CBE—Life Sciences Education
Vol. 15, 1–11, Spring 2016

Article

Examining Gender Differences in Written Assessment Tasks in Biology: A Case Study of Evolutionary Explanations

Meghan Rector Federer,* Ross H. Nehm,† and Dennis K. Pearl‡

*Department of Teaching and Learning and ‡Department of Statistics, Ohio State University, Columbus, OH 43210; †Center for Science and Mathematics Education, Stony Brook University, Stony Brook, NY 11794

Submitted January 29, 2014; Revised November 30, 2015; Accepted December 7, 2015
Monitoring Editor: Jennifer Knight

Understanding sources of performance bias in science assessment provides important insights into whether science curricula and/or assessments are valid representations of student abilities. Research investigating assessment bias due to factors such as instrument structure, participant characteristics, and item types are well documented across a variety of disciplines. However, the relationships among these factors are unclear for tasks evaluating understanding through performance on scientific practices, such as explanation. Using item-response theory (Rasch analysis), we evaluated differences in performance by gender on a constructed-response (CR) assessment about natural selection (ACORNS). Three isomorphic item strands of the instrument were administered to a sample of undergraduate biology majors and nonmajors (Group 1: $n = 662$ [female = 51.6%]; G2: $n = 184$ [female = 55.9%]; G3: $n = 642$ [female = 55.1%]). Overall, our results identify relationships between item features and performance by gender; however, the effect is small in the majority of cases, suggesting that males and females tend to incorporate similar concepts into their CR explanations. These results highlight the importance of examining gender effects on performance in written assessment tasks in biology.

INTRODUCTION

Despite more than 50 yr of research exploring how curriculum choices, pedagogical practices, and assessment techniques can be used to improve women's conceptual understanding of science content (Scantlebury and Baker, 2007), attitudes (Weinburgh, 1995), and participation in science education and science careers (Linn and Hyde, 1989; Clark Blickenstaff, 2005), women continue to be underrepresented in science, technology, engineering, and mathematics (STEM) disciplines (National Science Foundation, 2011). And while research in these areas has wide-reaching, practical implications for science teacher education, development of science curricula and assessments, and students attitudes toward learning and participating in science—to name a few—the gender problem persists.

There have been many factors linked to the potential gender gap in science, particularly in relation to achievement and measures of student achievement. For example, the format and type of assessment have been linked to limitations in ability to assess the depth of student knowledge or to evaluate practices of communicating scientific understanding (e.g., Lee *et al.*, 2011; Liu *et al.*, 2011; Messick, 1995), particularly for performance on verbal and written assessments (e.g., Weaver and Raptis, 2001; Penner, 2003). Similarly, differences in problem-solving strategies and achievement between males and females are well documented in a variety of disciplines (e.g., Hedges and Nowell, 1995; Halpern, 2000; Madsen *et al.*, 2013).

Whether or not a particular science assessment provides measures that are truly representative of students' achievement is another factor that may inform research and practice on gender in science education. A number of studies have examined whether or not the ways we assess student

CBE Life Sci Educ March 1, 2016 15:ar2
DOI:10.1187/cbe.14-01-0018
Address correspondence to: Meghan R. Federer (federer.21@osu.edu).

understanding present an inherent gender bias in the measures of student achievement. Studies comparing assessment format (i.e., multiple choice [MC] vs. constructed response [CR]) have found that females underperform compared with males on MC assessments in math and science (Murphy, 1980; Bolger and Kellaghan, 1990; Garner and Englehard, 1999; Weaver and Raptis, 2001). Other studies have investigated the cognitive mechanisms underlying differences in achievement between males and females (e.g., Bielinski and Davison, 1998; Scheuneman and Garritz, 1990; Harris and Carlton, 1993; Royer *et al.*, 1999; Penner, 2003).

However, despite the large number of factors linked to gender gaps, no single factor has been sufficient to explain the well-documented performance differences between males and females (Madsen *et al.*, 2013). In our study, we focus on two major factors that could influence measures of gender achievement in science education: assessment format and item difficulty.

Gender Achievement and Assessment Format

A number of researchers have questioned whether certain item types have an inherent bias for or against male or female students. Research on a variety of item types (i.e., MC, comparison, constructed response) in mathematics has provided support for differences in gender achievement that favored females for CR items and males for MC items (Burton, 1996; Lane *et al.*, 1996; Garner and Englehard, 1999). In science, similar results have been found, indicating that the inclusion of MC items in assessments creates an advantage for male students (Maccoby and Jacklin, 1974; Murphy, 1980, 1982; Bolger and Kellaghan, 1990; DeMars, 1998, 2000). Importantly, these results suggest that the inclusion of a variety of item types, including CR items, may make assessments more gender equitable (Lane *et al.*, 1996).

A variety of reasons have been proposed to explain the generally accepted finding that males perform better than females on MC- versus CR-type items. For example, some researchers have suggested that females have an advantage on CR-type questions because of superior verbal skills (Maccoby and Jacklin, 1974). Others have identified differences in learning styles as a reason for gender differences in performance. For example, males tend to exhibit greater risk-taking behaviors on MC-type items (Hanna, 1986) and when generalizing to unfamiliar items (Kimball, 1989). One additional hypothesis is that females have an advantage during episodic memory tasks (Herlitz *et al.*, 1997). While both MC and CR tasks can test episodic memory, they do so in different ways (i.e., supported recognition [MC] versus free recall [CR]), which may explain the noted differences in male and female student performances. Despite these findings, recent studies on gender differences in performance on science assessments were not able to attribute these performance differences to any particular cause (e.g., Lawrenz *et al.*, 2001; Weaver and Raptis, 2001). Nonetheless, these results highlight the importance of recognizing that assessment format *can* influence ability measures for particular groups of students in different ways.

Gender Achievement and Cognitive Complexity

Research examining gender and item difficulty interactions in mathematics has suggested that more complex (i.e., higher-difficulty) items favor males, particularly for items that favor problem-solving skills, geometry, and higher-order thinking (e.g., Doolittle and Cleary, 1987; Harris and Carlton, 1993; Bielinski and Davison, 1998). Bielinski and Davison (1998) also found that less complex mathematics items favor females over males. Similarly, in an international comparison of math and science achievement tests, Penner (2003) found that the gender gap is more pronounced on more difficult science items (compared with math items).

Gender differences have also been correlated with serial position in a test, with females and minorities more likely to perform better on items located at the beginning of the assessment and perform worse on (or not respond to) items at the end of an assessment (Becker, 1990; Bielinski and Davison, 1998; Boone, 1998). However, these results are confounded by the arrangement of less difficult items at the beginning of the test and more difficult items at the end. Even so, these results suggest that a comparison of overall performance could result in test bias between groups, as females would have lower scores for more difficult items located at the end of a test.

BIOLOGY ASSESSMENT TOOLS: FORMAT AND COGNITIVE COMPLEXITY

As the science education and research communities attempt to evaluate student performance using a fusion of core concepts and scientific practices (Pellegrino, 2013), it is imperative that the community carefully investigates assessment tools being used to guide instruction or to measure and evaluate students' performances. Recent efforts to understand how different assessment formats and features inform inferences about student understanding have led to new tools and practices that measure more authentic problem-solving performances (e.g., NRC, 2001, 2007; Gitomer and Duschl, 2007; Nehm *et al.*, 2012). However, it is unclear how gender may influence measures of performance on tasks related to scientific practices. The overarching purpose of our study is to explore the relationship between gender and CR explanation performance. Specifically, we use a recently published assessment of evolutionary knowledge involving written explanation tasks to examine three research questions:

1. Are there gender achievement differences on written explanation tasks?
2. To what extent are gender achievement patterns on written assessment influenced by a) item position and b) item features?
3. Do patterns of gender achievement on written explanation tasks predict performance on isomorphic oral interview tasks (i.e., does the format of the explanation task matter)?

METHODS

Participants

To address our research questions, we studied a large sample of undergraduate students enrolled in introductory nonmajors, introductory majors, and advanced majors biology courses at a large, public midwestern research university. In

each of these courses, evolution is presented as a unifying theme and is integrated throughout the content presented in the course. Specific instruction on evolution and natural selection is also addressed in the biology non–majors course, the second introductory biology course for majors (Intro Bio 2), and the advanced majors course (Evolution). The assessments for this study were given toward the beginning of the semester, before specific instruction on evolution in the introductory courses. In addition, while the sample populations for the introductory courses ranged from freshmen to seniors, the level of prior biology exposure was similar, as each introductory course served as a prerequisite for additional biology course work (nonmajors or majors level). Likewise, students enrolled in the advanced majors biology course (Evolution) had, at minimum, completed the majors-level introductory biology sequence.

Each subsample of participants responded to a unique sequence of isomorphic items from the ACORNS (Assessing Contextual Reasoning about Natural Selection) CR instrument (Nehm *et al.*, 2012) and provided consent for the researches to access their demographic information in the university database. While all participants were volunteers, knowledge of the treatment was restricted to a general description of the task they would complete (i.e., respond to a survey related to course content). While this does not eliminate the possibility of volunteer bias, the lack of treatment knowledge and large sample sizes do provide some limitations to this threat to validity (e.g., Brownell *et al.*, 2013).

Stereotype threat is another factor to consider when exploring performance differences between groups. As discussed by Steele (1997), negative stereotypes about intellectual and performance abilities can disrupt the performance of students targeted by the stereotype. This is of particular concern in studies of gender and scientific achievement, as women are outperformed by men in many STEM disciplines. While gender was of primary interest in the current study, participants were not aware of this study focus, nor were they asked to self-report any demographic information, thus limiting aspects of stereotype threat, such as awareness of gender as a research focus, that are well documented in recent studies of gender performance (e.g., Croizet *et al.*, 2001; Miyake *et al.*, 2010; Shapiro and Williams, 2012).

Table 1 summarizes participant characteristics and the item sequences used for each subsample. Demographics of the participant samples were generally representative of the larger student body of the university, with females comprising 56.2% of the sample (university: 51.5%) and an average reported age of 20.2 (±2.2) years. Ethnicity was not reported for all samples and thus was not explored in this study; however, the majority (76.3%) of participants identified as non-Hispanic whites (university: 70%). Multiple rounds of data collection occurred over the course of the 2009–2012 academic years to address the specific research questions discussed above.

Instrument

Explanations were collected using a published CR instrument, the ACORNS (Nehm *et al.*, 2012). The ACORNS is a short-answer, diagnostic instrument built on the work of Bishop and Anderson (1990) that assesses student reasoning about the construct of natural selection (Nehm *et al.*, 2012). The instrument consists of an open-ended, isomorphic framework that prompts: "How would biologists explain how a living species of X *with/without* Y evolved from an ancestral species of X species *without/with* Y?" The isomorphic nature of this instrument is of central importance for its use in the research presented here, as it allows for the construction of

Table 1. Description of item strands and sample populations[a]

		Item features						
Item strand	Item organism (trait)	Polarity	Taxon	Familiarity	Item number	Sequences	Course	*N* (% female)
ACORNS 1 (ELSP)	Snail (teeth)	G	A	F	1	1/2/3/4	Intro Bio 1 (majors)	342 (48.8)
	Prosimian (tarsi)	G	A	U	2	2/1/4/3		
	Elm (winged seeds)	G	P	F	3	3/4/1/2	Intro Bio 2 (majors)	318 (55.5)
	Labiatae (pulegone)	G	P	U	4	4/3/2/1		
ACORNS 2 (MSLG)	Snail (teeth)	G	A	F	1	1/2/3/4	Intro Bio 2 (Majors)	156[b] (55.9)
	Grape (tendrils)	G	P	F	2	2/1/4/3		
	Mouse (claws)	L	A	F	3	3/4/1/2		
	Lily (petals)	L	P	F	4	4/3/2/1		
ACORNS 3 (SRPE)	Snail (poison)	G	A	F	1	1/2/3/4	Intro Bio	273 (52.7)
	Penguin (flight)	L	A	F	2	2/1/4/3	Intro Bio 2 (Majors)	244 (57.3)
	Elm (winged seeds)	G	P	F	3	3/4/1/2		
	Rose (thorns)	L	P	F	4	4/3/2/1	Evolution	123 (58.9)
ACORNS 4 (BRLS)	Bacteria (resistance)	G	–	F	1	1/2/3/4	Evolution	126[b] (59.5)
	Rose (thorns)	L	P	F	2			
	Lily (petals)	L	P	F	3	4/3/2/1		
	Suricata (pollex)	G	A	U	4			

[a]Item features include the trait polarity (G = gain of trait; L = loss of trait), the type of taxa/trait (A = animal; P = plant), and the familiarity of the taxa/trait (F = familiar; U = unfamiliar). Please note that the specified courses for each item strand comprised the sample from which participants were selected for different item sequences. For full instrument details, please see Nehm et al. (2012) or visit www.evograder.org.

[b]Participants also completed the MC CINS.

multiple items that are conceptually similar while differing in the specific scenarios presented (i.e., X and Y represent variable surface features). Thus, it was possible to construct item prompts and sequences suited to the particular research questions addressed in this study. In addition, the conceptual similarity allows for the comparison of performance between different groups, (e.g., gender, ethnicity, academic achievement) without raising concerns about item-level performance stereotypes.

Data Collection

To fully investigate gender differences in performance on CR items, we examined participant responses to four different item strands. As outlined above, the items in each strand were isomorphic in structure but varied with respect to specific item features (e.g., familiarity of taxa trait, trait gain/loss; see Table 1). Therefore, gender differences in patterns of performance were examined for specific features of the items (see *Statistical Analyses* for more details). In addition, item strands were sequenced using a Latin square design (e.g., Holland and Dorans, 2006), allowing for the examination of gender performance patterns across changes in item position.

Items were presented electronically, one at a time, to participants via a survey link accessed through the university course-management system. After an item was completed it could not be viewed again (i.e., participants could not review their prior responses). Following the completion of the ACORNS items, a subset of study participants completed the Conceptual Inventory of Natural Selection (CINS; Anderson *et al.*, 2002), an MC instrument about the construct of natural selection. This allowed for a comparison of performance for CR and MC assessments across gender subgroups. While prior research has examined the effects of *section* ordering (i.e., MC sections before CR sections and vice versa; Weaver and Raptis, 2001), this was not a focus of the current study. Therefore, the ordering of item sections was controlled for in this study, with participants completing all CR items before completing MC items, to avoid influencing students' free recall on the CR explanation tasks.

Prior research on students' evolutionary reasoning has documented the variety of cognitive elements used when constructing evolutionary explanations (Nehm and Schonfeld, 2010). Given the centrality of causal reasoning to scientific explanation, each ACORNS response was scored by tabulating the frequency of normative and nonnormative causal elements using the rubrics of Nehm *et al.* (2010; Table 2). Elements included eight scientifically normative ideas (i.e., key concepts [KCs]) and six nonnormative ideas (i.e., naïve ideas [NIs]) about the construct of natural selection that are widely discussed in the evolution education literature. In addition to tabulating frequency scores, we tallied the number of *different* KCs and NIs used by an individual *across* the item sequence, which refers to the composite measures of key concept diversity (KCD) and naïve idea diversity (NID). All ACORNS responses were independently scored by a minimum of two expert raters; interrater reliabilities exceeded 0.81 (Cohen's Kappa).

Response Verbosity. In addition to measures of concept frequency and diversity, the verbosity of each response was also measured. Response verbosity was determined to be the total number of words included in the response.

Written and Oral Explanation Tasks. A subset of participants also participated in oral interview tasks using ACORNS CR items (see Table 1). Interviews were completed following participation in the written task and consisted of four isomorphic ACORNS items and follow-up questions to clarify student responses. Participant responses to the four ACORNS items were tallied for the number of scientific KCs (e.g., heredity, variation, differential survival) and NIs (e.g., needs, goals, use and disuse) using the scoring rubrics of Nehm *et al.* (2010). The performance scores are analyzed here to examine differences in gender performance patterns in relation to the type of assessment task (i.e., written vs. oral tasks). (For more information on the interview protocols, process, and scoring, see Beggrow *et al.*, 2013.)

Statistical Analyses

Mean raw scores (KCs and NIs), reliability coefficients, and correlation coefficients were calculated for the total sample for each gender subgroup. However, as differences in

Table 2. Sample scoring of ACORNS responses[a]

Causal elements		Sample responses	KC score	NI score
KC1	Causes of variation	Winged seeds probably gave the trees some sort of biological advantage that enabled them to *survive and reproduce more easily* (KC6). Over time the trees *developed these seeds because it aided them* (NI4) in survival and reproduction.	1	1
KC2	Heritability of variation			
KC3	Competition			
KC4	Biotic potential			
KC5	Limited resources			
KC6	Differential survival			
KC7	Change over time			
KC8	Nonadaptive reasoning			
NI1	Need as a goal (teleology)			
NI2	Use and disuse	Natural selection, *the need for* (NI1) more efficient seed dispersal	0	1
NI3	Intentionality			
NI4	Adapt (acquired traits)			
NI5	Energy reallocation			
NI6	Pressure			

[a]Responses were scored for the total number of normative causal (i.e., KCs) and nonnormative causal (i.e., NIs) elements.

performance may reflect a difference in abilities between genders or some feature of the item that causes it to function differently for the two genders, it is important to match the groups statistically with respect to the ability being measured by that item. This allows for the determination of whether the item functions differently between groups because of some property not related to ability. Therefore, to test the gender × item difficulty interactions, item difficulties were estimated on the total sample and then on males and females separately using both conventional item difficulties and the partial credit model (PCM).

The PCM is a unidimensional model in the Rasch family for the analysis of responses recorded in two or more ordered categories (Masters and Wright, 1997). As a member of the Rasch family of models, the PCM enables objective comparisons of person abilities and item difficulties. Rasch person ability and item difficulty estimates were obtained with WINSTEPS software (Linacre, 2013), which uses the joint maximum-likelihood estimate to calculate item parameter and ability estimates that maximize the likelihood of obtaining the observed item responses (for details, see Bond and Fox, 2007). To determine whether ACORNS CR explanation tasks function the same for both gender subgroups, we used differential test function (DTF) and differential item function (DIF) analyses to compare performance on the test overall (DTF) and on common individual items (DIF). Each method of analysis produces item difficulties that can then be statistically compared between gender subgroups. The degree to which DTF/DIF is present in the sample can suggest that inferences about test scores or item scores are more or less valid for a particular group.

There are several statistical models for DIF detection. One of the most widely used is the Mantel-Haenszel procedure (Mantel and Haenszel, 1959). With this procedure, indicators of DIF were used to identify whether each functioning set of items (i.e., strand of four ACORNS items) or an individual item was biased toward one gender group or the other (i.e., items were investigated for signs of interaction with gender). Mean difficulty measures were compared statistically using an independent *t* test. In addition, item position and surface features associated with differential performance patterns were correlated with item difficulties for the total sample and each gender subgroup.

RESULTS

Mean Raw Scores

There was no significant difference in mean KC scores between males and females on individual CR items or on strands of items (i.e., four-item sequence, grouped by item features; Table 3). While KC frequency and diversity tended to be higher for females, they were not significantly different from those produced by males for the same items/strands. A significant difference and moderate effect size in favor of females was found for NI score, but only for one item strand (MSLG; $p < 0.001$, Cohen's $d = -0.58$), with females using fewer types of NIs across their explanation sequence (see Table 1 for item strand definitions). However, similar results were not found for individual items or other item strands (Table 3). Similarly, item position within a strand did not appear to lead to significant differences in performance by gender.

Table 3. Summary statistics for raw KC and NI scores for ACORNS items[a]

	ELSP strand ($n = 680$)	MSLG strand ($n = 184$)	SRPE strand ($n = 640$)
Females			
KC	2.24 ± 1.36	2.66 ± 1.16	1.91 ± 1.33
NI	0.87 ± 0.90	1.30 ± 0.99	1.11 ± 0.95
t	–0.05	0.16	1.23
ES	0.00	0.02	0.10
Males			
KC	2.24 ± 1.30	2.63 ± 1.27	1.78 ± 1.23
NI	0.90 ± 0.90	2.01 ± 1.40	1.11 ± 0.91
t	–0.39	–3.86***	–0.23
ES	–0.03	–0.58	0.00

[a]Total scores represent the average diversity of KCs and NIs for responses to that item strand. *t* = independent *t* test of the group differences between male and female students; ES = effect size as calculated by Cohen's *d*.

***$p < 0.001$.

For further exploration of the potential impact of item-feature combinations on gender achievement differences in CR assessment, DTF and DIF indices were obtained for items in each of the three item strands of the ACORNS tests.

General Performance Trends

Analyses of overall test performance, as measured by the sum total of KCs and NIs for individual items, revealed significant differences in test functioning (DTF) between genders with regard to item familiarity, suggesting that the familiarity of the item scenario may be an important factor for explaining performance differences (Table 4). Specifically, it appears that when the test consists of both *familiar* and *unfamiliar* items (ELSP strand), males tended to incorporate fewer KCs when explaining evolutionary change in *unfamiliar* taxa (i.e., had higher difficulty scores), whereas females tended to incorporate fewer NIs when explaining evolutionary change in *familiar* taxa. However, when the test consisted of all *familiar* items (SRPE strand), the KC trend reversed, with females tending to use fewer KCs than males. Collectively, these results indicate that the few gender differences observed may be a result of relationships between item familiarities within a test and performance by gender.

Cognitive Complexity/Item Difficulty: DIF

Independent analysis of KC and NI scores across the 12 ACORNS items revealed that the vast majority of items did not function differently across genders. However, measures of KCs for two items suggested possible performance differences between males and females: Prosimian (ELSP strand) and Penguin (SRPE strand; see Table 5). Corresponding to the above results, difficulty scores for KCs were higher for males on these items. While both items are about evolutionary change in animal taxa, the item features are otherwise dissimilar. The Prosimian item prompts students to explain evolutionary trait *gain* in *unfamiliar* taxa, whereas the Penguin item prompts students to explain evolutionary trait

Table 4. Differential test functioning for male and female performance on CR item sequences

		KC measures (±SE)				NI measures (±SE)			
Strand	Item	Females	Males	*t*	Cohen's *d*	Females	Males	*t*	Cohen's *d*
ELSP	E	3.51 ± 0.07	3.68 ± 0.08	1.59	–	–3.19 ± 0.1	–3.63 ± 0.12	–2.81**	0.22
	L	3.67 ± 0.07	3.95 ± 0.08	2.63**	0.53	–3.62 ± 0.12	–3.83 ± 0.12	–1.23	–
	S	2.74 ± 0.06	2.75 ± 0.07	0.1	–	–2.85 ± 0.09	–3.15 ± 0.1	–2.22*	0.17
	P	3.42 ± 0.07	3.86 ± 0.08	4.13**	0.33	–3.67 ± 0.12	–3.63 ± 0.12	0.23	–
SRPE	S	2.52 ± 0.07	3.14 ± 0.07	–6.26**	0.50	–3.34 ± 0.12	–3.69 ± 0.1	2.24*	0.17
	P	3.82 ± 0.1	4.07 ± 0.08	–1.95	–	–2.52 ± 0.1	–2.87 ± 0.09	2.60**	0.20
	E	3.23 ± 0.09	3.68 ± 0.08	–3.73**	0.29	–3.91 ± 0.14	–4.49 ± 0.13	3.03**	0.24
	R	3.15 ± 0.08	3.7 ± 0.08	–4.86**	0.38	–2.95 ± 0.11	–3.54 ± 0.1	3.96**	0.31

t = independent *t* test of the group differences between male and female students; ES = effect size as calculated by Cohen's *d*.
*$p < 0.05$.
**$p < 0.01$.

loss in *familiar* taxa. Interestingly, despite males having more difficulty with KCs for these items, achievement differences were not found for NI measures. This suggests that, while the familiarity of taxa and the polarity of trait change may result in gender achievement differences for some CR explanation items, the effect of these item features on measures of performance is somewhat limited.

Corroborating the item feature results, examination of performance across item *positions* by gender revealed no significant differences in gender achievement for this sample of CR explanation tasks. However, this finding may be skewed by the rather inconsistent results presented above for item feature effects, as opposed to a lack of item-sequencing effects on gender achievement differences.

Assessment Format

Response Verbosity. Evaluation of response verbosity across gender subgroups revealed that, on average, explanation length was equivalent between males and females. However, corroborating the results from the preceding section, explanations for particular items (e.g., trait loss in Penguins) were significantly shorter for males compared with females (25.08 ± 0.95 vs. 29.18 ± 1.10 words, respectively; $F(1638) = 7.44$, $p = 0.007$), but these results were not consistent across items or item strands.

Comparison of Performance on MC and CR Items. Comparisons of mean performance scores on the CINS MC and ACORNS CR items revealed no significant gender effects for the KC (CR) and CINS (MC) scores (Table 6). In addition, while KCD scores were significantly correlated with performance on the MC instrument for both males and females, NID scores were not. This finding supports prior research that suggests the CINS MC instrument may not be a valid indicator of students' naïve conceptions about evolutionary change (e.g., Nehm *et al.*, 2012).

Pairwise comparisons of male and females scores were conducted to determine whether females were generally more likely than males to perform better on CR items than MC items (e.g., Garner and Englehard, 1999) and whether males were more likely than females to perform better on MC items than CR items (e.g., DeMars, 2000). Pairs were determined using a randomization algorithm, and our analysis represents all possible pairings of males and females in each sample. For example, for the majors-level biology sample there were 119 females (F) and 93 males (M), resulting in 11,067 pairs to examine. Comparison of CINS scores with ACORNS KCD scores revealed that 1420 of these pairs were of type 1 ($F_{CINS} < M_{CINS}$ but $F_{ACORNS} > M_{ACORNS}$), 1062 of type 2 ($F_{CINS} < M_{CINS}$ but $F_{ACORNS} = M_{ACORNS}$), and 361 of type 3 ($F_{CINS} = M_{CINS}$ but $F_{ACORNS} > M_{ACORNS}$), suggesting that females *are* more frequently ranked higher on CR items than males (i.e., types 1 and 3). However, this difference appears to be related to biology experience/expertise, with a strong majority of pairs for nonmajors, a slight majority of pairs for biology majors, and a slight minority of pairs for advanced biology majors (Table 7). In addition, among nonmajors, females were also ranked higher for a strong majority of pairs for NID on CR items. Overall, this suggests that while females were more frequently ranked higher on CR items (ACORNS)

Table 5. Differential item functioning for measures of male and female performance

			Mean item difficulties (±SE)			
Item strand	Item	Measure	Females	Males	*t*	ES
ELS**P**	Prosimian	KC	3.53 ± 0.07	3.78 ± 0.08	–2.31*	0.18
		NI	–3.78 ± 0.12	–3.56 ± 0.12	–1.34	–
SR**P**E	Penguin	KC	3.86 ± 0.08	4.14 ± 0.10	–2.16*	0.15
		NI	–2.69 ± 0.10	–2.82 ± 0.10	0.97	–

t = independent *t* test of the group differences between male and female students; ES = effect size as calculated by Cohen's *d*.
*$p < 0.05$.

Table 6. Significance test for gender achievement differences for MC and CR assessments[a]

Instrument/sample	Female mean	Score	Male mean	Score	*t* Test	Score
ACORNS	KCD	NID	KCD	NID	KCD	NID
MSLG strand	38.0	21.6	27.5	33.5	0.169	–3.863***
BRLS strand	47.4	11.8	47.0	11.5	0.113	0.130
CINS	Total score		Total score		Total score	
MSLG strand	75.8		71.8		1.474	
BRLS strand	67.5		66.1		0.340	
Correlation coefficients						
MSLG strand	0.295**	–0.026	0.337**	–0.179		
BRLS strand	0.342**	0.004	0.573**	–0.118		

[a]To allow comparison across the three measures of student performance, scores are presented as a percent of total possible. In addition, KCD and NID response scores are used here (as opposed to total KCs or total NIs) to provide a more equivalent measure of comparison with the CINS total scores. *t* = independent *t* test of the group differences between male and female students.

**$p < 0.01$.

***$p < 0.001$.

than MC items (CINS), it may be an artifact of differences in response length between genders on particular items (e.g. the Penguin ACORNS item), leading females to be ranked higher for both KCD and NID.

Written versus Oral Explanation Tasks. In alignment with previous results, there were no significant differences in mean KC or NI scores for written responses for the subsample who also completed oral interviews (grouped by item features; Table 8). Unlike previous subsamples, male respondents tended to use more KCs in their written explanations; however, they also tended to use more NIs. A significant difference and moderate effect size in favor of males was found for KC scores of oral interview items about *animal* taxa ($p < 0.05$, Cohen's $d = –0.40$), with males using more KCs in their explanations for these items (although not necessarily more *types* of KCs). For further exploration of the potential impact of this item feature on gender differences in explanatory practice, DIF indices were obtained for each item in the written and oral item subsets of the ACORNS instrument.

Analysis of the eight ACORNS items used across the written and oral assessments revealed that the majority of items *do not* function differently across gender subgroups and their use of KCs and NIs in evolutionary explanations (Table 9). However, DIF analyses indicated that one item appeared to be influencing the observed difference in gender achievement for oral explanations. Item difficulty measures indicated that females had significantly more difficulty (i.e., included fewer KCs) with the Opossum item than males ($p < 0.001$, Cohen's $d = 0.67$). This item prompted students to explain evolutionary trait *loss* in *familiar* taxa, similar to the Penguin item described above. Overall, these results suggest that item features, such as polarity of trait change (i.e., trait loss), may contribute to some differences in performance between genders on oral explanation tasks.

DISCUSSION

Exploring differences in performance in relation to gender provides many opportunities to improve science learning and assessment. While research over the past 50 yr has identified many patterns of gender performance differences, the source of these differences has not always been clear. Studies examining performance differences in science education have consistently documented that females tend to outperform males on written assessments and that males tend to

Table 7. Pairwise comparisons of male and female performance on CR and MC assessments[a]

	CINS (total) vs. ACORNS (KCD)						CINS (total) vs. ACORNS (NID)					
Comparison	T1	T2	T3	T4	T5	T6	T1	T2	T3	T4	T5	T6
Nonmajors	47	38	12	23	20	14	0	151	0	0	72	0
Majors	1420	1062	361	1133	1164	309	2084	881	376	2269	1167	365
Advanced	456	361	96	483	441	113	620	715	85	634	783	70

[a]Comparisons were made between all possible pairs of male and female performances on the ACORNS and CINS. Comparisons were made using KCD and NID scores for the ACORNS and total score for the CINS. The number of cases represents the number of pairs represented by each category: type 1: $F_{CINS} < M_{CINS}$ but $F_{ACORNS} > M_{ACORNS}$; type 2: $F_{CINS} < M_{CINS}$ but $F_{ACORNS} = M_{ACORNS}$; type 3: $F_{CINS} = M_{CINS}$ but $F_{ACORNS} > M_{ACORNS}$; type 4: $F_{CINS} > M_{CINS}$ but $F_{ACORNS} < M_{ACORNS}$; type 5: $F_{CINS} > M_{CINS}$ but $F_{ACORNS} = M_{ACORNS}$; type 6: $F_{CINS} = M_{CINS}$ but $F_{ACORNS} < M_{ACORNS}$. Shaded cells represents categories in which $F_{ACORNS} \geq M_{ACORNS}$ and $F_{CINS} \leq M_{CINS}$.

Table 8. Summary statistics for raw KC scores and NI for ACORNS items[a]

	Written: MSLG					Oral: OSLG				
	Total[b]	G[c]	L[c]	A[c]	P[c]	Total[b]	G[c]	L[c]	A[c]	P[c]
KCs										
Female	2.61 ± 1.38	3.08 ± 2.06	2.50 ± 1.79	2.50 ± 1.99	3.08 ± 2.17	4.74 ± 1.03	10.6 ± 2.48	6.21 ± 2.29	6.44 ± 2.38	6.94 ± 2.13
Male	2.67 ± 1.33	3.10 ± 2.13	3.05 ± 1.91	2.83 ± 1.98	3.31 ± 2.10	5.07 ± 1.17	10.10 ± 3.52	7.07 ± 2.18	7.38 ± 2.25	7.14 ± 2.36
t	–0.19	–0.35	–1.48	–0.83	–0.53	–1.50	0.989	–1.91	–2.02*	–0.46
ES	–0.04	0.00	–0.29	–0.16	–0.10	–0.29	0.16	–0.38	–0.40	–0.08
NIs										
Female	3.10 ± 0.39	5.35 ± 1.14	5.35 ± 1.05	5.55 ± 1.03	5.16 ± 1.24	3.35 ± 0.72	7.55 ± 1.77	5.26 ± 1.53	5.15 ± 1.32	5.52 ± 1.25
Male	3.14 ± 0.35	5.55 ± 0.70	5.52 ± 0.98	5.57 ± 0.73	5.50 ± 0.80	3.52 ± 0.77	7.67 ± 2.12	5.76 ± 1.24	5.55 ± 1.04	5.76 ± 1.28
t	–0.61	–0.97	–0.85	–0.12	–1.68	–1.13	–0.30	–1.76	–1.64	–0.97
ES	–0.10	–0.21	–0.16	–0.02	–0.32	–0.22	–0.06	–0.35	–0.33	–0.18

[a]Total scores represent the average diversity of KCs and NIs for responses to that item strand. Scores reported for gain (G), loss (L), animal (A), and plant (I) items represent the average total KCs and NIs used in students' explanations for these items. *t* = independent *t* test of the group differences between male and female students; ES = effect size as calculated by Cohen's *d*.

[b]Four items.

[c]Two items.

*$p < 0.05$.

outperform females on MC tasks (e.g., Maccoby and Jacklin, 1974; Murphy, 1982; DeMars, 1998; Weaver and Raptis, 2001). In explaining this generally observed pattern, some have suggested that females perform better on written tasks due to their stronger, on average, verbal skills (Maccoby and Jacklin, 1974). Others have speculated that MC assessments tend to favor the greater risk-taking tendencies and guessing behaviors that have been documented for male students (Rowley, 1974). However, the results of our study suggest that males are not always risk takers, particularly when faced with unfamiliar item scenarios. The reduction in evolutionary KCs in unfamiliar item contexts may be indicative of less risky behavior patterns for CR assessments.

Similarly, female students have been shown to make frequent answer changes on MC tasks, reducing the amount of testing time available for other activities (Skinner, 1983). However, it is likely that the gender gap is due to a multitude of small factors rather than a large single factor that can be controlled for (Madsen *et al.*, 2013).

In seeking to explain differential performance patterns, researchers have turned to differential item functioning (as opposed to mean raw-score comparisons) to examine patterns of strengths and weakness in the population subgroups. Item features associated with differential performance are thought to reflect some aspect of the task difficulty for different groups. Features such as item content,

Table 9. Differential item functioning for oral and written responses to ACORNS items[a]

			DIF measures (±SE)		
Item strand	Item	Measure	Females	Males	*t*
Written: MSLG	Mouse	KC	1.29 ± 0.16	1.11 ± 0.18	0.74
		NI	–1.41 ± 0.21	–1.01 ± 0.24	–1.27
	Snail	KC	1.21 ± 0.16	1.21 ± 0.18	0.00
		NI	–1.10 ± 0.19	–1.01 ± 0.24	–0.29
	Lily	KC	1.21 ± 0.16	0.91 ± 01.8	1.26
		NI	–0.70 ± 0.18	–0.90 ± 0.24	0.67
	Grape	KC	0.43 ± 0.15	0.75 ± 0.18	–1.35
		NI	–0.96 ± 0.19	–0.96 ± 0.24	0.00
Oral: OSLG	Opossum	KC	0.20 ± 0.15	–0.56 ± 0.17	3.29***
		NI	0.56 ± 0.16	0.56 ± 0.19	0.00
	Snail	KC	–0.79 ± 014	–0.68 ± 0.17	–0.51
		NI	0.68 ± 0.16	0.68 ± 0.19	0.00
	Lily	KC	–0.50 ± 0.15	–0.29 ± 0.18	–0.92
		NI	0.50 ± 0.15	0.36 ± 0.19	0.59
	Grape	KC	–0.65 ± 0.17	–0.79 ± 0.14	–0.64
		NI	0.17 ± 0.15	0.57 ± 0.19	–1.65

[a]Response measures are reported as total KCs and total (NIs) for each item. *t* = independent *t* test of the group differences between male and female students.

***$p < 0.001$.

item structure, and cognitive complexity are but a few that have been associated with differential achievement patterns for gender subgroups (Scheuneman and Garritz, 1990).

The present study explored gender differences in students' performance on CR explanation tasks using mean raw scores and DIF patterns. While prior research in science education has suggested that females tend to perform better on CR items, the results of this study revealed no overall differences in gender achievement. However, evaluation of performance patterns associated with specific item features suggested that female respondents may have a slight advantage when confronted with explanation tasks containing unfamiliar surface features. That is, male students tended to incorporate fewer scientifically normative concepts (i.e., KCs) than females for unfamiliar taxa. Conversely, females tended to incorporate more scientifically nonnormative ideas (i.e., NIs) than males for familiar taxa. This is counter to results of gender studies with MC items, for which males are thought to display higher risk-taking behaviors on novel or more difficult items (Rowley, 1974). It is possible that the less frequent incorporation of KCs in responses to unfamiliar taxa is indicative of *lower* risk-taking behavior by males on CR items. One hypothesis for this measured difference may be the level of participant investment in MC versus CR items. In a typical exam setting, there is less value placed on MC items compared with CR items. Therefore, students may actually perceive the level of risk to be *lower* for MC items relative to CR items, increasing the likelihood they will engage in risk-taking behaviors on novel or more difficult items.

Together, these results indicate that the few gender achievement differences detected using the ACORNS instrument may be a result of differences in how males and females interpret and respond to particular combinations of item features. Although there were a few items for which DIF indicated performance differences, differential gender achievement was not a consistent finding. This suggests that measures of explanation performance using the ACORNS instrument are generally comparable across genders. However, similar results may not be found for other measures of evolutionary thinking and deserve empirical scrutiny. Indeed, remarkably little work in biology assessment has investigated DIF in relation to gender.

CONCLUSIONS

In recent years, science educators, researchers, and policy documents have redefined what it means to be "proficient" in science, recognizing that teaching students to become scientifically literate individuals able to make informed decisions about contemporary scientific issues involves more than rote memorization of facts (e.g., NRC, 2001, 2007, 2012). In particular, this new conceptualization highlights the disjuncture that exists "between students' knowledge of science facts and procedures and their understanding of how that knowledge can be applied through the practices of scientific reasoning, argumentation, and inquiry" (Pellegrino, 2013, p. 320). This contemporary perspective of what it means to *know* and *do* science identifies challenges for developing assessments that are supportive of this new model of teaching and learning of science.

Understanding sources of performance bias in science assessment is a major challenge for science education reform. Prior research has documented significant differences in gender achievement on MC and CR items in science (e.g., Garner and Englehard, 1999; Weaver and Raptis, 2001; Madsen *et al.*, 2013), with consistent, slight mean advantages documented for males on MC items and females on CR items (e.g., Burton, 1996). Despite such recognition, the issue of the gender achievement gap has not been solved and much has yet to be determined for CR assessments in biology and their use for evaluating student understanding of core ideas and scientific practices. While the results of this study identified a few individual items for which gender differences occurred, both mean raw scores and DTF/DIF indices suggested that gender achievement differences may be of minimal significance for explanation tasks. Those performance differences that were documented indicate that it was the *features* of the particular items that drove DIF for gender subgroups.

In addition to gender differences in performance by item type, it has been suggested that differences in learning styles allow males to be more successful at generalizing knowledge to unfamiliar problems (Kimball, 1989). However, this is not supported by the results of this study, which found that, while gender effects were minimal, when present, they indicated that males had more difficulty than females on familiar (including more NIs) *and* unfamiliar items (including fewer KCs). However, other features of the items (i.e., trait polarity) may have contributed to these patterns of DIF. For example, both of the familiar items for which DIF was documented prompted students to explain evolutionary trait *loss*, a feature that has previously been identified as more "difficult" (i.e., lower KC scores and higher NI scores) relative to CR items about trait *gain* (e.g., Nehm and Ha, 2011).

LIMITATIONS

Overall, while this study is one of the first to empirically explore gender achievement differences for CR explanation tasks, the findings were not consistent across different item strands (e.g., ELSP versus MSLG) or items with similar features within an item sequence (e.g., taxa, familiarity, trait polarity); therefore, it is difficult to generalize these findings beyond the scope of this study. Furthermore, as participant features such as prior academic ability, course achievement, ethnicity, and language proficiency were not explored in our samples, there may be other contributing variables that explain the observed patterns of performance. Nevertheless, as the ACORNS is one of the few CR instruments that has been examined for gender effects, our study provides a starting point for future research on gender biases, as CR items are used with increasing frequency to evaluate core competencies relating to scientific practices.

ACKNOWLEDGMENTS

Support for M.R.F. was provided by National Science Foundation TUES 1322872 (R.H.N.) and REESE 0909999 (R.H.N.). Any opinions, findings, conclusions or recommendations expressed in this publication are those of the authors and do not necessarily reflect the views of the National Science Foundation.

REFERENCES

Anderson DL, Fisher KM, Norman GJ (2002). Development and evaluation of the conceptual inventory of natural selection. J Res Sci Teach *39*, 952–978.

Becker BJ (1990). Item characteristics and gender differences on the SAT-M for mathematically able youths. Am Educ Res J *27*, 65–87.

Beggrow EP, Ha M, Nehm RH, Pearl D, Boone WJ (2013). Assessing scientific practices using machine-learning methods: how closely do they match clinical interview performance? J Sci Educ Technol *23*, 160–182.

Bielinski J, Davison ML (1998). Gender differences by item difficulty interactions in multiple-choice mathematics items. Am Educ Res J *35*, 455–476.

Bishop B, Anderson C (1990). Student conceptions of natural selection and its role in evolution. J Res Sci Teach *27*, 415–427.

Bolger N, Kellaghan T (1990). Method of measurement and gender differences in scholastic achievement. J Educ Measure *27*, 165–174.

Bond TG, Fox CM (2007). Applying the Rasch model. Fundamental measurement in the human sciences, University of Toledo.

Boone W (1998). Assumptions, cautions, and solutions in the use of omitted test data to evaluate the achievement of under-represented groups in science—implications for long-term evaluation. J Women Minor Sci Eng *4*, 183–194.

Brownell SE, Kloser MH, Fukami T, Shavelson RJ (2013). Context matters: volunteer bias, small sample size, and the value of comparison groups in the assessment of research-based undergraduate introductory biology lab courses. J Microbiol Biol Educ *14*, 176–182.

Burton NW (1996). How have changes in the SAT affected women's math scores? Educ Res *15*, 5–9.

Clark Blickenstaff J (2005). Women and science careers: leaky pipeline or gender filter? Gender Educ *17*, 369–389.

Croizet J-C, Desert M, Dutrevis M, Leyens J-P (2001). Stereotype threat, social class, gender, and academic under-achievement: when our reputation catches up to us and takes over. Soc Psychol Educ *4*, 295–310.

DeMars CE (1998). Gender differences in mathematics and science on a high school proficiency exam: the role of response format. Appl Meas Educ *11*, 279–299.

DeMars CE (2000). Test stakes and item format interactions. Appl Meas Educ *13*, 55–77.

Doolittle AE, Cleary TA (1987). Gender-based differential item performance in mathematics achievement items. J Educ Measure *24*, 157–166.

Garner M, Engelhard G (1999). Gender differences in performance on multiple-choice and constructed response mathematics items. Appl Meas Educ *12*, 29–51.

Gitomer DH, Duschl RA (2007). Establishing multilevel coherence in assessment. In: Evidence and Decision Making. The 106th Yearbook of the National Society for the Study of Education, Part 1, ed. PA Moss, Chicago: National Society for the Study of Education, 288–320.

Halpern DF (2000). Sex Differences in Cognitive Abilities, Mahwah, NJ: Erlbaum.

Hanna G (1986). Sex differences in the mathematics achievement of eighth graders in Ontario. J Res Math Educ *17*, 231–237.

Harris AM, Carlton ST (1993). Patterns of gender differences on mathematics items on the scholastic aptitude test. Appl Meas Educ *6*, 137–151.

Hedges LV, Nowell A (1995). Sex differences in mental test scores, variability, and numbers of high-scoring individuals. Science *269*, 41–45.

Herlitz A, Nilsson L-G, Backman L (1997). Gender differences in episodic memory. Mem Cognit *25*, 801–811.

Holland PW, Dorans NJ (2006). Linking and equating. In: Educational Measurement, 4th ed., ed. RL Brennan, Westport, CT: American Council on Higher Education and Praeger Publishers, 187–220.

Kimball M (1989). A new perspective on women's math achievement. Psychol Bull *105*, 198–214.

Lane S, Wang N, Magone M (1996). Gender-related differential item functioning on a middle-school mathematics performance assessment. Educ Res *15*, 21–27.

Lawrenz F, Huffman D, Welch W (2001). Policy considerations based on a cost analysis of alternative test formats in large scale science assessments. J Res Sci Teach *37*, 615–626.

Lee H-S, Liu L, Linn MC (2011). Validating measurement of knowledge integration in science using multiple-choice and explanation items. Appl Meas Educ *24*, 115–136.

Linacre JM (2013). Winsteps® Rasch Measurement Computer Program, Beaverton, OR: Winsteps.

Linn MC, Hyde JS (1989). Gender, mathematics, and science. Educ Res *18*, 17–27.

Liu OL, Lee H-S, Linn MC (2011). An investigation of explanation multiple-choice items in science assessment. Educ Assessment *16*, 164–184.

Maccoby EE, Jacklin CN (1974). The Psychology of Sex Differences, Stanford, CA: Stanford University Press.

Madsen A, McKagan SB, Sayre EC (2013). Gender gap on concept inventories in physics: what is consistent, what is inconsistent, and what factors influence the gap? Phys Rev Spec Top Phys Educ Res *9*, 020121.

Mantel N, Haenszel MW (1959). Statistical aspects of the analysis of data from retrospective studies of disease. J Natl Cancer Inst *22*, 719–748.

Masters GN, Wright BD (1997). The Partial Credit Model, Handbook of Modern Item Response Theory, New York: Springer, 101–121.

Messick S (1995). Validity of psychological assessment. Am Psychol *50*, 741–749.

Miyake A, Kost-Smith LE, Finkelstein ND, Pollock SJ, Cohen GL, Ito TA (2010). Reducing the gender achievement gap in college science: a classroom study of values affirmation. Science *330*, 1234–1237.

Murphy RJL (1980). Sex differences in GCE examination entry statistics and success rates. Educ Stud *6*, 169–178.

Murphy RJL (1982). Sex differences in objective test performance. Br J Educ Psychol *52*, 213–219.

National Research Council (NRC) (2001). Knowing What Students Know: The Science and Design of Educational Assessment, Washington, DC: National Academies Press.

NRC (2007). Taking Science to School: Learning and Teaching Science in Grades K–8, Washington, DC: National Academies Press.

NRC (2012). A Framework for K–12 Science Education: Practices, Crosscutting Concepts, and Core Ideas, Washington, DC: National Academies Press.

National Science Foundation (2011). Women, Minorities, and Persons with Disabilities in Science and Engineering: 2011 (NSF 11-309), Arlington, VA.

Nehm RH, Beggrow E, Opfer J, Ha M (2012). Reasoning about natural selection: diagnosing contextual competency using the ACORNS instrument. Am Biol Teach *74*, 92–98.

Nehm RH, Ha M (2011). Item feature effects in evolution assessment. J Res Sci Teach *48*, 237–256.

Nehm RH, Ha M, Rector M, Opfer JE, Perrin L, Ridgway J, Mollohan K (2010). Scoring Guide for the Open Response Instrument (ORI) and Evolutionary Gain and Loss Test (ACORNS), Technical Report of National Science Foundation REESE Project 0909999.

Nehm RH, Schonfeld I (2010). The future of natural selection knowledge measurement. J Res Sci Teach *48*, 358–362.

Pellegrino JW (2013). Proficiency in science: assessment challenges and opportunities. Science *340*, 320–323.

Penner AM (2003). International gender × item difficulty interactions in mathematics and science achievement tests. J Educ Psychol *95*, 650–655.

Rowley GL (1974). Which examinees are most favoured by the use of multiple choice tests? J Educ Meas *44*, 423–430.

Royer JM, Tronsky LN, Chan Y, Jackson SG, Marchant HG (1999). Math fact retrieval as the cognitive mechanism underlying gender differences in math achievement test performance. Contemp Educ Psychol *24*, 181–266.

Scantlebury K, Baker D (2007). Gender issues in science education research: remembering where the difference lies. In: Handbook of Research on Science Education, ed. SK Abell and NG Lederman, Mahwah, NJ: Erlbaum, 257–286.

Scheuneman JD, Garritz K (1990). Using differential item functioning procedures to explore sources of item difficulty and group performance characteristics. J Educ Meas *27*, 109–131.

Shapiro JR, Williams AM (2012). The role of stereotype threats in undermining girls' and women's performance and interest in STEM fields. Sex Roles *66*, 175–183.

Skinner NF (1983). Switching answers on multiple-choice questions: shrewdness or shibboleth? Teach Psychol *10*, 220–222.

Steele CM (1997). A threat in the air: how stereotypes shape intellectual identity and performance. Am Psychol *52*, 613–629.

Weaver AJ, Raptis H (2001). Gender differences in introductory atmospheric and oceanic science exams: multiple choice versus constructed response questions. J Sci Educ Technol *10*, 115–126.

Weinburgh M (1995). Gender differences in student attitudes toward science: a meta-analysis of the literature from 1970–1991. J Res Sci Teach *32*, 387–389.