Sci & Educ (2017) 26:1161–1191
DOI 10.1007/s11191-017-9925-z

ARTICLE

Assessment of Genetics Understanding

Under What Conditions Do Situational Features Have an Impact on Measures?

Philipp Schmiemann[1] **· Ross H. Nehm**[2] **· Robyn E. Tornabene**[2]

Published online: 14 October 2017

Abstract 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. The goal of our study was to employ an experimental research design to quantify the degree to which situational features impact inferences about participants' understanding of Mendelian genetics. Two participant samples from different educational levels and cultural backgrounds (high school, $n = 480$; university, $n = 444$; Germany and USA) were used to test for context effects. A multi-matrix test design was employed, and item packets differing in situational features (e.g., plant, animal, human, fictitious) were randomly distributed to participants in the two samples. Rasch analyses of participant scores from both samples produced good item fit, person reliability, and item reliability and indicated that the university sample displayed stronger performance on the items compared to the high school sample. We found, surprisingly, that in both samples, no significant differences in performance occurred among the animal, plant, and human item contexts, or between the fictitious and "real" item contexts. In the university sample, we were also able to test for differences in performance between genders, among ethnic groups, and by prior biology coursework. 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.

Electronic supplementary material The online version of this article (https://doi.org/10.1007/s11191-017-9925-z) contains supplementary material, which is available to authorized users.

✉ Robyn E. Tornabene
robyn.tornabene@stonybrook.edu

[1] Faculty of Biology, Biology Education, University of Duisburg-Essen, 45141 Essen, Germany

[2] Institute for STEM Education, Stony Brook University, 092 Life Sciences Building, Stony Brook, NY 11747-5233, USA

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 et al. 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 and 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. (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 toward domain-specific models that may be leveraged to improve teaching and learning of core ideas (see Nehm and 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 and 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 et al. 2016; Nehm et al. 2012; Nehm and Ha 2011; Nehm and Reilly 2007; Nehm and Ridgway 2011; Opfer et al. 2012). Some of these situational effects have also been documented in international participants subjected to different educational experiences and cultural contexts (e.g., Ha and 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 and Ridgway 2011). Against this background, it seems reasonable to investigate the impact of situational features on students' reasoning and performance measures for diverse problems, particular (sub-)domains, and in different cultural contexts. Especially the latter is considered of importance to prove findings for stability (cf. Van Bavel et al. 2016).

1.1 Situational Features and the History of Science

In addition to studies of contemporary science learning, research on the history of science (HOS) has great potential for science education research on student conceptions, particularly the development of theoretical frameworks for understanding students' reasoning about biological phenomena (Kampourakis and Nehm 2014). Careful reading of HOS is paramount, as the science education and cognitive psychology communities have in some cases misinterpreted the history of science (for several examples involving Darwin and Lamarck,

see Kampourakis and Nehm 2014, Kampourakis 2015). One of the most valuable aspects of HOS for science educators may be the identification of cognitively challenging phenomena and the examination of how naturalists attempted to make sense of them. Examples from evolution—the degeneration and loss of traits—and genetics—the concept of dominance—both highlight how HOS pointed to special roles for situational features in reasoning about science past and present (e.g., Opfer et al. 2012).

Studying the history of evolutionary biology, Ha and Nehm (2014) focused on Darwin and his contemporaries' conceptual frameworks relating to the gain and loss of phenotypic traits within and among lineages. Their work revealed that naturalists often utilized natural selection to explain the evolutionary gain of traits, but use-disuse inheritance and other ideas to explain instances of trait degeneration and loss. Considerable and long-standing debates emerged concerning the range of phenomena that could be accounted for by the same underlying mechanisms (Ha and Nehm 2014). In many ways, the challenges that naturalists faced in their attempts to use natural selection to explain trait loss was a harbinger of modern-day students' difficulties with the same phenomenon. As expected, the unique challenges explaining patterns of trait loss documented in the historical literature are also evident in modern-day students from different cultures and different educational systems (Ha and Nehm 2014). Findings such as these reinforce the notion that particular biological phenomena are inherently difficult to understand (e.g., trait loss). This work suggests a broader point, namely that situational features can significantly impact accomplished problem solvers, and that particular features (e.g., trait loss) may be more challenging for all learners, regardless of cultural background, educational experiences, and expertise level.

In a study of historic explanations of biological inheritance, Jamieson and Radick (2013) described contrasting perspectives of the concept of dominance presented by Mendel and Weldon. Although Mendel's original German description of dominance was specific to the behavior of a particular trait in a particular context, his work is widely associated with the common perception of dominance as a fixed and universally applicable characteristic of a gene variant (e.g., the yellow pea seed color allele is dominant over green). Using Mendel's example of yellow and green pea seeds (among others), Weldon presented evidence of continuous phenotypic variation and various color outcomes in descendants depending on which parental combinations were crossed. Weldon clarified that what has been commonly attributed as Mendel's presentation of dominance was actually nothing more than the context-specific result of a particular cross. Depending on which organisms were bred and what environmental conditions existed, any version of a trait might appear to be dominant. Overall, applied to modern Mendelian problem solving, the results of Weldon's experiments present dominance as a contextually relative relationship rather than a permanent characteristic of an allele (Jamieson and Radick 2013). It is interesting to note the contrasting roles of context in reasoning about evolution and genetics as revealed by HOS perspectives: In evolution, HOS suggests situational features to may play a superficial role in reasoning about trait gain or loss, whereas in genetics, situational features were shown to be integral to conceptions of dominance yet have been traditionally neglected in common portrayals of Mendel's model as universal.

Viewing contemporary classical genetics instruction from the perspective of HOS shows that dominance, as conventionally portrayed, was drawn from mere snapshots of possible pea plant crosses (which, although reproduced in large numbers, were unnaturally constructed). In effect, a phenomenon which was situational has been misrepresented as universal. Rather than correcting this misrepresentation as additional evidence arose to further delineate the

situational applicability of dominance, new terminologies such as "codominance," "incomplete dominance," and "pleiotropy" have been invented to accommodate the original perspective in different contexts (Jamieson and Radick 2013). Various biologists and educators throughout the last century have cautioned against over-emphasis on dominance and under-emphasis on contextual features such as the environment or particular situations of organisms. Concerns have been raised that such treatment might encourage over-simplified conceptions of gene function and deterministic views of trait inheritance (e.g., Allchin 2005; Dougherty et al. 2011; Jamieson and Radick 2013, 2017; Kampourakis 2017; Smith and Gericke 2015). Despite these insights, common teaching practices and textbooks often present Mendelian concepts of dominance and its variations as universal (Allchin 2005; Castéra et al. 2008; Gericke et al. 2014; Jamieson and Radick 2013). Given the problems associated with simplistic treatment of dominance and the unique role HOS has suggested context might play in thinking across biological domains, it is valuable to examine the degree to which contemporary students consider Mendel's model universal.

Overall, historical studies of scientific reasoning, coupled with contemporary studies of student thinking, have the potential to provide insights into the cognitive challenges inherent to understanding the mechanisms responsible for particular situational features. The history of science has great potential for directing educational attention to challenging concepts and phenomena in the domain of evolution. Much still remains to be understood about how situational features impact student and expert reasoning in biological domains such as genetics.

1.2 Genetics Education

Although a large body of work has explored student difficulties with genetics problem solving (e.g., Collins and Stewart 1989; Shea et al. 2015; Smith 1983; Todd and 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 et al. (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 12- through 16-year-old 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 and 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 and Gelman 2014, p. 234). Like Kargbo et al.'s (1980) and Clough and 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 et al. (2011) found that participants offered more robust genetic explanations for tasks using human examples compared to those using plants and bacteria. Shea et al. (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.

1.3 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 and Gericke 2015). Stewart et al. (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 et al. 2009; Todd and Romine 2016). Despite several conceptual reorganizations, classical transmission genetic problem solving has been retained within these new frameworks, although reformulated to some extent (Todd and 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 and Radick 2013; Smith and Gericke 2015; Todd and 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 and Radick 2013; Smith and 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 and Romine 2016), students often struggle draw connections between Mendelian and molecular aspects of genetics (Lewis and Kattmann 2004; Todd and 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 and Radick 2013), feeding into existing confusion about the concept of dominance (Allchin 2005; Jamieson and Radick 2013), and promoting exaggerated deterministic perspectives on inheritance which can contribute to related social extensions of deterministic views (Gericke et al. 2014; Castéra and Clément 2014; Castéra et al. 2008; Jamieson and Radick 2013). 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 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 1), our study focused on the role of situational features on Mendelian problem solving performance.

1.4 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 and Lehman 1988; Cavallo 1994; Corbett et al. 2010; Gipson et al. 1989; Moll and Allen 1987; Simmons and Lunetta 1993; Slack and Stewart 1990; Smith and 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 et al. (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 our 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

Table 1 Recent assessments of genetic understanding that include the measurement of transmission genetics

Instrument	Target population	Number of items	Taxonomic context
Written Test of Argumentation in Genetics Dilemmas (Zohar and Nemet 2002)	Secondary (Grade 9)	3	Human
Test of Basic Genetics Concepts (Sadler 2003; Sadler and Zeidler 2005)	Undergraduate	7	Human and unspecified[b]
Genetics Concept Inventory (Elrod 2007)	Undergraduate	4	Unspecified[b]
Genetics Literacy Assessment (Bowling et al. 2008)	Undergraduate	3	Human
Genetics Concept Assessment (Smith et al. 2008)	Undergraduate	8	Human
Modern Genetics Learning Progression (Duncan et al. 2009)	Upper elementary–secondary (Grades 5–10)	Included[a]	Not applicable[a]
Genetics Diagnostic Instrument (Tsui and Treagust 2010)	Secondary (Grades 10 and 12)	8	Human, animal, and unspecified[bc]
Biology Concept Inventory (Klymkowsky et al. 2010)	Undergraduate	5	Human and unspecified[b]
Molecular Biology Capstone Assessment (Couch et al. 2015)	Undergraduate	1	Human
Learning Progression-based Assessment of Modern Genetics- Version 2 (Todd and Romine 2016)	Undergraduate	6	Human and plant

[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

theoretical perspective on the construct, our 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 USA 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 et al. 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 USA.

Likewise, in Germany, Mendelian transmission genetics and associated genetic crosses are part of the National Educational Standards for Biology (E11) (KMK 2004). The standards include a particular task relating to Mendelian genetics, specifically, identifying inheritance

patterns for a genetic disorder (galactosemia) using a multigenerational family tree (p. 55f.). Although individual regions (federal states) have some curricular autonomy, the general topic of Mendelian genetics is widespread. In Berlin, for example, students taking the intermediate examination (at the end of grade 9–10) should be able to complete a genetic cross (Senatsverwaltung 2006). In North Rhine-Westphalia, students should be able to apply Mendelian laws to biological examples (MSW NRW 2008). In sum, Mendelian transmission genetics is a widespread curricular topic in both countries we conducted our study (USA and Germany), providing a useful research context for the exploration of situational effects in different educational and cultural settings (for sample selection procedures please see Section 3.2).

1.5 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 and Brownell 2016; Kahle and Meece 1994; Lee and Luykx 2007; Linn and Hyde 1989; Peng et al. 1995; Scantlebury 2014; Scantlebury and Baker 2007; Weinburgh 1995). Differences in science achievement may be due to factors specific to a demographic group (Peng et al. 1995; Scantlebury and Baker 2007) or as a result of bias in curriculum, instructional practices, school climate, or assessment methods (Lee and 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 et al. 2002; Lauer et al. 2013; Schroeders et al. 2013; Shepardson and Pizzini 1994; Willoughby and 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 and Sweeder 2012) and test scores (Eddy et al. 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, our study disaggregates data by gender and ethnicity to examine any potential testing bias or performance discrepancies.

2 Research Question

Our 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, we ask the following research question: Do Mendelian inheritance problems that differ in taxon (animal, plant, human) or familiarity (real, fictitious) produce equivalent measures of student understanding in high school and university participants across genders and ethnic backgrounds?

3 Methods

3.1 Item Design

To answer our research question, we sought to quantify differences in item difficulty (dependent variable) of 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. We relied on prior research to guide our 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 our 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 and 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 and Gelman 2014, p. 233). We therefore focused on developing fictitious taxa and traits that would by definition be novel to participants, and real taxa that participants had been exposed to in their curricula. Given that taxa and familiarity have been shown to have strong influences on biological reasoning, they made sense as a starting point for our experimental work.

Using this framework, we developed a core collection of five multiple-choice item types addressing the Mendelian inheritance mechanisms of complete dominance and incomplete dominance. These topics were chosen given their (1) ubiquity in genetics education, and hence their relevance to educators worldwide (see Sections 1.2 and 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 1), earlier genetics research (e.g., Gipson et al. 1989; Kinnear 1983; Knippels et al. 2005; Slack and Stewart 1990; Smith and 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 and Stewart 1989). In line with the Mendelian inheritance problem types described in Section 1.4, our items represent both types of genetics reasoning described by Stewart (1988), 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 et al. (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 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, we also included fictitious taxa and traits (e.g., fur color of "Amalcho" animals). 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 our core of five types of inheritance problems, we generated a total of 81 items: 35 featuring animals (16 fictitious), 34 featuring plants (16 fictitious), and 12 featuring humans. An example of an item altered to feature

Table 2 Five types of Mendelian inheritance problems used in this study

Mendelian inheritance pattern	Information given	Question posed
Dominant-recessive	Homozygous parental generation (P)	Distribution of first filial generation (F1)
Dominant-recessive	Heterozygous parental generation (P)	Distribution of first filial generation (F1)
Dominant-recessive	Distribution of first filial generation (F1)	Parental generation (P) genotypes
Incomplete dominance	Homozygous parental generation (P)	Distribution of first filial generation (F1)
Incomplete dominance	Heterozygous parental generation (P)	Distribution of first filial generation (F1)

Table 3 An example of variation in situational features for a Mendelian inheritance problem. The core problem remained the same while situational features were altered

Animal (fictitious)	Plant (real)	Human
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?	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?	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?
·All descendants have black fur color	·All descendants have smooth seed shape.	·All descendants have pointed hairlines.
·All descendants have white fur color.	·All descendants have wrinkled seed shape	·All descendants have round hairlines.
·The descendants have an approximate ratio of 1:1 black fur color to white fur color.	·The descendants have an approximate ratio of 1:1 smooth seed shape to wrinkled seed shape.	·The descendants have an approximate ratio of 1:1 pointed hairline to round hairline.
·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 smooth seed shape to wrinkled seed shape.	·The descendants have an approximate ratio of 3:1 pointed hairline to round hairline.
·The descendants have an approximate ratio of 3:1 white fur color to black fur color.	·The descendants have an approximate ratio of 3:1 wrinkled seed shape to smooth seed shape.	·The descendants have an approximate ratio of 3:1 round hairline to pointed hairline.

different taxa and their respective traits is shown in Table 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 2). Further validity evidence is discussed in Sections 3.3 and 5.

Translation from originally German items to English was relatively straightforward given the constrained nature of the Mendelian genetics questions. Two bilingual (German/English) biology educators translated the items into English, and two American biology educators checked the translation for grammatical clarity. The final English version was then reviewed by the bilingual educators for fidelity to the original version.

3.2 Test Administration and Participant Samples

We administered item packets varying in situational features to two large participant samples that differed in educational level (high school, university undergraduate) and cultural backgrounds (American, German). This decision was driven by three main considerations, according to which participant samples should (1) 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 (see Section 1.5), (2) represent a broad range of abilities to allow for a high variance in performance patterns (different educational levels), and (3) be selected from different cultural backgrounds in order to test for the robustness of putative context effects on genetics reasoning. Consequently, our participants were drawn by what may be considered a type of quota sampling as, given the above considerations, random selection would not be appropriate. The sample selection from the American (USA) and European (Germany) background could allow for proving reproducibility in different cultural contexts.

The first participant sample consisted of 444 undergraduates from a large, comprehensive research university in the northeastern USA. The second participant sample consisted of 480 tenth grade participants from Germany. For both groups, an overlapping multi-matrix sampling was used to distribute the items among different test booklets (cf. Sirotnik and 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.

Extra credit was offered to the undergraduate sample for participation in the study, and participants were aware that their performance would not be reflected in their course grade. Participants were biology majors enrolled in the spring semester of an introductory biology course at a research-intensive public university in the northeastern USA. 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, we categorized it as follows: 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%), and 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).

The sample of high school participants was drawn from 20 tenth grade public school classes in Germany. All students had received instruction in genetics, including transmission genetics, during the same school year or the year prior. Students were informed that the test was for research purposes only and would not be part of the class grade. Because only 30 min were provided for the research study, and High School students required more time than the university sample, a subset of 34 genetics items was used, including four out of the five types of Mendelian inheritance problems (cf. Table 2): 16 items featuring animals (7 fictitious), 14 items featuring plants (4 fictitious), and 4 items featuring humans (for details, see the Appendix). The items were spread over

eight paper booklets each containing 13 items. The booklets were assigned randomly to the participants in every class. Overall, 480 participants (54.6% female, 1.3% missing or invalid) with an average age of 15.6 years (SD = 0.7) took part in the study. An average number of 120.2 participants (SD = 2.3) worked on each item. No race or ethnicity information could be collected due to strict privacy protections in Germany. However, participants' gender and self-reported grade in biology was collected in accordance with local research guidelines.

3.3 Rasch Analysis

We used Conquest (Adams et al. 2016) to analyze our data using the Rasch model (Rasch 1960). 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 our analysis, a person had a 50% chance of solving a particular item if that particular item measure is equal to the person measure (Bond and 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.

We estimated item parameters and person abilities 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 et al. 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 and 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). We used 0.7 and 1.3 as cutoff values for the MNSQ parameter to ensure an adequate match between the empirical data and the statistical model (Boone et al. 2014; Bond and 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 and Stone 1979).

3.4 Comparison of Item Difficulties

To compare item difficulties for the different question types and situational features, we used classical statistics and box plots of Rasch scores. The Kruskal-Wallis test (Kruskal and Wallis 1952) and Mann-Whitney *U* test (Mann and 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. We used ANOVAs and Pearson correlation tests of Rasch scores to analyze student performance by demographic group, gender, and associated contextual variables.

4 Results

4.1 Results of University Participants

4.1.1 Rasch Model Fit for University Participants

Our data showed good fit to the Rasch model. However, our initial analysis revealed four items with poor fit: three with so-called underfit (wMNSQ > 1.3) and one with overfit (wMNSQ < 0.7). Therefore, we removed these items 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). The Appendix 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 (Fig. 1). Items are represented by their item number on the right side of the scale (see the Appendix 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.

The Wright map in Fig. 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, we 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.

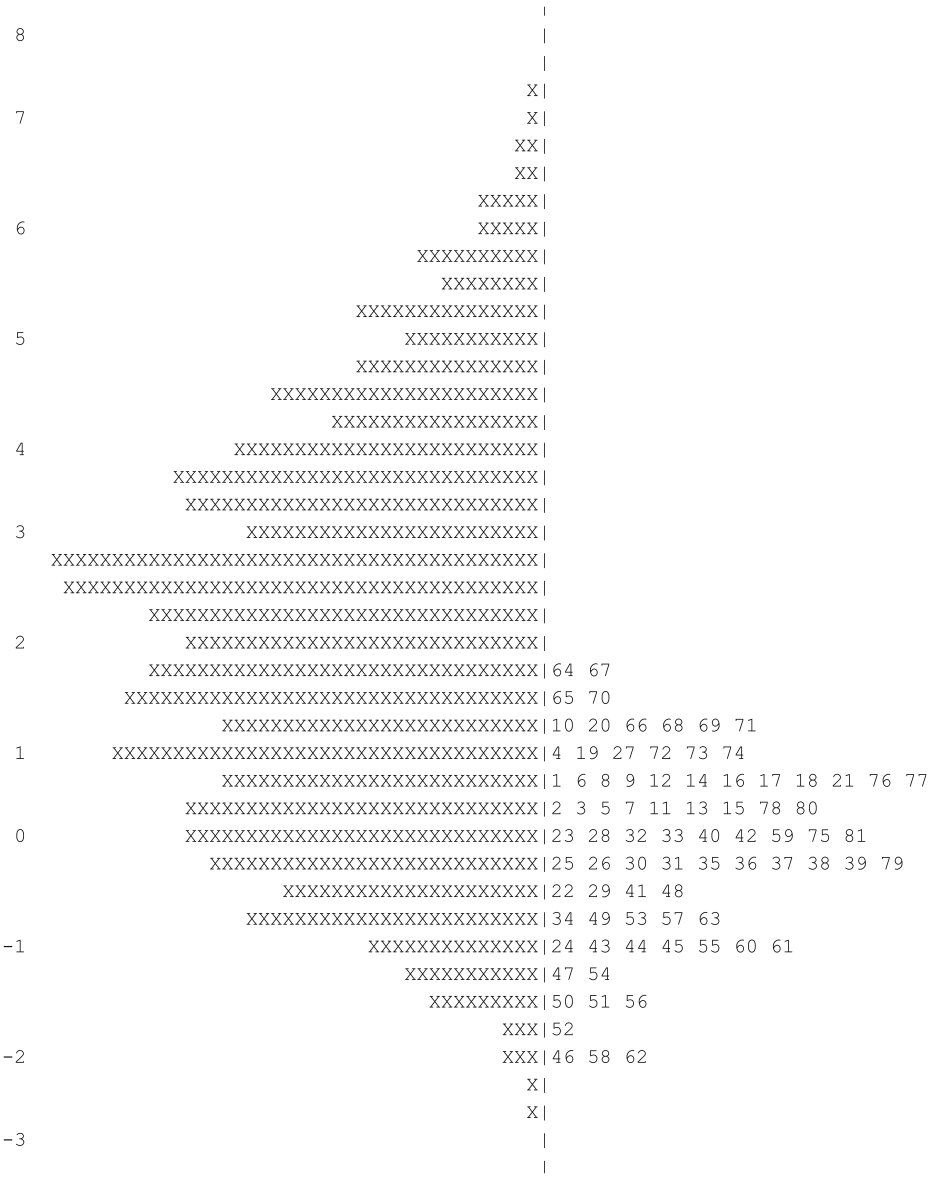

Fig. 1 A Wright map displaying item and person measures on the same logit scale for the US university participants. Each $X = 0.6$ cases. For detailed information about the items, see the Appendix

4.1.2 University Participant Item Difficulties

As expected, there were no significant differences in student performance among the eight test packets ($F(7436) = 0.534$, $p = 0.809$). Box plots (Fig. 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

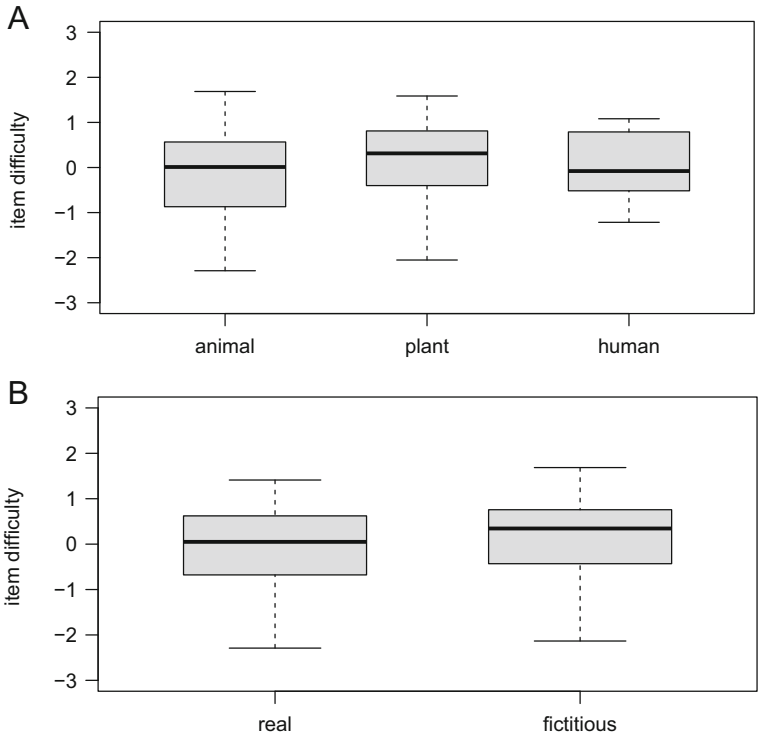

Fig. 2 Boxplots comparing item difficulty by **a** taxon featured **b** real or fictitious taxon featured for items used in US university student sample. The dark bars represent median item difficulty, the boxes represent lower and upper quartile boundaries, and the whiskers represent lowest and highest item measures

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 our undergraduate sample.

In addition to comparing item difficulties by taxon and familiarity, it is useful to compare performance across Mendelian problem types (see Table 2). The boxplots (Fig. 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).

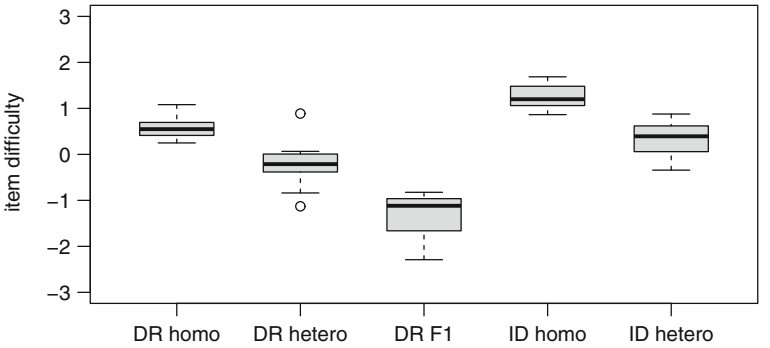

Fig. 3 Boxplots comparing item difficulty by types of Mendelian inheritance problems (cf. Table 2) included in the university student sample. For information about statistical differences, please refer to the text. DR, dominant-recessive inheritance; ID, incomplete dominance; homo, given parental generation is homozygous; hetero, given parental generation is heterozygous; F1, first filial generation given

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.

In addition to testing for situational effects across problem types (see above), we explored 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

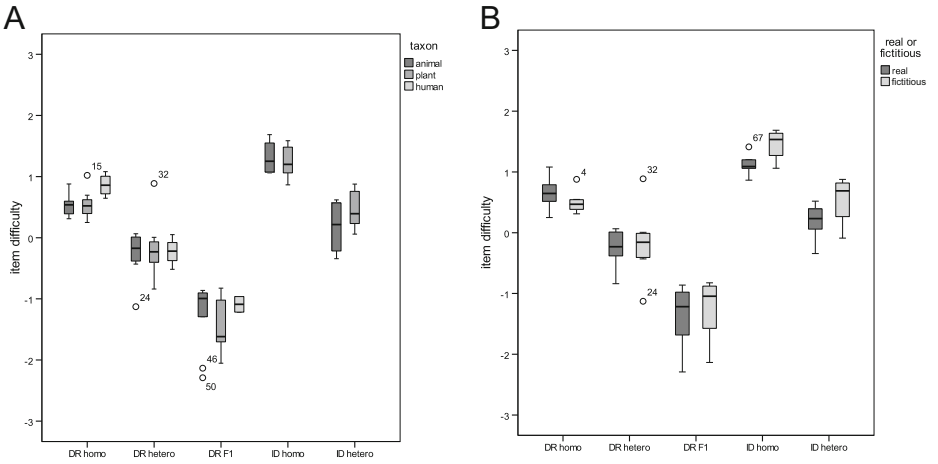

Fig. 4 Boxplots comparing item difficulties by problem types and situational features in the university student sample. **a** Item difficulties grouped by problem type, and shaded by taxon (animal, plant, human). **b** Item difficulties grouped by problem type, shaded by real or fictitious feature. For information about statistical differences, please refer to the text. DR, dominant-recessive inheritance; ID, incomplete dominance; homo, given parental generation is homozygous; hetero, given parental generation is heterozygous; F1, first filial generation given

item feature effects within each of the five item types. The boxplots (Fig. 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 Fig. 4a, item type DR F1 differed in animal, plant, and human features, but produced similar results. Likewise, in Fig. 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.

4.1.3 University Participant Demographics

We found no significant differences in performance ($F(1416) = 1.302$, $p = 0.255$) between male (mean = 1.63) and female (mean = 1.84) participants. Although an ANOVA revealed an overall difference in performance among demographic groups ($F(5412) = 3.155$, $p = 0.008$), post hoc tests did not produce any significant pairwise differences ($p \geq 0.094$). We found a very small negative correlation between performance and age ($r = -0.01$, $p \leq 0.05$). Finally, as one might expect, we found a significant association between performance and number of completed biology courses ($r = 0.27$, $p < 0.01$).

4.2 Results of High School Participants

4.2.1 Rasch Model Fit for High School Participants

Each of the 34 items used for the high school sample was completed by an average of 120.2 participants (SD = 2.3), which is sufficient for Rasch scaling (Hartig and Frey 2013). Regarding item fit statistics, there was no need to remove any of the 34 items ($0.7 \leq$ wMNSQ ≤ 1.3). The Appendix contains a detailed report of all item fit statistics for this sample. The item reliability (WLE reliability = 0.639, EAP/PV reliability = 0.776) is lower than that of the university student sample, but is within the acceptable range. Similar to the university sample, the separation reliability (0.836) and the mean item-total correlation (0.66) were robust. The variance was high (3.434) as well, which indicates that there was a wide range in answer patterns. In comparison to the university sample, the average performance of the high school participants was much lower (− 0.55 logits) although the values are not directly comparable. The average student performance below zero logits indicates that this subsample of items was, as expected, harder for the high school participants to solve. However, the test was not too difficult for this sample as the percentage of correct answers per item ranged from 22.2% and 62.4%.

The Wright map for the high school sample (Fig. 5) displays good person-item alignment, person ability spanning the item difficulty (− 1.42 to + 1.39 logits), and a majority of persons with measures below the mean item difficulty. Despite receiving detailed instructions about how to work on the paper-pencil test, some high school participants skipped single items (mean < 5% per item) or chose more than one answer option (mean ≤ 2% per item). Both were treated as wrong. Because this number was low, no systematic pattern was apparent, and participants had sufficient time to complete the test, this situation is not problematic for the analysis.

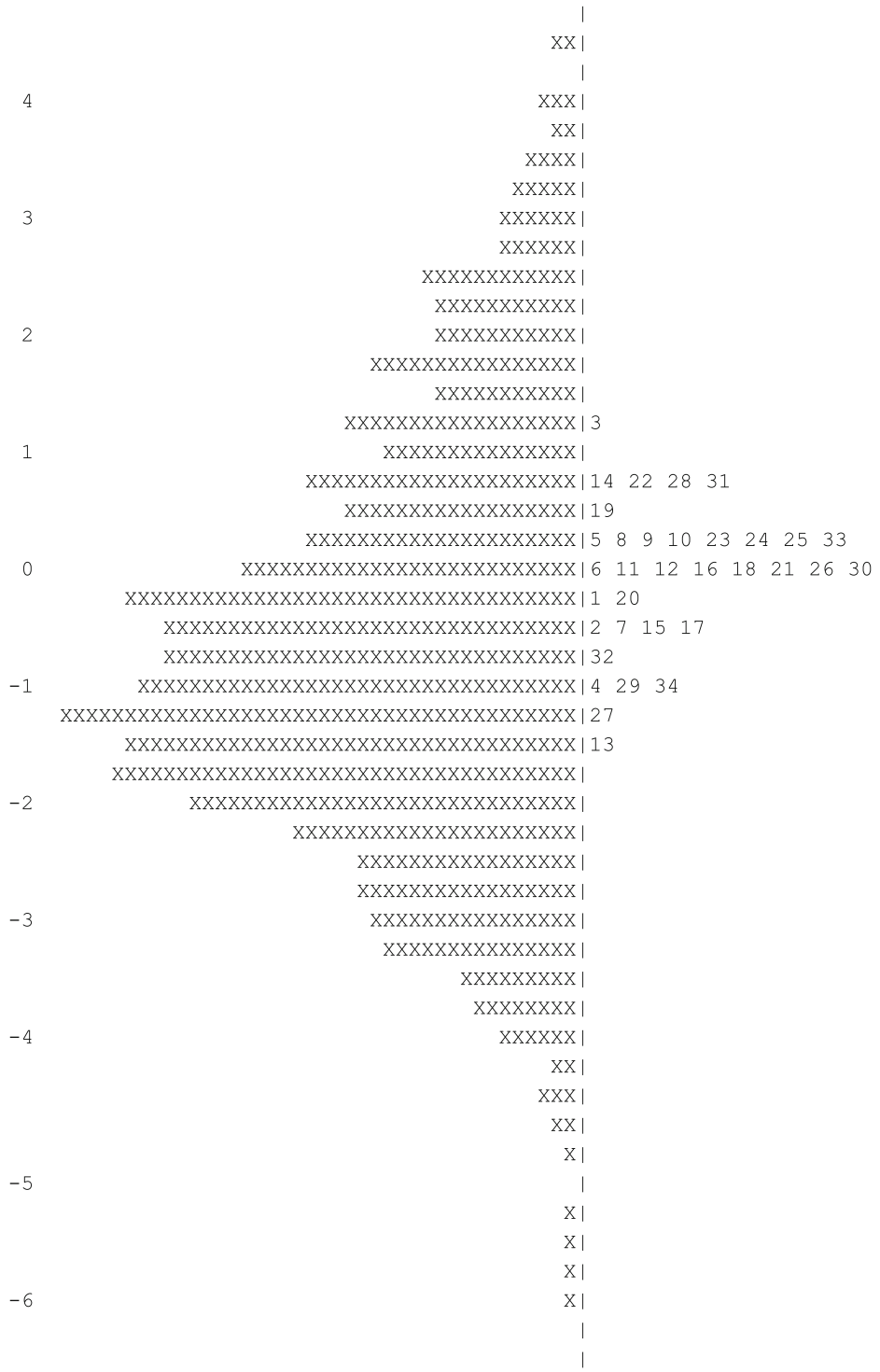
4
3
2
1
0
-1
-2
-3
-4
-5
-6
3
14 22 28 31
19
5 8 9 10 23 24 25 33
6 11 12 16 18 21 26 30
1 20
2 7 15 17
32
4 29 34
27
13

Fig. 5 A Wright map displaying item and person measures on the same logit scale for the German high school participants. Each $X = 0.8$ cases. For detailed information about the items, see the Appendix

4.2.2 High School Participant Item Difficulties

As expected, there were no significant differences in student performance among the eight test booklets administered to the high school participants ($F(7472) = 0.422$, $p = .888$). Boxplots (Fig. 6) and statistical tests revealed no significant differences ($H(2) = 3.68$, $p = 0.158$) in item difficulties among animal (median = 0.11), plant (median = 0.09), and human items (median = 0.34). Additionally, the Mann-Whitney test indicated that no significant differences occurred in item difficulty between real (median = 0.19) and fictitious taxa (median = −0.30), ($U = 103.0$, $p = 0.387$). As with the American university participants, German high school participants' reasoning about the types of inheritance problems posed herein was not significantly impacted by the contextual features of taxon (plant, animal, human) or familiarity (real, fictitious).

The high school student sample items featuring a dominant-recessive (DR) mechanism with a given homozygous ("homo") parental generation (median = 0.22) are more difficult than such items with a given heterozygous ("hetero") parental generation (median = −1.02) (Fig. 6). We found the same pattern for items featuring incomplete dominance (ID) for a given homozygous parental generation (median = 0.11) and heterozygous parental generation

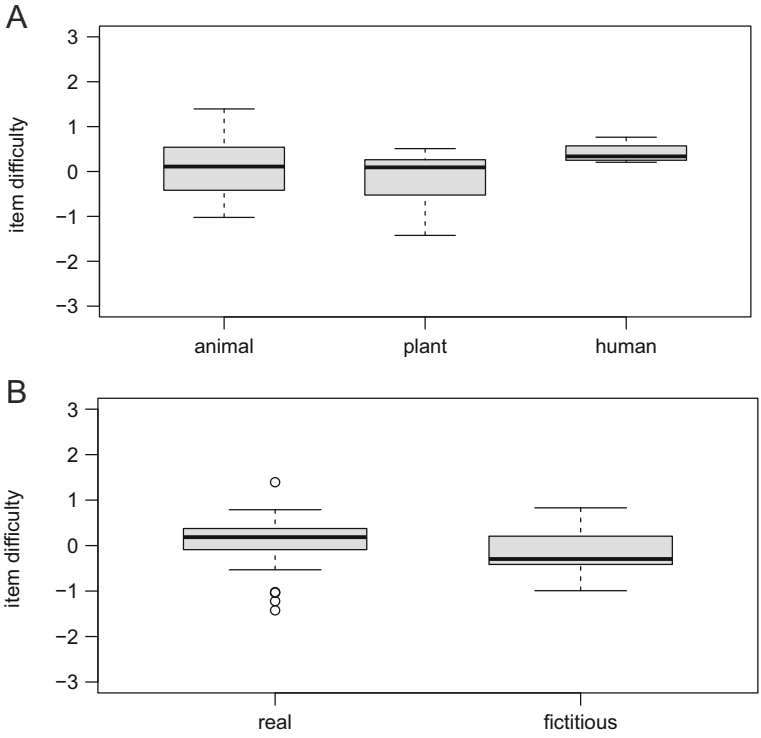

Fig. 6 Boxplots comparing item difficulty by **a** taxon featured **b** real or fictitious taxon featured for items used in German high school student sample. The dark bars represent median item difficulty, the boxes represent lower and upper quartile boundaries, the whiskers represent lowest and highest item measures, and the circles represent outliers

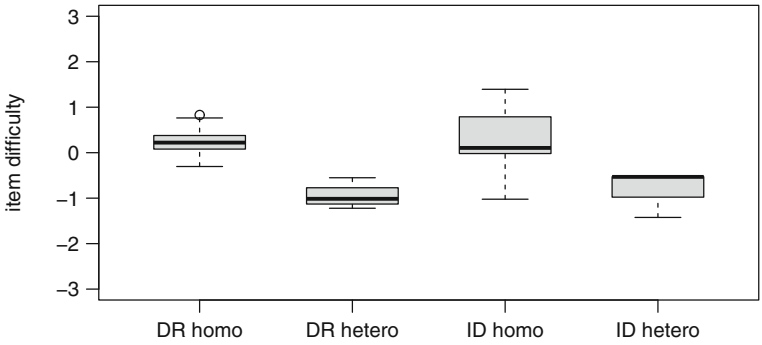

Fig. 7 Boxplots comparing item difficulty by types of Mendelian inheritance problems (cf. Table 2) included in the high school student sample. For information about statistical differences, please refer to the text. DR, dominant-recessive inheritance; ID, incomplete dominance; homo, given parental generation is homozygous; hetero, given parental generation is heterozygous

(median = −0.53). Items with the same given parental generation seemed to be nearly on the same difficulty level (Fig. 7). The Kruskal-Wallis test produced statistical differences overall ($H(3) = 14.96$, $p = 0.002$); however, only two significant differences were found in the pairwise comparisons: Items featuring dominant-recessive (DR) patterns with a given homozygous parental generation were significantly more difficult ($p \leq 0.001$) than items with a heterozygous parental generation irrespective of the pattern (DR hetero, ID hetero). Overall, although we did not find situational effects across the university and high school samples, we *did* find consistent difficulty patterns across Mendelian problem types.

4.2.3 High School Participant Demographics

Similar to the university sample, we found no significant differences in performance ($F(1472) = 2.80$, $p = 0.095$) between male (mean = − 0.56) and female (mean = − 0.29) participants. We found that prior performance in biology was related to performance on the genetics items ($r = -0.37$, $p < 0.001$; note the German grading system provides lower scores to higher performance, hence the negative association).

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 and Ha 2011; Nehm and Reilly 2007; Nehm and Ridgway 2011; Opfer et al. 2012). Similar to these findings in evolution, Clough and 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 et al. (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.

We 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 and 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, our experimental study failed to find situational effects on Mendelian genetics problem solving in large samples of high school and university students in two very different cultural contexts (i.e., USA and Germany). Contrary to Nehm and colleagues' work in the domain of evolution (e.g., Opfer et al. 2012), and Ware and 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.

Our findings raise the question of why we did not find context effects 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.

Our findings 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 our 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 we 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 et al. 1977; Silver 1979) and genetics (Collins 1986; Krajcik et al. 1988; Slack and 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 we

employed were familiar to our high-performing university sample, who had received genetics instruction during the semester and in secondary school. It is also possible that the high school sample was familiar enough with the problem types that they could bypass any impact of surface features, even if they sometimes lacked sufficient expertise to correctly solve the problem. A follow-up to our 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 and 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 and Schonfeld 2008). It is worth noting that Shea et al. (2015) and Kargbo et al. (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 our inferences about situational effects on biological reasoning. Multiple-choice questions and answer options, like the ones used in our 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 our study design cannot reveal the cause(s) of our finding of the 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).

Although our study explored the general topic of Mendelian transmission genetics, we presented participants with several different inheritance problems (see Table 2). Our results indicated that the type of problem impacted item difficulty to a greater extent than situational features (e.g., Figs. 3 and 4). Prior work has suggested that different inheritance problems elicit different cognitive demands. For example, Collins and 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 this prior work, our 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 American secondary schools than incomplete dominance problems, so familiarity may once again partially explain our 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 and Good 1984). For the high school student sample, we could not confirm that incomplete dominance problems were the most difficult. A very likely explanation relates to item familiarity; both problem types (dominant-recessive and incomplete dominance) are typically taught by the end of grade 9/10 (cf. Senatsverwaltung 2006; MSW NRW 2008). Even though traditional teaching sequences begin with dominant-recessive problems, incomplete dominance problems are commonly used thereafter.

There is another interesting pattern regarding the problem types. In both samples, the item with the homozygous parental generation provided are 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 and Lehman 1988; Moll and Allen 1987; Stewart and 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.

Finally, we found no significant influence of gender or ethnicity on Mendelian problem-solving performance. Few genetics studies have considered potential biases in measures of understanding, and none of those that we reviewed (see Table 1) have provided a cross-cultural or multi-level perspective as a source of generalization validity. Notably, our 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 (2008) 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; Schroeders et al. 2013; Shepardson and Pizzini 1994; Willoughby and Metz 2009). While few studies have examined racial or ethnic differences in biology achievement, those that have (Creech and Sweeder 2012; Dimitrov 1999) found no impact, which is in alignment with our findings.

In contrast to the lack of gender and ethnicity effects, we did find significant (but small to moderate) correlations between participants' performance, number of completed biology courses, and biology course marks. These findings provide some convergent validity evidence for our assessment.

6 Limitations and Further Research

Our findings should be viewed in light of several limitations. Considering Rasch analysis results stringently, we note the item reliability of the 34 items administered to the high school participants is acceptable but low (Boone et al. 2014). As our construct was very constrained and based on a linguistically limited set of only five item types with replicas that differed only in surface features, the resulting narrow span of item difficulties is not surprising. In comparison, the answer patterns elicited by the items was broad and was underpinned by the high variance measures. This could be explained by our sampling strategy aiming to gather a broad range of answer patterns and performances. Overall, given that all of the results of Rasch analysis were within acceptable ranges, our interpretation does not appear to be significantly impacted by this perspective.

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 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). In the subset of items used on the high school sample, there was an imbalance in the representation of the four Mendelian problem types, as priority was given to balancing item features consistent with our research question. This imbalance likely explains why we found just two significant differences in item difficulty by problem type. A more balanced distribution—not necessarily a higher number—of items would probably lead to clearer findings for this sample.

Although we used large samples (> 800 participants) and many items (81), we did not find significant differences 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, we did not detect context effects. 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 our test, we conducted supplemental power analyses using G*Power (Faul et al. 2007). To calculate the required effect size necessary to detect an effect with our sample, we used the following constraints: level of significance $\alpha = .05$ (a typical cut-off value in educational research), test power $(1\text{-}\beta) = 0.8$ (following Cohen 1988), and our particular sample sizes and numbers of groups. For our university student sample, for example, our test 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 we 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, our 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 differences we found with our test and, second, the data from our descriptive statistics. If we, for example, compare dominant-recessive items with given homozygous or heterozygous parental generation (DR homo vs. DR hetero; cf. Table 2), we find 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, we could reason that a change in item features would cause medium effects at minimum and would consequently be sensed by our test; still no significant differences for item features were detected. Descriptive data reported in the box plots (Figs. 2, 4, and 6) support this interpretation. The overlap in item difficulties for item groups with very different situational features is very large for both student samples 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 our 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 we have strong evidence to support the claim that the types of Mendelian inheritance problems that we studied are representative of common genetics problems, strictly speaking, our findings are limited to these five problems. We 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 our research focus was on item function and associated item difficulties rather than describing a population of subjects, we chose to utilize a type of quota sampling to obtain participants. Hence, our conclusions about subjects cannot claim global generalizability. Nevertheless, we can assume that our findings will be valid for populations representing similar genetics problem solving experience and similar cultural contexts.

All in all, our findings would be stronger with a larger sets of items per category, particularly in the high school sample. The five types of Mendelian inheritance problems (cf. Table 2) that we 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 our findings are restricted to particular types of problems. The addition of constructed response items to complement our 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 genetic problems students solve heuristically. One might assume that there may be a continuum from problems which can be solved heuristically (like those used in our study) to items that require a deeper application of content knowledge.

Acknowledgements We thank the reviewers for very thoughtful and helpful contributions that have strengthened our work. PS would like to thank Desiree Henning and Shareen Baumann for supporting item development and the participants of "Entwicklung und Evaluation 3" at Freie Universität Berlin for supporting data collection.

Funding Information Financial support for PS was provided by a German Research Foundation (DFG) grant (SCHM 2664/1-1). Financial support for RHN and RET was provided by a National Science Foundation TUES grant (1322872).

Compliance with Ethical Standards Any opinions, findings, conclusions, or recommendations expressed in this publication are those of the authors and do not necessarily reflect the views of the DFG or NSF.

Conflicts of interest The authors declare no conflicts of interest.

References

Adams, R., Wu, M., Macaskill, G., Haldane, S. A., & Sun, X. X. (2016). *ConQuest [computer software]*. Melbourne: Australian Council for Educational Research.

Allchin, D. (2005). The dilemma of dominance. *Biology and Philosophy, 20*(2), 427–451.

Bond, T. G., & Fox, C. M. (2007). *Applying the Rasch model: fundamental measurement in the human sciences* (2nd ed.). Mahwah: Lawrence Erlbaum Associates.

Boone, W. J., Staver, J. R., & Yale, M. S. (2014). *Rasch analysis in the human sciences*. Dordrecht: Springer.

Bowling, B. V., Acra, E. E., Wang, L., Myers, M. F., Dean, G. E., Markle, G. C., Moskalik, C. L., & Heuther, C. A. (2008). Development and evaluation of a genetics literacy assessment instrument for undergraduates. *Genetics, 178*(1), 15–22.

Browning, M. E., & Lehman, J. D. (1988). Identification of student misconceptions in genetics problem solving via computer program. *Journal of Research in Science Teaching, 25*(9), 747–761.

Castéra, J., & Clément, P. (2014). Teachers' conceptions about the genetic determinism of human behaviour: a survey in 23 countries. *Science & Education, 23*(2), 417–443.

Castéra, J., Clément, P., Abrougui, M., Nisiforou, O., Valanides, N., Turcinaviciene, J., … & Carvalho, G. (2008). Genetic determinism in school textbooks: a comparative study conducted among sixteen countries. *Science Education International*, 19(2), 163–184.

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

Chi, M. T., Feltovich, P. J., & Glaser, R. (1981). Categorization and representation of physics problems by experts and novices. *Cognitive Science, 5*(2), 121–152.

Clough, E. E., & Driver, R. (1986). A study of consistency in the use of students' conceptual frameworks across different task contexts. *Science Education, 70*, 473–496.

Cohen, J. (1988). *Statistical power analysis for the behavioral science*. New York: Erlbaum.

College Board. (2015). AP biology course and exam description. https://secure-media.collegeboard.org/digitalServices/pdf/ap/ap-biology-course-and-exam-description.pdf . Accessed 28 Sept 2017.

College Board. (2016). The SAT subject tests student guide. https://collegereadiness.collegeboard.org/pdf/sat-subject-tests-student-guide.pdf . Accessed 28 Sept 2017.

Collins, A. (1986). *Strategic knowledge required for desired performance in solving transmission genetics problems.* (Unpublished doctoral dissertation). University of Wisconsin-Madison, WI.

Collins, A., & Stewart, J. H. (1989). The knowledge structure of Mendelian genetics. *The American Biology Teacher, 51*(3), 143–149.

Corbett, A., Kauffman, L., Maclaren, B., Wagner, A., & Jones, E. (2010). A cognitive tutor for genetics problem solving: learning gains and student modeling. *Journal of Educational Computing Research, 42*(2), 219–239.

Couch, B. A., Wood, W. B., & Knight, J. K. (2015). The molecular biology capstone assessment: a concept assessment for upper-division molecular biology students. *CBE-Life Sciences Education, 14*(1), ar10.

Creech, L. R., & Sweeder, R. D. (2012). Analysis of student performance in large-enrollment life science courses. *CBE-Life Sciences Education, 11*(4), 386–391.

Dimitrov, D. M. (1999). Gender differences in science achievement: differential effect of ability, response format, and strands of learning outcomes. *School Science and Mathematics, 99*(8), 445–450.

Dogru-Atay, P., & Tekkaya, C. (2008). Promoting participants' learning in genetics with the learning cycle. *The Journal of Experimental Education, 76*(3), 259–280.

Dougherty, M. J., Pleasants, C., Solow, L., Wong, A., & Zhang, H. (2011). A comprehensive analysis of high school genetics standards: are states keeping pace with modern genetics? *CBE-Life Sciences Education, 10*(3), 318–327.

Duncan, R. G., Rogat, A. D., & Yarden, A. (2009). A learning progression for deepening participants' understandings of modern genetics across the 5th–10th grades. *Journal of Research in Science Teaching, 46*(6), 655–674.

Eddy, S. L., & Brownell, S. E. (2016). Beneath the numbers: a review of gender disparities in undergraduate education across science, technology, engineering, and math disciplines. *Physical Review Physics Education Research, 12*(2), 020106.

Eddy, S. L., Brownell, S. E., & Wenderoth, M. P. (2014). Gender gaps in achievement and participation in multiple introductory biology classrooms. *CBE-Life Sciences Education, 13*(3), 478–492.

Elrod, S. (2007). Genetics concept inventory. http://bioliteracy.colorado.edu/Readings/papersSubmittedPDF/Elrod.pdf . Accessed 28 Sept 2017.

ETS. (2015). The Praxis study companion-biology: content knowledge. https://www.ets.org/s/praxis/pdf/5235.pdf . Accessed 28 Sept 2017.

Faul, F., Erdfelder, E., Lang, A.-G., & Buchner, A. (2007). G*Power 3: a flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods, 39*, 175–191.

Federer, M. R., Nehm, R. H., & Pearl, D. K. (2016). Examining gender differences in written assessment tasks in biology: a case study of evolutionary explanations. *CBE-Life Sciences Education, 15*(1), ar2.

Franke, G., & Bogner, F. X. (2011). Conceptual change in participants' molecular biology education: tilting at windmills? *The Journal of Educational Research, 104*(1), 7–18.

Freidenreich, H. B., Duncan, R. G., & Shea, N. (2011). Exploring middle school students' understanding of three conceptual models in genetics. *International Journal of Science Education, 33*(17), 2323–2349.

Gericke, N. M., Hagberg, M., dos Santos, V. C., Joaquim, L. M., & El-Hani, C. N. (2014). Conceptual variation or incoherence? Textbook discourse on genes in six countries. *Science & Education, 23*(2), 381–416.

Gipson, M. H., Abraham, M. R., & Renner, J. W. (1989). Relationships between formal-operational thought and conceptual difficulties in genetics problem solving. *Journal of Research in Science Teaching, 26*(9), 811–821.

Ha, M., & Nehm, R. H. (2014). Darwin's difficulties and students' struggles with trait loss: cognitive-historical parallelisms in evolutionary explanation. *Science & Education, 23*(5), 1051–1074.

Hartig, J., & Frey, A. (2013). Sind Modelle der Item-Response-Theorie (IRT) das Mittel der Wahl für die Modellierung von Kompetenzen? [Are models of IRT the choice for the modeling of competencies?] *Zeitschrift für Erziehungswissenschaft* [*Journal of Educational Science*]*, 16*(1), 47–51.

Hickey, D. T., Wolfe, E. W., & Kindfield, A. C. (2000). Assessing learning in a technology-supported genetics environment: evidential and systemic validity issues. *Educational Assessment, 6*(3), 155–196.

Hinsley, D. A., Hayes, J. R., & Simon, H. A. (1977). From words to equations: meaning and representation in algebra word problems. *Cognitive Processes in Comprehension*, *329*.

Hott, A. M., Huether, C. A., McInerney, J. D., Christianson, C., Fowler, R., Bender, H., Jenkins, J., Wysocki, A., Markle, G., & Karp, R. (2002). Genetics content in introductory biology courses for non-science majors: theory and practice. *Bioscience, 52*(11), 1024–1035.

Huppert, J., Lomask, S. M., & Lazarowitz, R. (2002). Computer simulations in the high school: students' cognitive stages, science process skills and academic achievement in microbiology. *International Journal of Science Education, 24*(8), 803–821.

International Baccalaureate Organization. (2014). *Diploma programme biology guide*. Cardiff: Author.

Jamieson, A., & Radick, G. (2013). Putting Mendel in his place: how curriculum reform in genetics and counterfactual history of science can work together. In K. Kampourakis (Ed) *The philosophy of biology: A companion for educators* (pp. 577–595). Springer: Netherlands.

Jamieson, A., & Radick, G. (2017). Genetic determinism in the genetics curriculum. *Science & Education*, 1–30.

Kahle, J. B., & Meece, J. (1994). Research on gender issues in the classroom. In D. E. Gabel (Ed.), *Handbook of research on science teaching and learning* (pp. 542–557). New York: Simon & Schuster Macmillan.

Kampourakis, K. (2015). Distorting the history of evolutionary thought in conceptual development research. *Cognitive Science, 39*(4), 833-837.

Kampourakis, K. (2017). *Making sense of genes*. Cambridge: Cambridge University Press.
Kampourakis, K. and Nehm, R.H. (2014). History and philosophy of science and student explanations and conceptions. In Matthews, M. (ed.) Handbook of the history and philosophy of science in science and mathematics teaching (pp. 377–400). Springer.
Kargbo, D. B., Hobbs, E. D., & Erickson, G. L. (1980). Children's beliefs about inherited characteristics. *Journal of Biological Education, 14*(2), 137–146.
Kinnear, J. (1983). Identification of misconceptions in genetics and the use of computer simulations in their correction. In H. Helms & J. Novak (Eds.), *Proceedings of the international seminar on misconceptions in science and mathematics* (pp. 84–92). Ithaca: Cornell University.
Klymkowsky, M. W., Underwood, S., & Garvin-Doxas, K. (2010). The biological concepts instrument (BCI), a diagnostic tool to reveal student thinking.
KMK. (2004). *Bildungsstandards im Fach Biologie für den Mittleren Schulabschluss. Beschluss der Kultusministerkonferenz (KMK)*. [*National educational standards in biology for the intermediate leaving examination. Resolution of the standing conference of the ministers of education and cultural affairs*]. Munich: Wolters Kluwer.
Knippels, M. C. P., Waarlo, A. J., & Boersma, K. T. (2005). Design criteria for learning and teaching genetics. *Journal of Biological Education, 39*(3), 108–112.
Krajcik, J. S., Simmons, P. E., & Lunetta, V. N. (1988). A research strategy for the dynamic study of students' concepts and problem solving strategies using science software. *Journal of Research in Science Teaching, 25*(2), 147–155.
Kruskal, W. H., & Wallis, W. A. (1952). Use of ranks in one-criterion variance analysis. *Journal of the American Statistical Association., 47*(260), 583–621.
Lauer, S., Momsen, J., Offerdahl, E., Kryjevskaia, M., Christensen, W., & Montplaisir, L. (2013). Stereotyped: investigating gender in introductory science courses. *CBE-Life Sciences Education, 12*(1), 30–38.
Lee, O., & Luykx, A. (2007). Science education and student diversity: race/ethnicity, language, culture, and socioeconomic status. In S. K. Abell & N. G. Lederman (Eds.), *Handbook of research on science education, 1* (pp. 171–197). New York: Routledge.
Lewis, J., & Kattmann, U. (2004). Traits, genes, particles and information: re-visiting students' understandings of genetics. *International Journal of Science Education, 26*(2), 195–206.
Linn, M. C., & Hyde, J. S. (1989). Gender, mathematics, and science. *Educational Researcher, 18*(8), 17–27.
Mann, H. B., & Whitney, D. R. (1947). On a test of whether one of two random variables is stochastically larger than the other. *Annals of Mathematical Statistics, 18*(1), 50–60.
Mayer, R. (2013). Problem solving. In D. Reisberg (Ed.), *Oxford handbook of cognitive psychology* (pp. 769–778). New York: Oxford.
McElhinny, T. L., Dougherty, M. J., Bowling, B. V., & Libarkin, J. C. (2014). The status of genetics curriculum in higher education in the United States: goals and assessment. *Science & Education, 23*(2), 445–464.
Moll, M. B., & Allen, R. D. (1987). Student difficulties with Mendelian genetics problems. *The American Biology Teacher, 49*(4), 229–233.
MSW NRW. (2008). *Kernlehrplan für das Gymnasium. Sekundarstufe I in Nordrhein-Westfalen. Biologie.* [*Core curriculum for the gymnasium. Lower secondary level 1 in North Rhine-Westphalia. Biology*]. Frechen: Ritterbach.
National Research Council. (1996). *National science education standards*. Washington, DC: The National Academies Press.
National Research Council. (2012). *A framework for K-12 science education: practices, crosscutting concepts, and core ideas*. Washington, DC: The National Academies Press.
Nehm, R. H., & Ha, M. (2011). Item feature effects in evolution assessment. *Journal of Research in Science Teaching, 48*(3), 237–256.
Nehm, R. H., & Reilly, L. (2007). Biology majors' knowledge and misconceptions of natural selection. *Bioscience, 57*(3), 263–272.
Nehm, R. H., & Ridgway, J. (2011). What do experts and novices "see" in evolutionary problems? *Evolution Education and Outreach., 4*(4), 666–679.
Nehm, R. H., & Schonfeld, I. S. (2008). Measuring knowledge of natural selection: a comparison of the CINS, an open-response instrument, and an oral interview. *Journal of Research in Science Teaching, 45*(10), 1131–1160.
Nehm, R. H., Beggrow, E. P., Opfer, J. E., & Ha, M. (2012). Reasoning about natural selection: diagnosing contextual competency using the ACORNS instrument. *The American Biology Teacher, 74*(2), 92–98.
NGSS Lead States. (2013). *Next generation science standards: for states, by states*. Washington, DC: The National Academies Press.
Opfer, J., Nehm, R. H., & Ha, M. (2012). Cognitive foundations for science assessment design: knowing what students know about evolution. *Journal of Research in Science Teaching., 49*(6), 744–777.

Pearsall, N. R., Skipper, J. E. J., & Mintzes, J. J. (1997). Knowledge restructuring in the life sciences: a longitudinal study of conceptual change in biology. *Science Education, 81*(2), 193–215.

Peng, S. S., Wright, D., & Hill, S. T. (1995). *Understanding racial-ethnic differences in secondary school science and mathematics achievement (NCES 95-710)*. Washington, DC: U. S. Department of Education.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Copenhagen: Nielsen & Lydiche.

Sadler, T. D. (2003). *Informal reasoning regarding socioscientific issues: the influence of morality and content knowledge*. (Unpublished Doctoral Dissertation). University of South Florida, FL.

Sadler, T. D., & Zeidler, D. L. (2005). The significance of content knowledge for informal reasoning regarding socioscientific issues: applying genetics knowledge to genetic engineering issues. *Science Education, 89*(1), 71–93.

Scantlebury, K. (2014). Gender matters. In N. K. Lederman & S. K. Abell (Eds.), *Handbook of research on science education, 2* (pp. 187–203). New York: Routledge.

Scantlebury, K., & Baker, D. (2007). Gender issues in science education: remembering where the difference lies. In S. K. Abell & N. G. Lederman (Eds.), *Handbook of research on science education, 1* (pp. 31–56). New York: Routledge.

Schroeders, U., Penk, C., Jansen, M., & Pant, H. A. (2013). Geschlechtsbezogene Disparitäten. [Gender-specific disparities]. In H. A. Pant, P. Stanat, U. Schoeders, A. Ropplet, T. Siegele, & C. Pöhlmann (Eds.), *IQB-Ländervergleich 2012. Mathematische und naturwissenschaftliche Kompetenzen am Ende der Sekundarstufe I. [IQB-National assessment studies 2012. Competencies at the end of secondary level I in mathematics and science competencies]* (pp. 249–274). Münster: Waxmann.

Senatsverwaltung für Bildung, Jugend und Sport Berlin (2006). *Rahmenlehrplan für die Sekundarstufe I. Jahrgangsstufe 7–10. Biologie.* [*Core curriculum for lower secondary level. Grades 7 to 10. Biology.*] Berlin.

Settlage, J. (1994). Conceptions of natural selection: a snapshot of the sense-making process. *Journal of Research in Science Teaching, 31*(5), 449–457.

Shea, N. A., Duncan, R. G., & Stephenson, C. (2015). A tri-part model for genetics literacy: exploring undergraduate student reasoning about authentic genetics dilemmas. *Research in Science Education, 45*(4), 485–507.

Shepardson, D. P., & Pizzini, E. L. (1994). Gender, achievement, and perception toward science activities. *School Science and Mathematics, 94*(4), 188–193.

Silver, E. A. (1979). Student perceptions of relatedness among mathematical verbal problems. *Journal for Research in Mathematics Education, 10*(3), 195–210.ibo.

Simmons, P. E., & Lunetta, V. N. (1993). Problem-solving behaviors during a genetics computer simulation: beyond the expert/novice dichotomy. *Journal of Research in Science Teaching, 30*(2), 153–173.

Sirotnik, K., & Wellington, R. (1977). Incidence sampling: an integrated theory for matrix sampling. *Journal of Educational Measurement, 14*(4), 343–399.

Slack, S. J., & Stewart, J. (1990). High school participants' problem-solving performance on realistic genetics problems. *Journal of Research in Science Teaching, 27*(1), 55–67.

Smith, M. U. (1983). *A comparative analysis of the performance of experts and novices while solving selected classical genetics problems.* (Unpublished doctoral dissertation). Florida State University, FL.

Smith, M. U. (1992). Expertise and the organization of knowledge: unexpected differences among genetic counselors, faculty, and students on problem categorization tasks. *Journal of Research in Science Teaching, 29*(2), 179–205.

Smith, M. U., & Gericke, N. M. (2015). Mendel in the modern classroom. *Science & Education, 24*(1–2), 151–172.

Smith, M. U., & Good, R. (1984). Problem solving and classical genetics: successful versus unsuccessful performance. *Journal of Research in Science Teaching, 21*(9), 895–912.

Smith, M. K., Wood, W. B., & Knight, J. K. (2008). The genetics concept assessment: a new concept inventory for gauging student understanding of genetics. *CBE-Life Sciences Education, 7*(4), 422–430.

Soyibo, K. (1999). Gender differences in Caribbean participants' performance on a test of errors in biological labelling. *Research in Science & Technological Education, 17*(1), 75–82.

Stanger-Hall, K. F. (2012). Multiple-choice exams: an obstacle for higher-level thinking in introductory science classes. *CBE-Life Sciences Education, 11*(3), 294–306.

Stewart, J. (1983). Student problem solving in high school genetics. *Science Education, 67*(4), 523–540.

Stewart, J. (1988). Potential learning outcomes from solving genetics problems: a typology of problems. *Science Education, 72*(2), 237–254.

Stewart, J., & Dale, M. (1989). High school students' understanding of chromosome/gene behavior during meiosis. *Science Education, 73*(4), 501–521.

Stewart, J., Cartier, J. L., & Passmore, P. M. (2005). Developing understanding through model-based inquiry. In M. S. Donovan & J. D. Bransford (Eds.), *How students learn* (pp. 515–565). Washington D.C: National Research Council.

Todd, A., & Romine, W. L. (2016). Validation of the learning progression-based assessment of modern genetics in a college context. *International Journal of Science Education, 38*(10), 1673–1698.

Tolman, R. R. (1982). Difficulties in genetics problem solving. *American Biology Teacher, 44*(9), 525–527.

Tsui, C. Y., & Treagust, D. (2010). Evaluating secondary students' scientific reasoning in genetics using a two-tier diagnostic instrument. *International Journal of Science Education, 32*(8), 1073–1098.

Van Bavel, J. J., Mende-Siedlecki, P., Brady, W. J., & Reinero, D. A. (2016). Contextual sensitivity in scientific reproductiblity. *PNAS, 113*(23), 6454–6459.

Ware, E. A., & Gelman, S. A. (2014). You get what you need: an examination of purpose based inheritance reasoning in undergraduates, preschoolers, and biological experts. *Cognitive Science, 38*(2), 197–243.

Warm, T. A. (1989). Weighted likelihood estimation of ability in item response models. *Psychometrika, 54*(3), 427–450.

Weinburgh, M. (1995). Gender differences in student attitudes toward science: a meta-analysis of the literature from 1970 to 1991. *Journal of Research in Science Teaching, 32*(4), 387–398.

Willoughby, S. D., & Metz, A. (2009). Exploring gender differences with different gain calculations in astronomy and biology. *American Journal of Physics, 77*(7), 651–657.

Wright, B. D. (1984). Despair and hope for educational measurement. *Contemporary Education Review, 3*(1), 281–288.

Wright, B. D., & Stone, M. (1979). *Best test design. Rasch measurement*. Chicago: MESA Press.

Wright, C. D., Eddy, S. L., Wenderoth, M. P., Abshire, E., Blankenbiller, M., & Brownell, S. E. (2016). Cognitive difficulty and format of exams predicts gender and socioeconomic gaps in exam performance of students in introductory biology courses. *CBE-Life Sciences Education, 15*(2), ar23.

Zohar, A., & Nemet, F. (2002). Fostering participants' knowledge and argumentation skills through dilemmas in human genetics. *Journal of Research in Science Teaching, 39*(1), 35–62.