Measuring Student Understanding of Genetics:
Psychometric, Cognitive, and Demographic Considerations

A Dissertation Presented

by

Robyn Tornabene

to

The Graduate School

in Partial Fulfillment of the

Requirements

for the Degree of

Doctor of Philosophy

in

Science Education

Stony Brook University

August 2018
We, the dissertation committee for the above candidate for the Doctor of Philosophy degree, hereby recommend acceptance of this dissertation.

Ross H. Nehm, Ph.D.
Associate Professor, Department of Ecology and Evolution

Gregory Rushton, Ph.D.
Associate Professor, Department of Chemistry

John True, Ph.D.
Associate Professor, Department of Ecology & Evolution

Stephen Finch, Ph.D.
Professor, Department of Applied Mathematics and Statistics

This dissertation is accepted by the Graduate School

Charles Taber
Dean of the Graduate School
Abstract of the Dissertation

Measuring Student Understanding of Genetics:
Psychometric, Cognitive, and Demographic Considerations

by

Robyn Tornabene

Doctor of Philosophy

in

Science Education

Stony Brook University

2018

Genetics is universally recognized as a core aspect of biological and scientific literacy. Beyond genetics’ own role as a major unifying topic within the biological sciences, understanding genetics is essential for understanding other integral ideas such as evolution and development. Genetics understanding also underlies public decision making about modern advances in health sciences and biotechnology and broader socio-scientific issues. Consequently, educators have attempted to measure student and teacher understanding of this domain. Using Rasch modeling--a superior but underutilized framework for instrument evaluation--this dissertation explored psychometric, cognitive, and demographic aspects of educational measurement in the field of genetics education in order to generate evidence-based examples to illustrate how instruments can be more carefully developed and robustly evaluated.

The first study (Chapter 3) sought to expand the sources of evidence supporting validity and reliability inferences produced by a relatively new concept inventory (the Genetic Drift Inventory [GeDI]) designed for use in diagnosing undergraduate students’ conceptual understanding of genetic drift. Concept Inventories (CIs) are commonly used tools for assessing student understanding of normative (scientific) and non-normative (naïve) ideas, yet the body of empirical evidence supporting the inferences drawn from CI scores is often limited in scope and remains deeply rooted in Classical Test Theory (CTT) despite the availability of more robust Item Response Theory (IRT) and Rasch frameworks. Specifically, this study focused on: (1) GeDI instrument and item properties as revealed by Rasch modeling, (2) item order effects on response patterns, and (3) generalization to a new geographic sample. A sample of 336 advanced undergraduate biology majors completed one of four randomly assigned and equivalent versions of the GeDI that differed in presentation order of the GeDI item suites. Rasch analysis indicated the GeDI was unidimensional, with good fit to the Rasch model. Items had high reliability and were well matched to the ability of the sample. Person reliability was low. Rotating the GeDI’s vignette-based item suites had no significant impact on overall scores, suggesting each vignette functioned independently. Scores from this new sample from the Northeast United States were
comparable to those from other geographic regions and provide evidence in support of score generalizability. Suggestions for improvement include: (1) incorporation of additional items to differentiate high-ability persons and improve person reliability, and (2) re-examination of items with redundant or low difficulty levels. These results expand the range and quality of evidence in support of validity claims and illustrate changes that are likely to improve the quality of the GeDI (and other) evolution education instruments.

The second study (Chapter 4) sought to determine how situational features impact inferences about participants’ understanding of Mendelian genetics. Understanding how situational features of assessment tasks impact reasoning is important for many educational pursuits, notably the selection of curricular examples to illustrate phenomena, the design of formative and summative assessment items, and determination of whether instruction has fostered the development of abstract schemas divorced from particular instances. To test for context effects, an experimental research design was employed to measure differences in item difficulty among items varying only in situational features (e.g., plant, animal, human, fictitious) across five common genetics problem types. A multi-matrix test design was employed, and item packets were randomly distributed to a sample of undergraduate biology majors (n=444). Rasch analyses of participant scores produced good item fit, person reliability, and item reliability. Surprisingly, no significant differences in performance occurred among the animal, plant, and human item contexts, or between the fictitious and “real” item contexts. Also notable, incomplete dominance problems proved to be more difficult than dominant-recessive problems, and problems featuring homozygous parents were more difficult than those featuring heterozygous parents. Tests for differences in performance between genders, among ethnic groups, and by prior biology coursework revealed that none of these factors had a meaningful impact upon performance or context effects. Thus some, but not all, types of genetics problem solving or item formats are impacted by situational features.

Overall, substantial evidence was generated about how current knowledge in the field of genetics education is measured and how measurement in this domain may be improved. The studies included herein exemplify some ways in which new and existing instruments can be examined to amass robust evidence for the quality of inferences generated by an instrument. Only with rigorously evaluated instruments can the educational community be confident that inferences about student learning are accurate and that consequent decisions are evidence-based.
Dedication Page

This work is dedicated to the teachers whose influence ultimately led me to seek this degree: Olga Verzar who, through ballet, taught me the value of humble dedication and hard work; Doris Akins whose extensive knowledge and passion for biology ignited my own fascination; and Dr. Joanne M. Willey whose expertise in scientific research elevated my academic aspirations and who directly encouraged me to pursue a Ph.D.

Through their individual examples of talent, ambition, confidence, and independence, and through their high expectations and belief in my abilities, I leaned that I could set and achieve higher goals than I ever thought possible.
Table of Contents

Abstract of the Dissertation iii
Dedication Page v
Table of Contents vi
List of Tables viii
List of Figures ix
Preface x
Acknowledgments xi

Chapter 1. Literature- Genetics Education 1
  1.1 Introduction 1
  1.2 Challenges in learning genetics and solving genetics problems 1
  1.3 Modelling as a means of improving genetics understanding 7
  1.4 Argumentation in genetics education 8
  1.5 Theoretical and empirical limitations of genetics education research and areas for future research 9
  1.6 Conclusion 13
  1.7 References (Chapter 1) 14

Chapter 2. Psychometric Analysis Using Item Response Theory 26
  2.1 Introduction 26
  2.1 Instrument evaluation using Item Response Theory 26
  2.3 Advantages of Rasch and IRT 27
  2.4 Use of IRT in this project 31
  2.5 References (Chapter 2) 31

Chapter 3. Testing Validity Inferences for Genetic Drift Inventory Scores Using Rasch Modeling and Item Order Analyses 34
  3.1 Introduction 34
  3.2 Instrument evaluation using Item Response Theory 37
  3.3 Additional relevant considerations for instrument evaluation 39
  3.4 Methods 41
<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>3.5 Results</td>
<td>43</td>
</tr>
<tr>
<td>3.6 Discussion</td>
<td>49</td>
</tr>
<tr>
<td>3.7 Limitations and further research</td>
<td>51</td>
</tr>
<tr>
<td>3.8 References (Chapter 3)</td>
<td>52</td>
</tr>
</tbody>
</table>

**Chapter 4. Testing the impact of surface features on students’ Mendelian genetics understanding**

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>4.1 Introduction</td>
<td>57</td>
</tr>
<tr>
<td>4.2 Research question</td>
<td>62</td>
</tr>
<tr>
<td>4.3 Methods</td>
<td>62</td>
</tr>
<tr>
<td>4.4 Results</td>
<td>67</td>
</tr>
<tr>
<td>4.5 Discussion</td>
<td>71</td>
</tr>
<tr>
<td>4.6 Limitations and further research</td>
<td>75</td>
</tr>
<tr>
<td>4.7 References (Chapter 4)</td>
<td>76</td>
</tr>
</tbody>
</table>

**References**

85

**Appendix A** - Overview of items used for study “Testing the impact of surface features on students’ Mendelian genetics understanding” (Chapter 4) 102

**Appendix B**. Example Items Used for Study “Testing the impact of surface features on students’ Mendelian genetics understanding” (Chapter 4) 106
**List of Tables**

*Table 3.1* Instruments measuring knowledge of evolutionary processes: Potential to elicit normative and non-normative ideas about adaptive and non-adaptive evolution 36

*Table 3.2* GeDI vignettes and associated items 37

*Table 3.3* Summary of validity and reliability evidence for the GeDI. 41

*Table 3.4* Design of GeDI forms 42

*Table 3.5* GeDI Rasch fit properties 44

*Table 3.6* Mean GeDI item measures and standard error by vignette rotation position 46

*Table 3.7* Comparison of performance on GeDI by form, course, and region 48

*Table 3.8* GeDI item difficulty rank in initial and present administrations 48

*Table 4.1* Recent assessments of genetic understanding that include the measurement of transmission genetics 60

*Table 4.2* Five types of Mendelain inheritance problems used in the study 64

*Table 4.3* Example of variation in situational features for a Mendelian inheritance problem 65
List of Figures

**Figure 3.1** Wright map derived from Rasch analysis of GeDI responses 45

**Figure 3.2** Position effects on GeDI item difficulty values 47

**Figure 4.2** Boxplots comparing item difficulty by taxon featured in items 69

**Figure 4.3** Boxplots comparing item difficulty by types of Mendelian inheritance problems 70

**Figure 4.4** Boxplots comparing item difficulties by problem types and situational features 71
Preface

Genetics is universally recognized as a core aspect of biological and scientific literacy. Consequently, educators have attempted to measure student and teacher understanding in this domain. This dissertation explores psychometric, cognitive, and demographic aspects of educational measurement in the field of genetics education in order to generate evidence-based examples to illustrate how instruments can be more carefully developed and robustly evaluated.

Chapters 1 and 2 of this dissertation will serve to frame the work, and include, respectively, a literature review on challenges to genetics problem solving and a description of the Rasch methodology that was central to the psychometric evaluation of the instruments in the study. Chapters 3 and 4 present two journal articles that detail investigations into psychometric, cognitive, and demographic considerations for assessment of genetics understanding. Specifically, Chapter 3 contains a study using Rasch methodology for psychometric evaluation and item order analysis to explore cognitive aspects of design for a published genetics and evolution education instrument. Chapter 4 uses a project-specific instrument, for which Rasch-based validity and reliability evidence is provided, to explore possible interactions between gender or ethnicity and measurement, genetics problem type and item difficulty, and the impact of situational features on measures of genetics understanding.

Overall, substantial evidence was generated about how current knowledge in the field of genetics education is measured and how measurement in this domain may be improved. The outcome of this work was a series of recommendations (contained in Chapters 3 and 4) for the design and evaluation of genetics instruments to produce more accurate and useful measures of a universally recognized aspect of biological literacy.
Acknowledgments

There are many people whose support has enabled me to complete my doctoral studies, and to whom I am very grateful.

I have been extremely fortunate to work under the guidance of my advisor Dr. Ross H. Nehm, whose judicious professional vision, excellent organization, and unwavering optimism conveyed a steady sense of purpose that made this experience productive and enjoyable. I thank you for always keeping me motivated and for your myriad efforts that show your true dedication to your students’ success. Thank you to my dissertation committee members, Dr. Gregory Rushton, Dr. John True, and Dr. Stephen Finch, for insightful comments and thought-provoking questions. It has been a wonderful learning experience and a true pleasure working with all of you.

I am very grateful to Dr. Erik Lavington and Dr. Philipp Schmiemann for their important contributions to the work in Chapters 3 and 4, respectively. Thanks also to fellow Rasch enthusiasts Dr. Gena Sbeglia, Dr. Minsu Ha, Donna Shapiro, and Daniela Fielder for many valuable discussions.

I thank the faculty and staff at the Institute for STEM Education for excellent instruction and assistance, especially Dr. Keith Sheppard for guidance throughout the program. The collegiality of the committed, intellectually inquisitive, and expert educators of Cohorts 1-3 kept me motivated throughout this process and provided a continual resource for learning about both the field of science education and my own teaching. Special thanks to Caren Gough and Dr. Linda Padwa for perpetual wise advice and encouragement.

It is not without the combined efforts and patience of dear family and friends that I was able to accomplish this long-time goal. Deepest respect to my parents who instilled a strong sense of academic responsibility and encouraged all of their children in our broad endeavors. The exemplary work ethic and witty encouragement of my brothers helped motivate me to complete my work. I am incredibly thankful to my family and friends who saw that my children were loved and entertained--and to my children for their patience--while I worked. I am also deeply grateful to my husband, who picked up extra responsibility at home and encouraged and supported me.

The National Association of Research in Science Teaching Scholarship for Classroom Teachers and Informal Educators provided financial support for aspects of this dissertation. Drs. Nehm and Schmiemann graciously supported open access publication of the portions of this dissertation mentioned below.


Portions of Chapter 4 have been published in the paper Schmiemann, P., Nehm, R. H., & Tornabene, R. E. (2017). Assessment of genetics understanding: Under what conditions do situational features have an impact on measures? Science & Education, 26(10), 1161-1191., for which I am the corresponding author.

Permission to adapt portions of the publications for this dissertation has been granted by all coauthors, and copyright for both papers is retained by the respective authors. Both publications are distributed open access under the Creative Commons 4.0 License.
Chapter 1. Literature - Genetics Education

1.1 Introduction

Genetics is a major unifying theme across the domain of biology and a central aspect of biology education. Mastery of genetics concepts is essential for understanding other integral themes in biology such as evolution and development. Knowledge of genetics also underlies public decision making about broader health, technological, and socio-scientific issues such as reproductive technology, cloning, stem cells, genetic testing, and genetic modification of organisms. Despite the core position of genetics within the domain of biology, its complex nature causes many students to struggle with many aspects of genetics learning. Given the importance of genetics in both the scientific and social arenas, it is necessary for the education community to develop an understanding of how students learn and apply genetics knowledge and how to best design assessments, organize curricula, and tailor instruction to facilitate development of knowledge in genetics and a genetically literate society.

Stewart, Cartier, & Passmore (2005) developed a three-part model of genetics knowledge comprising genetic (e.g., classical, Mendelian, or transmission genetics), meiotic (e.g., processes relating to the production of gametes), and molecular (e.g., gene expression) understanding. Stewart (2005) asserts that genetics literacy requires understanding within each of the three model areas as well as understanding of the interactions among areas. This chapter will describe the better-understood barriers students encounter while learning genetics, with a focus on transmission genetics. Theoretical and empirical limitations of genetics problem solving studies and further areas for research will also be considered.

1.2 Challenges in learning genetics and solving genetics problems

A substantial literature has identified and attempted to explain the widespread struggles faced by secondary and undergraduate students attempting to master genetics concepts (e.g., Bahar, Johnstone, & Hansell 1999; Hackling & Treagust 1984; Lewis, Leach, & Wood-Robinson 2000; Pearson & Hughes 1988; Stewart 1982). Evidence has been drawn from diverse sources including student and teacher interviews, textbooks and curricula, student discourse, instructor discourse and practices, and mainly, studies of genetics problem solving. Several, often interrelated, reasons have emerged to explain why conceptual expertise in genetics is so difficult to attain.

1.2.1. Linguistic challenges

Success in genetics--especially problem solving--is largely dependent on mastering abundant and difficult domain-specific terminology. Despite this centrality, genetics terminology has been shown to be confusing, inconsistent, and misleading, as used within the scientific community (Beurton, Falk, & Rheinberger 2000; Gericke & Hagberg 2007; Morange 2001; Portin 2002; Portin & Wilkins 2017; Smith & Adkinson 2010; Venville & Donovan 2005), in textbooks and textbook-style problems (Dos Santos, Joaquim, & El-Hani, 2012; Gericke et al. 2014; Evans 1976; Knippels, Waarlo, & Boersma 2005; Pearson & Hughes 1988), and in teacher discourse during instruction (Thorne, Gericke & Hagberg 2013).

The concept of the gene, as used in the scientific community, has evolved as knowledge of gene structure, function, regulation, and interaction has expanded, and, the meaning of the
term “gene” is often context-dependent. Several historic models of the gene (e.g., Mendelian, classical, classical-biochemical, neoclassical, modern) have been clearly delineated in the literature (e.g., Beurton, Falk, & Rheinberger 2000; Gericke & Hagberg 2007; Morange 2001; Portin 2002), and still additional models and definitions have been proposed (e.g., Morange 2001, Portin & Wilkins 2017). Initially, new knowledge about genes, chromosomes, and biochemistry served to more precisely define the “gene” concept through the first half of the twentieth century. The late twentieth century (through present) has brought discoveries of overlapping genes, transposable elements, nested genes, interrupted genes, repeating genes, pseudogenes, multiple alternate promoters, polyprotein genes, functional RNAs and many other phenomena, which collectively have served to stretch the boundaries of conventional gene definitions (Beurton, Falk, & Rheinberger 2000; Morange 2001; Portin & Wilkins 2017). With a single term seeming insufficient to accurately address the diversity of gene structure and function, some have argued that the scientific community is in need of new terminology and new definitions of the gene (e.g., Portin & Wilkins 2017). Despite a deeper understanding of gene function, many areas of traditional genetics instruction still rely heavily on early models of gene function (Gericke & Hagberg 2007; Gericke et al. 2014).

Adding an additional layer of ambiguity to the gene concept is the tendency of different biological disciplines to utilize different operational gene definitions (Morange 2001, Portin 2002; Gericke & Hagberg 2007). Consider, for example, the differing roles a gene plays from the perspective of a population geneticist, evolutionary biologist, developmental biologist, molecular biologist, and a genomicist. Textbooks and teachers have been found to be inconsistent in their use of genetics terminology and gene models as well. For example, a sample of Swedish high school teachers who were observed during lessons were inconsistent in their use of the term gene and omitted the use of the term allele altogether (Thorpe, Gericke & Hagberg 2013). Textbooks often switch between gene models (and definitions) or present conflicting gene models without warning, or even present mixed or hybrid models of gene function (Gericke et al. 2014).

In addition to the complexity of the term “gene,” more than one research group has noted student confusion about the meanings of words such as “dominant” and “mutation,” particularly as meanings differ between the contexts of scientific research and everyday use (Pearson & Hughes 1988; Radford & Bird-Stewart 1982; Smith & Good 1984). Using a history of science perspective, Jamieson and Radick (2013) argued that some of the confusion about dominance can be attributed to widespread curricular representations that are actually misinterpretations of Mendel’s original model. Further, the conceptual closeness and physical proximity of structures such as alleles, genes, chromosomes, and chromatids make it unsurprising that students such as those in two United Kingdom studies reported being unable to distinguish among the terms (Bahar et al. 1999; Longden 1982). Similar sounding words--such as homologue, homologous, homozygous, homozygote--further complicate communication (Bahar et al. 1999). Indeed, Burian & Kampourakis (2013) and Smith & Adkinson (2010) have proposed replacement language phrases for teaching genetics in order to better reflect our modern molecular understandings and the complex nature of traits.

1.2.2 Reasoning across organizational levels and divisions

Adding to the complexity of learning genetics is the need for simultaneous thinking about processes that occur at different organizational levels (i.e., macro, micro, and molecular). For example, one must understand trait manifestation at the organismal (macro) level, the cellular
(micro) level, and chromosome, gene, DNA, or protein (molecular) level (e.g., Bahar et al. 1999; Duncan & Reiser 2007; Marbach-Ad & Stavy 2000; Marbach-Ad 2001). Beyond that, in genetics problem solving one must correctly apply symbols to represent one level (typically the molecular or micro level), manipulate the symbols according to specific rules that are cued by interpretations of domain-specific terminology and which employ probability and ratios, then reason back to another level (usually macro) to deduce an outcome (Bahar et al. 1999). Duncan & Reiser (2007) used the term “hybrid hierarchical” to describe the ways in which a change in a function at one level affects a function at another level. The complex nature of problem solving means that there is a high cognitive demand and multiple points at which student thinking might get derailed. Duncan, Freidenreich, Chinn, & Bausch (2011) designed and successfully implemented a “learning scaffold” depicting genes, proteins, cells, and traits to help students think across these organizational levels.

Because genetics draws on interrelated information from several subtopics (commonly referred to as the meiotic, Mendelian, and molecular aspects of genetics; Stewart, Cartier & Passmore 2005), confusion in any one area—or about the links among these areas—can impede the development of expertise. Many studies document student problems with meiosis or the meiotic mechanisms behind segregation and recombination, including the representation of these ideas in Punnett squares (Browning & Lehman 1988; Cavallo 1994, 1996; Kindfield 1991; Moll & Allen 1987; Smith & Good 1984; Stewart 1983; Stewart & Dale 1989; Stewart, Hafner & Dale 1990; Tolman 1982; Todd & Romine 2017). Still other studies document a disconnect between the molecular and Mendelian models (e.g., Stewart & Dale 1989; Lewis & Kattman 2004; Duncan & Reiser 2007; Todd & Romine 2017). Most recently, while developing their learning progression of molecular genetics, Todd and Romine (2017) confirmed that, even after instruction, most high school and college students have difficulty integrating the meiotic and Mendelian models and the Mendelian and molecular models. Until the missing links between these three facets of genetics learning can be clearly described (and methods can be devised and tested to address these conceptual breaks), genetics will remain a challenging subject. Research-informed, empirically-tested interventions such a Duncan et al.’s (2011) scaffolds are an important step in bridging the gaps among the three genetics subtopics.

1.2.3 Textbooks and curricula

Research from the nature, history, and philosophy of science perspective has identified inconsistencies in representations of genetics in educational research literature, textbooks, and teaching which may contribute to students’ confusion about genetics. Gericke and Hagberg (2007) analyzed historic models of the gene as described in genetics education literature and concluded that misrepresented, misunderstood, or hybridized representations of the phenomena of gene function in historic models can contribute to alternate, non-normative ideas about how genes function. Specific areas of incoherence included gene function within different organizational levels, consideration of definitions of the gene from: “top-down” versus “bottom-up” perspectives; the relationships between genotype and phenotype; and the relationships between genes and the environment. With colleagues, Gericke et al. (2014) also found that textbooks across six countries commonly presented hybridized versions and inconsistent representations of different historical gene models. Based on literature showing that textbooks largely shape the content, organization, and discussion of biological concepts within classrooms, Gericke et al. reasoned that hybridized and incoherent models of gene function also abound in
genetics classes. They suggest that, to avoid confusion, teachers and textbooks need to: (1) make explicit the idea that different sub-disciplines of genetics and/or research maintain different conceptual models of the gene relevant to the uses and goals of each sub-discipline, and (2) continually help students organize ideas along conceptual scaffolds during instruction.

Several researchers maintain that additional roadblocks to genetics learning lie in curricular organization. Temporally (Longden 1982; Kindfield 1991) or conceptually (Tolman 1982; Longden 1982) disjointed and potentially inappropriate (Cho et al. 1985) sequencing of genetics-related topics within the school curriculum are thought to exacerbate confusion about genetics. In particular, the common separation of units on meiosis and inheritance, especially without making explicit attempts to depict a conceptual connection (Cho et al. 1995), seem to add to troubles students have connecting these concepts (Knippels, Waarlo & Boersma 2005; Longden 1982). Recent results from Duncan, Castro-Faix, and Choi’s (2016) exploration of a middle school genetics learning progression demonstrated that an understanding of molecular genetics seemed to bootstrap understanding of Mendelian genetics, making a case for teaching molecular genetics before Mendelian genetics. Dougherty (2009) argued, given that our modern understanding of genetics acknowledges most traits are indeed “complex” (e.g., polygenic, quantitative, and multifactorial), and that our current “historic” approach (from the concepts of genes to dominance to alleles and then molecular aspects) is not very effective, curricula should be reorganized to teach quantitative and multifactorial traits before monogenic traits. (It should be noted that Dougherty’s proposal for an inverted curriculum includes the Mendelian and meiotic aspects of genetics but does not address the molecular components). Indeed, which treatment and sequence of genetics material is best to facilitate deep understanding has been and remains a topic of debate (e.g., Tolman 1982; Smith & Gericke 2015). Robust empirical measures will be central to generating evidence-based evaluations of these alternative perspectives.

1.2.4 Genetics problem types and cognitive demands

When considering difficulties in genetics problem solving, it is important to consider the wide array of problems that are commonly used in classrooms and which have been employed in research as it has been suggested that different inheritance problem types elicit different cognitive demands. Typical textbook and assessment genetics problems usually involve a cross, or mating, in which information is given about the parents, and questions are asked about the offspring, or vice versa. Symbols and Punnett squares are conventionally used to diagram the combinatorial or probabilistic thinking required in these problems, although they can be solved without construction of a Punnett square as well. A variety of genetics problems exist and can be classified according to several criteria (Collins & Stewart 1989; Hackling & Lawrence 1988; Hickey et al. 2000; Smith & Sims 1992; Stewart 1988):

1. Problem format [e.g., traditional cross (giving parental information asking for offspring information), backward thinking cross (predicting parental information given offspring information), pedigree (family tree using diagrams to represent traits)]
2. Inheritance pattern (e.g., complete dominance, codominance, sex linkage)
3. Number of traits involved (e.g., monohybrid, dihybrid)
4. Type of reasoning involved (cause to effect or effect to cause)
Some evidence from problem solving research and cognitive theory suggests that different genetics problem types elicit different cognitive demands (though more research is needed in this area). Traditional problems (which are the most common) are also considered the easiest because they often can be solved by rote algorithms with limited knowledge of inheritance processes (e.g., Cavallo 1994, 1996; Smith & Good 1984; Stewart 1983; Stewart, Hafner & Dale 1990; Browning & Lehman 1988). As the number of traits simultaneously studied increases, so does problem difficulty; with two or more traits, the Punnett square method becomes cumbersome and students must switch to a multiplication of ratios approach to combine the probabilities of individual trait inheritance. Smith & Sims (1992) note that the latter approach, while more efficient, requires higher-level thought. In contrast to traditional crosses, pedigree, or family tree problems, interpretation problems tend to be more open-ended and require more advanced reasoning and genetics knowledge because one must deduce information about inheritance modes and genotypes from a given diagram prior to solving any further questions (e.g., Hackling & Lawrence 1988; Smith & Sims 1992). With respect to inheritance pattern, Collins & Stewart (1989) considered incomplete dominance problems to be less demanding than simple dominance problems. Incomplete dominance features a 1:1 mapping of each genotype to phenotype, whereas in simple dominance both homozygous dominant and heterozygous genotypes are mapped to the dominant phenotype, which can be confusing for novice learners. Likewise, cause-to-effect problems have been shown to be less demanding for students than effect-to-cause problems (Hackling & Lawrence 1988; Hickey et al. 2000; Smith & Sims 1992; Stewart 1988). Cause-to-effect problems require 1:1 mapping of the genotype to phenotype whereas effect-to-cause problems require consideration of more than one possible genotypic antecedent for a given phenotypic effect (Hackling & Lawrence 1988; Smith & Sims 1992; Stewart 1988). It is clear that the wide variety of genetic problem types adds to the complexity of successfully understanding (and demonstrating competency in) transmission genetics. Students are tasked first with correctly identifying a problem type from a problem statement and then appropriately applying a complicated domain-specific, knowledge-dependent, mathematically-driven heuristic to reach a solution.

Several studies have documented challenges posed by the highly abstract and even mathematical nature of reasoning required in genetics. Genetics has been described as a highly theoretical field requiring a significant amount of abstract thought (Baker & Lawson 2001; Banet & Ayuso 2000) and hypothetico-deductive reasoning (Baker & Lawson 2001). Many studies have noted that genetics involves highly developed levels of reasoning at Piaget’s formal operational level (Cavallo 1996; Gipson, Abraham & Renner 1989; Smith & Sims 1992; Walker 1979, 1980). Smith & Sims (1992) clarified that formal operational reasoning ability—in the forms of proportional, probabilistic, and combinatorial reasoning specifically—was not an absolute requirement for genetics problem solving (as some students were able to solve problems by alternative routes), however formal reasoning was very helpful to increase success. As an elaborative example, Cavallo (1996) found reasoning ability was a better predictor of student success at correctly solving Punnett square problems, but not for understanding meiotic events underlying genetics (discussed below in the section on meaningful learning).

Probabilistic reasoning and computational skills (Browning & Lehman 1988; Kinnear 1983; Smith & Good 1984) as well as symbolic representation (Bahar et al. 1999; Thomson & Stewart 1985) are key requirements for solving genetics problems. Students sometimes lack the simple computational skills necessary to perform mathematical calculations of genetics problems (Browning & Lehman 1988; Smith & Sims 1992). Instructors and textbooks commonly employ
inconsistent systems of symbolism to represent common genetics topics (Bahar et al. 1999), and
students sometimes manipulate symbols and alter algorithms without proper consideration of
genetic laws (Thompson & Stewart 1985). Improper choice of symbols, use of symbols without
understanding what they are intended to represent, or manipulation of symbols in a manner
inconsistent with the rules relevant to the content represented by the symbols can lead to
incorrect conclusions. Such pitfalls are especially characteristic of students who attempt to solve
genetics problems by rote rather than understanding (e.g., Smith & Sims 1992).

1.2.5 Approaches to problem solving

The overwhelming tendency of students is to approach genetics by rote—relying on
simple algorithms to obtain fixed ratios (Cavallo 1994, 1996; Smith & Good 1984; Stewart 1983;
Stewart, Hafner & Dale 1990; Browning & Lehman 1988). This strategy promotes reliance on
superficial problem features instead of mechanistic processes (Smith & Good 1984; Smith &
Sims 1992; Stewart, Hafner & Dale 1990). This approach clearly limits deeper conceptual
understanding (Cho et al. 1985; Longden 1982, Novak 2002; Smith & Good 1984). Despite the
eventual learning constraints arising from rote approaches to problem solving, many students
persist in this approach because it efficiently rewards a correct answer (Cavallo 1994; Moll &
Allen 1987; Smith & Sims 1992; Stewart 1983). Students who avoid this pitfall are those whose
epistemic stance is oriented toward meaningful learning (i.e., deliberately connecting new ideas
to existing ideas while learning; Ausubel 1963). It has been suggested that students have an
inherent and identifiable orientation toward either meaningful or rote learning (Entwistle &
Ramsden 1983; Novak 2002) and some interventions in genetics learning have been aimed at
providing scaffolding to help students make meaning by connecting ideas. One example centers
on the observation that students who draw on knowledge of meiosis have greater success solving
problems (Cavallo & Shafer 1994; Moll & Allen 1987). When Cavallo & Shafer provided
students with parallel assessment tasks which differed only in one suggestion to consider
knowledge of meiosis while solving the problem, the additional prompt to connect knowledge
enabled more students to solve the task. Similar results have been noted by Duncan et al. (2011),

1.2.6 Content knowledge

Because genetics problem solving draws heavily on domain-specific content knowledge
while simultaneously requiring complex domain-specific heuristics, several problem-solving
barriers exist for students who have not amassed sufficient prior content knowledge.
Specifically, prior knowledge is accessed when successful problem solvers look for cues as to
characterize problem type (Hackling 1990; Hackling & Lawrence 1988; Slack & Stewart 1990;
Smith 1983; Smith & Good 1984), activate schemas (Slack & Stewart 1990), generate and test
hypotheses and test alternate hypotheses to disprove alternate modes of inheritance (Collins
1986; Hackling 1990; Hackling & Lawrence 1988), justify inheritance according to meiotic
models (Cavallo & Schafer 1994; Stewart and Dale 1981), employ domain vocabulary to
summarize data (Collins 1986; Collins & Stewart 1989; Simmons & Lunetta 1993), and check
for consistency with the problem and the larger body of genetics knowledge (Smith 1983; Smith
& Good 1984). Overall, a broader knowledge base affords a larger library of schema and
schema-cuing situations, and connections between cues and schemas are stronger; it thus
becomes easier to select clues in the problem, apply a heuristic, and work out the problem. Given that an appreciable portion of secondary students begin to study genetics having yet to develop the understandings that certain organisms are alive, and that they contain cells, chromosomes, and genes (Banet & Ayuso 2000), it is not surprising that many students remain ill-equipped to solve genetics problems. Hackling (1990; Hackling & Lawrence 1988) has experimented with problem-solving scaffolding by directing students to check for certain things and respond with certain behaviors during problem solving, much like experts do.

Given the large amount of content knowledge required to solve genetics problems, and the importance that baseline knowledge plays in selecting appropriate problem-solving heuristics, it is not surprising that Stewart (1983; Stewart & Van Kirk 1990) and Hafner (Hafner & Stewart 1995) recommend problem solving within the domain as a way of developing domain-specific knowledge as well as domain-specific and domain-general skills. On the contrary, Sweller (1988) presents evidence from cognitive psychology to suggest that, in some instances (such as less skilled individuals employing a means-end approach), high cognitive load demands may interfere with learning and acquisition of appropriate problem solving schema. It is possible that a threshold model applies in which problem solving can reinforce concepts and skills in learners who have attained at least a certain baseline level of prior knowledge, but problems that are too challenging create roadblocks to learning. While problem solving practice is a common learning vehicle in many domains, the impact of excessive cognitive load in genetics problem solving is not to be overlooked. Clearly, more work is needed on this topic.

1.2.7 Life experiences reinforce misconceptions

A simple but potentially important impediment to learning genetics may be that everyday life experiences work against the development of an adequate understanding of genetics. Some research suggests that the public is resistant to developing an accurate understanding of genetics because everyday social relationships and practices regularly reinforce the naive concept of kinship (Richards 1996). Kinship conceptions are typically held by young children as they begin to understand relatedness in families and animals. Children observe and consequently expect similar traits in organisms based on social constructs such as perceived degree of relatedness. This expectation may even extend to certain acquired traits. Kinship theory is problematic because it fails to consider the physical basis of inheritance, such as DNA, genes, or chromosomes; indeed students have been noted to apply molecular-deficient reasoning when they go so far as to conflate genes with traits (e.g., Lewis and Kattmann 2004). Social reinforcement of insufficient explanatory elements such a kinship may be thus associated with misconceptions about the true causes of traits (Venville, Gribble & Donovan 2005).

1.3 Modeling as a means of improving genetics understanding

Several researchers in genetics problem solving (in addition to Moll & Allen 1987) have documented how the inclusion of supports such as diagrammatic-, physical-, analogic-, or computer-generated models can enhance problem-solving ability. It is possible that these supports provide the scaffolding to help students attend to important cues or skills in order to work forward, thereby also decreasing cognitive load for struggling students. Kindfield (1994) noted that successful problem solvers constructed and used diagrams of meiosis to recall and accumulate relevant knowledge to support problem solving, then checked their own reasoning
against diagrams that they had created. Consequently, she advocated for the addition of instructional activities that involve creating and using diagrams as biological models for learning. Several other researchers have invented creative means for modeling meiosis, including using: students as chromosomes (Chinnici, Yue & Torres 2004), string and paper (Stencel 1995), socks (Chinnici, Neth & Sherman 2006), and cards (Krauskopf 1999), though most have not empirically demonstrated significant learning gains associated with these activities. Venville and Donovan (2008) described using an analogical model to support the language of genetics. Learning gains were modest and best revealed qualitatively, and were also dependent on students prior knowledge and interactions with the model. A large scale interactive computer modeling project, GenScope (a precursor to BioLogica), was also shown to produce only modest learning gains for most groups of students (Horwitz et al. 1998). However, the program had a very strong effect on improving the very low baseline knowledge of a group of urban at-risk participants. It would be interesting to explore whether other similar effects might be noted from similar computer interventions with populations such as this.

Tsui & Treagust (2007) performed an extensive analysis of student learning gains with GenScope’s successor program (the interactive computer modeling project BioLogica). Some (but not all) students showed gains, and some gains seemed to be linked to the students’ prior knowledge (or how intently they interacted with the program). These findings echo the results of Cavallo’s (1996; Cavallo & Schafer’s 1994) study on the intersections among meaningful learning orientation, prior knowledge, and genetics success. Regardless of overall assessment results, most students claimed that the interactive computer models helped them visualize the processes and understand what was happening (for example, an animation of meiosis labeled with alleles played while students answered questions about what they saw) although it did not translate into measurable results. In both Tsui & Treagust’s (2003, 2007) studies, BioLogica was most effective when used in conjunction with teacher scaffolding and cooperative student interaction to support the construction of knowledge. Essentially, Tsui & Treagust were describing the efficacy of a cooperative constructivist learning model, and thus these findings could be viewed as support for the importance of meaningful learning behaviors in genetics. A conclusion that can be drawn from this collective body of research is that it is the interactions with the model that seemed to be the important factor related to learning outcomes. Models can clearly support--but do not on their own account for or replace--deep, meaningful learning.

1.4 Argumentation in genetics education

In line with recent efforts to provide more authentic science education programs that couple scientific practices with scientific content to better reflect the nature and work of scientists (NGSS Lead States 2013), studies have begun to examine the relationships between genetics and argumentation. Argumentation in genetics may serve as an alternative to traditional “problem solving,” and may function as a way of learning as well as a means of assessing understanding in genetics (e.g., Jimenez-Aleixandre, Rodriguez & Duschl 2000; NGSS Lead States 2013). Research in genetics argumentation has shown that stronger genetics content knowledge was associated with higher quality argumentation when confronted with socioscientific issues in genetics (Sadler & Donnelly 2006; Sadler & Fowler 2006; Sadler & Zeidler 2005). Sadler & Donnelly proposed a threshold model in which the relationship between content knowledge and argumentation is not linear but does exist. In their model, individuals with expert-like knowledge were less impacted by contextual features than those with a lesser
knowledge base. Shea, Duncan & Stevenson (2015) expanded on the Sadler model to include the impact of context in argumentation quality. Specifically, individuals with strong content knowledge offered the most robust arguments regardless of context, while those with limited or partial subject knowledge produced arguments of varying strengths in relation to the contextual framing of the question prompt (discussed further below.) This is in alignment with findings on the role of context in other science content areas, (e.g., Chi et al.1981; Opfer, Nehm & Ha 2012; Ware & Gelman 2014). Overall, although it is clear that Mendelian genetics problems remain a central part of the curriculum (discussed further in Section 4.1.2), increasing attention in genetics education research is being focused on argumentation and modelling as means for measuring genetics understanding (e.g., Jiménez-Aleixandre 2014; Jiménez-Aleixandre et al. 2000; Sadler & Donnelly 2006; Shea, Duncan & Stevenson 2015).

1.5 Theoretical and empirical limitations of genetics education research and areas for future research

1.5.1 Samples and study types

While nearly forty years of research has produced a sizeable literature on genetics problem solving, there remain several substantial limitations of past research and several additional issues deserving attention. Most studies of genetics problem solving have been smaller-scale qualitative studies, which have yielded rich, detailed information of problem solving procedures and pitfalls, but lack evidence of generalizability to other populations. While overlapping findings from several descriptive studies in varied settings may serve to provide some degree of corroboration, the body of literature reviewed above remains subject to common theoretical limitations of qualitative studies (e.g., limits of subjects’ vocalizations during interviews or “think-alouds,” interactions with the researcher, small sample sizes lacking representativeness). In general, this work lacks large-scale, broad-context, empirically-designed, and more rigorously-validated studies capable of refuting, corroborating, or extending existing understandings.

For example, in addition to existing novice-expert studies, more studies performed on large samples drawn from middle school, high school, and early and late stage university levels might yield generalizable findings about how genetics problem solving may improve or what challenges are faced at each level. Todd and Romine’s and Duncan’s research groups have begun to lay early groundwork in this area. Additionally, while the body of genetics problem solving research represents a patchwork of national contexts, studies intentionally comparing different geographic areas might shed light on similarities and differences in how diverse students approach similar problems. Much work also has failed to consider whether demographic factors, such as gender, are significantly associated with measures of genetics learning (discussed below). Overall, experimental studies with greater attention to sample size, composition, representativeness and generalizability are needed in the field of genetics education research in order to generate robust evidence-based claims.

1.5.2. Genetics problem selection

Most of the studies reviewed above have relied on measures of genetics understanding. More rigorous statistical analyses employing (at minimum) Classical Test Theory (CTT) or
(ideally) Item Response Theory (IRT) perspectives in alignment with some of the guidelines set forth in the *Standards for Educational and Psychological Testing* (AERA, APA, NCME 2014) would offer higher-quality insights into the degree to which specific genetics tasks are able to generate valid and reliable inferences about student performance. The vast majority of studies in genetics problem solving, especially prior to the last decade, have involved such small item sets (and often few participants) that statistical validity and reliability analyses for tasks were not possible. Although the tendency in the research community has been to draw problems from popular textbooks or craft similar-style problems, there is really no indication—beyond items having met a textbook editor’s or the principal investigator’s approval—that any of the problems used in these studies are “good” or functioning in the manner the researcher has anticipated. Indeed, few studies even make an attempt to offer any evidence toward the content or substantive validity of their questions (see Chapter 3, below). Larger, more robust, and empirically substantiated experimental studies are needed to complement the array of descriptive studies in genetics problem solving.

The array of genetics problems employed in the extant body of research is neither evenly representative of all genetics problem types nor organized according to any rational principle. The lack of consistency in problem selection precludes this body of work from yielding more coherent insights into the cognitive demands elicited by each problem type. Most genetics problem-solving studies focus on monohybrid crosses featuring complete, and to a lesser extent, incomplete dominance. Dihybrid crosses, pedigrees, and sex-linkage problems constitute a smaller portion of the research, and other problem types are nearly absent from the literature. More studies are needed that are able to experimentally isolate and compare problems by type and represent a broader sample of problem types. For example, a study might examine whether genetics rules are applied consistently across problem types regardless of number of traits studied. This could be achieved by developing a carefully constrained set of traits (with attention to a consistent organism type, similar trait type and mode of trait inheritance) and examining student responses to monohybrid, dihybrid, and even trihybrid crosses systematically constructed from the constrained feature set. Another study might examine whether inheritance pattern impacts problem solving ability by comparing multiple parallel problems featuring otherwise comparable dominant, recessive, incompletely- and co-dominant traits. Other studies might compare “cause to effect” to “effect to cause” reasoning by employing a constrained problem set varying only by whether a problem demanded forward or backward reasoning. Such experimental rotation of problem types is likely to generate a large item set and may require a multi-matrix design to disperse smaller, overlapping groups of problems over a large participant set, requiring sophisticated IRT analysis as mentioned prior. Still additional studies are needed to explore advanced and underrepresented inheritance patterns such as sex-linkage, multiple alleles, lethal alleles, epistasis, and quantitative traits. Clearly, there is much work that remains to be done in genetics education relating to problem solving.

It has been well-documented that most types of genetics problems can be solved by rote without drawing on deeper knowledge and, consequently, are not always true indicators of student understanding (Browning & Lehman 1988; Cavallo 1994, 1996; Cho et al. 1985; Longden 1982; Smith & Sims 1992; Smith & Good 1984; Stewart 1983; Stewart, Hafner & Dale 1990). Yet the larger body of genetics problem-solving research has continued to employ algorithmic problems. It is essential that this field of research experiment with a variety of alternative and ill-structured problem types and formats to shed more light on reasoning, behaviors, and knowledge use during problem solving. More open-ended formats such as
pedigree-based problems (e.g., Hackling 1990; Hackling & Lawrence 1988; Smith 1988) are but one example. A variety of less-structured problem types may accompany traditional genetics problems and extend and clarify information gleaned from traditional problem types. For example, Cavallo accompanied Punnett square problems with additional, related questions (Cavallo & Schafer 1994) or with open-ended questions (Cavallo 1996), and Kindfield (1994) assessed knowledge using diagrams. More recently, Tsui & Treagust (2010) have used a CTT-validated, two-tiered instrument to measure genetics knowledge. The first tier offers a forced-choice question and the second tier asks for the student’s reasoning for the choice made in the first tier. Todd & Romine (2016) used a similar two-tiered IRT-validated genetics instrument in which the second tier asks students about their extent of guessing (on the first-tier, forced-choice items). Two-tier instruments represent a practical solution to the need to gather rich information while managing large sample sizes. Automated machine scoring of constructed response items, such as those employed by Nehm and colleagues in the domain of evolution learning (Ha, Nehm, Urban-Lurain & Merrill 2011; Nehm, Ha, & Mayfield 2012) offers another solution to the question-quality versus scoring-quantity dilemma. Overall, the field of genetics problem solving will benefit from exploration of more modern and robust problem types in combination with the technology-enabled approaches.

Beyond the eclectic assemblage of problem types employed in genetics problem solving research, remarkably little consideration has been given to whether the superficial features of genetics problems or the contexts problems are situated in impact problem solving ability or measures of student understanding. A substantial literature in cognitive psychology and science education, much dating back to the 1970s and 1980s, offers evidence that assessment task features can impact the retrieval of scientific knowledge and resulting measures of participants’ conceptual understanding (e.g., Chi, Felchtovich, & Glaser 1981; further reviewed in Nehm & Ha 2011). In biology this has been documented most extensively by Nehm and colleagues who found, in large international samples, effects of evolution task features varied predictably depending on (1) taxon featured (2) the scale of evolutionary comparisons, (3) the familiarity of the taxa and traits, and (4) polarity of evolutionary change, and manifest most strongly in participants with low expertise in the domain (Federer et al. 2016; Nehm & Reilly 2007; Nehm & Ha 2011; Nehm et al. 2012; Nehm & Ridgway 2011; Opfer, Nehm, & Ha 2012).

In genetics, Smith (1992), too, found that superficial features of genetics problems influence the manner in which novices categorize problems. Because problem categorization is a key first step in problem solving, it is reasonable to imagine that task context would impact genetics problem-solving ability. Indeed, some studies in inheritance have hinted that contextual features impact genetics knowledge formation and retrieval (Banet & Ayuso 2000; Clough & Driver 1986; Kargbo, Hobbs, and Erickson 1980; Ware and Gelman 2014; Freidenreich, Duncan, and Shea 2011; Shea, Duncan, and Stephenson 2015). However, remarkably few studies have experimentally explored impacts of task context on a large scale (Banet & Ayuso 2000 and Ware and Gelman 2014 are a notable exception). Equally important to understanding how surface features impact student reasoning is an understanding of which surface features impact reasoning and under what situations. In line with the findings of Nehm and colleagues, it is reasonable to expect that surface features such as taxon of organism (human, animal, plant), familiarity with organism, scale on which a trait is presented (macroscopic, microscopic, or molecular level), whether a trait is advantageous or deleterious, and whether an allele variant confers a gain or loss of function, may impact formation and retrieval of genetics knowledge.
Further research, especially from larger empirical studies, is clearly needed to elucidate which situational features impact particular types of genetics problems, and how these features impact measures of student learning. An understanding of how task cover stories, contexts, or situational features impact reasoning are important for many culturally relevant educational activities, including the selection of curricular examples to illustrate scientific phenomena, the design of formative and summative assessment items, and determination of whether instruction has fostered the development of abstract schemas divorced from particular instances (Mayer 2016). Indeed, research needs to move away from the general documentation of context effects and towards domain-specific models that may be leveraged to improve teaching and learning of core ideas.

1.5.3 Demographic factors

Most studies in genetics education have not explicitly considered the role of demographic factors (e.g., gender, race) in their research designs despite a half century of research demonstrating differences in attitudes, understanding, achievement, and participation (Linn & Hyde 1989; Kahle & Meece 1994; Weinburgh 1995; Scantlebury & Baker 2007; Scantlebury 2014; Eddy & Brownell 2016; Peng, Wright, & Hill 1995; Lee & Luykx 2007). Differences in science achievement may be due to factors specific to a demographic group’s experience (Peng, Wright, & Hill 1995; Scantlebury & Baker 2007), or as a result of bias in curriculum, instructional practices, school climate, or assessment methods (Lee & Luykx 2007).

Understanding how either of these reasons can impact performance are important for practicing culturally-relevant pedagogy. In biology education, the roles of gender and ethnicity on domain-specific performance remain unsettled. Some studies, for example, have documented the absence of significant gender effects (e.g., Shepardson & Pizzini 1994; Dimitrov 1999; Huppert et al. 2002; Lauer et al. 2013; Willoughby & Metz 2009) or impact of ethnicity (e.g., Creech and Sweeder 2012; Dimitrov 1999; Nehm and Schonfeld 2008) on biology performance. Other studies, in contrast, have found advantages for males in undergraduate biology course grades (Creech & Sweeder 2012) and test scores (Eddy, Brownell, & Wenderoth. 2014; Stanger-Hall 2012; Wright et al. 2016) whereas other studies have found that females outperformed males on concept maps (Pearsall et al. 1997) and on tests of labeling errors (Soyibo 1999). In genetics, gender effects were not found by Cavallo (1994) in high school participants’ written explanations of genetics and meiosis or by Dogru-Atay and Tekkaya’s (2008) study of eighth graders’ multiple-choice responses about inheritance and genetics crosses. Franke and Bogner (2011), however, showed a female advantage for retaining new conceptions about molecular genetics and genetics technology on a multiple-choice test. To ensure accurate measures of learning and appropriately-designed curriculum and instructional methods, more information is needed about how different assessment methods measure understanding in the various branches of biology across all demographic groups.

1.5.4 Alignment with current educational standards and goals

Lastly, it is important to consider the modernization of genetics problem-solving research to address next-generation scientific problems that tackle content and practices and are in line with recent science education standards. Since the wave of problem-solving studies in the 1980s and 1990s, the standards for U.S. science education such as the National Science Education
Standards (NRC 1996) and Next Generation Science Standards (NGSS Lead States 2013) have shifted much of the basic traditional genetics material to the middle school level, yet few studies (excepting Duncan and colleagues’ learning progression work) have examined how genetics problem solving is approached in young adolescents. To the contrary, a substantial body of genetics problem solving work has been carried out in undergraduates, who presumably exceed adolescents in reasoning ability, mathematical ability, prior biology knowledge, maturity, academic experience, and life experience. All of these areas have some bearing, whether explicit or implicit, on genetics problem solving success. Even some item surface features might be expected to resonate differently with a thirteen year old and a nineteen year old. Work is in order to determine the extent to which insights from the existing body of knowledge in genetics learning and problem solving can be applied to middle school students and what new challenges may exist. Beyond studying genetics learning at individual educational levels, multi-level studies (e.g., Schmiemann, Nehm, & Tornabene 2017), especially those that are learning progression aligned (e.g., Todd & Romine 2016; Todd, Romine, & Cook Whitt 2017) can offer valuable insights into how students interact with genetics concepts as they progress through their educational experience.

Kampourakis et al. (2014) argue that the current teaching in the field of genetics is in need of better alignment with the current knowledge in the field and the current socioscientific issues that exist. A growing awareness of the important role of the nature of science and the need to empower students to be scientifically literate citizens who can engage in scientific practices to learn about the world and navigate modern socioscientific issues is reflected in the new education Standards (NGSS Lead States 2013). While it is clear that Mendelian inheritance remains a part of the curriculum, new standards require students’ engagement in scientific practices such as the construction of domain-specific models to develop and demonstrate knowledge about scientific phenomena and employ scientific discourse and argumentation to support and evaluate claims about data and phenomena. These practices will serve both as a vehicle for learning and as a means of measuring of understanding in genetics. Research on learning in genetics must therefore continue to move away from problem-solving heuristics and towards model-use and argumentation. Some notable examples are an early work by Hafner & Stewart (1995) and Finkel (1996) on revision of meiotic models for dominant inheritance patterns in order to facilitate knowledge construction for additional inheritance patterns. More recent studies in problem-based learning and genetics argumentation have been conducted by the Duncan group (e.g., Hmelo-Silver, Duncan, Chinn 2007; Shea, Duncan & Stephenson 2015), Sadler and colleagues (Sadler & Donnelly 2006; Sadler & Fowler 2006; Sadler & Zeidler 2005) and Jiménez-Aleixandre (2014) and colleagues (Jimenez-Aleixandre et al. 2000). Work in these arenas (e.g., Shea, Duncan and Stephenson 2015) has identified the roles that task context and prior knowledge play in new measures of genetics literacy. Specifically, they have explored how task context, prior knowledge use, and learning approaches can be extended to contemporary measures of genetics understanding. Studies such as these are essential to inform teaching and measurement practices in line with the Next Generation Science Standards (NGSS).

1.6 Conclusion

Genetics is one of the most widely-researched divisions of biology education, and with good cause. Its conceptual centrality means that genetics is a foundation for understanding the very nature of biological characteristics, change, and diversity. Its abstract, theoretical, and
The ontologically broad nature means that genetics is particularly challenging to master, and especially difficult to teach. As a field, genetics offers a point of personal connection, as it provides some answers, although only partial, as to why we are the way we are. We have witnessed continuing technological advances that have facilitated exponential expansion of genetics knowledge and applications. The field has, in many cases, ‘grown up’ alongside us. Most of us will now have occasion to interact directly with diagnostic methods or treatments related to genetics.

Just as the field of genetics has changed, so have our goals for scientific literacy and consequent goals for science education. The field of genetics education needs to move beyond the confines of small, descriptive studies and employ more robust, empirically validated experimental studies. Within genetics problem solving research, variables such as problem type, pattern of inheritance, participant level of education, and impact of problem context must be explored in samples from diverse settings to generate robust, generalizable conclusions about genetics learning. Extending the field of problem solving, research much focus on newer, standards-aligned and philosophically-aligned methods of developing and measuring genetics understanding (such as the use of models and argumentation). The field of traditional genetics learning is transitioning away from the days of a teacher standing at the board and students solving textbook-style problems on their own, and towards students working in collaborative groups while the teacher facilitates discourse development and model revision to construct arguments and explanations. Attention to the success of all students, and aims for culturally-relevant instruction, require that research designs consider how the demographic constitution of participants may interact with instructional practices and assessment. Improved understanding of cognitive aspects of learning and assessment warrant consideration of how selection of curricular examples, assessment item contexts, and presentation order may interact with student learning and measures of such. New technologies in the form of computer automated testing, scoring, and psychometric analysis are available to support these improvements in research.

1.7 References (Chapter 1)


Longden, B. (1982). Genetics—are there inherent learning difficulties?. *Journal of Biological Education*, 16(2), 135-140.


Chapter 2. Psychometric Analysis Using Item Response Theory

2.1 Introduction

Accurate measurements of student understanding are essential for educators and researchers alike (Campbell & Nehm 2013). A growing array of assessments are being developed to generate data about undergraduate understanding in a variety of STEM disciplines, whether study-specific instruments (e.g., Cavallo 1996; Tsui & Treagust 2010), Concept Inventories (e.g., chemical bonding, physical forces, genetics, and natural selection) or broader learning-progression aligned instruments (e.g., Todd & Romine 2016). Although organizations such as the American Educational Research Association, American Psychological Association, National Council on Measurement in Education (AERA, APA & NCME 2014) and measurement experts such as Messick (1995) have outlined standards for construct validity and reliability of assessment scores, such evidence for the majority of these instruments remains limited, and rooted in classical test theory (CTT), despite the known advantages of Item Response Theory (IRT) and Rasch modeling (Boone, Staver & Yale 2014). For example, of ten recently-developed instruments assessing Mendelian inheritance (see Chapter 4), only six were supported by any form of validity or reliability evidence, only one of which used IRT/Rasch analysis (see Table 4.1). More robust evidence in support of the validity and reliability of inferences generated by measurement instrument scores is essential to support claims in science education. In a recent review of the status of genetics instruction in higher education, McElhinny et al. (2014) specifically identified the need for genetics measurement instruments evaluated using the more powerful IRT framework. The studies produced for this dissertation provide additional evidence to that end.

The IRT framework and Rasch modelling provide several advantages over CTT for instrument evaluation and the generation of accurate scores. This chapter provides background on IRT and Rasch, describes several advantages of the IRT framework, and explains how Rasch modeling via WINSTEPS (Linacre 2017a) or ConQuest (Adams et al. 2016) software is employed to empirically evaluate the validity and reliability of inferences derived from the scores of the instruments used in the studies herein (the Genetic Drift Inventory [GeDI; Price et.al. 2014] and a project-specific Mendelian inheritance instrument) for their respective populations (Chapters 3 and 4, respectively).

2.1 Instrument evaluation using Item Response Theory

The frameworks for developing and evaluating assessment instruments have changed substantially over the past few decades. Familiarity with modern instrument evaluation methods is essential in order to understand the strengths and weaknesses of the measures that are derived from science education instruments (AERA, APA, NCME 2014). Classical Test Theory (CTT) and Item Response Theory (IRT) are two conceptual and empirical frameworks commonly used for analyzing and evaluating measurement instruments.

Item response theory is a model-based psychometric approach centered on the premise that responses to an item set measuring a single trait are functions of both the test taker’s attributes (i.e., ability level on the trait) and the item’s attributes (i.e., difficulty). IRT posits a predictable response pattern (model) whereby easier items are correctly answered more frequently than difficult items, and more able persons correctly answer more items, including the
more difficult items. Parameters of person ability and item difficulty are estimated from a set of iterative comparisons of response patterns according to this premise. A variety of IRT models exist, varying in the type of instrument responses they accommodate (e.g., dichotomous or polytomous) and in the number of parameters considered (e.g., the 1 parameter logistic, or 1PL, model considers the parameter of item difficulty, while the two parameter logistic model, 2PL, considers both difficulty and discrimination; see Bond & Fox 2007 for more information).

Rasch methodologies share much in common with the IRT framework, and are often considered to be a form of IRT. The dichotomous Rasch model (used in the Chapters 3 and 4) is mathematically equivalent to the 1PL IRT model. A key philosophical and practical distinction between Rasch and other IRT analyses is that Rasch considers only the first IRT parameter (item difficulty) and does not alter the model (e.g., add parameters) to fit the data. As such, Rasch affords characterization of persons and items in a manner that is more robust, with greater inferential potential, than CTT or other IRT approaches (Bond & Fox 2007; Boone et al. 2014). Several of Rasch’s advantages that are discussed in the following paragraphs stem from this distinction.

2.3 Advantages of Rasch and IRT

2.3.1 Ratio-scale logit scores for persons and items

The vast majority of science education instruments have been developed and evaluated using CTT as a guiding framework. IRT/Rasch frameworks address many inherent limitations of CTT (Bond & Fox 2007; Nehm & Schonfeld 2008; Boone et al. 2014). One major advantage of IRT and Rasch methods is their conversion of raw (interval scale) test scores into linear, ratio-scaled scores. This feature is essential for addressing unequal difficulty intervals between raw test scores. Consider, for example, the ability difference between two low performing individuals whose raw scores differ by one point (e.g., scores of 70 and 71 out of 100) and the ability difference between two high performing individuals whose scores also differ by one point (e.g., scores of 99 and 100). It is unlikely that the items that separated the two high-achieving students have the same difficulty value as the items that separated the low achieving students, and yet for both pairs the difference between raw scores is equal (1 point). Because raw scores are calculated without consideration of item difficulty, they do not adequately represent the true ability difference between individuals. Put another way, the quantity “one point” does not seem to measure the same attribute in these four students; the true difference in ability between the two high achieving students would be much greater than the difference between the two lower-scoring students. Rasch ratio-scale scores are calculated with consideration of item difficulty and thus remedy raw score inconsistencies. Conversion to linear data is also crucial to satisfy the assumptions of parametric statistical analyses commonly conducted using test scores. In sum, IRT/Rasch methods address a fundamental problem with CTT scores: non-ratio-scaled data.

Rasch scores (or “measures”) for persons and items are reported as logit units and derive from a probability-based logarithmic equation that considers both item difficulty and person ability. Using the same logit scale to quantify both item difficulty and person ability facilitates comparison among items, persons, and items and persons. It also affords analyses capable of determining the probability that a particular person could solve a particular item. In typical Rasch analyses, mean item difficulty and mean person ability are set at 0 logits. More difficult items (or higher achievers) are given higher scores, while easier items (or lower achievers) are
given lower (more negative) scores. When logit values for person measure and item measure are equivalent, an individual has a 50% probability of correctly answering the item (Bond & Fox 2007, p. 38).

### 2.3.2 Instrument dimensionality

An important component of instrument evaluation is confirmation of the instrument’s dimensionality. Most instrument evaluation methods and parametric analyses of data generated by instruments assume unidimensionality, or that the instrument measures one (and only one) construct (Neumann et al. 2011). Attempting to capture more than one construct at a time, or probing distinct facets of a single construct, can introduce multidimensionality. Multidimensionality presents complications when reporting an individual’s instrument scores as a single value (e.g., Which portions of the total score represent which construct?) and for analyses—including Rasch—that inherently assume one construct is being measured. (Note that methodological extensions of Rasch do exist that can accommodate multidimensionality.) Thus, multidimensional instruments must either (1) be treated as multiple unidimensional instruments, with scores reported and analyzed as such (along with corresponding validity evidence), or (2) be analyzed with advanced psychometric methodologies specific to multidimensionality.

Traditional CTT-aligned approaches to ascertaining dimensionality (e.g., confirmatory factor analysis and principal component analysis) can be problematic; these analyses typically require data to be normally distributed and conform to an equal interval scale, which, as has been mentioned, is most often not the case. In evaluating the degree to which an instrument is multidimensional, principal components analysis of Rasch-scaled scores offers information on the response variance that can be attributed to the items (i.e., variance explained by the model; first contrast in Rasch-scaled principal component analysis) and the degree to which response variance is unexplained (i.e., the second contrast, and so on). A second dimension is hypothesized to exist if the unexplained variance is larger than what would be expected to be due to random noise in the data (for details, see Raiche 2005). Variance beyond the random noise threshold can be attributed to additional dimensions within the instrument, though other considerations such as construct structure, variance in responses, and the purpose of measurement afford some degree of flexibility in this interpretation (Linacre 2017b).

### 2.3.3 Wright maps

A display of all person and item measures for a unidimensional construct on a shared logit scale (commonly known as a Wright map, e.g., Figures 3.1 and 4.1) is another powerful and unique application of Rasch. This side-by-side comparison enables researchers to examine the alignment of test items to test taker “ability” and to identify possible measurement gaps (i.e. difficulty/ability ranges in which items are lacking). Items are represented by their respective number on the right side of the scale, while persons are represented by “X’s” on the left side of the scale. Given the probabilistic nature of the analysis, each person has a 50% chance of correctly answering an item with an equivalent measure. In a well-designed instrument, question difficulty should be aligned with test-taker ability, with items present that are able to differentiate among learners at all ability levels. Thus, instrument evaluation using a Wright map includes examining the match of the “spread” of test takers to the “spread” of test items. Items that are too easy appear below the lowest test takers (having been estimated to be correctly answered by
everyone, these items likely add little value to the measures), while items that are too difficult appear above the highest test takers (these items may be too challenging for the sample). If all of the persons are plotted above the highest item or below the lowest item, then the items lack alignment with ability level. Multiple items aligned at the same difficulty levels on the Wright map, and testing the same concept or misconception, add little to measurement and are candidates for elimination. Large clusters of persons at the same ability level indicate locations where additional items could be added to better separate their abilities. Overall, the Wright map is a useful visual tool for examining instrument properties and person-item relationships.

2.3.4 Item and person fit

Analyses of the degree to which the empirical data fit the statistical Rasch model are one approach for evaluating the quality of the test items, the test instrument, and overall evidence in support of validity claims (Boone et al. 2014). Rasch analysis includes several parameters to examine model fit. Overall item fit and person fit scores describe how well the collective item set and collective person sample fit the Rasch model respectively. These values provide insights into overall instrument function. Individual item and person fit statistics are useful for determining whether items and persons fit the Rasch model. Poor model fit reveals when items and persons behave unexpectedly (e.g., an item may be interpreted differently and elicit inconsistent responses, a person may guess, a high ability person may get some low difficulty items wrong). Accordingly, poorly functioning individual items or persons can be identified using these fit statistics.

In Rasch measurement, fit is expressed as weighted (“infit”) or unweighted (“outfit”) values for the mean square parameter (MNSQ), and calculation of fit is based on a chi-square test of how well the empirical data fit the Rasch model (Bond & Fox 2007, p. 238). For a standard multiple choice assessment, MNSQ values above 1.3 are considered to be “underfitting”, indicating that the response pattern for that item is erratic. Values below 0.7 are considered to be “overfitting”, indicating that the response pattern is overly predictable. Both overfit and underfit suggest that the item is not functioning properly (i.e., eliciting information consistent with test-taker ability). Cut off values of 0.7 and 1.3 are used for the MNSQ parameter to ensure an adequate match between the empirical data and the statistical model (Boone et al. 2014; Bond & Fox 2007). Z-Standard (ZSTD) scores are transformed t-test statistics that report the probability of MNSQ scores occurring by chance when the data fit the Rasch model (Linacre 2017b). Ideal ZSTD scores range from 0 to 2. However, as sample size increases, accumulation of random responses tends to elevate ZSTD scores (Smith et al. 2008). For this reason, and because ZSTD statistics are based on MNSQ statistics, ZSTD values are considered secondary to MNSQ scores. Depending upon measurement goals and sample sizes, ZSTD scores may be ignored if MNSQ values are acceptable (Linacre 2017b). With multiple indicators of fit that correspond to different causes of misfit as well as parameters to report the probability of fit statistics, Rasch and IRT provide a much more detailed characterization of item fit properties compared to CTT.

2.3.5 Item and person reliability

Further indicators of instrument quality include Rasch item and person (separation) reliability measures, which reflect internal consistency and can be interpreted analogously to
Cronbach’s alpha in CTT (cf. Wright & Stone 1979). Together, acceptable item reliability and person reliability indicate that the item set functions to differentiate the measured trait into a number of ability levels sufficient for measurement goals in manner that can be replicated in comparable samples. Specifically, item reliability addresses whether the persons sampled demonstrated sufficiently diverse abilities to support the calculated item difficulty structure, while person reliability addresses whether the item difficulty structure is sufficient to reliably produce person measures. Together these are again a more nuanced measurement of reliability than CTT affords.

Item reliability values <0.9 suggest that the participant sample is likely to be too small to confirm the apparent item-difficulty structure. Person reliability values <0.8 suggest that assessment items are insufficient to distinguish among test takers. This may also suggest that the Rasch person measure score (or how well each person performed based on the Rasch ratio-score model) may not be a reliable reflection of person ability (Boone et al. 2014). These values are guidelines for a “general” instrument and sample, and should be interpreted according to specific characteristics of an instrument including its format (e.g., number of items, number of response choices), and the stated goals of measurement (e.g., norm- or criterion-referenced) (Boone et al. 2014; Linacre 2017b).

2.3.6 Missing data

A key benefit of IRT and Rasch modeling is the ability to readily accommodate “missing” data. Because person estimates are based on the probability a person will correctly respond to a given item of a particular difficulty, failure to answer a few items among many others whose difficulty is known does not significantly impact person estimates; the model is able to predict how a person would likely have answered a skipped question based on responses to items of similar difficulty. Similarly, because item measures are estimated based on the probability that a person of a determined ability will select a correct answer for that item, item estimates are not impacted by the absence of a few individuals’ responses from among many responses of known ability. These properties ensure that Rasch person scores are item-independent and item scores are sample-independent, characteristics which afford researchers the widespread benefit of being able to confidently utilize partially completed student response sets. Accommodation of missing data is also essential for computer adaptive testing (Bond & Fox 2007) and multi-matrix studies in which participants are assigned only a subset of items from the total collection of questions (cf. Sirotnik and Wellington 1977; e.g., Schmiemann, Nehm & Tornabene 2017). Such designs allow testing of a wider variety of items while minimizing participant test fatigue. Accommodation of missing data is also important in strategies intended to minimize the impact of guessing, including multi-tiered instruments that incorporate answer confidence tiers (cf. Romine, Schaffer & Barrow 2015). In sum, Rasch and IRT hold considerable potential for expanding the body of empirical evidence on instrument quality, yet remain broadly underutilized. The analysis of the GeDI presented in Chapter 3 provides an example of how Rasch analysis can offer greater insight into the measurement capabilities and limitations of educational measurement instruments.
2.3.7 Rating scale optimization

Rasch converts ordinal rating scale survey responses into linear data affording more meaningful interpretation and analysis of rating scale surveys. Though not specific to this project, this point is worth mentioning because many biology education instruments include rating scales (e.g., Carver et al. 2017; Glynn et al. 2011; Nadelson & Southerland 2012; Rutledge & Warden 1999; Smith, Snyder & Devereaux 2016). Rasch analysis further provides information about how well each item’s rating scale category functions to capture respondents’ perspectives, which facilitates revision and optimization of rating scale categories. These represent notable advantages over CTT and traditional item analysis, which are unable to account for arbitrary and inconsistently spaced rating scale categories and consequently cannot yield conclusive information about the trait of interest (Boone et al. 2014). More studies are needed throughout the science education field, especially for rating scale instruments, to improve the quality of measurement in all science disciplines.

2.4 Use of IRT in this project

The present project employs Rasch analysis to generate evidence about the validity, reliability and generalizability of score inferences from two genetics education instruments— one relating to genetic drift (GeDI) and one on Mendelian inheritance. This aspect of the project will serve to (1) add to the body of empirical evidence underlying the quality of research instruments (2) suggest possible limitations of selected instruments and identify possible areas for instrument improvement, and (3) empirically support conclusions to additional research questions posed herein about measurement of genetics knowledge (cf. Chapter 4).

2.5 References (Chapter 2)


3.1 Introduction

The accurate measurement of student understanding is an essential feature of educational practice because it provides evidence-based insights into students’ conceptual ecologies, guides learning progression development, and permits empirical evaluation of the efficacy of alternative educational interventions (National Research Council 2001). A diverse array of assessment tools and types have been developed for evolution educators (Table 3.1). They range from static, multiple-choice formats (e.g., Price et al. 2014) to open-ended questions whose answers can be scored by computers (e.g., Moherrari et al. 2014). Available assessment tools cover many different evolutionary concepts, including natural selection, evo-devo, genetic drift, and macroevolution. These assessments vary significantly in the types of information that they can reveal about student understanding, in the situations in which they are most appropriately implemented, and in the robustness of the inferences that they are able to support (AAAS 2011; AERA, APS, NCME 2014; Nehm and Schonfeld 2008).

3.1.1 Concept Inventories as valid and reliable tools to generate inferences about student understanding

Concept Inventories (CIs) are a type of research-based educational assessment designed to rapidly reveal (through easy administration and scoring) students’ preferences for normative (i.e., scientifically accurate) or non-normative (e.g., preconceptions, misconceptions) facets of core ideas (e.g., natural selection, genetic drift) (Nehm & Haertig 2012, p. 56-57). Although CIs have become indispensable tools for assessing undergraduate students’ conceptual understandings of many core ideas in the sciences (e.g., force and motion, chemical bonding), few have been carefully evaluated in terms of (1) the forms of validity outlined in the Standards for Educational and Psychological Testing (AERA, APA, NCME 2014), (2) item order effects and associated response biases (Federer et al. 2015, 2016), or (3) item properties using ratio-scaled data (generated by Rasch or Item Response Theory [IRT] analyses; Boone, Staver & Yale 2014). Consequently, validity evidence--that is, evidence that the measures derived from CIs accurately reflect the construct of interest--remains limited. Given the centrality of accurate measurement to evidence-based educational practices, evolution education research must include the study of instrument quality. Such studies help to support instructional decisions firmly rooted in high-quality evidence.

Given the paucity of work on evolution education instrument quality (Nehm and Schonfeld 2010), our study examines the psychometric properties of a relatively new evolution education instrument known as the Genetic Drift Inventory (GeDI). As the only instrument designed to measure conceptions of non-adaptive evolutionary mechanisms, the GeDI fills a crucial gap in the array of evolution education instruments and holds potential to offer insights into a much neglected area of student thinking about evolution. To date, use of the GeDI for research on genetic drift conceptions and validity evidence for inferences generated by the GeDI both remain limited to CTT-based data (Price et al. 2014; Price et al. 2016; Price & Perez 2016), despite availability of more robust IRT and Rasch methodologies (Boone et al. 2014).
3.1.2 Research Questions

In order to build a larger body of validity evidence in support of evolution education assessments in general, and to empirically examine the strengths and weaknesses of the inferences that may be drawn from GeDI scores in particular, our study explores three research questions: (1) How well does the GeDI function when studied within the context of the Rasch model? (2) Does the presentation order of instrument vignettes (and associated item suites) impact measures of student understanding? And (3) Does the GeDI measure student knowledge in a manner that is generalizable across geographic regions of the United States (e.g., Northeast, Southeast, and Midwest) when administered to students of similar academic backgrounds? Prior to discussing our psychometric approach, we begin with a brief review of the position of genetic drift within evolution education, continue with an overview of Classical Test Theory and Item Response Theory frameworks for instrument evaluation, and end with a summary of GeDI instrument properties and prior validation work relative to these frameworks.

3.1.3 Genetic Drift and Evolution Education

A major goal of science education is to promote student understanding that is aligned with expert conceptions, practices, and dispositions. The scientific community recognizes both adaptive and nonadaptive causes of evolutionary change (reviewed in Beggrow & Nehm 2012, Masel 2012). While standards and textbooks vary in the extent to which they address non-adaptive evolutionary processes, genetic drift is recognized foremost among the various non-adaptive evolutionary factors (Beggrow & Nehm 2012; Price & Perez 2016). Genetic drift is included in college textbooks for biology majors (Beggrow & Nehm 2012), is a recommended topic in undergraduate biology curricula, and is also taught in Advanced Placement (AP) Biology (reviewed in Price & Perez 2016; The College Board 2015). International Baccalaureate (IB) Biology (a popular alternative to AP biology), however, fails to mention non-adaptive mechanisms for evolution (International Baccalaureate Organization 2014).

At the introductory high school biology level, the Next Generation Science Standards (NGSS Lead States 2013) also omit non-adaptive evolutionary mechanisms. Recent editions of popular high school textbooks, however, continue to include genetic drift (e.g., Miller & Levine 2014; Nowicki 2017), leaving the option to cover this topic in the hands of individual teachers, schools, or districts. While genetic drift is commonly taught in evolution courses (e.g., Masel 2012) or within evolution units of biology survey courses (e.g., The College Board 2015; Masel 2012; Urry et al. 2017), it may also be taught in genetics courses (e.g., Masel 2012; Stony Brook University 2017, p. 49). Overall, while there is consensus that nonadaptive causes of evolution are an essential component of biology education, inconsistent attention to genetic drift (and other non-adaptive evolutionary concepts) in high-school and college curricula makes it difficult to determine the extent to which students are exposed to instruction on non-adaptive evolutionary processes as well as the degree to which they are able to integrate it into their mental models of evolutionary change (Nehm 2018). The Genetic Drift CI was developed to address the latter issue and is an important advance in evolution assessment.
Table 3.1 Instruments measuring knowledge of evolutionary processes: Potential to elicit normative and non-normative ideas about adaptive and non-adaptive evolution

<table>
<thead>
<tr>
<th>Instrument</th>
<th>Formata and target population</th>
<th>Conceptions measuredb</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bishop &amp; Anderson's Diagnostic Instrument (Bishop &amp; Anderson 1990)</td>
<td>Combination MC and ORc: Undergraduates (introductory biology non-majors)</td>
<td>Intended</td>
</tr>
<tr>
<td>Concept Inventory of Natural Selection (CINS; Anderson, Fisher &amp; Norman 2002)</td>
<td>20 MC: Undergraduates</td>
<td>Intended</td>
</tr>
<tr>
<td>Assessing Contextual Reasoning about Natural Selection (ACORNS; Nehm, Beggrow, Opfer &amp; Ha 2012)</td>
<td>Flexible number OR: Undergraduates</td>
<td>Intended</td>
</tr>
<tr>
<td>Conceptual Assessment of Natural Selection (CANS; Kalinowski, Leonary &amp; Taper 2016)</td>
<td>24 MC: Undergraduates (introductory biology majors)</td>
<td>Intended</td>
</tr>
<tr>
<td>Daphne Assessment for Natural Selection (DANS; Furtak, Morrison, and Kroog 2014)</td>
<td>26 MC: High school</td>
<td>Intended</td>
</tr>
<tr>
<td>Genetice Drift Inventory (GeDI; Price et al. 2014)</td>
<td>22 TF: Undergraduates (upper-division biology majors)</td>
<td>Intended</td>
</tr>
<tr>
<td>Evo-Devo Concept Inventory (Perez et al. 2013)</td>
<td>11 MC: Undergraduates</td>
<td>Intended</td>
</tr>
<tr>
<td>Measure of Understanding of Macroevolution (MUM; Nadelson &amp; Southerland 2009)</td>
<td>27 MC and 1 OR: Undergraduates</td>
<td>OR:</td>
</tr>
</tbody>
</table>

a Multiple choice (MC), Open response (OR), True-false (TF)
b Natural selection, normative ideas (NS-N), Natural selection, non-normative ideas (NS-NN), Genetic drift, normative ideas (GD-N), Genetic drift, non-normative ideas (GD-NN); “Intended” indicates that the instrument intentionally targeted ideas of this type.
c Bishop & Anderson’s instrument includes 2 OR, 3 MC with OR explanation, and 1 question about belief in evolution.
d Open response format affords the possibility of capturing reasoning about genetic drift, although, in line with instrument’s intent, scoring guide addresses natural selection only.
e Includes one question (item 20) asking whether chance plays a role in whether a cactus will produce a seedling.
f MC items address macroevolution. OR item asks student to explain how two species might have arisen from one. Authors state that item does not address speciation by means beyond natural selection, though they include a student response mentioning genetic drift.
3.1.4 The Genetic Drift Inventory

The Genetic Drift Inventory (known as the GeDI; Price et al. 2014) is a 22-item CI designed to measure advanced undergraduate biology majors’ understanding of four key concepts and six alternative conceptions (or “misconceptions”) of genetic drift. To date, it is the only concept inventory to focus on non-adaptive evolutionary processes (Table 3.1). The GeDI features four vignettes, each followed by one to three question stems containing a number of associated agree-disagree statements (i.e., items; see Table 3.2 for details). The 22 items target an individual key concept (15 items) or a misconception (7 items). Misconceptions targeted by the GeDI are limited to those expected to be harbored by upper division majors whose knowledge of genetic drift is developing but often conflated with other evolutionary mechanisms (see Price et al. 2014 for more information on misconception delineation by expertise levels). For scoring, GeDI authors recommend that all items are given equal weight (e.g., 17/22 = 77%). To compensate for the high guessing rate for dichotomous questions, GeDI developers recommended: comparing raw scores before and after instruction, using higher than usual raw score cut-points to define success, or consideration of only the percentage correct above 50% (the score that could potentially be obtained by guessing alone) (Price et al. 2014). All of these scoring recommendations are grounded in Classical Test Theory (see below).

Table 3.2 GeDI vignettes and associated items

<table>
<thead>
<tr>
<th>Vignette</th>
<th>Items</th>
<th>Scenario featured</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1-8</td>
<td>Small subpopulation of land snails colonize a new island</td>
</tr>
<tr>
<td>2</td>
<td>9-11</td>
<td>Dung beetles geographically isolated by canals</td>
</tr>
<tr>
<td>3</td>
<td>12-18</td>
<td>Biologist randomly selects fruit flies to breed in captive populations</td>
</tr>
<tr>
<td>4</td>
<td>19-22</td>
<td>Nearsighted island population of humans before and after a devastating storm</td>
</tr>
</tbody>
</table>

3.2 Instrument evaluation using Item Response Theory

3.2.1 What is the Item Response Theory framework?

The frameworks for developing and evaluating assessment instruments have changed substantially over the past few decades, and faculty at all educational levels need to be familiar with these changes in order to understand the strengths and weaknesses of the measures that are derived from science education instruments (AERA, APA, NCME 2014). Classical Test Theory (CTT) and Item Response Theory (IRT) are two conceptual and empirical frameworks commonly used for analyzing and evaluating measurement instruments. This section briefly calls attention to some of the salient features of IRT/Rasch that are relevant to the present analysis of the GeDI instrument. A more comprehensive and technical discussion of these frameworks is offered in Bond and Fox (2007), Boone et al. (2014), and de Ayala (2009).

Item Response Theory is a superior yet underutilized framework for instrument evaluation. This model-based psychometric approach is centered on the premise that responses to an item set measuring a single trait are functions of both the test taker’s attributes (i.e., ability level on the trait) and the item’s attributes (i.e., difficulty). In contrast, CTT, the framework
used to evaluate the vast majority of science education measurement instruments, presumes responses are functions only of the test taker’s ability, along with an expected degree of measurement error. Various IRT models exist to accommodate diverse types of instrument responses (e.g., dichotomous or polytomous) and which consider different numbers of parameters during estimation (e.g., the 1 parameter logistic, or 1PL, model considers the parameter of item difficulty, while the two parameter logistic model, 2PL, considers both difficulty and discrimination; see Bond & Fox 2007 for more information).

Rasch methodologies share much in common with the IRT framework, and are considered by many to be a form of IRT (although strict Rasch enthusiasts maintain a firm distinction, e.g., Boone et al. 2014). The dichotomous Rasch model used in this study is mathematically equivalent to the 1PL IRT model. As such, Rasch affords characterization of persons and items in a manner that is more robust, with greater inferential potential, than Classical Test Theory or other IRT approaches (Bond & Fox 2007; Boone et al. 2014).

3.2.2 Advantages of IRT

Versatility

IRT and Rasch frameworks address many inherent limitations of CTT (Bond & Fox 2007; Nehm & Schonfeld 2008; Boone et al. 2014). A broad advantage is the existence of diverse IRT and Rasch models suitable for different types of data (unidimensional, multidimensional, dichotomous, polytomous, large and small data sets), permitting analyses to be more closely matched to study type. Also, missing data are readily accommodated in IRT/Rasch frameworks, adjusted for during the iterative comparisons of person and item response patterns used to estimate person and item scores. Not only does this enable researchers to utilise a greater proportion of survey responses, it is particularly useful for research designs which minimize test fatigue by intentionally assigning individuals only a subset of questions from wider variety of items such as in computer adaptive testing (Bond & Fox 2007) and multi-matrix studies (cf. Sirotnik and Wellington 1977; e.g., Schmiemann, Nehm & Tornabene 2017). Overall, Rasch and IRT are designed for adaptability to support a wide range of study formats.

Ratio scale scores

An integral and unparalleled advantage of IRT and Rasch approaches is production of person and item scores on a linear, ratio-scale (rather than ordinal raw data). Linear data is essential to accurately quantify person and item characteristics as well as to satisfy the assumptions of parametric statistical analyses commonly conducted using test scores. Dimensionality analyses, necessary to ascertain which theoretical model (unidimensional, multidimensional) used for evaluations is appropriate, are also properly conducted with linear ratio scale data. The distinction between ratio scale IRT/Rasch scores and ordinal CTT scores lies in CTT’s failure to consider the existence of varying item difficulties or their impact on test score when calculating score values. Overall, IRT and Rasch’s system of reporting item and person scores on a ratio scale fundamentally improves the quality of most inferences drawn from test results.
Common Logit Scale for Person and Item Score Reporting

IRT and Rasch report person and item measures on the same logit scale. This facilitates comparison among and within participants and items: When logit values for person measure and item measure are equivalent, an individual has a 50% probability of correctly answering the item (Bond & Fox 2007, p. 38). Graphical displays of persons and items on a common logit scale (known as “Wright maps”), are a unique feature of Rasch and IRT which enable researchers to explicitly examine the alignment of test items to test taker “ability” and to identify possible measurement gaps (i.e. difficulty/ability ranges in which items are lacking) or areas of redundancy.

Multiple Indicators of Fit and Reliability

Rasch and IRT feature multiple indicators of fit (e.g., infit and outfit mean square [MNSQ]) that correspond to different causes of misfit and also include parameters to report the probability of fit statistics (e.g., Z-standardized t scores [ZSTD]). Fit statistics are provided both at the item/person level and instrument/sample level. Together, these parameters provide a much more detailed characterization of item and person fit properties compared to CTT. Similarly, IRT/Rasch derived values for individual and overall item reliability and person reliability together afford a more nuanced measurement of reliability than CTT. Item reliability values convey whether the persons sampled demonstrated sufficiently diverse abilities to support the calculated item difficulty structure, and person reliability values address whether the item difficulty structure is sufficient to reliably produce person measures.

Overall, Rasch and IRT hold considerable potential for expanding the body of empirical evidence on instrument quality, yet remain broadly underutilized in science education. The present analysis of the GeDI provides an example of how Rasch analysis can offer greater insight into the measurement capabilities and limitations of measurement instruments.

3.3 Additional relevant considerations for instrument evaluation

3.3.1 Item order effects on student performance

An extensive body of work extending back to the 1950’s (e.g., MacNicol 1956; Mollenkopf 1950) has found that instrument scores may be influenced by interactions among (1) item position (that is, which questions students encounter first, second, third, etc.) and item difficulty, (2) item format (multiple choice, constructed response; qualitative or quantitative), and (3) test type (aptitude or achievement) (reviewed in Federer et al. 2015; Leary & Dorans 1985). For example, working with the ACORNS instrument, Federer et al. (2015) found an interaction between item order and taxon familiarity on student performance measures. The GeDI contains several separate vignettes with associated item suites that vary in task contexts (cf. Table 3.2) and item difficulty levels (Price et al. 2014). It is possible that these (or other unidentified) aspects of the items could influence student responses to subsequent items (cf. Federer et al. 2015). Hence, investigation of whether vignette order impacts student performance is a worthwhile step towards understanding the measurement properties of the GeDI.
3.3.2 Generalizability of instrument scores

Evidence for generalization validity is important to substantiate claims that an instrument measures a trait in the same manner across different populations and administration contexts. Instruments are designed to measure a specific construct under specific circumstances, such as a particular educational level (e.g., undergraduate biology majors, elementary students) under certain administration conditions (e.g., unproctored computerized testing, timed paper-and-pencil tests), and for particular purposes (e.g., formative evaluation of instructional interventions, employment screening). Explicit delineation of such contexts and evidence to support validity and reliability of inferences generated under these circumstances should accompany instruments (AERA, APA, NCME 2014). Under alternative administration contexts (e.g., sample populations, testing conditions), items are subject to differing interpretations or stress factors which may bias responses. For instance, a question may be beyond the comprehension level of a group, may be scrutinized more stringently by those with greater subject expertise, or may contain terms whose meaning differs according to the cultural or regional background of a sample. Accordingly, biased item responses compromise the validity of inferences about the latent trait (AERA, APA, NCME 2014). When an instrument is used in a new context, evidence is needed to support the validity and reliability of inferences generated in the new context.

The GeDi is intended to measure upper division biology majors’ conceptions of genetic drift across different institution types and in different courses. While development and initial validation sampled a broad array of students from different biology courses and institution types throughout the Midwest and Central United States regions, samples from the Northeast were not included (Price et al. 2014). Given that regions of the United States vary widely in demographic composition, religion, and evolution acceptance, additional information from a Northeastern population would further substantiate claims about the utility of the GeDI across geographic regions.

3.3.3 Summary of validity and reliability evidence for the GeDI

Now that we have reviewed some of the salient empirical evidence that may be used to evaluate the quality of the inferences derived from assessment scores, we can turn our attention to the GeDi. The GeDI has only been evaluated using Classical Test Theory methods despite many known limitations of using raw data to interpret item and instrument properties (as discussed in Chapter 2; Boone et al. 2014). A summary of the forms of evidence used to support validity inferences for the GeDI are shown in Table 3.3. The present study expands upon prior validity and reliability work by (1) employing Rasch Modeling, which produces more accurate ratio-scaled scores and can contribute evidence to examine dimensionality, construct validity, internal structure validity, item properties, and reliability, (2) examining item order effects, and (3) studying a participant population from a new geographic region of the country (Table 3.3).
### Table 3.3 Summary of validity and reliability evidence for the GeDI.

<table>
<thead>
<tr>
<th>Validity/Reliability Evidence Type and Description</th>
<th>CTT Framework (Price et al. 2014)</th>
<th>Rasch Framework (Present Study)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Construct validity: Instrument appropriately represents the specified knowledge domain</td>
<td>Textbook analysis, expert survey, student interviews, review of student work and literature review for misconceptions</td>
<td>Rasch model fit, Rasch dimensionality analysis, item fit, person reliability</td>
</tr>
<tr>
<td>Substantive validity: Participants use the thought processes that were anticipated for each item</td>
<td>Student interviews</td>
<td>(none)</td>
</tr>
<tr>
<td>Internal structure validity: Items capture a single construct</td>
<td>Cronbach’s alpha</td>
<td>Rasch dimensionality test, person &amp; item reliability</td>
</tr>
<tr>
<td>External structure validity: Scores are appropriately associated (positively or negatively) with an independent measure</td>
<td>(none)</td>
<td>(none)</td>
</tr>
<tr>
<td>Generalization validity: Score inferences hold true in different administrative contexts</td>
<td>Five campuses over two geographic areas (Southeast/Midwest)</td>
<td>New population (Northeast)</td>
</tr>
<tr>
<td>Consequential validity: Considers positive or negative consequences of score use</td>
<td>Not applicable</td>
<td>Not applicable</td>
</tr>
<tr>
<td>Reliability: Reproducibility of scores</td>
<td>Test-retest</td>
<td>Item and person reliabilities</td>
</tr>
<tr>
<td>Item properties: Individual item performance characteristics</td>
<td>Difficulty, discrimination</td>
<td>Item measures, item fit statistics, Wright map</td>
</tr>
<tr>
<td>Item order effects: Possible item interactions and associated sequence biases</td>
<td>(none)</td>
<td>ANOVA of Rasch-scaled scores from forms rotating item-suite order</td>
</tr>
</tbody>
</table>

*Based on Campbell & Nehm (2013); Messick (1995); Nitko & Brookhart (2010)*

### 3.4 Methods

#### 3.4.1 Item order

The GeDI features four vignettes, each followed by one to three question stems containing a number of associated agree-disagree statements (i.e., items; see Table 3.2). The GeDI’s vignettes differ in situational features (cf. Table 3.2) and difficulty, two factors which have been shown to demonstrate item-order effects in prior studies (reviewed in Federer et al. 2014). In order to determine whether the sequence of vignettes and related items within the GeDI instrument impacted student performance, four complete forms of the GeDI were generated,
which differed only in the presentation sequence of vignettes. A four-by-four Latin square design was used to rotate vignette sequence among the test forms (see Table 3.4). Each of the four vignettes (and related items) constituted a block in the square; the original order of the vignettes and items (Price et al. 2014) was used to seed the Latin square, and the original order of the items within a block was maintained throughout all forms (see Table 3.4).

**Table 3.4** Design of GeDI forms

<table>
<thead>
<tr>
<th>Position 1</th>
<th>Position 2</th>
<th>Position 3</th>
<th>Position 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Vignette 1</td>
<td>Vignette 2</td>
<td>Vignette 3</td>
<td>Vignette 4</td>
</tr>
<tr>
<td>Vignette 2</td>
<td>Vignette 3</td>
<td>Vignette 4</td>
<td>Vignette 1</td>
</tr>
<tr>
<td>Vignette 3</td>
<td>Vignette 4</td>
<td>Vignette 1</td>
<td>Vignette 2</td>
</tr>
<tr>
<td>Vignette 4</td>
<td>Vignette 1</td>
<td>Vignette 2</td>
<td>Vignette 3</td>
</tr>
</tbody>
</table>

The presentation sequence of the four vignettes (and accompanying item suites that comprise the GeDI) was rotated among four equivalent test forms using a Latin square design. A Latin square is an array of \( n \) rows and \( n \) columns, with each row and each column containing units 1 through \( n \) exactly once. By rotating experimental treatments in this manner, the researcher can generate data to determine whether an adjacent treatment (or, in this case, item suite placement) influences overall performance. For a description of vignettes and associated items see Table 3.2.

### 3.4.2 Sample and administration

The GeDI forms (Table 3.4) were administered online using a learning management system in the spring semester of an upper division (300-level) genetics class at a large, Northeastern Doctoral-granting university. This course was chosen because it aligns with the target population for GeDI use and is among the course types used in the development and initial validation studies of the GeDI (Price et al. 2014). Students were randomly assigned to one of four experimental groups, each of which had access to only one of the four forms of the assessment (Table 3.4). Students were provided a 60-minute time limit and allowed one attempt to complete the activity, with extra credit offered as an incentive for participation. Random student identification numbers were assigned to anonymize response data. The assessment was open for a period of one week beginning in the 10th week of the semester, prior to which no instruction relating to genetic drift had occurred. Of the 480 students enrolled, 336 (70%) completed the assessment in the following distribution: \( n_{\text{form 1}} = 91, n_{\text{form 2}} = 78, n_{\text{form 3}} = 80 \) and \( n_{\text{form 4}} = 87 \). All data collection qualified for IRB exemption status under IRB number 2013-2298-R1.

### 3.4.3 Data analysis

In order to empirically evaluate the validity and reliability inferences derived from GeDI scores, Rasch modeling was performed using WINSTEPS v 3.682 (Lincree 2017a).
Dimensionality was examined via a Principal Components Analysis (PCA) of Rasch residuals. The overall fit of items and persons to the unidimensional Rasch model were examined by infit and outfit mean square (MNSQ) values and Z standard (ZSTD) values. A Wright map was generated to visualize item difficulty relative to test-taker ability, and individual item fit values were considered. Item reliability was calculated to determine whether responses were varied enough to confirm the item difficulty structure, person reliability was calculated to determine whether the items differentiated among achievement levels sufficiently. To determine if item order impacted test performance, a one-way ANOVA was performed on Rasch-scaled scores for the four GeDI forms. Finally, total scores and item difficulty ranks were compared across administrations in order to examine score generalizability.

3.5 Results

3.5.1 Dimensionality

Principal Components Analysis (PCA) of Rasch residuals was used to compare the amount of variance explained by items and persons in relation to unexplained variance (which might correspond to additional dimensions). For our sample, items explained 13.3% of the variance while 6.6% remained unexplained, and person measures explained approximately as much. With an approximate 2:1 ratio of variance due to items versus unexplained variance, a high loading for the first dimension was apparent. High unexplained variance is common for samples demonstrating narrow ranges of ability (see Figure 1). For an instrument with 22 items, an Eigenvalue greater than two would suggest additional dimensions (Linacre 2017b). For our sample, the Eigenvalue was 1.8 in the first contrast. Thus, the analysis did not support additional dimensions for the GeDI.

3.5.2 Overall model fit

Overall item fit and person fit values are measures of how well a dataset fits the Rasch model. Values outside of the acceptable range (MNSQ = 0.7-1.3 and Z standard values < 2.0) suggest that test takers were responding in a manner that was either overly predictable, very erratic, or inconsistent with their ability. Excellent overall person fit values (infit MNSQ = 1.0, ZSTD = 0.0; outfit MNSQ = 0.99, ZSTD = 0.0) and item fit values (infit MNSQ = 1.0, ZSTD = 0.0; outfit MNSQ = 0.99, ZSTD = -0.1) were apparent (Table 3.5) and indicated that the participant sample responses fit the Rasch model very well.

3.5.3 Individual item fit

Rasch infit and outfit MNSQ and ZSTD values were used to assess how well individual GeDI items aligned with the student population and with the Rasch model. Infit and outfit MNSQ values for all test items were within acceptable ranges (Table 3.5) and are thus functioning to elicit responses consistent with test-taker ability. Five items (2, 9, 16, 19, and 22) had infit and/or outfit ZSTD values outside of the acceptable range. According to Linacre (2017b), misfitting ZSTD scores are very sensitive to sample size, and may be disregarded when samples are large (over 300 observations) and MNSQ scores are acceptable. This is because ZSTD values reflect how perfectly data fit the Rasch model rather than how usefully data fit the
model, and in large samples (over 300 observations), the accumulation of rare individual atypical responses can inflate ZSTD scores without having a bearing on the usefulness of the data.

Table 3.5 GeDI Rasch fit properties.

<table>
<thead>
<tr>
<th>Item</th>
<th>Infit MNSQ</th>
<th>Infit ZSTD</th>
<th>Outfit MNSQ</th>
<th>Outfit ZSTD</th>
</tr>
</thead>
<tbody>
<tr>
<td>Item</td>
<td>1.00</td>
<td>0.0</td>
<td>0.99</td>
<td>0.0</td>
</tr>
<tr>
<td>Person</td>
<td>1.00</td>
<td>0.0</td>
<td>0.99</td>
<td>-0.01</td>
</tr>
<tr>
<td>Item 1</td>
<td>1.09</td>
<td>1.28</td>
<td>1.11</td>
<td>0.94</td>
</tr>
<tr>
<td>Item 2</td>
<td>1.14</td>
<td>3.75</td>
<td>1.22</td>
<td>3.84</td>
</tr>
<tr>
<td>Item 3</td>
<td>0.96</td>
<td>-0.86</td>
<td>0.92</td>
<td>-1.06</td>
</tr>
<tr>
<td>Item 4</td>
<td>1.08</td>
<td>1.40</td>
<td>1.09</td>
<td>0.91</td>
</tr>
<tr>
<td>Item 5</td>
<td>0.93</td>
<td>-1.68</td>
<td>0.89</td>
<td>-1.94</td>
</tr>
<tr>
<td>Item 6</td>
<td>0.99</td>
<td>-0.21</td>
<td>0.96</td>
<td>-0.54</td>
</tr>
<tr>
<td>Item 7</td>
<td>1.02</td>
<td>0.45</td>
<td>1.07</td>
<td>0.91</td>
</tr>
<tr>
<td>Item 8</td>
<td>0.96</td>
<td>-0.64</td>
<td>0.96</td>
<td>-0.56</td>
</tr>
<tr>
<td>Item 9</td>
<td>0.92</td>
<td>-2.04</td>
<td>0.88</td>
<td>-1.67</td>
</tr>
<tr>
<td>Item 10</td>
<td>1.03</td>
<td>0.74</td>
<td>1.05</td>
<td>0.69</td>
</tr>
<tr>
<td>Item 11</td>
<td>1.02</td>
<td>0.41</td>
<td>1.03</td>
<td>0.43</td>
</tr>
<tr>
<td>Item 12</td>
<td>1.00</td>
<td>-0.07</td>
<td>1.00</td>
<td>-0.05</td>
</tr>
<tr>
<td>Item 13</td>
<td>1.00</td>
<td>0.13</td>
<td>0.98</td>
<td>-0.35</td>
</tr>
<tr>
<td>Item 14</td>
<td>0.98</td>
<td>-0.46</td>
<td>0.95</td>
<td>-0.80</td>
</tr>
<tr>
<td>Item 15</td>
<td>0.99</td>
<td>-0.12</td>
<td>0.96</td>
<td>-0.43</td>
</tr>
<tr>
<td>Item 16</td>
<td>1.20</td>
<td>4.05</td>
<td>1.27</td>
<td>4.24</td>
</tr>
<tr>
<td>Item 17</td>
<td>0.99</td>
<td>-0.20</td>
<td>1.00</td>
<td>-0.04</td>
</tr>
<tr>
<td>Item 18</td>
<td>0.98</td>
<td>-0.35</td>
<td>0.88</td>
<td>-1.05</td>
</tr>
<tr>
<td>Item 19</td>
<td>0.93</td>
<td>-2.12</td>
<td>0.88</td>
<td>-2.25</td>
</tr>
<tr>
<td>Item 20</td>
<td>0.98</td>
<td>-0.54</td>
<td>0.96</td>
<td>-0.63</td>
</tr>
<tr>
<td>Item 21</td>
<td>0.96</td>
<td>-0.79</td>
<td>0.92</td>
<td>-0.89</td>
</tr>
<tr>
<td>Item 22</td>
<td>0.91</td>
<td>-2.61</td>
<td>0.86</td>
<td>-2.54</td>
</tr>
</tbody>
</table>

Bold values refer to higher than expected values

3.5.6 Wright map

Wright map depicts item difficulty measures (on the right side) and person ability scores (on the left side) on the same logit scale (Figure 1). This side-by-side comparison enables one to understand how well test-takers are performing relative to item difficulty, and how well items are functioning relative to test-taker ability. Item numbers are plotted on the right side of the map, while persons are represented by the # symbols on the left side of the map. Mean item difficulty and mean person ability are set to zero, with the most difficult items and highest performers appearing toward the top of the map and the easiest items and lowest scorers appearing toward the bottom of the map. Typically, question difficulty should be well-matched with test-taker ability, with the presence of items that can differentiate among learners at all ability levels. A person has a fifty percent probability of correctly answering an item with an equivalent logit value.
Figure 3.1 Wright map derived from Rasch analysis of GeDI responses. The distribution of persons (on the left) and GeDI items (on the right) are illustrated on the same log interval (-2 to 4) scale. Each # = 4 persons.

The logit scores for test items and persons in Figure 1 demonstrate that the GeDI item difficulty is generally well matched to test-taker ability, with the exception of the top of the logit scale. About 12.5% of participants had logit scores above the most difficult item (item 8). Thus, the GeDI successfully differentiates most of this student population, but, from a strict perspective, requires additional (high difficulty) items to differentiate the highest scorers. Further, almost all test takers correctly answered item 1, indicating that it is “too easy” to differentiate students’ knowledge levels. The Wright map also illustrates three instances of test items displaying equivalent difficulty levels (see “Redundant difficulty” in Figure 1). Items of redundant difficulty are not functioning to discriminate among test-takers and may be candidates for removal in the interest of a removing uninformative items, unless such items are necessary for content validity, or some other aspect of construct validity. In this particular case (i.e., items 14 and 6, and items 10, 7, and 9), items with equivalent difficulty address different concepts or
“misconceptions.” Items 12 and 17, however, address the same misconception: “Natural selection is always the most powerful mechanism of evolution, and it is the primary agent of evolutionary change” (Price et al. 2014).

**Table 3.6** Mean GeDI item measures (Measure) and standard error (S.E.) by vignette rotation position

<table>
<thead>
<tr>
<th>Item</th>
<th>Position 1</th>
<th>Position 2</th>
<th>Position 3</th>
<th>Position 4</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Measure</td>
<td>S.E.</td>
<td>Measure</td>
<td>S.E.</td>
</tr>
<tr>
<td>1</td>
<td>-0.85</td>
<td>0.24</td>
<td>-1.04</td>
<td>0.26</td>
</tr>
<tr>
<td>2</td>
<td>0.19</td>
<td>0.22</td>
<td>0.64</td>
<td>0.23</td>
</tr>
<tr>
<td>3</td>
<td>-0.05</td>
<td>0.22</td>
<td>-0.78</td>
<td>0.25</td>
</tr>
<tr>
<td>4</td>
<td>-0.91</td>
<td>0.25</td>
<td>-0.78</td>
<td>0.25</td>
</tr>
<tr>
<td>5</td>
<td>0.75</td>
<td>0.23</td>
<td>0.75</td>
<td>0.24</td>
</tr>
<tr>
<td>6</td>
<td>1.14</td>
<td>0.24</td>
<td>0.81</td>
<td>0.24</td>
</tr>
<tr>
<td>7</td>
<td>-0.79</td>
<td>0.24</td>
<td>-0.6</td>
<td>0.24</td>
</tr>
<tr>
<td>8</td>
<td>1.2</td>
<td>0.25</td>
<td>1.1</td>
<td>0.25</td>
</tr>
<tr>
<td>9</td>
<td>-0.41</td>
<td>0.25</td>
<td>-0.73</td>
<td>0.24</td>
</tr>
<tr>
<td>10</td>
<td>-0.16</td>
<td>0.25</td>
<td>-0.46</td>
<td>0.23</td>
</tr>
<tr>
<td>11</td>
<td>-0.73</td>
<td>0.26</td>
<td>-0.57</td>
<td>0.23</td>
</tr>
<tr>
<td>12</td>
<td>1.11</td>
<td>0.26</td>
<td>1.13</td>
<td>0.26</td>
</tr>
<tr>
<td>13</td>
<td>1.11</td>
<td>0.26</td>
<td>1.06</td>
<td>0.26</td>
</tr>
<tr>
<td>14</td>
<td>1.18</td>
<td>0.27</td>
<td>1.2</td>
<td>0.27</td>
</tr>
<tr>
<td>15</td>
<td>-0.52</td>
<td>0.26</td>
<td>-0.87</td>
<td>0.27</td>
</tr>
<tr>
<td>16</td>
<td>1.18</td>
<td>0.27</td>
<td>0.61</td>
<td>0.25</td>
</tr>
<tr>
<td>17</td>
<td>0.92</td>
<td>0.29</td>
<td>1.63</td>
<td>0.25</td>
</tr>
<tr>
<td>18</td>
<td>-0.97</td>
<td>0.26</td>
<td>-1.52</td>
<td>0.31</td>
</tr>
<tr>
<td>19</td>
<td>0.32</td>
<td>0.23</td>
<td>-0.2</td>
<td>0.25</td>
</tr>
<tr>
<td>20</td>
<td>0.06</td>
<td>0.23</td>
<td>-0.72</td>
<td>0.26</td>
</tr>
<tr>
<td>21</td>
<td>-0.97</td>
<td>0.26</td>
<td>-1.17</td>
<td>0.28</td>
</tr>
<tr>
<td>22</td>
<td>0.06</td>
<td>0.23</td>
<td>-0.08</td>
<td>0.25</td>
</tr>
</tbody>
</table>

**3.5.7 Reliability**

Rasch person and item reliabilities reflect internal consistency reliability. Item reliability values <0.9 suggest that the test-taker sample is not large enough to confirm the apparent item-difficulty structure. Person reliability (separation) values <0.8 suggest that the items are insufficient for precisely and reproducibly distinguishing among the apparent abilities of test takers. Such values may also suggest that the Rasch person measure score (or how well each person performed based on the Rasch ratio-score model) may not be a reliable reflection of person ability (Linacre 2017b).

The overall item reliability value for the GeDI was 0.97 and the overall person reliability was 0.62. (As a point of comparison, Cronbach's alpha for this administration was 0.65.) The high Rasch item reliability value indicates that the student sample in this study is sufficient to support the item difficulty and item fit values. Low person reliability scores are commonly
associated with a narrow range of test-taker ability or an instrument with few items or few options for each item (which consequently elicits less varied responses than an instrument with many items and many answer options). In both cases, lack of variance in responses translates to fewer increments among which to delineate test taker ability (Linacre 2017b). This interpretation seems appropriate given the moderate number of items in the GeDI, the dichotomous response options, the presence of three items of redundant difficulty on the Wright map, the instrument’s failure to distinguish among the top 12.5% (n=42) of test takers in this administration, and the bulk of test takers clustered between -1 and 1 on the logit scale (Figure 3.1).

**Figure 3.2** Position effects on GeDI item difficulty values.
GeDI items appear on the X axis and mean item measures for each GeDI item are plotted on the Y axis. Symbols denote vignette position (1, 2, 3, or 4) in a counterbalanced rotation sequence of vignettes and accompanying item suites (see methods section for description). Error bars represent two standard errors of measurement about each mean item measure. As an example, item 4 showed little variation in item measure regardless of presentation order, and was also easier overall than item 13. In contrast, item 13 showed a slightly larger variation in item measure by position. Overall, no substantial differences were found between item difficulty and item position when controlling for the number of tests.

### 3.5.8 Item order effects

Raw score group means for all four forms of the GeDI were very similar, ranging between 12.02 to 12.20 (SD 3.30-3.98) out of a possible 22 (Table 3.7). A one-way ANOVA confirmed that there was no statistically significant difference in Rasch-scaled scores for each of the four GeDI forms ($F[3,332]= 0.038$, $p = 0.99$). This result indicates that the order of vignettes did not impact overall performance. Comparisons of mean item measures for the first, second, third, or fourth rotation position showed no apparent differences in item difficulty when controlling for the number of statistical tests (Figure 3.2). Detailed information on item measures for all items and rotation positions is available in the Table 3.6.
Table 3.7 Comparison of performance on GeDI by form, course, and region

<table>
<thead>
<tr>
<th>Course, Region (number tested)</th>
<th>Mean of items correct (SD)</th>
</tr>
</thead>
<tbody>
<tr>
<td>300-level genetics, Northeast (N=336)(^b)</td>
<td>12.11 (3.59)</td>
</tr>
<tr>
<td>form 1 (n=81)(^b)</td>
<td>12.02 (3.30)</td>
</tr>
<tr>
<td>form 2 (n= 78)(^b)</td>
<td>12.15 (3.61)</td>
</tr>
<tr>
<td>form 3 (n= 80)(^b)</td>
<td>12.20 (3.98)</td>
</tr>
<tr>
<td>form 4 (n=87)(^b)</td>
<td>12.09 (3.54)</td>
</tr>
<tr>
<td>300-level genetics, Southeast (N=318)(^c)</td>
<td>12.35 (3.29)</td>
</tr>
<tr>
<td>300-level genetics, Midwest (N=141)(^c)</td>
<td>11.94 (3.35)</td>
</tr>
<tr>
<td>300-level cell biology, Northwest (N=51)(^c)</td>
<td>13.35 (3.64)</td>
</tr>
<tr>
<td>300-level evolution, Northwest (N=91)(^c)</td>
<td>14.47 (3.78)</td>
</tr>
<tr>
<td>400-level evolution, Midwest (N=60)(^c)</td>
<td>16.66 (3.44)</td>
</tr>
</tbody>
</table>

\(^a\) Raw scores were used for comparison as Rasch-scaled data were not available from prior studies. \(^b\) Denotes present study. \(^c\) Denotes data from Price et al. 2014. All institutions were Doctoral-granting. Maximum number of correct items is 22.

Table 3.8 GeDI item difficulty rank in initial and present administrations

<table>
<thead>
<tr>
<th>Items addressing key concepts</th>
<th>Price et al. 2014</th>
<th>Present study</th>
</tr>
</thead>
<tbody>
<tr>
<td>16</td>
<td>16</td>
<td>8</td>
</tr>
<tr>
<td>13</td>
<td>13</td>
<td>6</td>
</tr>
<tr>
<td>10</td>
<td>3</td>
<td>12</td>
</tr>
<tr>
<td>4</td>
<td>10</td>
<td>17</td>
</tr>
<tr>
<td>5</td>
<td>15</td>
<td>5</td>
</tr>
<tr>
<td>3</td>
<td>4</td>
<td>14</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>11</td>
<td>19</td>
</tr>
<tr>
<td></td>
<td>18</td>
<td>22</td>
</tr>
<tr>
<td></td>
<td>20</td>
<td>20</td>
</tr>
<tr>
<td></td>
<td>9</td>
<td>9</td>
</tr>
<tr>
<td></td>
<td>7</td>
<td>7</td>
</tr>
<tr>
<td></td>
<td>19</td>
<td>11</td>
</tr>
<tr>
<td></td>
<td>22</td>
<td>21</td>
</tr>
<tr>
<td></td>
<td>21</td>
<td>18</td>
</tr>
</tbody>
</table>

Items listed from most challenging items (top) to least challenging items (bottom). Difficulty rank based on CTT difficulty (\(P\)) values for initial\(^a\) study and Rasch item measures for present study.

3.5.9 Comparisons with other undergraduate participant samples

Given that evolution acceptance, religion, and demographic variables differ across the United States, it is important to determine if instrument properties generalize. GeDI scores from our sample of undergraduates from the Northeastern United States were nicely aligned with the scores obtained by Price et al. (2014) from similar courses from other regions of the country (Table 3.7). In particular, no significant difference was found between raw scores from the 300-level genetics class in our sample (M= 12.35, SD = 3.59) and those of 300-level genetics classes in the Midwest (M=11.94, SD=3.35; t(475)=0.481, p=0.631) or the Southeast (M=12.35,
SD=3.29; t(652)=0.890, p=0.374). Similarity in scores across institutions indicates that the GeDI is functioning to elicit similar responses in comparable populations across the country and may suggest generalizability of score inferences (cf. Messick 1995). It should be noted that no Rasch-scaled scores are available from prior GeDI administrations so comparisons are limited to raw scores. Raw score similarity also provides limited evidence that Rasch-based validity measures obtained with our population may generalize to the GeDI as a whole, though this should be confirmed in future studies.

To examine whether individual items functioned similarly across administrations, item difficulty rank from our sample was compared to CTT-based item difficulty (P) rank from Price et al. 2014 (Table 3.8). Overall, most items maintained a similar or only slightly shifted difficulty order, though a few notable differences in item difficulty across administrations were found. Among items targeting key concepts, the hardest and easiest items maintained the same difficulty position and mid-level items showed only minor rearrangement. Item 3 (relating to a loss of variation associated with genetic drift), initially ranked as an easy item by Price et al. 2014, ranked among the more difficult key concept items in our administration. Among items targeting misconceptions, many items maintained a similar difficulty ranking, but items 14, 19 and 22 (all of which addressed “genetic drift is random mutation”) were notably more difficult in the present administration. Items 11 and 18 (addressing “genetic drift is gene flow”) were notably easier in the present administration. Overall, most of the GeDI questions maintained the general difficulty hierarchy across diverse samples, which may be used as evidence in support of generalization validity. Difficulty rank differences in items 3, 11, 14, 18, 19 and 22 should be investigated further.

3.6 Discussion

3.6.1 Insights about instrument quality from Rasch modelling

Rasch and IRT afford a more comprehensive and rigorous evaluation of instrument quality compared to CTT approaches (Boone et al. 2014). The present study has generated further evidence in support of the GeDI’s use as an instrument capable of generating valid and reliable inferences about upper-level undergraduates’ knowledge of genetic drift in American samples. The GeDI was found to be unidimensional, with item response patterns consistent with Rasch model expectations. The difficulty levels of items on the GeDI were generally well-calibrated for upper division students, with the exception of the highest scorers, for whom challenging items were lacking.

Rasch analysis is useful to help a test developer to improve test quality because it can provide information on how items function individually and as a whole. While the GeDI overall functioned very well within the IRT framework and Rasch modeling, we offer a few recommendations that may further improve the quality of measurement from a psychometric perspective. Item 1 was not difficult enough to differentiate students in this sample. Price et al.’s (2014) CTT analysis produced similar findings for item 1, however, they retained this item to satisfy validation criteria for earlier GeDI drafts. Future versions might revise, replace, or remove item 1. Further investigation is also needed to determine whether items 12 and 17, which test the same misconception at the same difficulty level, should both be retained in their present form or perhaps removed or revised. From an empirical perspective, inclusion of additional high-difficulty items or perhaps adjustment of a few current items would be beneficial to target the
highest-ability test-takers and would likely improve person reliability scores. Of course, any decision about test design must balance consideration of both empirical properties and theoretical concerns such as construct representation, so adjustment of items to improve psychometric properties is only appropriate if it continues to satisfy content validity criteria. Developers specified that the GeDI targets what they refer to as “stage 2” (mid-level) misconceptions, wherein drift is conflated with other evolutionary mechanisms. The true/false format of the GeDI precluded assessment of more nuanced “stage 3” (advanced-knowledge level) misconceptions, characterized by inappropriate constraints on the situations in which drift may occur. Further exploration is necessary to determine whether the GeDI might be modified to better measure the small group of high performers or whether the observed response pattern indeed represents the successful mastery of the upper bounds of the intended construct. As is always the case, any modifications of the existing instrument would require additional validation studies (cf. Table 3.3). Beyond these concerns about item difficulty, all items functioned appropriately in all other aspects of the analysis.

3.6.2 Impact of presentation order of instrument vignettes (and associated item suites) on measures of student understanding

The GeDI features four vignettes differing in taxonomic context and item difficulty, two factors which have been associated with item position effects in studies with other instruments (cf. Federer et al. 2014). Rearranging the order of GeDI vignettes and associated item suites has no significant impact on test scores, thus each vignette is functioning independently to assess student knowledge and does not appear to be impacting responses to subsequent items. Almost no other concept inventories in biology education have been tested for order effects.

3.6.3 Generalization validity

The ability of the GeDI to generate comparable scores and fairly similar item difficulty rank patterns among academically similar students from diverse institutions from different geographic regions could be used as a source of evidence in support of claims of generalization validity (AERA, APA, NCME 2014; Mesick 1995). The addition of evidence from a Northeast population is particularly important because evolution acceptance and associated factors vary widely across different U.S. geographic regions (which differ in religion and political party affiliations; see www.pewresearch.org).

3.6.4 Genetic drift, natural selection, and their interrelationships

Empirical studies on teaching, learning, and assessing non-adaptive contributors to evolution have been scarce in a vast body of evolution education research dominated by studies on natural selection (Andrews et al. 2012; Beggrow & Nehm 2012; Price & Perez 2016). How students conceptualize genetic drift and how genetic drift fits into the broader conceptual ecology of evolutionary thought are two areas that have only recently begun to be explored. Current research indicates that student thinking about genetic drift and understanding of genetic drift are both typically secondary to—and independent of—understanding of adaptive evolutionary change (Beggrow & Nehm 2012; Andrews et al. 2012). Students appear to conceptualize non-
adaptive mechanisms as *alternatives* to natural selection rather than co-occurring processes (Beggrow & Nehm 2012). When openly prompted to describe mechanisms for evolutionary change, students rarely suggest genetic drift (Beggrow & Nehm 2012), and, when specifically prompted to write about drift, many students still struggle to identify or explain drift (Andrews et al. 2012). Studying these responses, Andrews et al. (2012) developed a hypothetical framework describing how genetic drift conceptual development might progress: They suggest students may shift from (1) naive and limited conceptions of evolution and genetics to (2) a state where students are aware of various evolutionary processes (e.g., genetic drift) but still unclear on the differences between them, to (3) a state where students may distinguish between different evolutionary processes (e.g., genetic drift) but the new knowledge is still marked with inaccuracies specific to each process. Later, Price et al. (2016) noted that students developing expertise may exhibit elements of stage 2 and stage 3 conceptions simultaneously. Specifically, students with mid-level expertise in genetic drift often confuse drift with normative or non-normative ideas about natural selection or other evolutionary events such as bottlenecks/population boundaries, random mutation, migration/gene flow, or speciation in general (Andrews et al. 2012; Beggrow & Nehm 2012). In contrast, students with more advanced ideas about drift tend to place inaccurate constraints on the situations under which drift occurs (Andrews et al. 2012).

The development of the GeDI to target stage 2 (mid-level) misconceptions about genetic drift is an important addition to the body of evolution measurement tools because it both gauges understanding of a previously neglected evolutionary mechanism and holds potential to capture some simultaneous reasoning about natural selection (as it relates to drift). Given the incoherence of naive student thought about evolution (inappropriately both conflating and failing to recognize simultaneous adaptive and non-adaptive processes), attention toward developing instruments that can simultaneously capture thought on adaptive and nonadaptive mechanisms is warranted. Few instruments are capable of simultaneously eliciting thought about natural selection and genetic drift, and none were designed with the intent to robustly measure knowledge of both processes (Table 3.1). More fully capturing the array of student thought about diverse evolutionary mechanisms, including how thoughts on diverse mechanisms intersect, will better equip educators to develop appropriate instructional strategies and develop curricula.

This work has provided evidence in support of validity inferences for the GeDI using contemporary instrument evaluation methods, and identified a few areas that would improve measurement quality. These findings are significant given the very limited set of assessment tools available for exploring student understanding of non-adaptive processes.

3.7 Limitations and further research

A limitation to the analysis of the effects of item position on student performance was that the sample size for each test form was limited (n = 78-87); larger samples would afford more robust conclusions about possible item order effects (Linacre 1994). Specifically, more replicates generate more precise and stable item measures and increased statistical power to reduce the chance of a type II error. Further, we did not investigate whether possible item order effects might exist within question suites sharing a common vignette; our primary concern was whether vignette presentation order impacted responses to subsequent vignettes.

Although our study adds additional evidence in support of the validity and reliability of the inferences generated by GeDI scores, further work in line with the measurement *Standards* is
needed (AERA, APA, NCME 2014; Mesick 1995). For instance, although surface feature effects have been well-documented in evolution assessment (e.g., Federer et al. 2016; Nehm et al. 2012; Nehm and Ha 2011; Nehm 2018; Opfer et al. 2012), such effects have yet to be examined for the GeDI. Future work might also investigate how the GeDI functions when data are disaggregated by gender, ethnicity, or other demographic factors (cf. Federer, Nehm & Pearl 2016; Schmiemann et al. 2017). Additionally, because all of the GeDI’s items offer dichotomous answer choices, the impact of guessing bears more significantly on inferences about understanding than on a traditional multiple choice instrument. Thus, an exploration of the extent to which guessing impacts inferences generated by the GeDI would be a worthwhile step. Such an investigation might consider how the instrument functions if item responses were to be moderated by a paired question tier to indicate student confidence in their responses (cf. Romine, Schaffer & Barrow 2015) or examined for guessing using Rasch or IRT (e.g., Andrich, Marais & Humphry 2012; Boone et al. 2014; Gershon 1992; Linacre 2017b). Overall, while the GeDI now stands among the more robustly evaluated evolution instruments, additional work remains to comprehensively characterize the validity and reliability of inferences generated by this (and many other) evolution education instrument(s). Attention should also be given to whether the array of measurement instruments available can adequately gauge normative and non-normative ideas about adaptive and nonadaptive evolution.

3.8 Conclusions

Validity evidence for the vast majority of instruments in biology education is based on CTT, and most instruments have only one form of validity evidence (i.e., content validity) (e.g., Campbell & Nehm 2013). The evolution education research community must place greater emphasis on the analysis of ratio-scaled data and expand its efforts to include studies of a more diverse array of forms of validity evidence to support the inferences derived from assessment scores (cf. AERA, APA, NCME 2014). The present study provides further evidence that the inferences derived from the GeDI are valid indicators of student understanding while identifying areas of improvement. The methodological approach we introduced provides a template for future studies of other evolution instruments that were validated using CTT methods.

3.8 References (Chapter 3)


Chapter 4. Testing the impact of surface features on students’ Mendelian genetics understanding

4.1 Introduction

A substantial literature in cognitive psychology has produced clear and convincing evidence that assessment task features -- such as the framing, context, or situation in which problems are posed -- can impact the retrieval of scientific knowledge and resulting measures of participants’ conceptual understanding (e.g., Chi, Feltovich, & Glaser 1981). Studies of student reasoning in many science domains, including chemistry, physics, earth science, and biology have demonstrated the ubiquity of what have been termed assessment “context effects” or item surface features (reviewed in Nehm & Ha 2011). Understanding of how task cover stories, contexts, or situational features impact reasoning is important for many educational activities, including the selection of curricular examples to illustrate scientific phenomena, the design of formative and summative assessment items, and determination of whether instruction has fostered the development of abstract schemas divorced from particular instances (Mayer 2013). Although the recognition of context effects on scientific reasoning has been widespread following Chi et al.’s (1981) seminal study, remarkably few experimental studies have been performed that clarify which contexts meaningfully impact measures of particular types of problem solving in relation to scientific ideas. Indeed, research needs to move away from the general documentation of context effects and towards domain-specific models that may be leveraged to improve teaching and learning of core ideas (see Nehm & Ha 2011).

In biology education, studies of the impact of situational features on student reasoning have been most extensively explored for the concept of natural selection (e.g., Ha & Nehm 2014; Settlage 1994). Nehm and colleagues used large samples of American participants’ constructed-response answers to carefully manipulated items to show that the measurement of student understanding was significantly and meaningfully impacted by the taxon included in the item (e.g., plant, animal, human), by the scale of evolutionary comparisons (e.g., within- vs. between species), by the familiarity of the taxa and traits (e.g., penguin vs. prosimian), and by the polarity of evolutionary change (e.g., the gain or loss of a trait) (Federer, Nehm, & Pearl 2016; Nehm, Beggrow, Opfer, & Ha 2012; Nehm & Ha 2011; Nehm & Reilly 2007; Nehm & Ridgway 2011; Opfer, Nehm, & Ha 2012). Some of these situational effects have also been documented in international participants subjected to different educational experiences and cultural contexts (e.g., Ha & Nehm 2014). This body of work shows that novice participants’ evolutionary reasoning is strongly influenced by situational features, and that as expertise in the domain of evolution increases, the impact of situational features on problem solving decreases (Nehm & Ridgway 2011). Against this background it seems reasonable to investigate the impact of situational features on students’ reasoning and performance measures for diverse problems and particular (sub-)domains.

4.1.1 Genetics education

Although a large body of work has explored student difficulties with genetics problem solving (e.g., Collins & Stewart 1989; Shea, Duncan, & Stephenson 2015; Smith 1983; Todd & Romine 2016), much less work has explored the roles that situational features or contexts play in the measurement of genetics understanding. An important early study on elementary school
children by Kargbo, Hobbs, and Erickson (1980) found that while students held intuitive notions that environmentally-acquired characteristics could be transmitted to offspring, the belief did not transfer uniformly across questions featuring representatives from familiar plant and animal taxa. Humans were presumed to inherit acquired traits more often than dogs, and trees were rarely presumed to inherit such traits. This work was an early indication that situational features could bias genetic reasoning processes.

In a longitudinal study of the consistency of twelve through sixteen year olds’ conceptions about inheritance of acquired characteristics and other scientific phenomena, Clough and Driver (1986) found that task context was most significant for students who had yet to develop normative scientific understanding. Interviews of students with lower knowledge levels revealed conceptions about acquired characteristics that displayed less consistency across parallel tasks (e.g., taillessness in mice, athletic ability in humans, and rough skin caused by gardening in humans) compared to interviews of students with higher knowledge levels. Clough and Driver concluded that students have multiple alternative frameworks which are employed to varying degrees depending on the context of the question. The authors went on to suggest the “hopeful finding …that once students learn and use a correct scientific explanation in one context, they are more likely to employ it in others” (Clough & Driver 1986: 489).

More recently, Ware and Gelman (2014) examined the degree to which animals’ phenotypic trait properties impacted student reasoning about inheritance. Specifically, inheritance prompts were manipulated to highlight the functional properties (function-predictive, e.g., “She uses her sharp claws to catch fish”) or habitat-relevant properties (habitat-predictive, e.g., “Animals with bumpy skin live in the desert”) relative to a null condition (“non-predictive”). Their empirical work showed that undergraduates believed that it was possible for an animal to acquire a physical property in its lifetime provided that it had a useful function or was a good fit with environmental conditions (Ware & Gelman 2014: 234). Like Kargbo et al.’s (1980) and Clough & Driver’s (1986) studies, student ideas about inheritance were impacted by item features, although in this case aspects of animal trait functions.

In a study of middle school students, Freidenreich, Duncan, and Shea (2011) found that participants offered more robust genetic explanations for tasks using human examples compared to those using plants and bacteria. Shea, Duncan, and Stephenson (2015) also found that situational features play a significant role in some aspects of undergraduate participants’ genetics reasoning. Specifically, a problem featuring human albinism elicited higher quality arguments among early career biology majors than an equivalent task featuring genetically modified corn, despite similar knowledge use across both tasks. Based on their findings, Shea et al. (p. 4) argued that “Expanding the definition of genetics literacy to include the role of situational features is critical, as the research literature suggests that [an] [individual’s] ability to generate and support arguments about authentic genetics issues relies on their capacity to consider how issues are framed by unique situational features.” The question remains as to which situational features impact particular types of genetics problems, and how these features impact measures of student learning. One possible starting point is Mendelian genetics.

4.1.2 Mendelian genetics

Although in recent years science education researchers have reconceptualized genetics literacy, Mendelian transmission genetics remains a central component of biology education (criticisms notwithstanding; see Smith & Gericke 2015). Stewart, Cartier, and Passmore (2005),
for example, developed a three-part model comprising genetic (e.g., classical, Mendelian, or transmission genetics), meiotic (e.g., processes relating to the production of gametes), and molecular (e.g., gene expression) understanding. Stewart’s model more recently was refined to encompass a broader range of more carefully delineated genetic constructs and was situated within a learning progression framework (see, for example, Duncan, Rogat and Yarden 2009; Todd & Romine 2016). Despite several conceptual reorganizations, classical transmission genetic problem solving has been retained within these new frameworks, although reformulated to some extent (Todd & Romine 2016, p. 1678).

With improved understanding of genomics and molecular genetics, the limitations of traditional inheritance-centered approaches to genetics education have become clearer. Increasing consideration from both theoretical and empirical perspectives has been given to the proper role of Mendelian genetics in the curriculum and as a component of genetics literacy (e.g. Duncan et al. 2009; Jamieson & Radick 2013; Smith & Gericke 2015; Todd & Romine 2016). From a biological perspective, it has been long recognized that attempts to characterize complex traits within a Mendelian framework are insufficient; even acrobatic adaptations of Mendelian “rules” in cases such as pleiotropy, epistasis, multiple alleles, and incomplete penetrance fail to explain the observed phenotypic patterns of many traits. Indeed, most human traits are multifactorial and can only be fully explained by addressing the molecular link between genotype and phenotype. Key molecular considerations beyond the scope of Mendelian genetics are the roles of variation in genetic code and protein structure, the interaction of genes and gene products (including RNA) with other gene products and the environment, the various mediators of gene expression, and epigenetics (Dougherty et al. 2011; Jamieson & Radick 2013; Smith & Gericke 2015). While these have been included to varying extents within the molecular model of genetics (Stewart et al. 2005; Duncan et al. 2009; Todd & Romine 2016), students often struggle to draw connections between Mendelian and molecular aspects of genetics (Lewis & Kattman 2004; Todd & Romine 2016). From a pedagogical perspective, Mendelian-centric presentations of genetics have been associated with fueling students’ tendency to accept the simplest explanation over more accurate but complex explanations for trait variant (Dougherty et al. 2011; Gericke et al. 2014; Jamieson & Radick 2013), feeding into existing confusion about the concept of dominance (Allchin 2005; Jamieson & Radick 2013), and promoting exaggerated deterministic perspectives on inheritance which can contribute to related social extensions of deterministic views (Gericke et al. 2014; Castéra & Clément 2014; Castéra et al. 2008; Jamieson & Radick 2013, 2017). Together, these concerns underpin the importance of research toward clearly delineating the place of Mendelian genetics within learning progressions and associated curricula. Careful study of how students respond to genetics problems situated in different contexts will further help to refine educators’ understanding of thinking and learning in genetics toward that end.

New assessments developed for genetics learning progressions (e.g., Duncan et al. 2009) and genetics learning in undergraduate settings (e.g., Bowling et al. 2008) continue to include items that fall under the umbrella of “Mendelian transmission genetics” (see Table 4.1). These assessments are variable in terms of the contexts or situational features that are used to measure student understanding, and disproportionately use animal (including human) contexts. Given the relative stability of “Mendelian transmission” questions in historical and contemporary educational research on genetics learning, and the continued use of assessments that differ in situational features (see Table 4.1), this study focused on the role of situational features on Mendelian problem solving performance.
<table>
<thead>
<tr>
<th>Instrument</th>
<th>Target population</th>
<th>Number of items</th>
<th>Taxonomic context</th>
</tr>
</thead>
<tbody>
<tr>
<td>Written Test of Argumentation in Genetics Dilemmas</td>
<td>Secondary (Grade 9)</td>
<td>3</td>
<td>Human</td>
</tr>
<tr>
<td>(Zohar and Nemet 2002)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Test of Basic Genetics Concepts</td>
<td>Undergraduate</td>
<td>7</td>
<td>Human and unspecified(^b)</td>
</tr>
<tr>
<td>(Sadler 2003; Sadler and Zeidler 2005)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Genetics Concept Inventory</td>
<td>Undergraduate</td>
<td>4</td>
<td>Unspecified(^b)</td>
</tr>
<tr>
<td>(Elrod 2007)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Genetics Literacy Assessment(^d)</td>
<td>Undergraduate</td>
<td>3</td>
<td>Human</td>
</tr>
<tr>
<td>(Bowling et al. 2008)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Genetics Concept Assessment(^d)</td>
<td>Undergraduate</td>
<td>8</td>
<td>Human</td>
</tr>
<tr>
<td>(Smith, Wood, &amp; Knight 2008)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Modern Genetics Learning Progression</td>
<td>Upper elementary - secondary (Grades 5-10)</td>
<td>Included(^a)</td>
<td>Not applicable(^a)</td>
</tr>
<tr>
<td>(Duncan, Rogat &amp; Yarden 2009)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Genetics Diagnostic Instrument(^d)</td>
<td>Secondary (Grades 10 &amp; 12)</td>
<td>8</td>
<td>Human, animal, and unspecified(^b)</td>
</tr>
<tr>
<td>(Tsui and Treagust 2010)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Biology Concept Inventory(^d)</td>
<td>Undergraduate</td>
<td>5</td>
<td>Human and unspecified(^b)</td>
</tr>
<tr>
<td>(Klymkowsky, Underwood, &amp; Garvin-Doxas 2010)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Molecular Biology Capstone Assessment(^d)</td>
<td>Undergraduate</td>
<td>1</td>
<td>Human</td>
</tr>
<tr>
<td>(Couch, Wood, &amp; Knight 2015)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Learning Progression-based Assessment of Modern Genetics- Version 2(^e)</td>
<td>Undergraduate</td>
<td>6</td>
<td>Human and plant</td>
</tr>
<tr>
<td>(Todd and Romine 2016)</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

\(^a\) Learning progression structure features components of “big ideas” rather than individual items.
\(^b\) The denotation “unspecified” indicates that item(s) tested knowledge of transmission genetics outside of the context of a particular taxon.  
\(^c\) Although all item types were identified, exemplars were provided for odd items only.
\(^d\) Instrument evaluated for validity/reliability (Classical Test Theory framework) 
\(^e\) Instrument evaluated for validity/reliability (Item Response Theory/Rasch framework); see Chapter 2 for further discussion of instrument evaluation.
4.1.3 Genetics Problem types

The field of genetics problem solving research has employed a variety of problem structures and types. Monohybrid crosses featuring simple dominance and, to a lesser degree, incomplete or codominance, have been used widely in genetics problem-solving research (e.g., Browning & Lehman 1988; Cavallo 1994; Corbett, Kauffman, Maclaren, Wagner, & Jones 2010; Gipson, Abraham & Renner 1989; Moll & Allen 1987; Simmons & Lunetta 1993; Slack & Stewart 1990; Smith & Good 1984; Stewart 1983). Simple dominance and codominance represent two of the four types or “classes” of genetics problems (simple dominance, codominance, sex linkage and multiple alleles) identified by Collins and Stewart (1989) in their categorization of Mendelian genetics knowledge structure. It should be noted that the distinction between incomplete dominance and codominance can be blurry, and, since the transmission pattern is the same, they are often considered together. Tasks involving sex linkage and multiple alleles are considered to be more complex and have been employed less frequently in education research.

Stewart (1988) also classified genetics problems according to whether they require the more commonly used “cause to effect” reasoning or the more cognitively demanding “effect to cause” reasoning. Hickey (2000) and Tsui and Treagust (2010) expanded Stewart’s categorization into six types of genetics problems. These authors proposed that Stewart’s reasoning types (plus a third type, process reasoning, not relevant to the present study) constitute a domain-general thought dimension which intersects with the domain-specific dimension of within-generation (simpler) or between-generation (complex) thought. In line with this theoretical perspective on the construct, this study includes problems testing knowledge of transmission between generations (which subsumes knowledge of the simpler within-generation mechanisms) with both “cause to effect” and “effect to cause” examples.

Although the United States lacks a national science curriculum, Mendelian transmission genetics and associated genetic crosses are a commonly encountered topic and problem type from upper elementary through undergraduate classrooms. The subject is included in (1) the K-12 science education standards (National Research Council 1996, 2012), (2) the Next Generation Science Standards (NGSS 2013), (3) the American Society for Human Genetics recommended content for the collegiate level (Hott et al. 2002), and (4) nearly all college biology textbooks (e.g., Hott et al. 2002; McElhinny, Dougherty, Bowling, & Libarkin 2014). In addition, biology teacher certification exams, such as Praxis (ETS 2015), include items on Mendelian inheritance, and genetic crosses remain in the most recent versions of high school Advanced Placement Biology (College Board 2015) and International Baccalaureate Biology curricula (International Baccalaureate Organization 2014). In sum, transmission genetics is a core aspect of genetics learning in the United States.

4.1.4 Demographic factors and genetics education

Many studies in genetics education have not explicitly considered the role of demographic factors (e.g., gender, race) in their research designs despite a half century of research demonstrating differences in attitudes, understanding, achievement, and participation (Eddy & Brownell 2016; Kahle & Meece 1994; Lee & Luykx 2007; Linn & Hyde 1989; Peng, Wright, & Hill 1995; Scantlebury 2014; Scantlebury & Baker 2007; Weinburgh 1995). Differences in science achievement may be due to factors specific to a demographic group.
(Peng, Wright, & Hill 1995; Scantlebury & Baker 2007) or as a result of bias in curriculum, instructional practices, school climate, or assessment methods (Lee & Luykx 2007).

In biology education, the roles of gender and ethnicity on domain-specific performance remain unsettled. Some studies, for example, have documented the absence of significant gender effects on biology performance (e.g., Dimitrov 1999; Huppert, Lomask & Lazarowitz 2002; Lauer et al. 2013; Shepardson & Pizzini 1994; Willoughby & Metz 2009). Dimitrov (1999) and Creech and Sweeder (2012) found no impact of ethnicity on biology performance, and Nehm and Schonfeld (2008) found similar types of alternative conceptions in underrepresented students as documented in other demographic groups. Other studies, in contrast, have found advantages for males in undergraduate biology course grades (Creech & Sweeder 2012) and test scores (Eddy, Brownell, & Wenderoth 2014; Stanger-Hall 2012, Wright et al. 2016), particularly on multiple-choice (Stanger-Hall 2012) and high difficulty (Wright et al. 2016) items. Other studies have found that females outperformed males on concept maps (Pearsall et al. 1997) and on tests of labeling errors (Soyibo 1999). Overall, gender and race/ethnicity have been shown to play significant roles in some studies and in some item formats, but not others (Federer et al. 2016).

Many studies in genetics education have failed to consider the roles that demographic factors might play on measures of performance and inferences about genetics learning challenges. An absence of gender effects was noted by Cavallo (1994) in high school participants’ written explanations of genetics and meiosis, and by Dogru-Atay and Tekkaya (2008) in eighth graders’ multiple-choice responses about inheritance and genetics crosses. However, Franke and Bogner (2011) showed a female advantage for retaining new conceptions about molecular genetics and genetics technology on a multiple-choice test. To ensure accurate measures of learning and appropriately designed curriculum and instructional methods, more information is needed about how different assessment methods measure understanding in the various branches of biology across all demographic groups. Because of the importance of gender and race/ethnicity to science education, and the paucity of work in genetics education in particular, this study disaggregates data by gender and ethnicity to examine any potential testing bias or performance discrepancies.

4.2 Research question

The study employs an experimental research design in order to investigate the degree to which situational features of genetics problems impact measures of student understanding. Using a suite of Mendelian inheritance problems about complete dominance and incomplete dominance, the following research question is posed: Do Mendelian inheritance problems that differ in taxon (animal, plant, human) or familiarity (real, fictitious) produce equivalent measures of student understanding in university participants across genders and ethnic backgrounds?

4.3 Methods

4.3.1 Item design

To understand whether Mendelian inheritance problems differing in taxa produce equivalent measures of understanding, this study quantified differences in item difficulty (dependent variable) for Mendelian inheritance problems featuring real and fictitious examples from different taxa such as animals, plants, and humans (independent variables). The universe of
possible situational features to choose from is quite large. Prior research to guided the choice of situational features (i.e., taxon: plant/animal/human, familiarity: real/fictitious). Many studies in cognitive developmental psychology have shown that plant/animal/human distinctions are a fundamental feature of early cognitive frameworks (so-called “naive biology”) and serve to organize biological reasoning in young children and many adults (reviewed in Opfer et al. 2012). The plant/animal/human distinctions have also been shown to be highly relevant to how children and adults think about biological processes such as evolution and genetics (Opfer et al. 2012; Shea et al. 2015). Thus, much work in psychology and education motivated the choice of taxon as a situational feature worthy of interest.

Controlling for familiarity using fictitious properties has been a central design feature of cognitive studies for decades, and has recently emerged as an important consideration in studies of biological reasoning (Opfer et al. 2012; Ware & Gelman 2014). In both genetics and evolution education, research has shown that “...reasoning deviates from accepted scientific ideas more so when considering novel categories” (Ware & Gelman 2014, p. 233). Therefore, fictitious taxa and traits were developed that would by definition be novel to participants, and real taxa were used which participants had been exposed to from their curricula. Given that taxa and familiarity have been shown to have strong influences on biological reasoning, these choices made sense as a starting point for this experimental work.

Using this framework, a core collection of five multiple-choice item types addressing the Mendelian inheritance mechanisms of complete dominance and incomplete dominance was developed. These topics were chosen given their (1) ubiquity in genetics education, and hence their relevance to educators worldwide (see Sections 4.1.2 and 4.1.3) and (2) presence in the enacted curriculum, ensuring that the sample had received basic instruction in the topic. This should help to make sure students have sufficient knowledge to solve the problems successfully and prevent statistical bottom effects. All items consisted of simple monohybrid crosses, resembling item types that are common in recent research instruments (cf. Table 4.1), earlier genetics research (e.g., Gipson, Abraham & Renner 1989; Kinnear 1983; Knippels 2005; Slack & Stewart 1990; Smith & Good 1984; Tolman 1982) and which continue to be popular in biology textbooks (Hott et al. 2002) and high-stakes international high school tests such as the SAT Subject Tests (College Board 2016), Advanced Placement Biology Exam (College Board 2015), and International Baccalaureate Biology Exam (International Baccalaureate Organization 2014). While these do not represent all possible Mendelian inheritance problem types, they are among the most widely used and form the basis for more complex genetics problems (Collins & Stewart 1989). In line with the Mendelian inheritance problem types described in Section 4.1.3 the items used represent both types of genetics reasoning described by Stewart (1998), the two more common types of dominance relationships described by Collins and Stewart (1989), and include knowledge of both between-generation and within generation reasoning as outlined by Hickey (2000) and Tsui and Treagust (2010).

Each item stem presented a particular taxon (e.g., pea plant), a particular trait (e.g., seed shape), and an inheritance pattern for that trait (e.g., round seed is dominant). The items then described a specific crossing experiment (e.g., homozygous pea plants with round and wrinkled seed shapes were crossed). Item tasks included predicting the phenotypic distribution of the first filial generation (F1) given information about the parental (P) genotypes, or predicting parental (P) genotypes given the phenotypes of first filial offspring (F1). Five multiple-choice options (1 attractor, 4 distractors) were given. An overview of the five types of items are given in Table 4.2.
The five item types were used as templates to generate alternate versions that differed only in the taxon featured (i.e. animal, plant, or human) and its corresponding trait (e.g., body color in fruit flies, seed shape in peas). To test for the impact of familiarity or prior knowledge, fictitious taxa and traits (e.g., fur color of “Amalcho” animals) were also included. To ensure participants correctly identified taxa as plants or animals—especially fictitious examples—item text included the words such as “plant” in all instances where plants were referred to (e.g., “pea plants” instead of “peas”). Each item also included a small picture of the “taxon”. By rotating different situational features among the core of five types of inheritance problems, a total of 81 items was generated: 35 featuring animals (16 fictitious), 34 featuring plants (16 fictitious), and 12 featuring humans. An example of an item altered to feature different taxa and their respective traits is shown in Table 4.3. Additional item information can be found in the Appendix.

Evidence of content validity was generated by four experts in biology education (university degrees in biology and biology education). They reviewed all item stems and answer options, and rated all items as appropriate to the domain of Mendelian genetics and correctly placed items within their expected problem categories (i.e. Table 4.2). Further validity evidence is discussed in Section 4.5.1. As items were initially developed for use with a similar study performed in Germany (see Schmiemann, Nehm, and Tornabene 2017 for more details), two bilingual (German/English) biology educators translated the original German items into English, and two American biology educators gave feedback on item content and checked the translation for grammatical clarity. The final English version was then reviewed by the bilingual German educators.

<table>
<thead>
<tr>
<th>Mendelian inheritance pattern</th>
<th>Information given</th>
<th>Question posed</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dominant-Recessive</td>
<td>Homozygous parental generation (P)</td>
<td>Distribution of first filial generation (F1)</td>
</tr>
<tr>
<td>Dominant-Recessive</td>
<td>Heterozygous parental generation (P)</td>
<td>Distribution of first filial generation (F1)</td>
</tr>
<tr>
<td>Dominant-Recessive</td>
<td>Distribution of first filial generation (F1)</td>
<td>Parental generation (P) genotypes</td>
</tr>
<tr>
<td>Incomplete dominance</td>
<td>Homozygous parental generation (P)</td>
<td>Distribution of first filial generation (F1)</td>
</tr>
<tr>
<td>Incomplete dominance</td>
<td>Heterozygous parental generation (P)</td>
<td>Distribution of first filial generation (F1)</td>
</tr>
</tbody>
</table>
Table 4.3  Example of variation in situational features for a Mendelian inheritance problem. The core problem remained the same while situational features were altered.

<table>
<thead>
<tr>
<th>Animal (fictitious)</th>
<th>Plant (real)</th>
<th>Human</th>
</tr>
</thead>
<tbody>
<tr>
<td>Amalchos can have black or white fur color. Fur color is inherited for amalchos in a dominant-recessive manner, where black fur color is dominant and white fur color recessive. Amalchos that have black fur color are crossed with amalchos that have white fur color. Both are homozygous regarding fur color. Which distribution is reflected in their offspring (F1 generation) with respect to fur color?</td>
<td>Corn plants can have smooth or wrinkled seed shape. Seed shape is inherited for corn plants in a dominant-recessive manner, where smooth seed shape is dominant and wrinkled seed shape recessive. Corn plants that have smooth seed shape are crossed with corn plants that have wrinkled seed shape. Both are homozygous regarding seed shape. Which distribution is reflected in their offspring (F1 generation) with respect to seed shape?</td>
<td>Humans can have a pointed or round hairline. Hairline is inherited for humans in a dominant-recessive manner, where pointed hairline is dominant and round hairline recessive. A human that has a pointed hairline is having children with a human that has a round hairline. Both are homozygous regarding hairline. Which distribution is reflected in their offspring (F1 generation) with respect to hairline?</td>
</tr>
<tr>
<td>· All descendants have black fur color</td>
<td>· All descendants have smooth seed shape.</td>
<td>· All descendants have pointed hairlines.</td>
</tr>
<tr>
<td>· All descendants have white fur color.</td>
<td>· All descendants have wrinkled seed shape.</td>
<td>· All descendants have round hairlines.</td>
</tr>
<tr>
<td>· The descendants have an approximate ratio of 1:1 black fur color to white fur color.</td>
<td>· The descendants have an approximate ratio of 1:1 smooth seed shape to wrinkled seed shape.</td>
<td>· The descendants have an approximate ratio of 1:1 pointed hairline to round hairline.</td>
</tr>
<tr>
<td>· The descendants have an approximate ratio of 3:1 black fur color to white fur color.</td>
<td>· The descendants have an approximate ratio of 3:1 smooth seed shape to wrinkled seed shape.</td>
<td>· The descendants have an approximate ratio of 3:1 pointed hairline to round hairline.</td>
</tr>
<tr>
<td>· The descendants have an approximate ratio of 3:1 white fur color to black fur color.</td>
<td>· The descendants have an approximate ratio of 3:1 smooth seed shape to wrinkled seed shape.</td>
<td>· The descendants have an approximate ratio of 3:1 round hairline to pointed hairline.</td>
</tr>
</tbody>
</table>

4.3.2 Test administration and participant samples

Item packets varying in situational features were distributed to large introductory biology class at a large public university in the Northeast. The sample was chosen because it represented students who have had prior exposure to Mendelian inheritance problems via relevant curricula and/or textbook content to ensure that the subjects had the potential to solve genetics problems.
An overlapping multi-matrix sampling was used to distribute the items among different test booklets (cf. Sirotnik & Wellington 1977). This allows for two important features of the study design: First, items constructed from the same template did not appear next to each other, ensuring that participants were required to think about each item separately. Second, not every student had to work on all 81 items, minimizing test fatigue, but maximizing study coverage.

All data collection qualified for IRB exemption status under IRB number 2013-2298-R1. Extra credit was offered for participation in the study, and participants were aware that their performance would not be reflected in their course grade. All participants were biology majors enrolled in the spring semester of an introductory biology course at a research-intensive public university in the northeastern United States. All participants had received genetics instruction (including Mendelian genetics) earlier in the semester. Participants differed in the amount of prior biology preparation, as is typical for American undergraduates. In order to control for background content preparation, the following categorizations were used: No college-level biology coursework other than the current course (24.7%), High School Advanced Placement biology only (10.5%), one college biology course (16.4%), two or more college biology courses (43%). No information about prior coursework was provided for 5.2% of the participants. Overall, 444 participants (62.4% female) with an average age of 19.6 years (SD=2.4) took part in the study. The sample included Asian (38.3%), White (34.0%), Hispanic (7.9%), African American (5.6%) and other (e.g., mixed background, 8.3%) participants. For 5.9% of the participants, no race/ethnicity data were available. One hour was provided to participants to complete the tasks, allowing all 81 items to be used across booklets. Eight test booklets containing 20-21 items were randomly assigned to these participants via Survey Monkey software. The software prevented participants from returning to prior questions once answers were submitted. Each item was worked on by an average of 114 participants (SD=20.5).

4.3.3 Rasch analysis

Conquest (Adams et al. 2016) was used to analyze data according to the Rasch model. Rasch modeling is ideal for educational measurement because it converts ordinal data into linear data and provides item and person measures as “logit” scores on the same equal-interval ratio scale. Generating item and person measures on the same scale makes it possible to determine the probability that a particular person could solve a particular item. In this analysis, a person had a fifty percent chance of solving a particular item if that particular item measure is equal to the person measure (Bond & Fox 2007, p. 38). Rasch modeling can also accommodate “missing” data, which is essential in multi-matrix designs in which participants are assigned only a subset of items from the total collection of questions. Such designs allow testing of a wider variety of items while minimizing participant test fatigue.

Item parameters and person abilities were estimated using the 1PL model. Therefore, correct answers were coded as one and incorrect answers (including skipped items and items with more than one option chosen) as zero using the key command of Conquest. Consideration of how well the empirical data fit the statistical Rasch model is one approach for evaluating the quality of the test items, the test instrument, and overall evidence in support of validity claims (Boone, Staver, & Yale 2014). Therefore, item fit statistics were examined (Wright 1984). Fit statistics indicate how well the empirical data meet the model requirement using a chi-square test (Bond & Fox 2007, p. 238). Fit is expressed as weighted (“infit”) or unweighted (“outfit”) values for the mean square parameter (MNSQ). For a standard multiple choice assessment, MNSQ
values above 1.3 are considered to be “underfitting”, indicating that the response pattern for that item is erratic. Values below 0.7 are considered to be “overfitting”, indicating that the response pattern is overly predictable. Both overfit and underfit suggest that the item is not functioning properly (i.e. eliciting information consistent with test-taker ability). Cut-off values of 0.7 and 1.3 for the MNSQ parameter were used to ensure an adequate match between the empirical data and the statistical model (Boone et al. 2014; Bond & Fox 2007). Further indicators of test quality include item and person reliability measures which can be interpreted similarly to Cronbach’s alpha in classical test theory (cf. Wright & Stone 1979).

4.3.4 Comparison of item difficulties

To compare item difficulties for the different question types and situational features, classical statistics and box plots of Rasch scores were used. The Kruskal-Wallis test (Kruskal & Wallis, 1952) and Mann-Whitney U test (Mann & Whitney, 1947) were used to test for significant differences in item parameter (dependent variable) for different groups of items (independent variables: animal vs. plant vs. human; fictitious vs. real). These non-parametric tests were chosen due to the small number of items in each group. ANOVAs and Pearson correlation tests of Rasch scores were used to analyze student performance by demographic group, gender, and associated contextual variables.

4.4 Results

4.4.1 Rasch model fit

Overall, the data showed good fit to the Rasch model. However, the initial analysis revealed four items with poor fit: three with so-called underfit (wMNSQ >1.3) and one with overfit (wMNSQ < 0.7). Therefore, those items were removed from further analysis. The final estimation, and all further analyses, were based on the set of 77 items demonstrating acceptable fit values (final deviance 7050.55). Appendix A contains a detailed report of all item fit statistics. An average number of 114 participants (SD=20.5) worked on each item. Warms Likelihood Estimates (WLE) were used as person measures (Warm 1989). The overall item reliability (WLE reliability = 0.726, EAP/PV reliability = 0.823) and the separation reliability (0.877) were robust. Similar to Cronbach’s alpha in classical test theory, item reliability and item separation are reported on a 0-1 scale and reflect internal consistency of the item set. Acceptable item reliability and separation reliability indicate that the items functioned together to hierarchically differentiate the measured trait into sufficient levels in a manner that can be replicated in comparable samples. This is underpinned by an average item-total correlation of 0.59 for the items. The test variance was very high (4.280), indicating there is a broad range of answer patterns.

A Wright map, or person-item map, may be used to compare how well-matched item difficulty is to person ability on the same logit scale (Figure 4.1). Items are represented by their item number on the right side of the scale (see the Appendix A for item details). Mean item difficulty is set at 0 logits; higher logit scores indicate more difficult items and lower (negative) logit scores indicate easier items. Persons are represented by “X’s” on the left side of the scale and are plotted so that each person has a 50% probability of correctly answering an item with an equivalent measure.
Figure 4.1 A Wright map derived from Rasch analysis of Mendelian inheritance item responses. The distribution of persons (on the left) and GeDI items (on the right) are illustrated on the same log interval (-3 to 8) scale. Each X=0.6 cases. For detailed information about the items see the Appendix.

The Wright map in Figure 4.1 displays acceptable item distribution for the university student sample, as the spread of person ability spans the difficulty of the items. The distribution of more than half of the persons with logit scores above the most difficult item indicates that the items were easy for this sample, which is reflected in the average person ability of 2.04 logits and a percentage of right answers per item between 54.4% and 96.1%. As there is a normal distribution of item difficulty and no ceiling effect, one can assume that the variance is not restricted artificially. Hence a further analysis of the item difficulties (Min = -2.29 logits, Max = +1.68 logits) seems to be reasonable (Section 4.5.2).

4.4.2 Item difficulties

As expected, there were no significant differences in student performance among the eight test packets (F(7,436)=0.534, p=0.809). Box plots (Figure 4.2) illustrate similar item
difficulties across animals (Median=0.01), plants (Median=0.31), and humans (Median=-0.08). This is supported by statistical findings (H(2)=0.809, p=0.667). There was also no significant difference (U=639, p=0.491) between fictitious (Median=0.35) and real taxa (Median=-0.05). These results indicate that the situational features of taxon and familiarity did not impact student problem-solving ability for the types of inheritance problems posed to this undergraduate sample.

In addition to comparing item difficulties by taxon and familiarity, it is useful to compare performance across Mendelian problem types (see Table 4.2). The boxplots (Figure 4.3) show clear differences in item difficulty between most of the five types of problems:

1. Items featuring an incomplete dominance (ID) Mendelian inheritance pattern with a given homozygous (“homo”) parental generation were the most difficult (median = +1.20).
2. Items featuring the same pattern (ID) with a given heterozygous (“hetero”) parental generation (median = +0.39).
3. Items featuring a dominant-recessive (DR) mechanism with a given homozygous parental generation (median = +0.59) on a nearly equivalent level of difficulty.
4. Items featuring a dominant-recessive mechanism with a given heterozygous (DR hetero) are easier than the previous three types (median = -0.21).
5. Items asking for the parental generation genotypes with a given distribution of the first filial generation using a dominant-recessive mechanism (DR F1) are the easiest (median = -1.12).
The patterns apparent in the figure are supported by the Kruskal-Wallis test indicating statistical differences overall (H(4)=63.87, p < 0.001). A post-hoc pairwise comparison of the item difficulties using Mann-Whitney tests confirmed these findings (all p < 0.01 except p = 0.629 for the two problem types with the second highest difficulties [DR homo and ID hetero]). These results indicate that the type of Mendelian inheritance problems represented by the items unsurprisingly has an impact on student problem-solving ability.

Figure 4.3 Boxplots comparing item difficulty by types of Mendelian inheritance problems (cf. Table 4.2). DR: dominant-recessive inheritance; ID: incomplete dominance; homo: given parental generation is homozygous; hetero: given parental generation is heterozygous; F1: first filial generation given. For information about statistical differences please refer to the text.

In addition to testing for situational effects across problem types (see above), consideration was given to whether situational features impacted performance within the five different problem types. Given that the problem types displayed different difficulties, it is important to test for potential item feature effects within each of the five item types. The boxplots (Figure 4.4A) illustrate that despite differences in item difficulties among types, different item features (plant, animal, human) produced similar performances within problem types (Kruskal Wallis test, p > 0.05 in all cases). For example, in Figure 4.4A, item type DR F1 differed in animal, plant, and human features, but produced similar results. Likewise, in Figure 4.4B, different item features (“real” and fictitious) produced similar performances within problem types (Mann-Whitney U test, p > 0.10 in all pairwise comparisons). In sum, situational features did not impact performance within problem types or among problem types.
Figure 4.4 Boxplots comparing item difficulties by problem types and situational features. 

A. Item difficulties grouped by problem type, and shaded by taxon (animal, plant, human). 

B. Item difficulties grouped by problem type, shaded by real or fictitious feature. 

DR: dominant-recessive inheritance; ID: incomplete dominance; homo: given parental generation is homozygous; hetero: given parental generation is heterozygous; F1: first filial generation given. For information about statistical differences please refer to the text.

4.4.3 Participant demographics

No significant differences in performance (F(1,416)=1.302, p=0.255) between male (Mean = 1.63) and female (Mean = 1.84) participants was found. Although an ANOVA revealed an overall difference in performance among demographic groups (F(5,412)=3.155, p=0.008), post-hoc tests did not produce any significant pairwise differences (p ≥ 0.094). A very small negative correlation between performance and age (r= -0.01, p ≤ 0.05) was found. Finally, as one might expect, there was a significant association between performance and number of completed biology courses (r=0.27, p < 0.01).

4.5 Discussion

Recent studies in biology education have documented significant and, in some cases, predictable impacts of situational features or contexts on measurements of student understanding. In the domain of evolution, for example, Nehm and colleagues have shown that measures of undergraduates’ understanding are impacted by the taxon included in the item, the scale of evolutionary differences, the familiarity of the taxa and traits, and the polarity of evolutionary change (e.g., the gain or loss of a trait) (Nehm & Ha 2011; Nehm & Reilly 2007; Nehm & Ridgway 2011; Opfer et al. 2012). Similar to these findings in evolution, Clough & Driver (1986), Kargbo et al. (1980), Freidenreich et al. (2011), Shea et al. (2015), and Ware and Gelman (2014) found that situational features can play a significant role in genetic reasoning and
argumentation. As a result of these findings, Shea, Duncan, and Stephenson (2015) developed a genetics-reasoning model that explicitly highlights the role of context effects. However, much of the prior work on genetics reasoning has been based on small samples and has not used an experimental research design capable of isolating the precise factors responsible for performance differences (Ware and Gelman’s 2014 design is a notable exception). Therefore, an important question in genetics education is which situational features impact particular types of problem solving, and how our understanding of these factors can be leveraged to improve teaching, learning, and assessment.

4.5.1 Impact of situational features on item difficulty

This study tested whether item difficulty was associated with situational features, which was expected given extensive prior work in cognitive psychology and a growing body of work in genetics education (Chi et al. 1981; Kargbo et al. 1980; Mayer 2013; Opfer et al. 2012; Ware & Gelman 2014). In all of these studies, attending to the situational features was not required for successful problem solving, but such features nevertheless impacted participants’ scientific reasoning. Surprisingly, this experimental study failed to find situational effects on Mendelian genetics problem solving. Similar results were also found in a partner study using a similar item set and study design in a sample of German high school students (Schmiemann, Nehm, and Tornabene 2017). Contrary to Nehm and colleagues’ work in the domain of evolution (e.g., Opfer, Nehm & Ha 2012), and Ware & Gelman’s (2014) work in the domain of genetics, different taxa (animal, plant, human) and familiarities (fictional, real) appear to have no significant impact on students’ genetic problem solving performance. These findings suggest that providing sets of genetics examples using mixtures of these contextual features will be unlikely to impact measures of student understanding, although studies of additional populations should be examined to test the generalizability of this claim.

The present findings raise the question of why no context effects were found given that (1) widespread work in cognitive psychology suggests that situational features impact human reasoning—even when such features are irrelevant to successfully solving the problems (Chi et al. 1981) and (2) prior work in genetics reasoning has found such effects (Shea et al. 2015). Several explanations are possible that could guide future work in this area.

The findings of this study suggest that very well-structured and constrained genetic problem formats might facilitate the recruitment of algorithmic problem-solving scripts (cf. Smith 1983), whereas more ill-structured problems (like those posed by Shea et al. 2015) may require deeper consideration of what the problem is about (e.g., deep structure), greater scrutiny of the situational features, and subsequent activation of a wider array of knowledge elements and problem-solving strategies. While such algorithmic problem solving may not advance a researcher's goal of ascertaining deep knowledge of genetics or other fields, it is nevertheless a commonly used method to solve widely employed domain-specific problems. Given the goal of evaluating the extent to which situational features impact problem solving, and the commonality of problems used in this study, it is worthwhile to know whether even algorithmic genetics problems are sensitive to feature-specific variation.

The transmission genetics problems that were posed had a consistent linguistic structure and constrained range of (forced-choice) answer options. It is possible that recognition of, or familiarity with, the type of problem and activation of a known problem-solving script was the key feature of the participants’ problem-solving strategy. Thus, familiarity with the problem
type could have driven the problem-solving procedure, thereby minimizing the impact of situational features on performance. This idea is supported by studies in mathematics (e.g., Hinsley, Hayes & Simon 1977; Silver 1979) and genetics (Collins 1986; Krajcik, Simmons & Lunetta 1988; Slack & Stewart 1990) which have found that low difficulty and/or familiar problem types quickly trigger an appropriate problem-solving strategy, often before the problem is fully read. Further, Chi et al. (1981), Nehm and Ridgway (2011) and, in genetics, Smith (1992), have found that experts categorize problems according to the methods or concepts required to solve the problem, whereas novices identify problems by their surface details. The problems employed were familiar to this high-performing university sample, who had received genetics instruction during the semester and in secondary school. A follow-up to this study could measure the magnitude of student familiarity with different genetics problem types and examine the association of this variable with student problem-solving success and situational impacts. As familiarity with problem type decreases, situational effects might increase. This prediction would be in line with Clough & Driver’s (1986) and Ware and Gelman’s (2014) studies of inheritance, and Opfer et al.’s (2012) study of natural selection. Indeed, familiarity with problem types clearly plays some role in the problem-solving process, as indicated by greater performance of the university students (who had been explicitly taught transmission genetics in both secondary school and university and hence had more opportunity to become familiar with these types of problems).

The role of assessment item format on the measurement of domain-specific concepts in biology is not well understood (Nehm & Schonfeld 2008). It is worth noting that Shea, Duncan, and Stephenson’s (2015) and Kargbo et al.’s (1980) studies documenting situational effects in genetics, and Nehm and colleagues’ work documenting situational effects in evolution, both employed open-ended tasks. It is possible that task format is contributing to present inferences about situational effects on biological reasoning. Multiple choice questions and answer options, like the ones used in the current study, may limit the range of cognitive resources elicited and problem solving strategies employed. However, Ware and Gelman (2014) used a forced-choice design, and uncovered context effects. An important aspect of their study was that it included misconception distractors, which makes the design more similar to the open-ended prompts of Kargbo et al. (1980). More detailed qualitative studies of problem solving strategies across a greater diversity of genetics problem types and formats (e.g., multiple choice vs. constructed response; arguments vs. explanations; normative vs. misconception distractors) are clearly in order. Overall, while the present study design cannot reveal the cause(s) of these findings showing an absence of situational effects in transmission genetics performance, it clearly indicates that situational features will not impact all types of genetics problems (Shea et al. 2015).

4.5.2 Impact of problem type on item difficulty

Although this study explored the general topic of Mendelian transmission genetics, participants were presented with several different inheritance problems (see Table 4.2). The results indicated that the type of problem impacted item difficulty to a greater extent than situational features (e.g., Figures 4.2 and 4.3). Prior work has suggested that different inheritance problems elicit different cognitive demands. For example, Collins & Stewart (1989) considered incomplete dominance problems to be less demanding than simple dominance problems. Incomplete dominance features a 1:1 mapping of each genotype to phenotype, whereas in simple
dominance both homozygous dominant and heterozygous genotypes are mapped to the dominant phenotype, which can be confusing for novice learners. Likewise, cause-to-effect problems have been considered less demanding than effect-to-cause problems (Stewart 1988). Cause-to-effect problems require 1:1 mapping of the genotype to phenotype whereas effect to cause problems require considering more than one possible genotypic antecedent for a given phenotypic effect.

Contrary to prior work, this university sample found incomplete dominance problems to be the most difficult and the effect-to-cause simple dominance problems to be the easiest. One possible explanation may lie in students’ familiarity with the problem types and subsequent recognition and activation of known problem-solving scripts. Despite greater putative cognitive complexity, it is conventional for simple dominance problems to be introduced first by instructors (and in textbooks) because they illustrate the traditional Mendelian concept of dominance. Incomplete dominance problems are typically taught later and treated as a more advanced variation on the basic rule. Simple dominance problems are also more widely taught in U.S. secondary schools than incomplete dominance problems, so familiarity may once again partially explain these findings. Familiarity may also afford a perception of this problem type as “easier” and students may be more committed to persist until an acceptable answer is reached. Persistence and checking answers were traits identified in expert problem solvers (Collins 1986, Smith & Good 1984).

Another interesting pattern notes was that items with the homozygous parental generation provided were more difficult than those with the heterozygous parental generation (irrespective of dominant-recessive or incomplete dominance). This seems to be contradictory, since problems with a given purebred homozygous parental generation are expected to be less difficult. The first filial generation is uniform, the Punnett square is quite simple, and the law of dominance is easy to understand. In contrast, a problem with a given heterozygous parental generation (which is the same as asking for an F2 generation for homozygous parents) seems to be more challenging.

One explanation is that teachers could have spent more time and effort on this kind of problem. In particular, because the idea of segregation - which is so important in all genetics contexts and is often difficult for students to understand (Browning & Lehman 1988; Moll & Allen 1987; Stewart & Dale 1989; Tolman 1982) - becomes very obvious in the characteristic phenotypic pattern. This may lead to students having more experience with this kind of problem to anticipate a “typical mixed phenotype pattern” (e.g., 3:1 or 1:2:1). This might impact success with other problem types. Thus, one explanation for this unexpected finding may relate to instructional focus and consequent problem perception. Further research is clearly necessary in order to confirm such a speculation.

### 4.5.3 Interaction of demographic factors and performance

Finally, the present study found no significant influence of gender or ethnicity on Mendelian problem solving performance. Notably, these findings differ from other American studies documenting a male advantage in biology at the undergraduate level (Eddy et al. 2014; Stanger-Hall 2012; Wright et al. 2016), but are similar to Dogru-Atay and Tekkaya’s (2002) study of middle schoolers, which also showed no gender advantage on multiple choice inheritance items, and several other studies finding no gender bias (Dimitrov 1999; Huppert et al. 2002; Lauer et al. 2013; Shepardson & Pizzini 1994; Willoughby & Metz 2009). While few studies have examined racial or ethnic differences in biology achievement, those that have
(Creech & Sweeder 2012; Dimitrov 1999) found no impact, which is in alignment with the findings herein.

In contrast to the lack of gender and ethnicity effects, significant (but small to moderate) correlations between participants’ performance, number of completed biology courses, and biology course marks were found. These findings provide some convergent validity evidence for the assessment items used.

4.6 Limitations and further research

The present findings should be viewed in light of several limitations. Though it does not impact statistical tests used for group comparisons, the number of items representing each independent variable group (taxon, real or fictitious) and Mendelian problem type (cf. Table 4.2) was not balanced. It seemed impossible to generate items representing a fictitious human being in order to balance items featuring fictitious animals or plants. Implementing fictitious taxa in items remains an interesting option for future studies as it allows one to control for potential effects of participant familiarity with taxa or traits (cf. Opfer et al. 2012).

Although large samples (> 400 participants) and many items (81) were used, significant differences were not found in certain cases that one might have expected. First, there seem to be no differences in students’ performances between male and female students or ethnic groups. Second, no context effects were detected. This lack of statistically significant differences does not guarantee that there are no such differences, as the power of a statistical test is in part reliant on employing a sample of adequate size to detect even small effect sizes. To get an impression about the sensitivity of the statistical tests used, a supplemental power analyses was performed using G*Power (Faul, Erdfelder, Lang & Buchner 2007). To calculate the required effect size necessary to detect an effect with this sample the following constraints were used: level of significance \( \alpha = .05 \) (a typical cut-off value in educational research), test power \( 1 - \beta = 0.8 \) (following Cohen 1988), and the particular sample size and numbers of groups. For this university sample, statistical tests would have detected differences between male and female students with an effect size higher than \( f \geq 0.133 (= d \geq 0.267) \) and between ethnic groups with an effect size of \( f \geq 0.164 (= d \geq 0.330) \). Both effect sizes are considered small effects (Cohen 1988). Therefore, there might be differences in performance within these groups, but one can assume that the effects will be small at most. To further reduce the possibility of failing to detect a small effect, additional research is required with much larger groups of participants (to increase statistical power). Focusing on item feature effects for this sample, the test was sensitive for effect sizes higher than \( d \geq 0.600 \) differentiating between items with real or fictitious organisms or for effect sizes higher than \( d \geq 0.629 \) between items with plants or animals. Both effect sizes are typically interpreted as medium. Thus there might be situational feature effects with small to lower-medium effect size.

The limitation of statistical power should be viewed in light of two considerations. First, the effect sizes of the differences found and, second, the data from descriptive statistics. If, for example, one compares dominant-recessive items with given homozygous or heterozygous parental generation (DR homo vs. DR hetero; cf. Table 4.2) one finds a very large effect \( (d = 1.572) \). This is notable because from a theoretical perspective the two problems seem to be very similar. One might interpret this to suggest that even small changes in items can have a strong impact on item difficulty. This circumstance may hold true for item feature effects, too. Thus, one could reason that a change in item features would cause medium effects at minimum and
would consequently be sensed by the tests used; still no significant differences for item features were detected. Descriptive data reported in the box plots (Figures 4.2, 4.4) support this interpretation. The overlap in item difficulties for item groups with very different situational features is very large and remains so even when disaggregated by problem type. This might be interpreted as a (non-inferential statistical) hint that there are no such item feature effects even though the test is not sensitive for small effects. To further clarify whether such a small effect of item features might exist, further research with larger number of items would be beneficial, and larger participant samples as well.

Although a major goal of educational research is the generalizations of findings, such generalization is often difficult or impossible in a single study. Indeed, the limits of generalizability are almost always a concern in empirical research. Even though there is have strong evidence to support the claim that the types of Mendelian inheritance problems studied are representative of common genetics problems, strictly speaking, the findings herein are limited to these five problems. One can assume that they will be valid for other problems in the context of Mendelian inheritance in which one has to apply a particular heuristic (e.g., problems on independent assortment) and in comparable participant samples.

Since the research focus was on item function and associated item difficulties rather than describing a population of subjects, a type of quota sampling was used to obtain participants. Hence conclusions about subjects cannot claim global generalizability. Nevertheless, one can assume that these findings will be valid for populations representing similar genetics problem solving experience and similar cultural contexts.

All in all, these findings would be stronger with a larger sets of items per category. The five types of Mendelian inheritance problems (cf. Table 4.2) that were developed could serve as blueprints for the development of larger item sets with a greater diversity of taxa. Moreover, the item design and situational features could be expanded to cover a greater array of genetics problems to determine if the findings are restricted to particular types of problems. The addition of constructed response items to complement the forced-choice items could help elucidate a possible interaction between context effects and item format. Further investigation is also needed to understand which kinds of genetics problems students solve heuristically. One might assume that there may be a continuum from problems which can be solved heuristically (like those used in this study) to items that require a deeper application of content knowledge.

4.7 References (Chapter 4)


Cavallo, A. M. (1994). Do females learn biological topics by rote more than males?. The American Biology Teacher, 56(6), 348-352.


References


Longden, B. (1982). Genetics—are there inherent learning difficulties?. *Journal of Biological Education*, 16(2), 135-140.


Appendix A - Overview of items used for study “Testing the impact of surface features on students’ Mendelian genetics understanding” (Chapter 4)

Table A1. Overview of Mendelian inheritance problems varying in surface features

<table>
<thead>
<tr>
<th>Type</th>
<th>Organism</th>
<th>Taxon group</th>
<th>Real or Fictitious</th>
<th>Trait</th>
<th>Item #</th>
<th>Item diffic.</th>
<th>wMNSQ</th>
<th>t-value</th>
<th>Item Abbreviation</th>
</tr>
</thead>
<tbody>
<tr>
<td>DR homo</td>
<td>amalcho</td>
<td>animal</td>
<td>fictitious</td>
<td>fur color</td>
<td>1</td>
<td>0.538</td>
<td>1.13</td>
<td>1.2</td>
<td>am_fellfaf1</td>
</tr>
<tr>
<td>DR homo</td>
<td>amalcho</td>
<td>animal</td>
<td>fictitious</td>
<td>fur texture</td>
<td>2</td>
<td>0.391</td>
<td>0.89</td>
<td>-0.6</td>
<td>am_fellstrf1</td>
</tr>
<tr>
<td>DR homo</td>
<td>beringo</td>
<td>animal</td>
<td>fictitious</td>
<td>plumage color</td>
<td>3</td>
<td>0.311</td>
<td>0.89</td>
<td>-0.9</td>
<td>berf1</td>
</tr>
<tr>
<td>DR homo</td>
<td>beringo</td>
<td>animal</td>
<td>fictitious</td>
<td>plumage pattern</td>
<td>4</td>
<td>0.878</td>
<td>0.84</td>
<td>-1.3</td>
<td>ber_strf1</td>
</tr>
<tr>
<td>DR homo</td>
<td>dulvitt</td>
<td>plant</td>
<td>fictitious</td>
<td>seed color</td>
<td>5</td>
<td>0.412</td>
<td>0.86</td>
<td>-1.1</td>
<td>dp_safaf1</td>
</tr>
<tr>
<td>DR homo</td>
<td>dulvitt</td>
<td>plant</td>
<td>fictitious</td>
<td>seed shape</td>
<td>6</td>
<td>0.525</td>
<td>0.92</td>
<td>-0.5</td>
<td>dp_safof1</td>
</tr>
<tr>
<td>DR homo</td>
<td>briscinum</td>
<td>plant</td>
<td>fictitious</td>
<td>flower color</td>
<td>7</td>
<td>0.378</td>
<td>0.78</td>
<td>-1.8</td>
<td>bris_flfaf1</td>
</tr>
<tr>
<td>DR homo</td>
<td>briscinum</td>
<td>plant</td>
<td>fictitious</td>
<td>seed length</td>
<td>8</td>
<td>0.548</td>
<td>0.80</td>
<td>-1.5</td>
<td>bris_sala1</td>
</tr>
<tr>
<td>DR homo</td>
<td>pea</td>
<td>plant</td>
<td>real</td>
<td>seed color</td>
<td>9</td>
<td>0.695</td>
<td>0.94</td>
<td>-0.5</td>
<td>ep_safaf1</td>
</tr>
<tr>
<td>DR homo</td>
<td>pea</td>
<td>plant</td>
<td>real</td>
<td>seed shape</td>
<td>10</td>
<td>1.020</td>
<td>0.97</td>
<td>-0.2</td>
<td>ep_safof1</td>
</tr>
<tr>
<td>DR homo</td>
<td>fruit fly</td>
<td>animal</td>
<td>real</td>
<td>body color</td>
<td>11</td>
<td>0.448</td>
<td>0.81</td>
<td>-1.5</td>
<td>ff_koerfaf1</td>
</tr>
<tr>
<td>DR homo</td>
<td>fruit fly</td>
<td>animal</td>
<td>real</td>
<td>wing shape</td>
<td>12</td>
<td>0.598</td>
<td>0.90</td>
<td>-0.8</td>
<td>ff_fluefof1</td>
</tr>
<tr>
<td>DR homo</td>
<td>corn</td>
<td>plant</td>
<td>real</td>
<td>seed color</td>
<td>13</td>
<td>0.249</td>
<td>0.92</td>
<td>-0.5</td>
<td>mp_safaf1</td>
</tr>
<tr>
<td>DR homo</td>
<td>corn</td>
<td>plant</td>
<td>real</td>
<td>seed shape</td>
<td>14</td>
<td>0.516</td>
<td>0.95</td>
<td>-0.4</td>
<td>mp_safof1</td>
</tr>
<tr>
<td>DR homo</td>
<td>guinea pig</td>
<td>animal</td>
<td>real</td>
<td>fur color</td>
<td>15</td>
<td>0.349</td>
<td>0.74</td>
<td>-2.2</td>
<td>ms_fellfaf1</td>
</tr>
<tr>
<td>DR homo</td>
<td>guinea pig</td>
<td>animal</td>
<td>real</td>
<td>fur texture</td>
<td>16</td>
<td>0.682</td>
<td>0.83</td>
<td>-1.4</td>
<td>ms_fellf2</td>
</tr>
<tr>
<td>DR homo</td>
<td>human</td>
<td>human</td>
<td>real</td>
<td>albinism</td>
<td>17</td>
<td>0.789</td>
<td>0.95</td>
<td>-0.3</td>
<td>me_albf1</td>
</tr>
<tr>
<td>DR homo</td>
<td>human</td>
<td>human</td>
<td>real</td>
<td>brachydactyly</td>
<td>18</td>
<td>0.646</td>
<td>0.89</td>
<td>-0.9</td>
<td>me_fingf1</td>
</tr>
<tr>
<td>DR homo</td>
<td>human</td>
<td>human</td>
<td>real</td>
<td>hairline</td>
<td>19</td>
<td>0.929</td>
<td>0.83</td>
<td>-1.4</td>
<td>me_nbf1</td>
</tr>
<tr>
<td>DR homo</td>
<td>human</td>
<td>human</td>
<td>real</td>
<td>night-blindness</td>
<td>20</td>
<td>1.081</td>
<td>0.75</td>
<td>-2.0</td>
<td>me_haarf1</td>
</tr>
<tr>
<td>DR homo</td>
<td>dove</td>
<td>animal</td>
<td>real</td>
<td>plumage pattern</td>
<td>21</td>
<td>0.594</td>
<td>0.74</td>
<td>-2.2</td>
<td>tauf1</td>
</tr>
<tr>
<td>DR hetero</td>
<td>amalcho</td>
<td>animal</td>
<td>fictitious</td>
<td>fur color</td>
<td>22</td>
<td>-0.431</td>
<td>1.05</td>
<td>0.4</td>
<td>am_fellfaf2</td>
</tr>
<tr>
<td>DR hetero</td>
<td>amalcho</td>
<td>animal</td>
<td>fictitious</td>
<td>fur texture</td>
<td>23</td>
<td>-0.028</td>
<td>0.94</td>
<td>-0.4</td>
<td>am_fellstrf2</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>-------</td>
<td>----</td>
<td>-----</td>
<td>--------</td>
<td>---------</td>
<td>---------</td>
<td>---------</td>
<td>---------</td>
<td>---------</td>
<td>---------</td>
</tr>
<tr>
<td>DR</td>
<td>hetro</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>beringo</td>
<td>animal</td>
<td>fictitious</td>
<td>plumage color</td>
<td>24</td>
<td>-1.128</td>
<td>1.03</td>
<td>0.2</td>
<td>ber_faf2</td>
<td></td>
</tr>
<tr>
<td>beringo</td>
<td>animal</td>
<td>fictitious</td>
<td>plumage pattern</td>
<td>25</td>
<td>-0.172</td>
<td>1.24</td>
<td>1.6</td>
<td>ber_strf2</td>
<td></td>
</tr>
<tr>
<td>dulvitt</td>
<td>plant</td>
<td>fictitious</td>
<td>seed color</td>
<td>26</td>
<td>-0.141</td>
<td>1.09</td>
<td>0.6</td>
<td>dp_safaf2</td>
<td></td>
</tr>
<tr>
<td>dulvitt</td>
<td>plant</td>
<td>fictitious</td>
<td>seed shape</td>
<td>27</td>
<td>0.887</td>
<td>0.92</td>
<td>-0.6</td>
<td>dp_safof2</td>
<td></td>
</tr>
<tr>
<td>briscinum</td>
<td>plant</td>
<td>fictitious</td>
<td>flower color</td>
<td>28</td>
<td>0.007</td>
<td>0.93</td>
<td>-0.5</td>
<td>bris_flfaf2</td>
<td></td>
</tr>
<tr>
<td>briscinum</td>
<td>plant</td>
<td>fictitious</td>
<td>seed length</td>
<td>29</td>
<td>-0.383</td>
<td>1.11</td>
<td>0.7</td>
<td>bris_salaf2</td>
<td></td>
</tr>
<tr>
<td>pea</td>
<td>plant</td>
<td>real</td>
<td>seed color</td>
<td>30</td>
<td>-0.419</td>
<td>1.03</td>
<td>0.2</td>
<td>ep_safaf2</td>
<td></td>
</tr>
<tr>
<td>pea</td>
<td>plant</td>
<td>real</td>
<td>seed shape</td>
<td>31</td>
<td>-0.212</td>
<td>0.94</td>
<td>-0.4</td>
<td>ep_safof2</td>
<td></td>
</tr>
<tr>
<td>fruit fly</td>
<td>animal</td>
<td>real</td>
<td>body color</td>
<td>32</td>
<td>0.065</td>
<td>1.14</td>
<td>1.0</td>
<td>ff_koerfaf2</td>
<td></td>
</tr>
<tr>
<td>fruit fly</td>
<td>animal</td>
<td>real</td>
<td>wing shape</td>
<td>33</td>
<td>0.011</td>
<td>1.16</td>
<td>1.1</td>
<td>ff_fluefof2</td>
<td></td>
</tr>
<tr>
<td>corn</td>
<td>plant</td>
<td>real</td>
<td>seed color</td>
<td>34</td>
<td>-0.839</td>
<td>0.84</td>
<td>-0.9</td>
<td>mp_safaf2</td>
<td></td>
</tr>
<tr>
<td>corn</td>
<td>plant</td>
<td>real</td>
<td>seed shape</td>
<td>35</td>
<td>-0.251</td>
<td>0.92</td>
<td>-0.5</td>
<td>mp_safof2</td>
<td></td>
</tr>
<tr>
<td>guinea pig</td>
<td>animal</td>
<td>real</td>
<td>fur color</td>
<td>36</td>
<td>-0.381</td>
<td>0.79</td>
<td>-1.4</td>
<td>ms_fellfaf2</td>
<td></td>
</tr>
<tr>
<td>guinea pig</td>
<td>animal</td>
<td>real</td>
<td>fur texture</td>
<td>37</td>
<td>-0.274</td>
<td>0.81</td>
<td>-1.4</td>
<td>ms_fellstrf2</td>
<td></td>
</tr>
<tr>
<td>human</td>
<td>human</td>
<td>real</td>
<td>albinism</td>
<td>38</td>
<td>-0.231</td>
<td>0.93</td>
<td>-0.5</td>
<td>me_albf2</td>
<td></td>
</tr>
<tr>
<td>human</td>
<td>human</td>
<td>real</td>
<td>brachydactyly</td>
<td>39</td>
<td>-0.208</td>
<td>1.23</td>
<td>1.4</td>
<td>me_fingf2</td>
<td></td>
</tr>
<tr>
<td>human</td>
<td>human</td>
<td>real</td>
<td>night-blindness</td>
<td>40</td>
<td>0.050</td>
<td>1.03</td>
<td>0.3</td>
<td>me_nb2f2</td>
<td></td>
</tr>
<tr>
<td>human</td>
<td>human</td>
<td>real</td>
<td>hairline</td>
<td>41</td>
<td>-0.517</td>
<td>0.88</td>
<td>-0.7</td>
<td>me_haarf2</td>
<td></td>
</tr>
<tr>
<td>dove</td>
<td>animal</td>
<td>real</td>
<td>plumage pattern</td>
<td>42</td>
<td>0.013</td>
<td>0.90</td>
<td>-0.7</td>
<td>tauf2</td>
<td></td>
</tr>
<tr>
<td>amalcho</td>
<td>animal</td>
<td>fictitious</td>
<td>fur color</td>
<td>43</td>
<td>-0.879</td>
<td>1.09</td>
<td>0.5</td>
<td>am_fellfaf_iv</td>
<td></td>
</tr>
<tr>
<td>amalcho</td>
<td>animal</td>
<td>fictitious</td>
<td>fur texture</td>
<td>44</td>
<td>-0.971</td>
<td>1.21</td>
<td>1.0</td>
<td>am_fellstrf_iv</td>
<td></td>
</tr>
<tr>
<td>beringo</td>
<td>animal</td>
<td>fictitious</td>
<td>plumage color</td>
<td>45</td>
<td>-1.118</td>
<td>0.97</td>
<td>-0.1</td>
<td>berf_fa_iv</td>
<td></td>
</tr>
<tr>
<td>beringo</td>
<td>animal</td>
<td>fictitious</td>
<td>plumage pattern</td>
<td>46</td>
<td>-2.134</td>
<td>0.69</td>
<td>-1.0</td>
<td>ber_str_iv</td>
<td></td>
</tr>
<tr>
<td>dulvitt</td>
<td>plant</td>
<td>fictitious</td>
<td>seed color</td>
<td>47</td>
<td></td>
<td></td>
<td></td>
<td>dp_safaf_iv</td>
<td></td>
</tr>
<tr>
<td>dulvitt</td>
<td>plant</td>
<td>fictitious</td>
<td>seed shape</td>
<td>48</td>
<td></td>
<td></td>
<td></td>
<td>dp_safof_iv</td>
<td></td>
</tr>
<tr>
<td>briscinum</td>
<td>plant</td>
<td>fictitious</td>
<td>flower color</td>
<td>49</td>
<td>-0.825</td>
<td>1.03</td>
<td>0.2</td>
<td>bris_flfaf2</td>
<td></td>
</tr>
<tr>
<td>briscinum</td>
<td>plant</td>
<td>fictitious</td>
<td>seed length</td>
<td>50</td>
<td>-1.573</td>
<td>1.09</td>
<td>0.4</td>
<td>bris_sala_iv</td>
<td></td>
</tr>
<tr>
<td>pea</td>
<td>plant</td>
<td>real</td>
<td>seed color</td>
<td>51</td>
<td>-1.662</td>
<td>1.13</td>
<td>0.5</td>
<td>ep_safaf_iv</td>
<td></td>
</tr>
</tbody>
</table>

103
<table>
<thead>
<tr>
<th>ID</th>
<th>Species</th>
<th>Type</th>
<th>Trait</th>
<th>Value 1</th>
<th>Value 2</th>
<th>Value 3</th>
<th>Value 4</th>
<th>Value 5</th>
</tr>
</thead>
<tbody>
<tr>
<td>DR F1</td>
<td>pea plant</td>
<td>plant</td>
<td>seed shape</td>
<td>52</td>
<td>-2.052</td>
<td>0.96</td>
<td>-0.1</td>
<td>ep_safof iv</td>
</tr>
<tr>
<td>DR F1</td>
<td>fruit fly</td>
<td>animal</td>
<td>body color</td>
<td>53</td>
<td>-0.994</td>
<td>1.13</td>
<td>0.7</td>
<td>ff_koerfaf iv</td>
</tr>
<tr>
<td>DR F1</td>
<td>fruit fly</td>
<td>animal</td>
<td>wing shape</td>
<td>54</td>
<td>-1.292</td>
<td>0.80</td>
<td>-1.1</td>
<td>ff_fluefop iv</td>
</tr>
<tr>
<td>DR F1</td>
<td>corn-plant</td>
<td>plant</td>
<td>seed color</td>
<td>55</td>
<td>-1.021</td>
<td>0.97</td>
<td>-0.1</td>
<td>mp_safaf iv</td>
</tr>
<tr>
<td>DR F1</td>
<td>corn-plant</td>
<td>plant</td>
<td>seed shape</td>
<td>56</td>
<td>-1.702</td>
<td>0.89</td>
<td>-0.3</td>
<td>mp_safaf iv</td>
</tr>
<tr>
<td>DR F1</td>
<td>guinea pig</td>
<td>animal</td>
<td>fur color</td>
<td>57</td>
<td>-0.903</td>
<td>0.85</td>
<td>-0.8</td>
<td>ms_fellfop iv</td>
</tr>
<tr>
<td>DR F1</td>
<td>guinea pig</td>
<td>animal</td>
<td>fur texture</td>
<td>58</td>
<td>-2.290</td>
<td>0.78</td>
<td>-0.6</td>
<td>ms_fellstrf iv</td>
</tr>
<tr>
<td>DR F1</td>
<td>human</td>
<td>human</td>
<td>albinism</td>
<td>59</td>
<td>1.586</td>
<td>1.14</td>
<td>1.1</td>
<td>me_albf iv</td>
</tr>
<tr>
<td>DR F1</td>
<td>human</td>
<td>human</td>
<td>brachydactyly</td>
<td>60</td>
<td>-1.216</td>
<td>1.08</td>
<td>0.4</td>
<td>me_fingf iv</td>
</tr>
<tr>
<td>DR F1</td>
<td>human</td>
<td>human</td>
<td>night-blindness</td>
<td>61</td>
<td>-0.963</td>
<td>1.09</td>
<td>0.5</td>
<td>me_nbf iv</td>
</tr>
<tr>
<td>DR F1</td>
<td>human</td>
<td>human</td>
<td>hairline</td>
<td>62</td>
<td>-0.861</td>
<td>1.06</td>
<td>0.4</td>
<td>me_haarf iv</td>
</tr>
<tr>
<td>DR F1</td>
<td>dove</td>
<td>animal</td>
<td>plumage pattern</td>
<td>63</td>
<td>-0.861</td>
<td>1.06</td>
<td>0.4</td>
<td>tauf iv</td>
</tr>
<tr>
<td>ID homo</td>
<td>klasron plant</td>
<td>plant</td>
<td>fictitious flower color</td>
<td>64</td>
<td>1.586</td>
<td>1.14</td>
<td>1.1</td>
<td>klpf1</td>
</tr>
<tr>
<td>ID homo</td>
<td>briscinum plant</td>
<td>plant</td>
<td>fictitious flower color</td>
<td>65</td>
<td>1.481</td>
<td>0.91</td>
<td>-0.7</td>
<td>brisf1</td>
</tr>
<tr>
<td>ID homo</td>
<td>strommler</td>
<td>animal</td>
<td>fictitious fur color</td>
<td>66</td>
<td>1.061</td>
<td>1.29</td>
<td>2.2</td>
<td>strf1</td>
</tr>
<tr>
<td>ID homo</td>
<td>theromosa</td>
<td>animal</td>
<td>fictitious plumage color</td>
<td>67</td>
<td>1.686</td>
<td>0.97</td>
<td>-0.2</td>
<td>thef1</td>
</tr>
<tr>
<td>ID homo</td>
<td>andalusian chicken</td>
<td>animal</td>
<td>real plumage color</td>
<td>68</td>
<td>1.089</td>
<td>1.05</td>
<td>0.4</td>
<td>ahf1</td>
</tr>
<tr>
<td>ID homo</td>
<td>snapdragon plant</td>
<td>plant</td>
<td>real flower color</td>
<td>69</td>
<td>1.060</td>
<td>1.07</td>
<td>0.6</td>
<td>glf1</td>
</tr>
<tr>
<td>ID homo</td>
<td>rabbit</td>
<td>animal</td>
<td>real fur color</td>
<td>70</td>
<td>1.412</td>
<td>1.16</td>
<td>1.3</td>
<td>kanf1</td>
</tr>
<tr>
<td>ID homo</td>
<td>primrose plant</td>
<td>plant</td>
<td>real flower color</td>
<td>71</td>
<td>1.199</td>
<td>1.14</td>
<td>1.0</td>
<td>prpf1</td>
</tr>
<tr>
<td>ID homo</td>
<td>four o’clock plant</td>
<td>plant</td>
<td>real flower color</td>
<td>72</td>
<td>0.864</td>
<td>1.09</td>
<td>1.0</td>
<td>wbf1</td>
</tr>
<tr>
<td>ID hetero</td>
<td>klasron plant</td>
<td>plant</td>
<td>fictitious flower color</td>
<td>73</td>
<td>0.877</td>
<td>1.18</td>
<td>1.4</td>
<td>klpf2</td>
</tr>
<tr>
<td>ID hetero</td>
<td>briscinum plant</td>
<td>plant</td>
<td>fictitious flower color</td>
<td>74</td>
<td>0.758</td>
<td>0.83</td>
<td>-1.4</td>
<td>brisf2</td>
</tr>
<tr>
<td>ID hetero</td>
<td>strommler</td>
<td>animal</td>
<td>fictitious fur color</td>
<td>75</td>
<td>-0.089</td>
<td>1.05</td>
<td>0.4</td>
<td>strf2</td>
</tr>
<tr>
<td>ID hetero</td>
<td>theromosa</td>
<td>animal</td>
<td>fictitious plumage color</td>
<td>76</td>
<td>0.619</td>
<td>1.17</td>
<td>1.3</td>
<td>thef2</td>
</tr>
<tr>
<td>ID hetero</td>
<td>andalusian chicken</td>
<td>animal</td>
<td>real plumage color</td>
<td>77</td>
<td>0.519</td>
<td>1.02</td>
<td>0.2</td>
<td>ahf2</td>
</tr>
<tr>
<td>ID hetero</td>
<td>snapdragon plant</td>
<td>plant</td>
<td>real flower color</td>
<td>78</td>
<td>0.393</td>
<td>0.76</td>
<td>-1.9</td>
<td>glf2</td>
</tr>
<tr>
<td>ID hetero</td>
<td>rabbit</td>
<td>animal</td>
<td>real fur color</td>
<td>79</td>
<td>-0.343</td>
<td>0.92</td>
<td>-0.5</td>
<td>kanf2</td>
</tr>
<tr>
<td>ID</td>
<td>Organism</td>
<td>Taxon group</td>
<td>Real or Fictitious</td>
<td>Item #: item number represented in Wright maps (cf. Figure 4.1)</td>
<td>wMNSQ: item fit measure</td>
<td>t-value: additional item fit measure</td>
<td>Item Abbreviation: project specific abbreviation of the item for re-identification irrespective of item number</td>
<td></td>
</tr>
<tr>
<td>-------</td>
<td>--------------------</td>
<td>-------------</td>
<td>--------------------</td>
<td>-----------------------------------------------------------------</td>
<td>--------------------------</td>
<td>--------------------------------------</td>
<td>-----------------------------------------------------------------------------------</td>
<td></td>
</tr>
<tr>
<td>hetero</td>
<td>primrose plant</td>
<td>plant</td>
<td>real</td>
<td>flower color 80 0,232 1,19 1,3 prpf2</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>eight o'clock plant</td>
<td>plant</td>
<td>real</td>
<td>Flower color 81 0,059 0,96 -0,4 wbf2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Explanation of terms used in Table A1:

- **Type**: description of five different types of Mendelian inheritance problems used in this study (cf. Table 4.2)
  - DR: dominant-recessive inheritance
  - ID: incomplete dominance
  - homo: given parental generation is homozygous
  - hetero: given parental generation is heterozygous
  - F1: first filial generation given.
- **Organism**: name of the organism mentioned in the item
- **Taxon group**: assignment of living being to the three compared taxon groups (animal, plant, human)
- **Real or Fictitious**: indication whether living being does exist in reality (real) or is made up for this study (fictitious)
- **Trait**: trait of the living being inhered in the item
- **Item #: item number represented in Wright maps (cf. Figure 4.1)**
- **wMNSQ**: item fit measure
- **t-value**: additional item fit measure
- **Item Abbreviation**: project specific abbreviation of the item for re-identification irrespective of item number

Items without statistical values highlighted in grey (# 47, 48, 59, 62) have been excluded for further analyses after first analyses due to poor fit statistics. For details please refer to the main text.
Appendix B. Example Items Used for Study “Testing the impact of surface features on students’ Mendelian genetics understanding” (Chapter 4)

On the following pages are tables containing item examples for each of the five types of Mendelian inheritance problems used in this study. Each header contains information about the problem type followed by item examples representing all used features regarding taxon (plant, animal, human) and familiarity (real, fictitious). To match these examples to other information given (Wright maps, item overview), the item number (#) is included at the beginning of the item text.
Problem Type 1: Dominant-recessive homozygous parental generation (DR homo)
Mendelian inheritance mechanism: Dominant-recessive
Information given: Homozygous parental generation (P)
Question posed: Distribution of first filial generation (F1)

<table>
<thead>
<tr>
<th>Plant</th>
<th>Real</th>
<th>Fictitious</th>
</tr>
</thead>
<tbody>
<tr>
<td>(#9) Pea plants can have yellow or green seed color. Seed color is inherited for pea plants in a dominant-recessive manner, where yellow seed color is dominant and green seed color recessive. Pea plants that have yellow seed color are crossed with pea plants that have green seed color. Both are homozygous regarding seed color. Which distribution is reflected in their offspring (F1 generation) with respect to seed color?</td>
<td>All descendants have yellow seed color. All descendants have green seed color. The descendants have an approximate ratio of 1:1 yellow seed color to green seed color. The descendants have an approximate ratio of 3:1 yellow seed color to green seed color. The descendants have an approximate ratio of 3:1 green seed color to yellow seed color.</td>
<td>All descendants have blue seeds. All descendants have grey seeds. The descendants have an approximate ratio of 1:1 blue seeds to grey seeds. The descendants have an approximate ratio of 3:1 blue seeds to grey seeds. The descendants have an approximate ratio of 3:1 grey seeds to blue seeds.</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Animal</th>
<th>Real</th>
<th>Fictitious</th>
</tr>
</thead>
<tbody>
<tr>
<td>(#15) Guinea pigs can have black or white fur color. Fur color is inherited for guinea pigs in a dominant-recessive manner, where black fur color is dominant and white fur color recessive. Guinea pigs that have black fur color are crossed with guinea pigs that have white fur color. Both are homozygous regarding fur color. Which distribution is reflected in their offspring (F1 generation) with respect to fur color?</td>
<td>All descendants have black fur color. All descendants have white fur color. The descendants have an approximate ratio of 1:1 black fur color to white fur color. The descendants have an approximate ratio of 3:1 black fur color to white fur color. The descendants have an approximate ratio of 3:1 white fur color to black fur color.</td>
<td>All descendants have black fur color. All descendants have white fur color. The descendants have an approximate ratio of 1:1 black fur color to white fur color. The descendants have an approximate ratio of 3:1 black fur color to white fur color. The descendants have an approximate ratio of 3:1 white fur color to black fur color.</td>
</tr>
</tbody>
</table>

| Human: | |
| (#18) Humans can have shortened fingers or normal finger length. Brachydactyly (short fingers and toes) is inherited for humans in a dominant-recessive manner, where shortened fingers is dominant and normal finger length recessive. A human that has shortened fingers is having children with a human that has normal finger length. Both are homozygous regarding brachydactyly. Which distribution is reflected in their offspring (F1 generation) with respect to brachydactyly? All descendants have shortened fingers. All descendants have normal finger length. The descendants have an approximate ratio of 1:1 shortened fingers to normal finger length. The descendants have an approximate ratio of 3:1 shortened fingers to normal finger length. The descendants have an approximate ratio of 3:1 normal finger length to shortened fingers. | |
### Type 2: Dominant-recessive heterozygous parental generation (DR hetero)

**Mendelian inheritance mechanism:** Dominant-recessive  
**Information given:** Heterozygous parental generation (P)  
**Question posed:** Distribution of first filial generation (F1)

<table>
<thead>
<tr>
<th></th>
<th>Real</th>
<th>Fictitious</th>
</tr>
</thead>
</table>
| **Plant** | (#34) Corn plants can have purple or yellow seed color. Seed color is inherited for corn plants in a dominant-recessive manner, where purple seed color is dominant and yellow seed color recessive. Two corn plants are crossed, that both have purple seed color. Both are **heterozygous** regarding seed color. Which distribution is reflected in their offspring (F1 generation) with respect to seed color?  
All descendants have purple seed color.  
All descendants have yellow seed color.  
The descendants have an approximate ratio of 1:1 purple seed color to yellow seed color.  
The descendants have an approximate ratio of 3:1 purple seed color to yellow seed color.  
The descendants have an approximate ratio of 3:1 yellow seed color to purple seed color. | (#24) Beringos can have dark brown or white plumage color. Plumage color is inherited for beringos in a dominant-recessive manner, where dark brown plumage color is dominant and white plumage color recessive. Two beringos are crossed, that both have dark brown plumage color. Both are **heterozygous** regarding plumage color. Which distribution is reflected in their offspring (F1 generation) with respect to plumage color?  
All descendants have dark brown plumage color.  
All descendants have white plumage color.  
The descendants have an approximate ratio of 1:1 dark brown plumage color to white plumage color.  
The descendants have an approximate ratio of 3:1 dark brown plumage color to white plumage color.  
The descendants have an approximate ratio of 3:1 white plumage color to dark brown plumage color. |
| **Animal** | (#32) Fruit flies can have grey or ivory colored body color. Body color is inherited for fruit flies in a dominant-recessive manner, where grey body color is dominant and ivory colored body color recessive. Two fruit flies are crossed, that both have grey body color. Both are **heterozygous** regarding body color. Which distribution is reflected in their offspring (F1 generation) with respect to body color?  
All descendants have grey body color.  
All descendants have ivory colored body color.  
The descendants have an approximate ratio of 1:1 grey body color to ivory colored body color.  
The descendants have an approximate ratio of 3:1 grey body color to ivory colored body color.  
The descendants have an approximate ratio of 3:1 ivory colored body color to grey body color. | (#28) Briscinum plants can have white or yellow flowers. Flower color is inherited for briscinum plants in a dominant-recessive manner, where white flower color is dominant and yellow flower color recessive. Two briscinum plants are crossed, that both have white flowers. Both are **heterozygous** regarding flower color. Which distribution is reflected in their offspring (F1 generation) with respect to flower color?  
All descendants have white flowers.  
All descendants have yellow flowers.  
The descendants have an approximate ratio of 1:1 white flowers to yellow flowers.  
The descendants have an approximate ratio of 3:1 white flowers to yellow flowers.  
The descendants have an approximate ratio of 3:1 yellow flowers to white flowers. |
Human:
(#41) Humans can have a pointed or round hairline. Hairline is inherited for humans in a dominant-recessive manner, where pointed hairline is dominant and round hairline recessive.
Two humans that both have pointed hairlines are having children. Both are heterozygous regarding hairline.
Which distribution is reflected in their offspring (F1 generation) with respect to hairline?
All descendants have pointed hairlines.
All descendants have round hairlines.
The descendants have an approximate ratio of 1:1 pointed hairline to round hairline.
The descendants have an approximate ratio of 3:1 pointed hairline to round hairline.
The descendants have an approximate ratio of 3:1 round hairline to pointed hairline.
Problem Type 3: Dominant-recessive filial generation (DR F1)
Mendelian inheritance mechanism: Dominant-recessive
Information given: Distribution of first filial generation (F1)
Question posed: Parental generation (P) genotypes

<table>
<thead>
<tr>
<th></th>
<th>Real</th>
<th>Fictitious</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Plant</strong></td>
<td>(#56) The seed shape of corn plants can be smooth or wrinkled. The seed shape is inherited in a dominant-recessive manner. The smooth seed shape is dominant. The offspring have smooth seed shape. Which parental combination would be least likely to produce these offspring? Homozygous parents, each with wrinkled and smooth seed shape. Heterozygous parents with smooth seed shape. Homozygous parents with wrinkled seed shape. Homozygous parents with smooth seed shape.</td>
<td>(#47) The seed color of dulvitt plants can be blue or grey. The seed color is inherited in a dominant-recessive manner. The blue seed color is dominant. The offspring have blue seed color. Which parental combination would be least likely to produce these offspring? Homozygous parents, each with grey and blue seed color. Heterozygous parents with blue seed color. Homozygous parents with grey seed color. Homozygous parents with blue seed color.</td>
</tr>
<tr>
<td><strong>Animal</strong></td>
<td>(#63) The plumage pattern of doves can be spotted or plain. The plumage pattern is inherited in a dominant-recessive manner. The spotted plumage pattern is dominant. The offspring have spotted plumage pattern. Which parental combination would be least likely to produce these offspring? Homozygous parents, each with plain and spotted plumage pattern. Heterozygous parents with spotted plumage pattern. Homozygous parents with plain plumage pattern. Homozygous parents with spotted plumage pattern.</td>
<td>(#44) The fur texture of amalchos can be curly haired or smooth haired. The fur texture is inherited in a dominant-recessive manner. The curly haired fur texture is dominant. The offspring have curly haired fur texture. Which parental combination would be least likely to produce these offspring? Homozygous parents, each with smooth haired and curly haired fur texture. Heterozygous parents with curly haired fur texture. Homozygous parents with smooth haired fur texture. Homozygous parents with curly haired fur texture.</td>
</tr>
</tbody>
</table>

**Human:**
(#61) Humans can have impaired vision at twilight or normal vision at twilight. The night-blindness is inherited in a dominant-recessive manner. The impaired vision at twilight is dominant. The offspring have impaired vision at twilight. Which parental combination would be least likely to produce these offspring? Homozygous parents, each with normal vision at twilight and impaired vision at twilight. Heterozygous parents with impaired vision at twilight. Homozygous parents with normal vision at twilight. Homozygous parents with impaired vision at twilight.
### Problem Type 4: Incomplete dominant-recessive homozygous parental generation (ID homo)

Mendelian inheritance mechanism: Incomplete dominance

Information given: Homozygous parental generation (P)

Question posed: Distribution of first filial generation (F1)

<table>
<thead>
<tr>
<th>Real</th>
<th>Fictitious</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Plant</strong></td>
<td><strong>Plant</strong></td>
</tr>
<tr>
<td>(#69) Snapdragon plants can have red, white or pink flower color. Flower color is inherited for snapdragon plants in an incompletely dominant (intermediate) manner, where pink flower color is the incompletely dominant (intermediate) manifestation/characteristic. A snapdragon plant, which has red flower color, is crossed with a snapdragon plant which has white flower color. Both are homozygous regarding flower color. Which distribution is reflected in their offspring (F1 generation) with respect to flower color? All descendants have red flower color. All descendants have white flower color. All descendants have pink flower color. The descendants have an approximate ratio of 1:2:1 red flower color to pink flower color to white flower color.</td>
<td>(#65) Briscinum plants can have red, white or pink flower color. Flower color is inherited for briscinum plants in an incompletely dominant (intermediate) manner, where pink flower color is the incompletely dominant (intermediate) manifestation/characteristic. A briscinum plant, which has red flowers, is crossed with a briscinum plant which has white flowers. Both are homozygous regarding flower color. Which distribution is reflected in their offspring (F1 generation) with respect to flower color? All descendants have red flower color. All descendants have white flower color. All descendants have pink flower color. The descendants have an approximate ratio of 1:2:1 red flower color to pink flower color to white flower color.</td>
</tr>
<tr>
<td><strong>Animal</strong></td>
<td><strong>Animal</strong></td>
</tr>
<tr>
<td>(#70) Rabbits can have dark brown, white or light brown fur color. Fur color is inherited for rabbits in an incompletely dominant (intermediate) manner, where light brown fur color is the incompletely dominant (intermediate) manifestation/characteristic. A rabbit, which has dark brown fur color, is crossed with a rabbit which has white fur color. Both are homozygous regarding fur color. Which distribution is reflected in their offspring (F1 generation) with respect to fur color? All descendants have dark brown fur color. All descendants have white fur color. All descendants have light brown fur color. The descendants have an approximate ratio of 1:2:1 dark brown fur color to light brown fur color to white fur color.</td>
<td>(#67) Thermomosas can have dark brown, white or light brown plumage color. Plumage color is inherited for thermomosas in an incompletely dominant (intermediate) manner, where light brown plumage color is the incompletely dominant (intermediate) manifestation/characteristic. A thermomosa, which has dark brown plumage color, is crossed with a thermomosa which has white plumage color. Both are homozygous regarding plumage color. Which distribution is reflected in their offspring (F1 generation) with respect to plumage color? All descendants have dark brown plumage color. All descendants have white plumage color. All descendants have light brown plumage color. The descendants have an approximate ratio of 1:2:1 dark brown plumage color to light brown plumage color to white plumage color.</td>
</tr>
</tbody>
</table>
Problem Type 5: Incomplete dominant-recessive heterozygous parental generation (ID hetero)
Mendelian inheritance mechanism: Incomplete dominance
Information given: Heterozygous parental generation (P)
Question posed: Distribution of first filial generation (F1)

<table>
<thead>
<tr>
<th></th>
<th>Real</th>
<th>Fictitious</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Plant</strong></td>
<td>(#80) Primrose plants can have red, white or pink flower color. Flower color is inherited for primrose plants in an incompletely dominant (intermediate) manner, where pink flower color is the incompletely dominant (intermediate) manifestation/characteristic. A primrose plant, which has pink flower color, is crossed with a primrose plant which has pink flower color. Both are heterozygous regarding flower color. Which distribution is reflected in their offspring (F1 generation) with respect to flower color? All descendants have red flower color. All descendants have white flower color. All descendants have pink flower color. The descendants have an approximate ratio of 1:2:1 red flower color to pink flower color to white flower color.</td>
<td>(#73) Klasron plants can have red, yellow or orange flower color. Flower color is inherited for klasron plants in an incompletely dominant (intermediate) manner, where orange flower color is the incompletely dominant (intermediate) manifestation/characteristic. A klasron plant, which has orange flowers, is crossed with a klasron plant which has orange flowers. Both are heterozygous regarding flower color. Which distribution is reflected in their offspring (F1 generation) with respect to flower color? All descendants have red flower color. All descendants have yellow flower color. All descendants have orange flower color. The descendants have an approximate ratio of 1:2:1 red flower color to orange flower color to yellow flower color.</td>
</tr>
<tr>
<td><strong>Animal</strong></td>
<td>(#77) Andalusian chickens can have black, white or bluish plumage color. Plumage color is inherited for Andalusian chickens in an incompletely dominant (intermediate) manner, where bluish plumage color is the incompletely dominant (intermediate) manifestation/characteristic. An Andalusian chicken, which has bluish plumage color, is crossed with an Andalusian chicken which has bluish plumage color. Both are heterozygous regarding plumage color. Which distribution is reflected in their offspring (F1 generation) with respect to plumage color? All descendants have black plumage color. All descendants have white plumage color. All descendants have bluish plumage color. The descendants have an approximate ratio of 1:2:1 black plumage color to bluish plumage color to white plumage color.</td>
<td>(#75) Strommlers can have dark brown, white or light brown fur color. Fur color is inherited for strommlers in an incompletely dominant (intermediate) manner, where light brown fur color is the incompletely dominant (intermediate) manifestation/characteristic. A strommler, which has light brown fur color, is crossed with a strommler which has light brown fur color. Both are heterozygous regarding fur color. Which distribution is reflected in their offspring (F1 generation) with respect to fur color? All descendants have dark brown fur color. All descendants have white fur color. All descendants have light brown fur color. The descendants have an approximate ratio of 1:2:1 dark brown fur color to light brown fur color to white fur color.</td>
</tr>
</tbody>
</table>