct my study was granted on the condition that students not take my test until the very end of the school year, after the season of state-mandated, standardized end-of-course tests. This is a time of year when students may be ill-disposed to take another test, possibly diminishing their motivation or concentration in tackling my own test's often challenging questions.

There was no practical way of avoiding this threat, nor the drawback of paper testing with one school. My study's conclusions regarding instrument validity should therefore be received with appropriate caution.

Experimental mortality. The threat of participant attrition normally bears on protracted studies, not a brief single-session study such as this. But there is a sense in which attrition was a threat here, too. After voluntarily starting the test, some students may have found it more demanding than anticipated and consequently withdrawn midstream. I dealt with this threat in part by discarding from the sample those who did not finish the test. But that may have introduced a new bias in the sample against students who find the questions difficult or prone to lose focus, especially compromising data drawn from later portions of the test. Moreover, it is likely that some of these students, rather than withdraw, rushed through the questions making selections with little or no attention to content. The time stamps provided by the survey software did indeed show that many students completed the test with surprising swiftness. In short, a systematic relationship could exist between the test's difficulty or length and de facto attrition, calling once again for caution in accepting causal conclusions about the instrument's ability to reveal student conceptions of evolution.

External Validity

A study's external validity comprises both "population validity" and "ecological validity" – the degree to which its findings may be legitimately generalized to new audiences and new settings respectively (Gall et al., 2007). I discuss each below.

Population validity. This study's gravest limitation is that it was impossible to draw a random, representative sample from the ideal target population – namely, all high school biology students in the United States. It is thus impossible to generalize results and conclusions cleanly to future audiences. Since I used a non-random sample of convenience, even the most tentative conclusions about instrument validity and reliability will generalize only to the population represented by the sample, and that population is unknown and unknowable. To some extent I did succeed in recruiting students from diverse schools: some large, some small, and some in between; some urban, some suburban, and some rural; some in working class communities, some middle class, some affluent, and some low SES. Students of diverse racial/ethnic heritage were represented, though not in proportion to their percentages in the overall student populace of the United States. Participants came from grades 9 through 12, ages 14 to 18. Despite this breadth, population validity is uncertain, and if the instrument were used with other student populaces, its estimates of student understanding and conceptual change with respect to the theory of evolution would have to be viewed with caution.

One critical audience in particular was gravely underrepresented in the sample: students who had not yet studied the theory of evolution in a science class. Nearly all participants (94%) indicated that they had already studied evolution, with only twenty saying they had not. With data collection dependent on teachers volunteering to participate – most of them biology teachers – and with four of the six school districts requiring that the test be administered at the end of the school year, it was not practicable to adequately include this important audience. Consequently, findings cannot be generalized to students new to the theory of evolution, and thus the instrument cannot yet be safely used for pre-testing in before/after intervention studies designed to assess student conceptual change.

To address these problems of population validity, the next step should be a cross-validation study: field-testing with a new audience that includes students who have not yet studied evolution, and comparison of results, including goodness of fit indices for the same factor analysis models that were tested in this study (Blunch, 2013; Klem, 2000; Thompson, 2000). I discuss this and other recommendations for future research below.

Ecological validity. As with internal validity, most major threats to ecological validity pertain to protracted studies with more than one treatment, typically an intervention study with pre- and post-testing, and so do not weigh as heavily upon a brief, single-session study such as this. For example, the mere novelty of an intervention, or the active involvement of the researcher herself in implementing it, can affect student and teacher performance in ways that will later vanish once the intervention becomes standard practice and the researcher is no longer involved (Gall et al., 2007). Still, something akin to a “novelty effect” or “experimenter effect” could have affected this study’s ecological validity. Students knew they were participating in special research, and this very awareness could have enhanced effort or performance beyond what it would typically be, as could the novel nature of the test and its online testing environment. Likewise, the involvement and attitude of a familiar teacher could have an effect, positive or negative, that would not pertain in future settings. I attempted to dilute such effects by providing teachers with a careful protocol for administering the test and asking them to read a standard statement at the start of testing.

In protracted intervention studies, a pre-test may inadvertently alter a student’s disposition toward the upcoming intervention (“pre-test sensitization”), while the post-test can be an occasion for learning above and beyond the intervention itself (“post-test sensitization”) – effects that will vanish if the intervention is later implemented without pre- and post-testing (Gall et al., 2007). My study should be immune to such threats.

In this study, the main threats to ecological validity have to do with context and timing. While the “online environment” of testing would presumably remain similar in the future (my use of paper tests with one school notwithstanding), the time of year would probably differ, as would students’ perception of the test’s relationship to their actual science classes. Many participants took the test near the very end of the school year, when classes were virtually over. In the future the instrument would likely be used mid-year and would be tied more tightly and transparently to coursework. It might be used in research to evaluate a classroom intervention to help students learn the theory of evolution. Or it might be used by a classroom teacher as a diagnostic test in connection with an instructional unit on evolution. Or it might be used as a benchmark test or data-gathering device employed by school districts in the midst of ongoing biology courses. These are environmental conditions – absent during this pilot study – that might influence student preparation, disposition, and performance. Once again, the findings and conclusions here cannot be generalized cleanly to future audiences.

Missing Comparisons: Other Instruments and Alternate Versions

Besides the limitations imposed by the study's structure and execution (internal validity) and by its sampling scheme (external validity), conclusions are also circumscribed by the absence of any comparison to already published instruments or to alternate versions of the same instrument. I originally proposed to evaluate the instrument's "convergent validity" (Bryant, 2000; Gall et al., 2007) by triangulating the responses of a small sub-sample of test-takers against their scores on two existing instruments: the Conceptual Inventory of Natural Selection (CINS; Evans & Anderson, 2013) and the Open Response Instrument (ORI, Nehm & Schonfeld, 2008). Both of these have already been validated as measures of student understanding of evolution (Anderson et al., 2002; Anderson, Fisher, & Smith, 2010; Nehm & Schonfeld, 2008). Thus if students who score highly on CINS and ORI were to also score highly on my experimental instrument, there would be reason to judge it a good gauge of correct scientific understanding. Confidence would be further boosted if students were to reveal similar misconceptions on all three instruments. Unfortunately, logistical hurdles prevented my teacher volunteers from administering the CINS and ORI to their students as originally proposed.

Confidence in conclusions might also be higher had the experimental instrument's questions not been limited to a single scenario about a single species. To make the test simpler, shorter, faster, and accessible for students, the instrument was deliberately built around a single concrete scenario: the microevolution and speciation of ghost crab populations on isolated islands – specifically, their acquisition of darker coloration in the wake of a volcanic eruption that has turned their once white beaches black. It is possible, however, that some of the students' intuitions were species-specific. For example, some students might believe that crabs can change color in real time to match their surroundings, as chameleons and octopi do. Such a belief might make them more prone *under this scenario* to fall for the common misconception that evolution occurs through individual organisms "adapting" to local circumstances within their lifetimes (see Table 1). On the other hand, perhaps students were *less* prone to fall for this fallacy than they would have been had the scenario centered on an intelligent mammalian species, such as bears or dolphins, since these animals are plainly capable of "adapting" to environmental changes through learning and behavioral flexibility. Some students might find it more difficult to adopt and express a correct scientific understanding when the species of concern is a plant or microbe instead of an animal, or to suppress intuition when the scenario involves the *loss* of a trait rather than the acquisition of a new trait, such as cave salamanders losing their eyes or whales losing their hind limbs (Nehm et al., 2012).

This is a generalizability issue of a different sort: not whether findings can be generalized to other students, but whether they can be generalized to other evolutionary scenarios that students might think about in class or elsewhere. In other words, the understanding of evolution uncovered by my instrument might not be so "abstract" that the student can readily “transfer” it to diverse scenarios and species; her understanding of how crab camouflage evolves may differ from her understanding of other evolutionary processes (Nehm et al., 2012). This study’s conclusions would be stronger if an alternate version of the instrument had been developed and piloted alongside the ghost crab version, parallel in structure and content but centered on another species and scenario.

Comparisons to other instruments like CINS and ORI, or to an alternate version of the same instrument, are candidates for future studies, as discussed in the next section.

Directions for Future Research

This is a pilot study, a work in progress with working conclusions, the initial steps in developing a novel instrument with an atypical structure, inspired and informed by an uncommon orientation: a Darwinian perspective on human psychology and learning. The limitations discussed in the preceding section flag the way to next steps.

Reanalysis

Before revising items and retesting with a new audience, I will conduct some additional analyses with my existing data. First, I will recalculate all the statistics reported in Chapter 4 and rerun the confirmatory factor analysis models after striking from the sample the 6% of participants who indicated that they had never studied the theory of evolution in a science class. This will remove only 20 students from a field of 340, so it is unlikely that the numerical results will shift much. Nevertheless, students who have and have not studied evolution really represent two distinct populations, and the instrument should be field-tested with both independently. Testing with both audiences is especially important because the instrument was designed not just for assessing student understanding (and misunderstanding) of evolution, but for assessing pre- versus post-instructional *conceptual change*.

Next, having rerun the CFA models on the pruned dataset, I will subject it to exploratory factor analysis (EFA) or principle components analysis (PCA). My initial decision to use CFA rather than EFA or PCA was a principled one: I had built the instrument from an a priori theory and conceptual framework, and so did not deem it appropriate to adopt an exploratory search for dimensionality in the data (Anderson, 2008; Blunch, 2013; Bryant & Yarnold, 1995). A consequence, however, is that I may be leaving other empirically plausible and informative factor models unexplored. The emergence during CFA of a quasi-Lamarckian factor (see Chapter 4) shows that students' intuitions may cluster in ways not anticipated by the instrument's initial blueprint. Now that a priori analysis is over, there is warrant for a looser a posteriori approach. EFA and PCA have potential to extract latent factors that might provide insight into participants' reading of test items and their mental model(s) of evolution, and by extension may suggest ways to revise the instrument before the next round of field-testing.

Revisions

As described in Chapter 4, I will revise, replace, or remove test items that performed problematically. I will also attempt to rework test items – especially the isolated statements in the non-forced choice section – so that each chiefly represents only one of the three core intuitions: the essentialist, teleological, and intentional stances. Many well-documented misconceptions about evolution appear to embody a combination of more than one of these three core intuitions (see Table 1; Anderson et al., 2002; Bishop & Anderson, 1990; Evans, 2008; Jensen & Finley, 1996; Mayr, 1991; Nehm &

Schonfeld, 2008; Shtulman, 2006). Nevertheless, it should be possible to slant each statement so that it voices one intuitive stance more strongly than the other two. In effect, the test items will then conform to an essentialist/teleological/intentional typology in addition to – or perhaps instead of – the immutabilist/transformationist/adaptationist typology. There were clues from the original factor analyses that such a revised scheme might strengthen construct validity (see Chapter 4).

Now that the instrument shows promise as a valid measure in the making, its items should also be subjected to a readability analysis and their wording adjusted as needed to render them at the appropriate reading level for the target audience. Finally, prior to the next round of field-testing, this may be an apt occasion to develop an alternate version of the instrument to be field-tested alongside the ghost crab version, parallel in structure and content but centered on another species and scenario. Not only might this make results more generalizable as discussed above, but could also provide researchers, teachers, and instructional leaders with a way to bookend instruction with similar but different assessments – a safeguard against practice effects.

Retesting

The newly revised instrument(s) should be field-tested with two distinct audiences: those who have already studied evolution and those who have not. Ideally testing would occur in the midst of the school year during a biology course's normal flow, either immediately before each class embarks on its study of evolution, or soon after. To provide evidence on the instrument's "convergent validity," a sub-sample of test-takers should also take the already validated CINS or ORI instruments, as originally proposed for the present study. Moreover, a sub-sample of participants should be interviewed after testing to solicit feedback on test items and to gain insight into their reading and interpretation of them: a form of “content validation” from the test-taker’s perspective, complementing that already provided by scientific experts. Finally, the new test responses should be subjected to the same statistical analysis used in the present study, including the same CFA models, and revisions made again as needed.

Once the instrument has been reworked into an acceptably valid and reliable tool, I hope to put it to use in the conceptual change research for which it was tailor-made: classroom intervention studies designed to compare the effectiveness of rival pedagogies representing the three hypotheses elaborated in Chapter 2 to explain the ancestral origins of our cognitive capacity for conceptual change – the “individual constructivist,” “cultural re-constructivist,” and “social strategic” hypotheses.

Implications for Educational Planning, Policy, and Leadership

Despite the present study’s limitations, the tentativeness of its conclusions about instrument validity and reliability, and the plain need for revision and retesting, I believe it already brushes with broad implications for science education, instructional planning, and instructional leadership, especially given the enduring emphasis in state and federal policy on learning standards and standardized assessments. In this section, I first discuss this study’s findings in connection with the policy stance and curricular standards sanctioned by the science education community with regard to evolution. In particular, I propose that the persistence of intuitive misconceptions even after students have learned the theory of evolution – plainly evident in students’ responses on my pilot test – indicates that learning standards ought to explicitly communicate to teachers the need to help students “unlearn” such misconceptions. Thereafter I recommend ways that the instrument itself might be used by professional educators charged with implementing policy and rendering science education standards into successful learning experiences, from curriculum leaders at the state and school district levels to classroom teachers. In particular, I propose its use for data gathering, diagnostic assessment, and even as a classroom teaching tool.

Evolution Education: Policy Stance and Science Education Standards

The eminent geneticist Theodosius Dobzhansky famously wrote, “Nothing in biology makes sense except in the light of evolution” (1973, p. 125). He continued:

> Seen in the light of evolution, biology is, perhaps, intellectually the most satisfying and inspiring science. Without that light it becomes a pile of sundry facts – some of them interesting or curious but making no meaningful picture as a whole. (p. 129)

For biologists, the theory of evolution is not merely one interesting topic alongside others, but the grand unifying framework that pulls all those other topics together into an organic narrative. In biological science, it is paradigmatic in the Kuhnian sense (Kuhn, 1962/1970): a coherent interpretive lens through which biologists habitually look at all biological phenomena, an explanatory model that makes singular sense of the extraordinary variety of life on Earth – from its tiniest molecular machinery to the most sophisticated multicellular organisms to the intricate interrelationships that comprise entire ecosystems (Dawkins, 1986; Dennett, 1995; Dobzhansky, 1973; Smith & Szathmáry, 1995, 1999).

Arguably, then, as a matter of educational policy and standard practice, biological science should be *taught* in high school the way biological science actually *is:* thoroughly cast within the Darwinian paradigm. Evolutionary thinking should saturate the biology curriculum, always already present in how textbooks, teachers, and students make sense of biological phenomena. It should be an omnipresent backdrop, the way that Newtonian mechanics infuses almost every physics lesson and lab activity, the way atomic theory and the periodic table run through every chemistry course. Such a curricular policy would be consistent with the official position of the National Science Teachers Association (2003):

> Evolution has not been emphasized in science curricula in a manner commensurate to its importance….Science curricula, state science standards, and teachers should emphasize evolution in a manner commensurate with its importance as a unifying concept in science and its overall explanatory power….Science textbooks shall emphasize evolution as a unifying concept. (pp. 1-2)

The National Association of Biology Teachers (2008), the National Academy of Sciences (1998), and the American Association for the Advancement of Science (1993) have likewise stressed the centrality of evolution and advocated its explicit incorporation into secondary biology classes as a major organizing pattern and explanatory model. And both the older National Science Education Standards (National Research Council, 1998) and the new Next Generation Science Standards (Achieve, Inc., 2013) emphasize evolution as one of a small handful of essential unifying concepts in natural science. Without a sound understanding of Darwinian theory, a synthetic grasp of diverse biological phenomena and principles is impossible.

Yet while the NSTA, NGSS, NSES, NABT, NAS, and AAAS all emphasize the importance of evolution's *inclusion* within biology courses, it is not clear that they prescribe its *integration throughout*. They ask teachers to teach it, and to teach it as a central unifying theme, but there is nothing that precludes teachers from teaching it only once, in relative isolation, and then neglecting it elsewhere and thereafter. Truer to real biological science would be a policy position that calls for pervasively rendering all (or most) secondary biology curricula – all topics, all content – within the Darwinian paradigm. That position would be reflected in the science education standards sanctioned by the community of researchers, practitioners, and policymakers. Under such a policy stance, a supreme goal would be for students to learn to *think* like professional biologists, to *habitually* interpret the living world through an evolutionary lens.

And that would require, as the highest priority, cultivating conceptual change that is thoroughgoing and enduring. Such a subtle but substantive shift in position and standards within the science education community would prompt researchers and curriculum scholars to rework – and sometimes build anew – the instructional interventions that they strive to disseminate to classroom practitioners within such staple content areas as cellular biology, human anatomy, genetics, and population ecology. Those topics would become vehicles for reinforcing conceptual change with respect to evolution by encouraging students to extend their understanding of it in new directions, while providing recurrent conceptual practice with it. Crucially, an initial study of evolution would have to move to the front end of every biology course, and students would need a solid start on conceptual change, lest they carry their intuitive preconceptions into the other topic areas.

The findings from this doctoral study suggest several implications pertinent to our science education standards and the assessment and instruction that follow therefrom. Given the science education community's strong policy stance in recent decades on evolution as an essential unifying theory that all students should learn, these implications already merit consideration. They would become even more important if the policy stance were to shift as I have just proposed that it should: in effect making conceptual change with respect to evolution *the key* to an authentic understanding of biological science across all major topics. Below I first discuss implications for evolution education specifically, and then I suggest that these implications might even extend to other science content areas as well.

Inclusion of common misconceptions in science standards. As this study and many others have shown (see Table 1), students are prone to develop certain intuitive but unscientific notions about how evolution works, and often continue to harbor them after formal instruction. Almost all the participants in this study had already studied evolution, and on the non-forced choice section of the test, they correctly endorsed a great majority of all eight scientifically valid statements. Yet they were still unable to resist the intuitive draw of the unscientific statements on over half the opportunities to do so. The very purpose of the non-forced choice section was to allow for the possibility that students' pre-instructional intuitive notions may continue to coexist alongside their post-instructional scientific understanding of evolution. The test results were consistent with that possibility.

One implication is that our science education standards, in addition to expecting students to learn the scientific model of evolution by natural selection, should perhaps also call for them to *unlearn* any common misconceptions that they might hold. Standards should perhaps include language that communicates to teachers the importance of explicitly addressing any essentialist, teleological, or intentional intuitions to which students might be susceptible. As two evolution education scholars put it: “Science instruction that explicitly targets students’ naïve epistemology...as well as encouraging metacognitive awareness of implicit folk concepts, is likely to have a much greater chance of successfully producing an educated lay-population and a new generation of scientists” (Poling & Evans, 2004, p. 517). Including language about common intuitions in the standards sanctioned by the science education community would encourage state government agencies to adopt similar language in their own standards, and by extension, to build items into their end-of-course state assessments that seek to uncover lingering misconceptions. And that in turn would encourage teachers to address them in class.

Lamarckian intuitions. The results of this study also point to one constellation of misconceptions that may merit special attention in our science education standards. As I described in Chapter 4, the structural equation model which provided the best fit to the student’s test responses introduced a new latent factor that I dubbed “Lamarckian.” It fused together two of the hypothesized latent factors that originally informed test design: (1) a “transformationist” view of evolution as a progressive process which slowly transforms entire species in the direction of increasing complexity and sophistication, and (2) a “within-lifetime adaptationist” stance that attributes evolutionary change to individual organisms “adapting” to local circumstances during their own lifetimes. The emergence of this new latent factor may warrant its explicit inclusion in our science education standards, for it echoes a mistaken model of evolution that has plagued the science of evolution for centuries and that may have special intuitive appeal for students.

Charles Darwin was not the first to propose that wild populations of plants and animals evolve over the long haul of natural history. Numerous others – including his own grandfather, Erasmus Darwin – had maintained that the gradual transmutation of species over time is demonstrated by the fossil record, geographical distribution of species, and embryological and morphological homologies between otherwise dissimilar organisms. Darwin himself, as a strategic prelude to the unveiling of his own theory, documented many of these predecessors in the opening moves of *The Origin of Species* (Darwin, 1859; Desmond & Moore, 1994). Where Darwin differed dramatically from his predecessors, however, was in his explanation of *how* evolution occurs. Its driving mechanism, natural selection, is a “blind,” branching, non-linear process that changes populations over time without foresight or direction, and that leads species to extinction far more often than not (Darwin, 1859; Dawkins, 1982; Dennett, 1995). Prior to Darwin, most evolutionists instead saw evolution as linear, directional, and progressive, with species inexorably steered – somehow – toward ever more sophisticated and successful forms (Mayr, 1991).

The most famous of these – to whom Darwin paid homage in *The Origin* – was the French naturalist Jean Baptiste Lamarck. According to Lamarck, life had a natural tendency to evolve from simple to complex, with all species past and present steadily climbing the “ladder of life,” driven by unseen alchemical forces (Darwin, 1859; Desmond & Moore, 1994; Mayr, 1991). A primary mechanism by which that happened, he proposed, was behavioral: Individual organisms change their own bodies through use (or disuse) of body parts, and then pass their newly acquired traits to their offspring (Darwin, 1859; Desmond & Moore, 1994; Mayr, 1991). The evolution of long giraffe necks is the ubiquitous textbook example: Early giraffes acquired slightly longer necks by persistently stretching for distant leaves on high, perhaps especially during periods of drought and food shortage. Their offspring inherited the additional length, then stretched their own necks even more, and the cycle repeated for many generations until eventually the species had evolved an altitude worthy of zoo attractions and safari pursuits.

Lamarck's explanation of evolution embraces all three of the intuitive, arguably innate tendencies that I elaborated in Chapter 2. The essentialist and teleological stances appear in the progressive, extinction-free ascent of whole species up the ladder of life. The teleological stance and shades of the intentional stance appear in the behavioral efforts of organisms to acquire resources for survival. The two major components of Lamarck's theory correspond roughly to what I have called transformationist and within-lifetime adaptationist misconceptions respectively. Lamarckian fallacies and kindred intuitions have long plagued both public and professional perceptions of evolution, even after Darwin corrected them and even among accomplished biologists (Dawkins, 1976/1989, 1982; Dennett, 1995; Mayr, 1991; Williams, 1966).

It is perhaps unsurprising, then, that in my own study, a quasi-Lamarckian factor appeared in the structural equation models that performed best during confirmatory factor analysis. Lamarckian propositions may hold special allure for the evolved human mind, and as a result, Lamarckian explanations of evolution may be especially likely to recur both in the history of science and in the mental models constructed by science students. For that reason, our science education standards should perhaps explicitly address

Lamarckism, not merely as a historical curiosity (as it is often treated), but as a genuine conceptual pitfall to avoid. Again, state standards might then follow suit and channel the concern to classroom instruction and end-of-course assessments.

Assessing standards through a "non-forced choice" format. If, as I have just suggested, the science education community should include in its curricular standards the need to help students escape Lamarckian and other intuitive pitfalls, with state education agencies following in kind, then a further implication follows from my study: Science assessments such as state end-of-course biology tests should include items to gauge students' adherence to common misconceptions, and moreover, should sometimes do so in a way that allows those misconceptions to surface side-by-side with correct scientific conceptions. Again, this study has shown that even when students are adept at correctly identifying scientifically valid statements about evolution, they may nonetheless continue to find unscientific intuitions appealing. Open response items are the usual way to invite scientific and unscientific ideas to surface simultaneously, but such a format is probably not feasible for statewide assessments. My experimental instrument shows a potentially more practicable strategy that test developers should consider: the use of a "non-forced choice" format that presents each scientific and unscientific idea in isolation, deliberately downplaying the sense of having to choose between them that the conventional multiple choice format evokes.

Curricular implications beyond the theory of evolution. The focus of this study was evolution education, but as I made clear in Chapter 2, I chose it as a "test case" for investigating the nature of scientific conceptual change more broadly. Although the implications that I tendered above were specific to the theory of evolution, they may extend to other challenging concepts as well. Cultivating conceptual change can be notoriously difficult, and for decades it has been among the most intensely researched areas in science education across all the sciences (Southerland et al., 2007; Vosniadou, 2008b). Often the difficulty is not just that the new concepts are unfamiliar and inherently challenging – too abstract, too abstruse, too complicated – but also that they do not neatly mesh with students' preexisting mental models about the natural phenomena of interest; those models may be deeply entrenched in a student's conceptual habits and resistant to revision (Carey, 2000, 2009, Chi, 1992; Chinn & Brewer, 1993; Driver et al., 1994; Posner et al., 1982; Vosniadou, 2008a). My Chapter 2 literature review showed that for students confronting the theory of evolution, those pre-instructional understandings are plausibly fueled by natural intuitions that are innate, hence especially hard to dismiss. It was that very possibility which motivated me to develop my experimental instrument and informed its design. The same could be true for other challenging concepts in natural science. Learning them may require students to "unlearn" (or at least suppress) preconceptions that are anchored in powerful, perhaps even innate, intuition.

Besides evolution, then, there may be other major ideas for which the curricular standards sanctioned by the science education community should explicitly address common student preconceptions/misconceptions. The new Next Generation Science Standards (Achieve, Inc., 2013) have placed special emphasis on the core concepts, fundamental principles, and unifying theories within each scientific discipline, as well as broad "cross-cutting" concepts such as "system" and "scale" that span more than one discipline. The NGSS architects deliberately kept these concepts few in number, and the

NGSS explicitly advocate instruction for deep understanding and fluency that enables students to apply the concepts flexibly to different situations – in effect, cultivating them as enduring habits of thinking, much as I argued above for the theory of evolution. Given this (welcome) emphasis, it is more important than ever for science educators to recognize that mastering the big ideas often entails "unlearning" certain common misconceptions, and for that reason those misconceptions may merit explicit articulation in curricular standards.

Further, as this study showed with evolution, students might continue to harbor misconceptions even after successfully acquiring these major science concepts, and for this reason, assessment tools may be needed that gauge the persistence of pre-instructional intuition in addition to scientific grasp. Here, too, portions should probably be designed to permit intuitive and scientific conceptions to surface side-by-side without conflict, while others should oblige students to discriminate and choose between them.

Instructional Planning and Leadership

The implications discussed above stem from this study's actual findings – that is, my analysis of student test responses – especially as they intersect with policy and standards sanctioned within the science education community. This section now suggests potential uses of the instrument itself (once revised and reevaluated) by professionals in the field charged with translating policy and standards into action, from curriculum leaders at the state and school district levels to classroom teachers.

The policy backdrop, of course, is the movement in recent decades to mandatory statewide learning standards coupled with high-stakes end-of-course assessments. Until the 1980's educators were generally respected by the U.S. public, and their expertise trusted; but public perception then shifted dramatically to a view that education was in crisis, inciting demands for accountability and top-down assertion of authority (Fowler, 2009). State-mandated standards and tests proliferated in the 1990's, and then became a national norm with the passing of the No Child Left Behind Act of 2001 (NCLB, 2002). Each state's central education agency is charged with translating the broad-stroke policies handed down by state legislators under NCLB into regulations, standards, incentives, and assessments to guide implementation, while implementation itself happens primarily at the district, school, and classroom levels (Fowler, 2009). A persistent challenge is that the pipeline from state agency to classroom is very long with multiple intervening levels, such that policy and standards may be *transformed* as much as transferred and translated en route to implementation (Clune, 1993; Weick, 1976). To maximize continuity and fidelity of implementation, curriculum and instructional leaders need useful tools and resources – consonant with their own reading of policy and its objectives – that they can sanction and share with practitioners. They also need diagnostic and data-gathering devices to help them survey the instructional terrain.

Below I suggest ways that my instrument might serve such functions (at least in those states whose science standards echo the science education community's emphasis on evolution as an essential unifying theory, and thus prioritize conceptual change). It may prove even more useful in the years to come with the advent of the new Every Student Succeeds Act (ESSA, 2015), which supplants NCLB and promises to grant state agencies more flexibility to assess student learning via multiple measures.

Diagnostic testing and data gathering. For science coordinators in public school districts who wish to gauge secondary biology students' understanding of the theory of evolution – e.g., in advance of the next round of state assessments – my instrument could be especially useful, since it is designed for large-scale online administration with automatic transfer of student responses to a central database. It is easily scored, and because multiple test items target each sub-component of Darwinian theory as well as the various misconceptions (see Tables 1, 3, and 4), it can provide instructional leaders with a granular profile of where students' grasp of evolution is strong and where further instruction or remediation may be in order. Results can be readily communicated to teachers, and through them to their students. Moreover, with the new Every Student Succeeds Act (2015) promising to broaden states' options for having students exhibit their learning – including active input by professionals at the local level – it may soon become more reasonable for school districts to employ mid-year benchmark assessments that do not slavishly mimic the state's standardized end-of-course test items. That might free instructional leaders to use instruments like this one to focus more on deep understanding and authentic conceptual change.

Likewise, for science education specialists in state level departments of education, the test offers a relatively simple online means of sampling student conceptions of evolutionary theory statewide. Such data gathering might influence the strategic dissemination of resources, from instructional materials to professional development initiatives to grant opportunities. Developers of state assessments might even borrow items, ideas, wording, or structures from the instrument, perhaps even the unconventional non-forced choice format. (This is not to suggest that the test could be used as a summative end-of-course assessment itself. It is too long for that, and was neither designed nor field-tested with that function in mind. I also want to repeat that the instrument will need revision and another round of field-testing before it is ready for any of the above applications.)

Conceptual change in the classroom. Less formally, instructional leaders at the state and school district levels might share the instrument with biology teachers and encourage them to put it to use within the walls of their own classrooms. Apart from its intended use in intervention research, I believe this is where the instrument could have its greatest value. As I maintained earlier, it is critical for classroom practitioners to recognize that conceptual change with difficult concepts like natural selection and macroevolution is not a merely additive process. Students are not blank slates, and they will likely have to “unlearn” – or at least learn to suppress – intuitive but unscientific preconceptions that they bring to the classroom. Instructional leaders should help make teachers aware, as this study and others have demonstrated, that those preconceptions can persist even after a student appears to have a handle on the scientifically normative explanation. As I hypothesized in detail in Chapter 2, student ideas about evolution may not constitute a unified mental model; instead, intuitive interpretations of biological phenomena may coexist shoulder-to-shoulder in the same mind beside scientific ones. On tests and elsewhere, students may thus be prone to construct and communicate “synthetic” or “mixed” models (Vosniadou et al., 2008) that hybridize pre- and post-instructional conceptions. Instructional leaders should help make teachers aware that traditional multiple choice, true/false, and short answer test questions may miss something important, that conceptual change might not be enduring, and that relapse may be likely.

The very structure and content of my instrument has potential to help teachers bear all this in mind when designing and delivering instruction. That structure and content also has potential to help students discern their own intuitive tendencies, and hopefully rein them in. Toward these ends, I suggest that instructional leaders recommend my instrument to teachers for two uses *in tandem:* as both a diagnostic assessment and a teaching tool. For example, teachers could use it as a formative assessment *after* formal instruction to gauge their students' budding grasp of evolutionary theory, and then immediately thereafter make its puzzling questions and vivid central scenario (ghost crab camouflage on volcanic island beaches) the subject of class dialogue and debate. During such a forum, teachers would be able to provide feedback and help students develop metacognitive awareness of their own lingering intuitive inclinations. Alternatively, a teacher might administer the test as a pre-assessment *before* breaking into an instructional unit on the topic, and then launch that unit with a public forum where students wrestle with the test's scenarios and many rival propositions about evolution.

In the interest of bringing things full circle, I will stretch this a little further. The idea of using my instrument as a diagnostic assessment and teaching tool in tandem fits nicely with the "social strategic" hypothesis that I championed in Chapter 2. According to that hypothesis, our 21st century capacity for conceptual change with respect to modern science's non-intuitive, even counterintuitive truth claims traces back to an ancestral social environment in which our hominin forebears were regularly confronted by deceptive or distorted truth claims put forth by others in the social group. In response, the human mind evolved an ability to suppress and suspend intuition while remaining receptive to – though healthily skeptical about – rival propositions proclaimed in the public sphere. Also according to that hypothesis, teachers can create contexts in the modern science classroom that will mobilize the mind's "social strategic" mechanisms to enable students to inhibit the intuitive temptation of essentialist, intentional, and teleological misconstruals of evolution, while being receptive to the correct but less intuitive scientific model. In this case, teachers might stage the classroom as a public forum where students confront rival explanations of the same natural phenomenon – the evolution of ghost crab camouflage – and oblige students to make decisions and take positions of their own in a publicly visible manner. In short, I am suggesting that my instrument might fruitfully be used by instructional leaders and classroom practitioners not only to assess student conceptual change, but to help it happen as well.

Summary: Implications for Educational Policy, Planning, and Leadership

This doctoral study was set in motion by a Darwinian perspective on the important educational issue of scientific conceptual change. I have revisited that perspective throughout this chapter. But even if the evolutionary accounts that I have postulated about human cognition and student learning are, in the end, erroneous, this research – and the instrument that it spawned – still has value if the following two conditions are met: (1) that students can simultaneously hold more than one competing mental model of the same natural phenomenon, and (2) that changes in classroom context and/or instructional practices can differentially evoke or bolster those competing models. Student responses during field-testing of my instrument strongly suggest that the first condition does indeed hold, at least for students learning the theory of evolution. Pre-instructional intuitions may continue to surface even after students have learned to correctly identify and voice the scientific theory. The second condition is ripe for testing through classroom intervention studies, and it was for such research that my instrument was primarily developed.

Beyond its use in research, however, I have proposed several other useful functions that my instrument (once revised and reevaluated) has potential to serve in the domain of actual educational practice – especially for professional educators charged with implementing policy and rendering science education standards into successful learning experiences for students. Because it was designed for large-scale online administration with automatic transfer of student responses to a central database, science specialists in state education agencies and local school districts may find it useful for diagnostic testing and/or data gathering. Classroom teachers, too, may find it useful as a diagnostic assessment, and I have proposed that they might even put it to good use as a teaching tool by stimulating class discussion and debate around its vivid scenarios and challenging questions.

I also suggested, based on the findings in this study, that the learning standards sanctioned within the science education community, in addition to articulating the scientific concepts that students are to master, should also explicitly identify common misconceptions that students may need to "unlearn." In the case of evolution, this study yielded signs that students may be especially susceptible to Lamarckian misconceptions, which have plagued evolutionary science for centuries. Furthermore, science assessments such as state end-of-course biology tests should perhaps include items that allow misconceptions to surface side-by-side with scientific conceptions. My instrument's unconventional "non-forced choice" format shows one potential strategy that might be practicable for statewide assessments.

There is one more connection to curricular policy, instructional planning, and classroom practice that I wish to draw. I will close this doctoral thesis with that.

Concluding Remarks: Less is More

An effect of the No Child Left Behind Act and the rise of high stakes testing – however unintentional – was to prompt many states to compose science education standards that attempted to exhaustively encompass the enormous scope of scholarship in each of the major scientific disciplines. This in turn encouraged many science teachers to prioritize content coverage over conceptual depth:

> If learning science was simply a process of accreting information or adding knowledge to already well-structured conceptual frameworks, such broad curricula might be congruent with the reforms. However, the difficulty of these more detailed and extensive state standards is that the shear amount of material to address in the school year often serves to prohibit a robust, clear, intensive treatment of foundational ideas….Focusing on too many details prevents students from grasping the foundational ideas, or the "big picture," of the science they are studying….Current learning theory focuses on understanding broad conceptual ideas and enculturation into the practices and discourses of science. From this perspective, learning is understood to be a long, complex process requiring an engaged learner. This conception of learning is at odds with the wide scope of the content curricula found in many states, their associated assessments, and the common pedagogical response to these extensive curricula…i.e., drilling months before the examination, wide content coverage to ensure student recognition of

> maximum amount of material, approaching science as a vocabulary exercise. (Southerland et al., 2007, p. 62)

Mastery of "big picture" explanatory models like Newtonian mechanics, atomic theory, the periodic table of elements, and the theories of relativity, plate tectonics, and evolution arguably constitute the most important content goals in science education (Clement, 2008). Consequently, many policy advocates in the science education community have protested the influence of NCLB and espoused against it a "less is more" philosophy. For example, the Carnegie Commission on Mathematics and Science Education (2009) urged that "standards be reshaped to counteract the tendency in American education to cover too much material in too little depth" (p. 26), and declared that science learning should instead revolve around a tight suite of unifying themes, key concepts, core principles, and essential skills, with students getting abundant practice in applying all these across multiple contexts via rich, authentic scientific inquiries. Among its highest priorities, the Commission urged the adoption nationwide of "common standards in math and science that are fewer, clearer, and higher" (p. viii). That would-be policy was soon given expression in the new Next Generation Science Standards (Achieve, Inc., 2013), which organized learning expectations around a small set of "disciplinary core ideas" and "cross-cutting concepts."

This "less is more" orientation also accords well with the Darwinian stance that I have taken toward scientific conceptual change throughout this doctoral study. And it accords especially well with the "social strategic" hypothesis and the hypothetical model of the evolved human mind (Figure 1) that I advanced in Chapter 2, which together were my motivation for developing a new instrument and for giving it its unconventional design. If that hypothesis and model are valid, an important implication is that cultivating genuine and enduring conceptual change requires patience on the part of teachers, for deep evolutionary countercurrents are ever at work. If conceptual change is not a process of "restructuring" or "replacing" a student's intuitive theories, but instead requires that she "override" or "suppress" them, then whenever salient cues, contexts, or content present themselves, *both* her folk intuitions *and* newly acquired scientific understanding may be aroused simultaneously. Only through conscious, effortful *practice* can she learn to habitually give the latter the upper hand in her communications, reasoning, and decision-making. And that takes time. Many conceptual change researchers have stressed that the cultivation of enduring conceptual change will necessarily be a protracted process (e.g., Carey, 2009; Chi, 1992; Chinn & Brewer, 1993; Clement, 2008; diSessa, 2008; Dole & Sinatra, 1998; Vosniadou et al., 2008). It will not take root unless science educators limit the number of big new concepts that they ask students to master, stitch them into the curriculum as pervasive and recurrent interpretive themes, and allow plenty of time for empirical experiences, experimental investigations, active applications, dialogue, debate, and the conscious construction by students of new conceptual lenses for making sense of the natural world. To plow students through a "mile wide, inch deep" body of facts and ideas is to fundamentally misunderstand the young mind and its long natural history.

Appendix A: Final Instrument

On the next page is the instrument that was field-tested in this doctoral study: 10 "forced choice" (multiple choice questions) and 24 "non-forced choice" (yes/no questions). Each forced-choice item obliges the test-taker to select a scientifically normative statement amidst a field of two or more intuitively appealing distracters. Each non-forced choice item is an isolated statement that embodies *either* a scientifically normative concept *or* an intuitively attractive but unscientific one, and the student is asked to judge whether it could be part of a biologist's explanation of the phenomenon at hand. Each non-forced choice statement and forced choice option is designated below as either scientific (SCI) or belonging to one of three families of unscientific-but-intuitive alternative conceptions: an "immutabilist" conception that species cannot change (IMB), a "transformationist" conception that evolution progressively transforms species as a whole (TFM), or a "within-lifetime adaptationist" conception that evolution occurs when individual organisms adapt to their environments during their own lifetimes and then pass their new adaptations to their offspring (see Tables 1 and 3). During online test administration, non-forced choice statements appeared one at a time in random sequence to dilute any order effects. Forced-choice items appeared in the sequence given here.

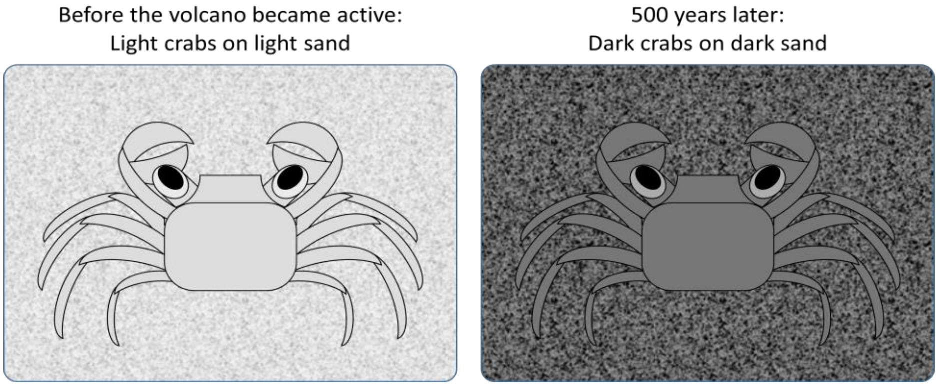
(continued)

Root Scenario: Ghost Crabs on Beaches

Beaches all over the world are inhabited by ghost crabs. They are called ghost crabs because they are so well camouflaged that it creates an illusion of ghost-like "transparency" against the sand. But they are not really transparent. Rather, each population has markings that closely resemble the color and composition of the local beach. This protects them from visual predators such as sea gulls flying high above. Beaches with light sand have light crabs, dark beaches have dark crabs, and speckled beaches have speckled crabs.

Images: Dorothy Pugh, Chuck Elzinga, Ray Farm (right).

Now suppose there's an island way out in the Pacific Ocean, with beaches of very light-colored sand. As expected, the ghost crabs here have a light color that, on average, closely matches the sand. There is a volcano on this island, but it has been dormant for many centuries. Then one day the volcano becomes active again and begins spewing huge clouds of black ash and dust into the sky. As the dust slowly settles back to Earth, it gradually makes the local beaches darker. 500 years later, the ghost crabs on these beaches are much darker than before, similar in shade to the newly darkened sand.



Before the volcano became active:
Light crabs on light sand

500 years later:
Dark crabs on dark sand

PART ONE: Non-Forced Choice Items

Main Prompt: How do you think a scientist – like a professional biologist who studies crabs – would explain this change? You will be shown a series of 24 statements. Please indicate whether each statement could be a part of the biologist's explanation, or if it would not be part of her explanation.

(ex) "The crabs in each generation are able to change colors to match the current background."

☐ Yes, this could be part of the biologist's explanation

☐ No, this would not be part of the biologist's explanation

Correct/Scientifically Normative (SCI)

"Even before the volcano began spreading black dust on the beaches, some crabs were already slightly darker than others."

"In each generation some crabs are darker and some are lighter, but the average darkness of the population is changing from one generation to the next."

"Coloration varies quite a bit from crab to crab, and those whose markings best match the beach are able to reproduce most successfully."

"As generations go by, genes for lighter colors get weeded out of the population."

"Once in a while, a chance mutation gives an offspring a lighter or darker shade."

"Because only a small percentage of young crabs can make it to adulthood, coloration may decide who survives and who doesn't."

"In the competition to avoid predators, darker offspring have an advantage over lighter offspring."

"Offspring who happen to be born darker tend to outlive the others, and so pass on their better genes, mutations, and traits."

Natural Kinds/Essences are Immutable (IMB) *(potentially intuitive but unscientific)*

"Since one kind of crab cannot turn into a different kind of crab, the original population probably went extinct or migrated away in search of a more suitable habitat. A new and darker kind of crab must have moved in and colonized the volcanic beach."

"There must have been two separate species of crabs in the area from the beginning – one light, one dark. After the volcano began blackening the beach, the dark species must have moved in, outcompeted the light species, and taken over the beach habitat."

"The dark species must have been on the island all along, but before the eruption, the light species dominated the white sand beaches. But as the beach grew darker, most of the light crabs were eaten by predatory birds. Now the dark species came to dominate the beach."

"The light crabs mated with a species of dark crab, creating a hybrid species that was better camouflaged against the newly darkened beach."

Evolution as Progressive Whole-Species Transformation (TFM) *(intuitive but unscientific)*

"Now that the population has adapted to the newly blackened beach, all the crabs are equally dark and well camouflaged."

"Although the crabs in each generation are basically the same color, their offspring are born a bit darker."

"Evolution is the process by which nature improves itself over time, and in this case it will push the crab population toward a perfect fit with the environment."

"The species of crab became darker so that it would not go extinct."

Evolution via Within-Lifetime Adaptation (WLA) *(intuitive but unscientific)*

"Mutations for darker coloration arose in the crabs because they needed camouflage to avoid predators."

"Exposure to the volcanic dust caused helpful mutations in the crabs' DNA."

"Some crabs learned that eating certain foods would make their shells darker, and passed this learning to their offspring."

"To grow, crabs periodically 'molt' or shed their exoskeleton (outer 'shell'), then grow a new one. Within limits, a crab can make its shell a little darker when it molts. Over many generations, this adapts the species to its changing environment."

Nonsense/False/Illogical (NON) *(neither scientific nor intuitive)*

"Predatory birds would be just as likely to spot a darker crab as a lighter crab, no matter what the current color of the beach."

"The crab population was already in the process of becoming darker, so the volcano erupted in order to give them a more suitable beach habitat."

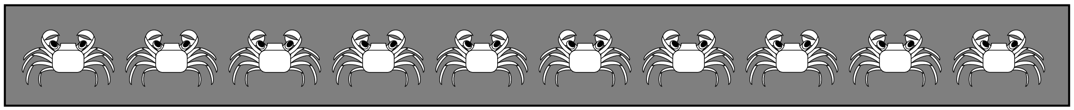
"The volcanic dust raining down from the sky stuck to the crab's shells like paint, making them better camouflaged."

"Exposure to the volcanic dust caused harmful mutations that prevented the crabs from reproducing."

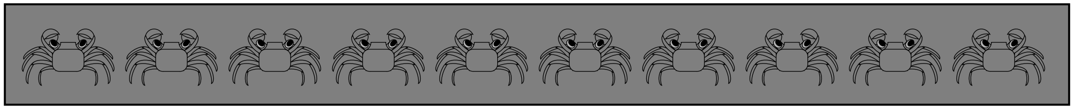
PART TWO: Forced Choice

Multiple Choice (10 Questions): Please choose the answer that a biologist would select.

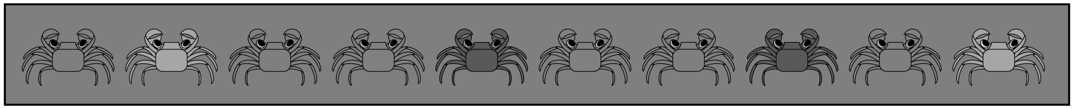
A



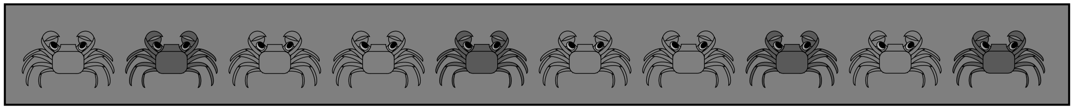
B



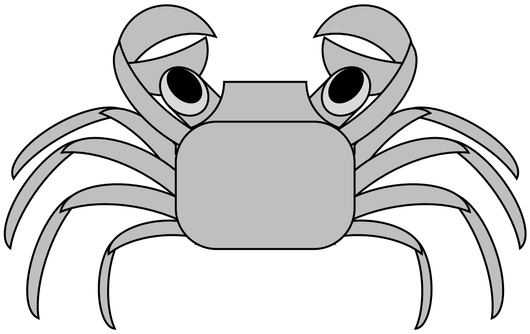
C

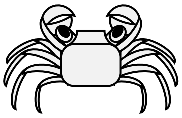


D

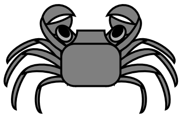


Suppose that after 250 years of volcanic activity, the once white beach has become medium gray, as shown in the background above. If a biologist took a random sample of 10 crabs from the population, which of the above would she expect her sample to be like? Choose the best answer: A, B, C, or D. *[A = IMB, B = TFM, C = SCI, D = TFM]*





Slightly lighter



Slightly darker

Volcanic ash has slowly been dusting the beach for many years now, and the beach is steadily getting darker. Shown above is a big mother crab. To her right are two very young crabs. Their color is the same as the day they were born. Which might be the offspring of the mother crab?

A. Only the lighter gray crab *[NON]*
B. Only the darker gray crab *[TFM]*
C. Neither crab *[IMB]*
D. Both crabs *[SCI]*

How did darker colors first appear in the crab population?

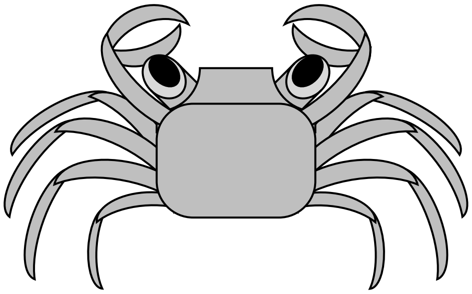
A. Darker markings arose because the crabs needed to blend in with the ever darker sand. *[WLA]*
B. Random changes in DNA gave some crabs a darker color and some crabs a lighter color. *[SCI]*
C. The new environment stimulated helpful changes in the crabs' genes. *[WLA]*
D. Members of the original light-colored species mated with crabs from a different, darker species. *[IMB]*
E. The crabs deliberately changed color to blend in with the background. *[WLA]*

A population of crabs has hundreds of individuals of a single species. Which sentence best describes the group of crabs?

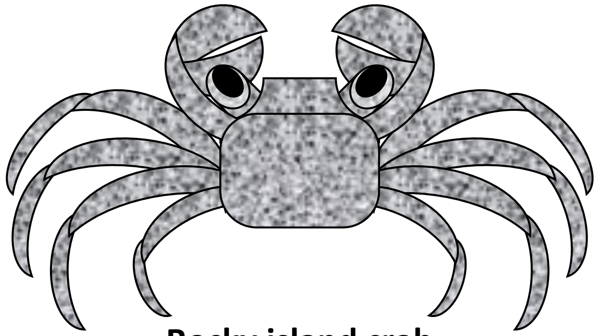
A. The crabs share all the same traits and are identical to each other. *[TFM]*
B. The crabs share all of the most important traits, and the small differences between them do not affect how well they reproduce or how long they live. *[TFM]*
C. The crabs share all of the most important traits, but also have differences that affect how well they reproduce or how long they live. *[SCI]*
D. The crabs share very few traits and are completely different from each other. *[NON]*

500 years after the volcano first erupted, the crabs are much better adapted for a darker beach. What's the best way to explain this adaptation?

A. The changing environment caused beneficial mutations to meet the crabs' new needs. *[WLA]*
B. Beneficial traits became more common because crabs that had them reproduced more. *[SCI]*
C. Each crab gradually adapted to the slowly changing environment. *[WLA]*
D. The original species must have been outcompeted by a better adapted species. *[IMB]*



Sandy island crab



Rocky island crab

Augmented Scenario

About 50 miles away there's a second island. Instead of soft, sandy beaches, this other island has rough, rocky, rugged shorelines. At first, this rocky island had NO crabs of its own. But then one day – many, many years ago – a big storm transported some crabs from the sandy island over to the rocky island on pieces of floating driftwood. Today, the crabs on the rocky island are rather different from crabs on sandy island. The rocky island crabs now have longer, stronger legs for clambering over the craggy terrain. Also, the rocks here are speckled, and so are the crabs. Finally, the main food here is hard-shelled mussels and oysters growing on the wet rocks, and the rocky island crabs have sturdy, shell-crushing claws to get to the soft meat inside. This is different from the sandy island, where the crabs have tweezer-like claws for plucking prey from sand and seaweed.

Today, many people would say the sandy island crabs and rocky island crabs are two different species, even though they came from the same ancestor many years ago. How could this have happened?

A. Over time, many genetic changes happened in each group until they could no longer breed with each other. This made them different species. *[SCI]*
B. They shouldn't be classified as two different species. Because the rocky island crabs descended from sandy island crabs, they still belong to the same species. *[IMB]*
C. The rocky island population changed into a new species suited to its new environment in order to prevent the population from going extinct. *[TFM]*
D. The rocky environment caused crabs to develop the traits they needed for survival. The new environment turned them into a new species. *[WLA]*

On the rocky island, how did crabs come to have stronger legs and claws?

A. Every animal has the ability to adapt to new or changing conditions that it faces during its life. So over many generations, the species' legs and claws got better adapted for the new habitat. *[WLA]*
B. The crabs' legs got stronger by climbing up and down the rugged terrain, and their claws got stronger from pinching the tough shells of their prey. Their offspring then inherited these stronger legs and claws. *[WLA]*
C. The new high protein food source (mussels and oysters) made their muscles stronger, and their offspring inherited this stronger musculature. *[WLA]*
D. Crabs with stronger legs and claws got more food and had more healthy offspring than crabs with weaker legs and claws. *[SCI]*

Suppose a scientist now conducts an experiment: She transplants 50 sandy island crabs to the rocky island. She predicts that "natural selection" will occur – sometimes called "survival of the fittest." What does she mean?

E. Because the species is not fit for this harsh new habitat and cannot change, the species will be weeded out. *[IMB]*
F. Crabs that keep themselves physically fit will have a chance to survive in this harsh new habitat. Less fit crabs will perish. *[WLA]*
G. Some crabs will outlive others because they happen to have traits that are better suited for the new habitat. *[SCI]*
H. Because the native rocky island species is already well adapted for this habitat, the native crabs will quickly outcompete the new crabs. *[IMB]*

After the scientist releases the 50 sandy island crabs on the rocky island, they immediately start climbing up and down the rugged shoreline. All this exercise makes their leg muscles stronger. They all survive long enough to mate with each other. When their offspring are born, how will the average strength of their legs compare to newborn crabs back on the SANDY island?

A. These offspring will probably be born with stronger legs than those back on the sandy island. *[WLA]*
B. These offspring will probably be born with weaker legs than those back on the sandy island. *[NON]*
C. These offspring will probably be born with legs of similar strength to those back on the sandy island. *[SCI]*

Before releasing the 50 sandy island crabs on the rocky island, the scientist measures the size of their claws. The only food here is hard-shelled oysters and mussels that are difficult to break into. The scientist predicts that future crab populations could have stronger claws than now. For that to happen, what would probably have to be true?

A. The 50 crabs start with small differences in claw size that give some an advantage in feeding on mussels and oysters. *[SCI]*
B. The 50 crabs start with the same size claws, but the protein-rich oyster and mussel meat makes their claws more muscular. *[WLA]*
C. The 50 crabs start with the same size claws, but their offspring have stronger claws, and each generation thereafter grows stronger claws than the previous generation. *[TFM]*
D. The 50 crabs start with the same size claws, but some mate with rocky island native crabs and have hybrid offspring with stronger claws. *[IMB]*

Demographics / Moderating Variables

That's the end of the test. Thank you! The following questions are for statistical purposes only. The researchers who designed the test want to try it out with a healthy diversity of students. These questions are optional but your answers are appreciated.

What is your age?

Gender? (male, female)

Race/Ethnicity? (American Indian or Alaska Native, Asian or Pacific Islander, Black or African American, Hispanic or Latino, White, Other: _____________)

Do you attend a public or private school?

What grade are you in? (7, 8, 9, 10, 11, 12, college)

Are you currently taking biology? If so, what grade do you think you will receive for the course? (A, B, C, or D or lower) If you already took biology, what grade did you get? (A, B, C, or D or lower)

Have you already studied the theory of evolution in biology class? (yes, not yet)

References

Achieve, Inc. (2013). *Next generation science standards.* Retrieved from http://www.nextgenscience.org

American Association for the Advancement of Science (1993). *Benchmarks for science literacy*. New York: Oxford University Press.

Anderson, D. L., Fisher, K. M., & Norman, G. J. (2002). Development and evaluation of the conceptual inventory for natural selection. *Journal of Research in Science Teaching, 39,* 952-978.

Anderson, D. L., Fisher, K. M., & Smith, M. U. (2010). Support for the CINS as a diagnostic conceptual inventory: Response to Nehm and Schonfeld (2008). *Journal of Research in Science Teaching, 47,* 354-357.

Anderson, D. R. (2008). *Model-based inference in the life sciences: A primer on evidence*. New York, NY: Springer.

Arbuckle, J. L. (2014). IBM® SPSS® AMOS™ 23 User's Guide. Crawfordville, FL: AMOS Development Corporation.

Ariely, D. (2008). *Predictably irrational*. New York, NY: HarperCollins.

Ariely, D. (2012). *The (honest) truth about dishonesty*. New York, NY: HarperCollins.

Asterhan, C. S. C., & Schwarz, B. B. (2007). The effects of monological and dialogical argumentation on concept learning in evolutionary theory. *Journal of Educational Psychology, 99,* 626-639.

Atran, S. (1994). Core domains versus scientific theories: Evidence from systematic and Itza-Maya folkbiology. In L. A. Hirschfield & S. A. Gelman (Eds.), *Mapping the* *mind: Domain specificity in cognition and culture* (pp. 316-340). Cambridge, UK: Cambridge University Press.

Atran, S. (1995). Causal constraints on categories and categorical constraints on biological reasoning across cultures. In D. Sperber, D. Premack, & A. J. Premack (Eds.), *Causal cognition: A multidisciplinary debate* (pp. 205-233). Oxford, UK: Clarendon Press.

Atran, S. (1998). Folk biology and the anthropology of science: Cognitive universals and cultural particulars. *Behavioral and Brain Sciences 21*, 547-609.

Atran, S. (2002). *In gods we trust: The evolutionary landscape of religion.* Oxford, UK: Oxford University Press.

Bahcall, J. N. (1969). Neutrinos from the sun. *Scientific American 221*(1), 28-37.

Baillargeon, R., Kotovsky, L., & Needham, A. (1995). The acquisition of physical knowledge in infancy. In D. Sperber, D. Premack, & A. J. Premack (Eds.), *Causal cognition: A multidisciplinary debate* (pp. 79-116). Oxford, UK: Clarendon Press.

Baillargeon, R., Spelke, E. S., & Wasserman, S. (1985). Object permanence in five-month-old infants. *Cognition, 20*, 191-208.

Barkow, J. H., Cosmides, L., & Tooby, J. (Eds.) (1992). *The adapted mind: Evolutionary psychology and the generation of culture.* New York, NY: Oxford University Press.

Baron-Cohen, S. (1995). *Mindblindness: An essay on autism and theory of mind.* Cambridge, MA: MIT Press.

Berlin, B., Breedlove, D. E., & Raven, P. H. (1973). General principles of classification and nomenclature in folk biology. *American Anthropologist, 74*, 214-242.

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

Bjorklund, D. F., & Pellegrini, A. D. (2002). *The origins of human nature: Evolutionary developmental psychology*. Washington, DC: American Psychological Association.

Blackmore, S. (1999). *The meme machine*. New York, NY: Oxford University Press.

Blunch, N. J. (2013). *Introduction to structural equation modeling using IBM SPSS Statistics and AMOS* (2nd ed.). Thousand Oaks, CA: Sage.

Boyer, P. (2001). *Religion explained: The evolutionary origins of religious thought.* New York, NY: Basic Books.

Bransford, J. D., Brown, A. L., & Cocking, R. (Eds.) (1999). *How people learn: Brain, mind, experience, and school.* Washington, D.C.: National Academies Press.

Brumby, M. N. (1984). Misconceptions about the concept of natural selection by medical biology students. *Science Education, 68,* 493-503.

Bryant, F. B. (2000). Assessing the validity of measurement. In L. G. Grimm & P. R. Yarnold (Eds.), *Reading and understanding more multivariate statistics* (pp. 99-146). Washington, DC: American Psychological Association.

Bryant, F. B., & Yarnold, P. R. (1995). Principle-components analysis and exploratory and confirmatory factor analysis. In L. G. Grimm & P. R. Yarnold (Eds.), *Reading and understanding multivariate statistics* (pp. 99-136). Washington, DC: American Psychological Association.

Bubner, R. (1981). *Modern German philosophy* (E. Matthews, Trans.). Cambridge, UK: Cambridge University Press.

Buss, D. M. (2004). *Evolutionary psychology: The new science of the mind* (2nd ed.). Boston, MA: Pearson Education.

Carey, S. (1985). *Conceptual change in childhood.* Cambridge, MA: MIT Press.

Carey, S. (2000). Science education as conceptual change. *Journal of Applied Developmental Psychology, 21,* 13-19.

Carey, S. (2009). *The origin of concepts.* Oxford, UK: Oxford University Press.

Carey, S., & Gelman, R. (Eds.) (1991). *The epigenesis of mind: Essays on biology & cognition.* Hillsdale, NJ: Lawrence Erlbaum Associates.

Carey, S., & Spelke, E. (1994). Domain-specific knowledge and conceptual change. In L. A. Hirschfield & S. A. Gelman (Eds.), *Mapping the mind: Domain specificity in cognition and culture* (pp. 169-200). Cambridge, UK: Cambridge University Press.

Carnegie Corporation of New York and Institute for Advanced Study, Commission on Mathematics and Science Education. (2009). *The opportunity equation: Transforming mathematics and science education for citizenship and the global economy.* Retrieved from http://opportunityequation.org/report

Chi, M. T. H. (1992). Conceptual change within and across ontological categories: Examples from learning and discovery in science. In R. N. Giere (Ed.), *Minnesota studies in the philosophy of science (Vol. XV), Cognitive models of science* (pp. 129-186). Minneapolis, MN: University of Minnesota Press.

Chi, M. T. H. (2008). Three types of conceptual change: Belief revision, mental model transformation, and categorical shift. In S. Vosniadou (Ed.), *International* *handbook of research on conceptual change (pp. 61-82).* New York, NY: Routledge.

Chinn, C. A. (1998). A critique of social constructivist explanations of knowledge change. In B. Guzzetti & C. Hynd (Eds.), *Perspectives on conceptual change: Multiple ways to understand knowing and learning in a complex world* (pp. 77-115). Mahwah, NJ: Lawrence Erlbaum.

Chinn, C. A., & Brewer, W. F. (1993). The role of anomalous data in knowledge acquisition: A theoretical framework and implications for science instruction. *Review of Educational Research, 63,* 1–49.

Chinn, C. A., & Buckland, L. A. (2012). Model-based instruction: Fostering change in evolutionary conceptions and in epistemic practices. In K. S. Rosengren, S. K. Brem, E. M. Evans, G. M. Sinatra (Eds.), *Evolution challenges: Integrating research and practice in teaching and learning about evolution.* Oxford, UK: Oxford University Press.

Clement, J. (2008). The role of explanatory models in teaching for conceptual change. In S. Vosniadou (Ed.), *International handbook of research on conceptual change* (pp. 417-452). New York, NY: Routledge.

Clune, W. H. (1993). The best path to systemic educational policy: Standard/centralized or differentiated/decentralized? *Educational Evaluation and Policy Analysis, 15,* 233-254.

Cosmides, L., & Tooby, J. (1987). From evolution to behavior: Evolutionary psychology as the missing link. In J. Dupré (Ed.), *The latest on the best essays on evolution and optimality* (pp. 277-306). Cambridge, MA: MIT Press.

Cosmides, L., & Tooby, J. (1992). Cognitive adaptations for social exchange. In J. H. Barkow, L. Cosmides, & J. Tooby (Eds.), *The adapted mind: Evolutionary psychology and the generation of culture* (pp. 163-228). New York, NY: Oxford University Press.

Cosmides, L., & Tooby, J. (1994). The origins of domain specificity: The evolution of functional organization. In L. A. Hirschfield & S. A. Gelman (Eds.), *Mapping the mind: Domain specificity in cognition and culture* (pp. 85-116). Cambridge, UK: Cambridge University Press.

Cosmides, L., & Tooby, J. (1997). Evolutionary psychology: A primer. Retrieved from http://www.psych.ucsb.edu/research/cep/primer.html

Cronin, H. (1991). *The ant and the peacock: Altruism and sexual selection from Darwin to today*. Cambridge, UK: Cambridge University Press.

Crotty, M. (1998). *The foundations of social research: Meaning and perspective in the research process.* London, UK: SAGE Publications.

Darwin, C. (1859). *The origin of species.* New York, NY: Random House.

Dawkins, R. (1976/1989). *The selfish gene* (new ed.). Oxford, UK: Oxford University Press.

Dawkins, R. (1982). *The extended phenotype: The long reach of the gene.* Oxford, UK: Oxford University Press.

Dawkins, R. (1986). *The blind watchmaker.* New York, NY: Norton.

Dawkins, R. (1996). *Climbing mount improbable*. New York, NY: Norton.

de Waal, F. (1996). *Good natured: The origins of right and wrong in humans and other animals.* Cambridge, MA: Harvard University Press.

de Waal, F. (2006). Morally evolved: Primate social instincts, human morality, and the rise and fall of "veneer theory." In J. Ober & S. Macedo (Eds.), *Primates and philosophers: How morality evolved* (pp. 1-81). Princeton, NJ: Princeton University Press.

Dennett, D. C. (1989). *Consciousness explained.* New York, NY: Back Bay Books.

Dennett, D. C. (1995). *Darwin's dangerous idea: Evolution and the meanings of life.* New York, NY: Simon & Schuster.

Dennett, D. C. (2002). The new replicators. In M. Pagel (Ed.), *Encyclopedia of evolution* (pp. E83-E92). Oxford, UK: Oxford University Press.

DeScioli, P., & Kurzban, R. (2009). Mysteries of morality. *Cognition, 112*, 281-299.

Descombes, V. (1980). *Modern French philosophy* (L. Scott-Fox & J. M. Harding, Trans.).Cambridge, UK: Cambridge University Press.

Desmond, A., & Moore, J. (1994). *Darwin: The life of a tormented evolutionist*. New York, NY: W. W. Norton.

diSessa, A. A. (2008). 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* (pp. 35-60). New York, NY: Routledge.

Dobzhansky, T. (1973). Nothing in biology makes sense except in the light of evolution. *American Biology Teacher 35*, 125- 129.

Dole, J. A., & Sinatra, G. M. (1998). Reconceptualizing change in the cognitive construction of knowledge. *Educational Psychologist, 33,* 109–128.

Driver, R., Newton, P., & Osborne, J. (2000). Establishing the norms of scientific argumentation in classrooms. *Science Education 84*, 287-312.

Driver, R., Squires, A., Rushworth, P., & Wood-Robinson, V. (1994). *Making sense of secondary science: Research into children's ideas.* New York, NY: RoutledgeFalmer.

Duit, R. (2009). *Bibliography: Students' and teachers' conceptions and science education*. Retrieved from http://www.ipn.uni-kiel.de/aktuell/stcse/stcse.html.

Duschl, R. (2008). Science education in three-part harmony: Balancing conceptual, epistemic, and social learning goals. *Review of Research in Education, 32*, 268-291.

Emlen, S. (1975). The stellar orientation system of a migratory bird. *Scientific American, 233,* 102-111.

Evans, E. M. (2008). Conceptual change and evolutionary biology. In S. Vosniadou (Ed.), *International handbook of research on conceptual change* (pp. 263-294). New York, NY: Routledge.

Evans, P. L., & Anderson, D. L. (2013, April). *The Conceptual Inventory of Natural Selection a decade later: Development and pilot testing of a middle school version leads to a revised college/high school version.* Paper presented at the Annual Meeting of the National Association for Research in Science Teaching, Rio Grande, Puerto Rico.

Every Student Succeeds Act. (2015). Pub. L. No. 114-95.

Ferris, T. (1988/2003). *Coming of age in the Milky Way*. New York, NY: Perennial.

Ferris, T. (1997) *The whole shebang: A state-of-the-universe(s) report.* New York, NY: Simon & Schuster.

Field, A. (2009). *Discovering statistics using SPSS* (3rd ed.). Thousand Oaks, CA: Sage.

Fischer, K. W., & Bidell, T. (1991). Constraining nativist inferences about cognitive capacities. In S. Carey & R. Gelman (Eds.), *The epigenesis of mind: Essays on biology and cognition* (pp. 199-235). Hillsdale, NJ: Lawrence Erlbaum Associates.

Fodor, J. (1983). *Modularity of mind.* Cambridge, MA: MIT Press.

Fowler, F. C. (2009). *Policy studies for educational leaders: An introduction* (3rd ed.). Boston, MA: Pearson.

Frank, R. H. (1988). *Passions within reason: The strategic role of the emotions.* New York, NY: W. W. Norton & Company.

Freeman, S., & Herron, J. C. (2001). *Evolutionary analysis* (2nd ed.). Upper Saddle River, NJ: Prentice Hall.

Futuyma, D. J. (1998). *Evolutionary biology* (3rd ed.). Sunderland, MA: Sinauer Associates.

Gall, M. D., Gall, J. P., Borg, W. R. (2007). *Educational research: An introduction* (8th ed.). Boston, MA: Pearson.

Gallistel, C. R., Brown, A. L., Carey, S., Gelman, R., & Keil, F. C. (1991). Lessons from animal learning for the study of cognitive development. In S. Carey & R. Gelman (Eds.), *The epigenesis of mind: Essays on biology and cognition* (pp. 3-36). Hillsdale, NJ: Lawrence Erlbaum Associates.

Gamble, C., Gowlett, J., & Dunbar, R. (2014). *Thinking big: How the evolution of social life shaped the human mind*. New York, NY: Thames & Hudson.

Garcia, J., & Koelling, R. A. (1966). The relation of cue to consequence in avoidance learning. *Psychonomic Science, 4,* 123-124.

Gazzaniga. M. S. (2011). *Who's in charge?: Free will and the science of the brain*. New York, NY: Ecco.

Geary, D. C. (2002). Principles of evolutionary educational psychology. *Learning and Individual Differences, 12,* 317-345.

Geary, D. C. (2007). Educating the evolved mind: Conceptual foundations for an evolutionary educational psychology. In J. S. Carlson & J. R. Levin (Eds.), *Educating the evolved mind: Conceptual foundations for an evolutionary educational psychology* (pp. 1-99). Charlotte, NC: Information-Age Publishing.

Geary, D. C., & Bjorklund, D. F. (2000). Evolutionary developmental psychology. *Child Development, 71,* 57-65.

Gelman, R. (1990). First principles organize attention to learning about relevant data: Number and the animate-inanimate distinction as examples. *Cognitive Science, 14,* 79-106.

Gelman, R. (1991). Epigenetic foundations of knowledge structures: Initial and transcendent constructions. In S. Carey & R. Gelman (Eds.), *The epigenesis of mind: Essays on biology and cognition* (pp. 293-322). Hillsdale, NJ: Lawrence Erlbaum Associates.

Gelman, R., Durgin, F., & Kaufman, L. (1995). Distinguishing between animates and inanimates: Not by motion alone. In D. Sperber, D. Premack, & A. J. Premack

(Eds.), *Causal cognition: A multidisciplinary debate* (pp. 150-184). Oxford, UK: Clarendon Press.

Gelman, S. A. (2003). *The essential child: Origins of essentialism in everyday thought.* Oxford, UK: Oxford University Press.

Gelman, S. A., & Kremer, K. E. (1991). Understanding natural causes: Children's explanations of how objects and their properties originate. *Child Development, 62,* 396-414.

Gelman, S. A., & Wellman, H. M. (1991). Insides and essences: Early understandings of the non-obvious. *Cognition, 38*, 213-244.

Gelman, S. A., Coley, J. D., & Gottfried, G. M. (1994). Essentialist beliefs in children: The acquisition of concepts and theories. In L. A. Hirschfield & S. A. Gelman (Eds.), *Mapping the mind: Domain specificity in cognition and culture* (pp. 341-365). Cambridge, UK: Cambridge University Press.

Gopnik, A., & Meltzoff, A. N. (1997). *Words, thoughts, and theories.* Cambridge, MA: MIT Press.

Gopnik, A., & Wellman, H. M. (1994). The theory theory. In L. A. Hirschfield & S. A. Gelman (Eds.), *Mapping the mind: Domain specificity in cognition and culture* (pp. 257-293). Cambridge, UK: Cambridge University Press.

Haidt, J. (2007). The new synthesis in moral psychology. *Science, 316,* 998-1002.

Haidt, J. (2012). *The righteous mind: Why good people are divided by politics and religion.* New York, NY: Pantheon.

Hair, J. F., Jr.,, & Black, W. C. (2000). Cluster analysis. In L. G. Grimm & P. R. Yarnold (Eds.), *Reading and understanding more multivariate statistics* (pp. 147-206). Washington, DC: American Psychological Association.

Hatano, G., & Inagaki, K. (1994). Young children's naïve theory of biology. *Cognition, 50,* 171-188.

Hatano, G., & Inagaki, K. (2003). When is conceptual change intended? A cognitive-sociocultural view. In G.M. Sinatra & P. R. Pintrich (Eds.), *Intentional conceptual change* (pp. 407-427). Mahwah, NJ: Lawrence Erlbaum Associates.

Hickling, A. K., & Gelman, S. A. (1995). How does your garden grow? Early conceptualization of seeds and their place in the plant growth cycle. *Child Development, 66,* 856-876.

Hirschfield, L. A., & Gelman, S. A. (Eds.) (1994a). *Mapping the mind: Domain specificity in cognition and culture.* Cambridge, UK: Cambridge University Press.

Hirschfield, L. A., & Gelman, S. A. (1994b). Toward a topography of the mind: An introduction to domain specificity. In L. A. Hirschfield & S. A. Gelman (Eds.), *Mapping the mind: Domain specificity in cognition and culture* (pp. 3-35). Cambridge, UK: Cambridge University Press.

Hofstadter, D., & Sander, E. (2013). *Surfaces and essences: Analogy as the fuel and fire of thinking.* New York, NY: Basic Books.

Humphrey, N. (1976). The social function of intellect. In P. P. G. Bateson & R. A. Hinde (Eds.), *Growing points in ethology* (pp. 303-317). Cambridge, UK: Cambridge University Press.

Humphrey, N. (1983). *Consciousness regained: Chapters in the development of mind.* Oxford, UK: Oxford University Press.)

Humphrey, N. (1986). *The inner eye*. Oxford, UK: Oxford University Press.

Hynd, C. (1998). Conceptual change in a high school physics class. In B. Guzzetti & C. Hynd (Eds.), *Perspectives on conceptual change: Multiple ways to understand knowing and learning in a complex world* (pp. 27-38). Mahwah, NJ: Lawrence Erlbaum.

Hynd, C. (2003). Conceptual change in response to persuasive messages. In G.M. Sinatra & P. R. Pintrich (Eds.) *Intentional conceptual change* (pp. 291-315). Mahwah, NJ: Lawrence Erlbaum Associates.

Inagaki, K., & Hatano, G. (1993). Young children's understanding of the mind-body problem. *Child Development, 64,* 1534-1549.

Inagaki, K., & Hatano, G. (1996). Young children's recognition of commonalities between animals and plants. *Child Development, 67,* 2823-2840.

Inagaki, K., & Hatano, G. (2002). *Young children's naïve thinking about the biological world*. New York, NY: Psychology Press.

Inagaki, K., & Hatano, G. (2008). Conceptual change in naïve biology. In S. Vosniadou (Ed.), *International handbook of research on conceptual change* (pp. 240-262). New York, NY: Routledge.

Jensen, M. S., & Finley, F. N. (1995). Teaching evolution using historical arguments in a conceptual change strategy. *Science Education, 79,* 147-166.

Jensen, M. S., & Finley, F. N. (1996). Changes in students' understanding of evolution resulting from different curricular and instructional strategies. *Journal of Research in Science Teaching, 33,* 879-900.

Joyce, B., Weil, M., & Calhoun, E. (2009). *Models of teaching* (8th ed.). Boston, MA: Pearson Education.

Kahneman, D. (2011). *Thinking, fast and slow.* New York, NY: Farrar, Strauss, and Giroux.

Kaplan, S. (1992). Environmental preference in a knowledge-seeking, knowledge-using organism. In J. H. Barkow, L. Cosmides, & J. Tooby (Eds.), *The adapted mind: Evolutionary psychology and the generation of culture* (pp. 581-598). New York, NY: Oxford University Press.

Karmiloff-Smith, A. (1991). Beyond modularity: Innate constraints and developmental change. In S. Carey & R. Gelman (Eds.), *The epigenesis of mind: Essays on biology and cognition* (pp. 171-197). Hillsdale, NJ: Lawrence Erlbaum Associates.

Karmiloff-Smith, A. (1992). *Beyond modularity: A developmental perspective on cognitive science.* Cambridge, MA: MIT Press.

Keil, F. C. (1989). *Concepts, kinds, and cognitive development*. Cambridge, MA: MIT Press.

Keil, F. C. (1991). The emergence of theoretical beliefs as constraints on concepts. In S. Carey & R. Gelman (Eds.), *The epigenesis of mind: Essays on biology and cognition* (pp. 237-256). Hillsdale, NJ: Lawrence Erlbaum Associates.

Keil, F. C. (1994). The birth and nurturance of concepts by domains: The origins of concepts of living things. In L. A. Hirschfield & S. A. Gelman (Eds.), *Mapping the mind: Domain specificity in cognition and culture* (pp. 234-254). Cambridge, UK: Cambridge University Press.

Keil, F. C. (1995). The growth of causal understandings of natural kinds. In D. Sperber, D. Premack, & A. J. Premack (Eds.), *Causal cognition: A multidisciplinary debate* (pp. 234-267). Oxford, UK: Clarendon Press.

Keil, F., & Newman, G. E. (2008). Two tales of conceptual change: What changes and what remains the same. In S. Vosniadou (Ed.), *International handbook of research on conceptual change* (pp. 83-101). New York, NY: Routledge.

Kelly, G. J., & Green, J. (1998). The social nature of knowing: Toward a sociocultural perspective on conceptual change and knowledge construction. In B. Guzzetti & C. Hynd (Eds.), *Perspectives on conceptual change: Multiple ways to understand knowing and learning in a complex world* (pp. 145-181). Mahwah, NJ: Lawrence Erlbaum.

Kirkpatrick, L. A. (2009). From genes to memes: Psychology at the nexus. In M. Schaller, A. Norenzayan, S. J. Heine, T. Yamagishi, & T. Kameda (Eds.), *Evolution, Culture, and the Human Mind* (pp. 71-79). New York, NY: Psychology Press.

Klem, L. (2000). Structural equation modeling. In L. G. Grimm & P. R. Yarnold (Eds.), *Reading and understanding more multivariate statistics* (pp. 227-259). Washington, DC: American Psychological Association.

Krauss, L. M. (2012). *A universe from nothing: Why there is something rather than nothing.* New York, NY: Simon & Schuster.

Krebs, J. R., & Davies, N. B. (1993). *An introduction to behavioural ecology* (3rd ed.). Malden, MA: Blackwell.

Kuhn, D. (1992). Thinking as argument. *Harvard Educational Review, 62,* 155-178.

Kuhn, D. (1993). Science as argument: Implications for teaching and learning scientific thinking. *Science Education, 77,* 319-337.

Kuhn, D. (2010). Teaching and learning science as argument. *Science Education, 94,* 810-824.

Kuhn, T. (1962/1970). *The structure of scientific revolutions*. Chicago, IL: University of Chicago Press.

Kurzban, R. (2010). *Why everyone (else) is a hypocrite: Evolution and the modular mind*. Princeton, NJ: Princeton University Press.

Kurzban, R., & Aktipis, C. A. (2007). Modularity and the social mind: Are psychologists too self-ish? *Personality and Social Psychology Review, 11,* 131-149.

Leach, J. T., & Scott, P. H. (2008). Teaching for conceptual understanding: An approach drawing on individual and sociocultural perspectives. In S. Vosniadou (Ed.), *International handbook of research on conceptual change* (pp. 647-675). New York, NY: Routledge.

Leslie, A. M. (1994). ToMM, ToBy, and Agency: Core architecture and domain specificity. In L. A. Hirschfield & S. A. Gelman (Eds.), *Mapping the mind: Domain specificity in cognition and culture* (pp. 119-148). Cambridge, UK: Cambridge University Press.

Leslie, A. M. (1995). A theory of agency. In D. Sperber, D. Premack, & A. J. Premack (Eds.), *Causal cognition: A multidisciplinary debate* (pp. 121-141). Oxford, UK: Clarendon Press.

Lively, C. M. (1986). Predator-induced shell dimorphism in the acorn barnacle *Chthamalus anispoma*. *Evolution, 40,* 232-242.

Lombardi, D., Sinatra, G. M., & Nussbaum, E. M. (2013). Plausibility reappraisals and shifts in middle school students' climate change conceptions. *Learning and Instruction, 27,* 50-62.

Marler, P. (1991). The instinct to learn, by Peter Marler. In S. Carey & R. Gelman (Eds.), *The epigenesis of mind: Essays on biology and cognition* (pp. 37-66). Hillsdale, NJ: Lawrence Erlbaum Associates.

Massey, C. M., & Gelman, R. (1988). Preschooler's ability to decide whether a photographed unfamiliar object can move itself. *Developmental Psychology, 24,* 307-317.

Maynard Smith, J., & Szathmáry, E. (1995). *The major transitions in evolution.* New York, NY: Oxford University Press.

Mayr, E. (1991). *One long argument: Charles Darwin and the genesis of modern evolutionary thought.* Cambridge, MA: Harvard University Press.

Mithen, S. (1996). *The prehistory of the mind: A search for the origins of art, religion and science.* London, UK: Thames & Hudson Ltd.

Miyake, N. (2008). Conceptual change through collaboration. In S. Vosniadou (Ed.), *International handbook of research on conceptual change* (pp. 453-478). New York, NY: Routledge.

Mortimer, E. F., & Scott, P. H. (2003). *Meaning making in secondary science classrooms*. Maidenhead, UK: Open University Press.

National Academy of Sciences, Working Group on Teaching Evolution. (1998). *Teaching about evolution and the nature of science.* Retrieved from the National Academies Press website: http://www.nap.edu/catalog.php?record_id=5787

National Association of Biology Teachers. (2008). *NABT position statement on teaching evolution.* Retrieved from http://www.nabt.org/websites/institution/index.php?p=92

National Research Council, National Committee on Science Education Standards and Assessment. (1998). *National science education standards.* Retrieved from the National Academies Press website: http://www.nap.edu/catalog.php?record_id=4962

National Science Teachers Association. (2003). *NSTA position statement: The teaching of evolution.* Retrieved from http://www.nsta.org/pdfs/positionstatement_evolution.pdf

Nehm, R. H., Beggrow, E. P., Opfer, J. E., and Ha, M. (2012). Reasoning about natural selection: Diagnosing contextual competency using the ACORNS instrument. *The American Biology Teacher*, *74,* 92-98.

Nehm, R. H., & Reilly, L. (2007). Biology majors' knowledge and misconceptions of natural selection. *Bioscience, 57*, 263-272.

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*, 1131-1160.

Nersessian, N. J. (2008). Mental modeling in conceptual change. In S. Vosniadou (Ed.), *International handbook of research on conceptual change* (pp. 391-416). New York, NY: Routledge.

No Child Left Behind Act of 2001, 20 U.S.C. § 6319 (2002).

Nussbaum, E. M. (2008). Collaborative discourse, argumentation, and learning: Preface and literature review. *Contemporary Educational Psychology, 33,* 345-359.

Orians, G. H., & Heerwagen, J. H. (1992). Evolved responses to landscapes. In J. H. Barkow, L. Cosmides, & J. Tooby (Eds.), *The adapted mind: Evolutionary psychology and the generation of culture* (pp. 555-579). New York, NY: Oxford University Press.

Pinker, S. (1994). *The language instinct: How the mind creates language.* New York, NY: HarperCollins.

Pinker, S. (1997). *How the mind works.* New York, NY: Norton.

Pinker, S. (2002). *The blank slate: The modern denial of human nature.* New York, NY: Penguin Group.

Pinker, S. (2007). *The stuff of thought: Language as a window into human nature.* New York, NY: Penguin.

Pintrich, P. R., Marx, R. W., & Boyle, R. A. (1993). Beyond cold conceptual change: The role of motivational beliefs and classroom contextual factors in the process of conceptual change. *Review of Educational Research, 63,* 167–199.

Plato. (trans. 1974). *The republic* (Rev. ed.) (D. Lee, Trans.). Middlesex, UK: Penguin Books.

Poling, D. A., & Evans, E. M. (2002). Why do birds of a feather flock together? Developmental change in the use of multiple explanations: Intention, teleology, and essentialism. *British Journal of Developmental Psychology, 20,* 89-112.

Poling, D. A., & Evans, E. M. (2004). Religious belief, scientific expertise, and folk ecology. *Journal of Cognition and Culture, 4,* 485-524.

Posner, G. J., Strike, K. A., Hewson, P. W., & Gertzog, W. A. (1982). Accommodation of a scientific conception: Toward a theory of conceptual change. *Science Education, 66,* 211-227.

Premack, D. (1990). The infant's theory of self-propelled objects. *Cognition 36,* 1-13.

Premack, D., & Premack, A. J. (1994). Moral belief: Form versus content. In L. A. Hirschfield & S. A. Gelman (Eds.), *Mapping the mind: Domain specificity in cognition and culture* (pp. 149-168). Cambridge, UK: Cambridge University Press.

Premack, D., & Premack, A. J. (1995). Intention as psychological cause. In D. Sperber, D. Premack, & A. J. Premack (Eds.), *Causal cognition: A multidisciplinary debate* (pp. 185-199). Oxford, UK: Clarendon Press.

Randall, J. H., Jr. (1926/1976). *The making of the modern mind* (50th anniversary ed.). New York, NY: Columbia University Press.

Richerson, P. J., & Boyd, R. (2005). *Not by genes alone: How culture transformed human evolution.* Chicago, IL: The University of Chicago Press.

Ridley, M. (2004). *Evolution* (3rd ed.). Oxford, UK: Blackwell.

Ridley, M. (1993). *The red queen: Sex and the evolution of human nature.* New York, NY: Harper Perennial.

Ridley, M. (2003). *The agile gene: How nature turns on nurture.* New York, NY: Perennial.

Rosengren, K. S., Gelman, S. A., Kalish, C. W., & McCormick, M. (1991). As time goes by: Children's early understanding of growth in animals. *Child Development, 62,* 1302-1320.

Samarapungavan, A., & Wiers, R. W. (1997). Children's thoughts on the origin of species: A study of explanatory coherence. *Cognitive Science, 21,* 147-177.

Settlage, Jr.. J. (1994). Conceptions of natural selection: A snapshot of the sense-making process. *Journal of Research in Science Teaching, 31,* 449-457.

Shtulman, A. (2006). Qualitative differences between naïve and scientific theories of evolution. *Cognitive Psychology, 52,* 170-194.

Shtulman, A., & Calabi, P. (2008). Learning, understanding, and acceptance: The case of evolution. *Proceedings of the 30th Annual Conference of the Cognitive Science Society,* 235-240.

Shtulman, A., & Calabi, P. (2012). Cognitive constraints on the understanding and acceptance of evolution. In K. S. Rosengren, S. Brem, E. M. Evans, & G. Sinatra (Eds.), *Evolution Challenges: Integrating Research and Practice in Teaching and Learning about Evolution* (pp. 47-65). Cambridge, UK: Oxford University Press.

Shtulman, A., & Schulz, L. (2008). The relation between essentialist beliefs and evolutionary reasoning. *Cognitive Science*, *32*, 1049-1062.

Shtulman, A., & Valcarcel, J. (2012). Scientific knowledge suppresses but does not supplant earlier intuitions. *Cognition, 124,* 209-215.

Sinatra, G. M., & Dole, J. A., Case studies in conceptual change: A social psychological perspective. In B. Guzzetti & C. Hynd (Eds.), *Perspectives on conceptual change: Multiple ways to understand knowing and learning in a complex world* (pp. 39-53). Mahwah, NJ: Lawrence Erlbaum.

Sinatra, G. M., & Pintrich, P. R. (2003). The role of intentions in conceptual change learning. In G.M. Sinatra & P. R. Pintrich (Eds.), *Intentional conceptual change* (pp. 1-18). Mahwah, NJ: Lawrence Erlbaum Associates.

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,* 189-195.

Smith, J. P., diSessa, A. A., & Roschelle, J. (1993). Misconceptions reconceived: A constructivist analysis of knowledge in transition. *The Journal of the Learning Sciences, 3,* 115–163.

Smith. M. U. (2009). Current status of research in teaching and learning of evolution: I. Philosophical/epistemological issues. *Science & Education, 19,* 523-538.

Smith. M. U. (2010). Current status of research in teaching and learning of evolution: II. Pedagogical issues. *Science & Education, 19,* 539-571.

Smith, N. K. (1923/1992). *A commentary to Kant's "critique of pure reason"* (Rev. 2nd ed.). Atlantic Highlands, NJ: Humanities Press International.

Solomon, R. C. (1983). *In the spirit of Hegel.* Oxford, UK: Oxford University Press.

Southerland, S. A., Smith, L. K., Sowell, S. P., & Kittleson, J. M. (2007). Resisting unlearning: Understanding science education's response to the United States' national accountability movement. *Review of Research in Education, 31,* 45-77.

Spelke, E. S. (1991). Physical knowledge in infancy: Reflections on Piaget's theory. In S. Carey & R. Gelman (Eds.), *The epigenesis of mind: Essays on biology and cognition* (pp. 133-169). Hillsdale, NJ: Lawrence Erlbaum Associates.

Spelke, E. (1994). Initial knowledge: Six suggestions. *Cognition, 50*, 431-445.

Spelke, E. S., Phillips, A., & Woodward, A. L. (1995). Infants' knowledge of object motion and human action. In D. Sperber, D. Premack, & A. J. Premack (Eds.), *Causal cognition: A multidisciplinary debate* (pp. 79-116). Oxford, UK: Clarendon Press.

Sperber, D. (1994). The modularity of thought and the epidemiology of representations. In L. A. Hirschfield & S. A. Gelman (Eds.), *Mapping the mind: Domain specificity in cognition and culture* (pp. 39-67). Cambridge, UK: Cambridge University Press.

Sperber, D. (1996). *Explaining culture: A naturalistic approach.* Oxford, UK: Blackwell.

Sperber, D., Premack, D., & Premack, A. J. (Eds.) (1995). *Causal cognition: A multidisciplinary debate.* Oxford, UK: Clarendon Press.

Sterelny, K. (2012). *The evolved apprentice: How evolution made humans unique.* Cambridge, MA: MIT Press.

Strike, K. A., & Posner, G. J. (1992). A revisionist theory of conceptual change. In R. A. Duschl & R. J. Hamilton (Eds.), *Philosophy of science, cognitive psychology, and educational theory and practice* (pp. 147-174). Albany, NY: State University of New York Press.

Tavris, C., & Aronson, E. (2007). *Mistakes were made (but not by me): Why we justify foolish beliefs, bad decisions, and hurtful acts*. Orlando, FL: Harcourt.

Thagard, P. (1992). *Conceptual revolutions.* Princeton, NJ: Princeton University Press.

Thompson, B. (2000). Ten commandments of structural equation modeling. In L. G. Grimm & P. R. Yarnold (Eds.), *Reading and understanding more multivariate statistics* (pp. 261-283). Washington, DC: American Psychological Association.

Tomasello, M. (1999). *The cultural origins of human cognition.* Cambridge, MA: Harvard University Press.

Tomasello, M. (2014). *A natural history of human thinking.* Cambridge, MA: Harvard University Press.

Tomasello, M., Kruger, A. C., & Ratner, H. H. (1993). Cultural learning. *Behavioral and Brain Sciences, 16,* 495-552.

Tooby, J., & Cosmides, L. (1992). The psychological foundations of culture. In J. H. Barkow, L. Cosmides, & J. Tooby (Eds.), *The adapted mind: Evolutionary psychology and the generation of culture* (pp. 19-136). New York, NY: Oxford University Press.

Toulmin, S. (1972). *Human understanding: The collective use and evolution of concepts.* Princeton, NJ: Princeton University Press.

Trivers, R. (1971). The evolution of reciprocal altruism. *The Quarterly Review of Biology, 46,* 35-37.

Trivers, R. (1972). Parental investment and sexual selection. In B. Campbell (Ed.), *Sexual selection and the descent of man 1871-1971* (pp. 136-179). Chicago, IL: Aldine.

Trivers, R. (1974). Parent-offspring conflict. *American Zoologist, 14,* 249-264.

Trivers, R. (1976). Foreword. In R. Dawkins, *The selfish gene*. Oxford, UK: Oxford University Press.

Trivers, R. (1985). *Social evolution.* Menlo Park, CA: Benjamin/Cummings.

Trivers, R. (2000). The elements of a scientific theory of self-deception. *Annals of the New York Academy of Sciences, 907,* 114-131.

Trivers, R. (2002). *Natural selection and social theory: Selected papers of Robert Trivers*. New York, NY: Oxford University Press.

Vosniadou, S. (1994). Universal and culture-specific properties of children's mental models of the earth. In L. A. Hirschfield & S. A. Gelman (Eds.), *Mapping the mind: Domain specificity in cognition and culture* (pp. 412-430). Cambridge, UK: Cambridge University Press.

Vosniadou, S. (Ed.) (2008a). *International handbook of research on conceptual change*. New York, NY: Routledge.

Vosniadou, S. (2008b). Conceptual change research: An introduction. In S. Vosniadou (Ed.), *International handbook of research on conceptual change* (pp. xiii-xxviii). New York, NY: Routledge.

Vosniadou, S., & Brewer, W. F. (1987). Theories of knowledge restructuring in development. *Review of Educational Research, 57,* 51–67.

Vosniadou, S., Vamvakoussi, X., & Skopetliti, I. (2008). A framework theory approach to the problem of conceptual change research. In S. Vosniadou (Ed.), *International handbook of research on conceptual change* (pp. 3-34). New York, NY: Routledge.

Vygotsky, L.S. (1978). *Mind in society: The development of higher psychological processes.* (M. Cole, V. John-Steiner, S. Scribner, & E. Souberman, Eds.). Cambridge, MA: Harvard University Press.

Weick, K. E. (1976). Educational organizations as loosely coupled systems. *Administrative Science Quarterly, 21*, 1-19.

Wellman, H. M. (1990). *The child's theory of mind.* Cambridge, MA: MIT Press.

Wellman, H. M., & Gelman, S. A. (1992). Cognitive development: Foundational theories of core domains. *Annual Review of Psychology, 43,* 337-375.

Whiten, A., & Byrne, R. W. (1988). *Machiavellian intelligence: Social expertise and the evolution of intellect in monkeys, apes and humans*. New York, NY: Oxford University Press.

Williams, G. C. (1966). *Adaptation and natural selection.* Princeton, NJ: Princeton University Press.

Willis, J. W. (2007). *Foundations of qualitative research: Interpretive and critical approaches.* Thousand Oaks, CA: Sage Publications.

Wilson, E. O. (1998). *Consilience: The unity of knowledge.* New York, NY: Random House.

Wilson, T. D. (2002). *Strangers to ourselves: Discovering the adaptive unconscious.* Cambridge, MA: Harvard University Press.

Zar, J. H. (1999.) *Biostatistical analysis* (4th ed.). Upper Saddle River, NJ: Prentice Hall.

Zhou, G. (2010). Conceptual change in science: A process of argumentation. *Eurasia Journal of Mathematics, Science & Technology Education, 6,* 101-110.

Zohar, A, & Nemet, F. (2002). Fostering students' knowledge and argumentation skills through dilemmas in human genetics. *Journal of Research in Science Teaching, 39*, 35-60.

Vita

Kevin David Goff

Date of Birth:	November 12, 1964	
Place of Birth:	Cumberland, Maryland	
Education:	2009-2017	The College of William & Mary, School of Education Doctor of Education Educational Policy, Planning, and Leadership / Curriculum Leadership
	2004-2006	Virginia Commonwealth University Virginia Teaching Endorsement in Gifted Education
	1999-2002	Virginia Institute of Marine Science, College of William & Mary Master of Science Marine Science / Fisheries Biology
	1990-1991	James Madison University Virginia Professional Teaching License Secondary Science / Biology
	1983-1987	The College of William & Mary Bachelor of Arts Biology / Religion
Experience:	2015-2017	Director of River Program St. Margaret's School, Virginia
	2013-2015	Marine Education Specialist Virginia Institute of Marine Science
	2010-2014	Adjunct Faculty / University Supervisor College of William & Mary, School of Education
	2009-2013	Program Coordinator / Recruitment Coordinator William & Mary Noyce Scholars Program
	2002-2009	Science Instructor / Director of Curriculum Chesapeake Bay Governor's School, Virginia
	1991-2002	Science Teacher / Athletic Director Essex County Public Schools