CBE—Life Sciences Education
Vol. 7, 422–430, Winter 2008

Article

The Genetics Concept Assessment: A New Concept Inventory for Gauging Student Understanding of Genetics

Michelle K. Smith, William B. Wood, and Jennifer K. Knight

Department of Molecular, Cellular, and Developmental Biology, University of Colorado, Boulder, CO 80309-0347; and Science Education Initiative, University of Colorado, Boulder, CO 80309

Submitted August 5, 2008; Revised September 19, 2008; Accepted September 22, 2008
Monitoring Editor: Diane Ebert-May

We have designed, developed, and validated a 25-question Genetics Concept Assessment (GCA) to test achievement of nine broad learning goals in majors and nonmajors undergraduate genetics courses. Written in everyday language with minimal jargon, the GCA is intended for use as a pre- and posttest to measure student learning gains. The assessment was reviewed by genetics experts, validated by student interviews, and taken by >600 students at three institutions. Normalized learning gains on the GCA were positively correlated with averaged exam scores, suggesting that the GCA measures understanding of topics relevant to instructors. Statistical analysis of our results shows that differences in the item difficulty and item discrimination index values between different questions on pre- and posttests can be used to distinguish between concepts that are well or poorly learned during a course.

INTRODUCTION

Physics instruction has been improved by the use of carefully developed multiple-choice tests (concept inventories) that examine student conceptual understanding on a narrow set of topics (Hestenes, 1992; Chabay and Sherwood, 1997; Thornton and Sokoloff, 1998; Ding *et al*., 2006). Data collected using these assessments in a variety of physics courses clearly indicate that student learning of these concepts is greater in interactive courses than in traditional courses (Hake, 1998; Crouch and Mazur, 2001; Hake, 2002). Biology educators also have begun to compare the effectiveness of different instructional approaches in their courses. Limited evidence suggests that, as in physics, replacement of traditional lectures with more interactive approaches can result in higher student learning gains (e.g., Udovic *et al*., 2002; Knight and Wood, 2005; Freeman *et al*., 2007). Although some assessment tools are available that could be suitable for widespread comparison of different instructional approaches in biology (Anderson *et al*., 2002; Klymkowsky *et al*., 2003; Garvin-Doxas *et al*., 2007; Bowling *et al*., 2008), more are needed to evaluate the effectiveness of teaching reforms in different subdisciplines of the life sciences, including genetics.

DOI: 10.1187/cbe.08–08–0045
Address correspondence to: Michelle K. Smith (michelle.k.smith@colorado.edu).

For this purpose, we developed and validated the Genetics Concept Assessment (GCA). The GCA consists of 25 multiple-choice questions, designed to be clear, concise, and as free of jargon as possible. The questions assess understanding of a set of basic genetics concepts likely to be taught in courses for both majors and nonmajors. The GCA is designed to be administered at the start of a course as a pretest and at the end of the course as a posttest, to measure student learning gains (Hake, 1998).

In this article, we describe our validation of the GCA through student interviews, pilot testing, and expert review. Statistical analysis of test answers from >600 students at three institutions demonstrates that the GCA has an acceptable range of question difficulty and shows high reliability when taken by two similar populations of students in subsequent semesters. We also describe how the GCA can be used to evaluate which concepts students have learned well and which still cause them persistent difficulties after taking a genetics course.

METHODS

Development of GCA

We are in the process of restructuring our majors and nonmajors genetics courses as part of the Science Education Initiative (SEI; www.colorado.edu/sei) at University of Colorado, Boulder (CU). The SEI is a 5-yr project designed to "transform" the core majors courses in five science departments by introducing proven teaching

practices such as formulation of specific learning goals, increased use of class time for interactive and group learning activities, increased use of formative assessments including "clicker" questions, and pre/post assessment to measure learning gains. We developed the GCA to have a validated, reliable, multiple-choice assessment instrument for evaluating how these changes would affect student learning and conceptual understanding in genetics. Following several examples of assessment development (Hestenes, 1992; Anderson *et al*., 2002; Hufnagel, 2002; Ding *et al*., 2006; Bowling *et al*., 2008), we established a set of learning goals for undergraduate genetics courses by interviewing course instructors and other genetics experts. We then created an assessment consisting of questions that address these learning goals using student-provided distracters, validated these questions through student interviews and expert reviews, and further refined the assessment based on pilot study results. When we began, at least four other assessments designed to measure genetics knowledge were available or under development. However, two of these were not validated by student interviews or input from multiple faculty members (Zohar and Nemet, 2002; Sadler and Zeidler, 2005); and the third, although validated by experts and student focus groups, was designed primarily for testing of nonscience majors (Bowling *et al*., 2008). Furthermore, the statistical evaluation of the third instrument was largely limited to data collected before genetics instruction and was not used to measure learning gains. A fourth assessment reported to be in development (Garvin-Doxas *et al*., 2007) was not yet available for use.

We used a multistep process to develop the GCA (Table 1). To identify question topics, we began by reviewing literature that highlighted common misconceptions in genetics (Venville and Treagust, 1998; Lewis *et al*., 2000a,b,c,d; Marbach-Ad and Stavy, 2000; Wood-Robinson *et al*., 2000; Marbach-Ad, 2001; Tsui and Treagust, 2004; Chattopadhyay, 2005; Orcajo and Aznar, 2005). We also interviewed faculty who teach genetics at CU, asking them to list major concepts in their genetics courses and to provide examples of multiple-choice questions they thought were effective at addressing common student misunderstandings. From this information, we created a series of 25 new questions, each specifically designed to address one or more of the nine major learning goals associated with the CU genetics courses (Table 2). The questions were intended to assess conceptual understanding of the learning goals rather than simple factual recall. Supplemental Material 1 contains two examples of GCA questions. The full set of questions is available upon request (see *Discussion*).

Table 1. Overview of the GCA development process

Multistep process used to develop the GCA
1. Review literature on common misconceptions in genetics
2. Interview faculty who teach genetics, and develop a set of learning goals that most instructors would consider vital to the understanding of genetics
3. Develop and administer a pilot assessment based on known and perceived misconceptions
4. Reword jargon, replace distracters with student supplied incorrect answers, and rewrite questions answered correctly by >70% of students on the pretest
5. Validate and revise the GCA through 33 student interviews and input from 10 genetics faculty experts at several institutions
6. Give current version of the GCA to a total of 607 students in both majors and nonmajors genetics courses at three different institutions
7. Evaluate the GCA by measuring item difficulty, item discrimination, and reliability

Table 2. Course learning goals for the CU majors genetics course and corresponding questions on the GCA

Course learning goal	Question no.[a]
1. Analyze phenotypic data and deduce patterns of inheritance from family histories	1, 11, 13
2. Describe the molecular anatomy of genes and genomes	9, 10, 15, 24
3. Describe the mechanisms by which an organism's genome is passed on to the next generation	7, 8, 16, 17, 25
4. Describe the phenomenon of linkage and how it affects assortment of alleles during meiosis	21, 23
5. Extract information about genes, alleles, and gene functions by analyzing the progeny from genetic crosses	4, 14, 18
6. Describe the processes that can affect the frequency of phenotypes in a population over time	3, 12
7. Compare different types of mutations and describe how each can affect genes and the corresponding mRNAs and proteins	2, 5, 6, 22
8. Apply the results of molecular genetic studies in model organisms to understanding aspects of human genetics and genetic diseases	20
9. Interpret results from molecular analyses to determine the inheritance patterns and identities of human genes that can mutate to cause disease	19

[a] The learning goals associated with each question are those intended by the authors. These associations are supported by expert responses (see Table 3) but have not been further verified through student interviews or other means.

We gave a pilot version of the GCA in spring 2007 to 358 students in the CU Molecular, Cellular, and Developmental Biology (MCDB) majors genetics course at the beginning and end of the semester. To identify jargon that might be unfamiliar, we asked a group of these students to circle any words they did not understand when taking the GCA at the beginning of the course, and we substituted more familiar words in subsequent versions. To make the GCA appropriately difficult for measurement of learning gains, we rewrote questions answered correctly by >70% of students on the pretest, and we reworded rarely chosen distracters (incorrect answers) to make them more plausible.

Validation

During the 2007–2008 academic year, we validated the GCA through 33 student interviews (described in detail below), as well as input from 10 genetics faculty experts at several institutions. After revision based on this feedback and the pilot testing described above, we and several cooperating faculty at other institutions gave the GCA to a total of 607 students in five majors and nonmajors genetics courses in fall 2007 and spring 2008 semesters. Eight instructors were involved in these courses. One of us (J.K.K.) taught the CU nonmajors genetics course; another (M.K.S.) helped to write materials for both CU majors courses. Except for distributing the

GCA, we had no influence on the design or teaching of the two other courses. We describe the analysis of our results below.

Student Interviews

We interviewed 33 volunteers from three different courses at CU, taking care to obtain a diverse group of students. Twenty-one students had completed the majors genetics course within the past year and had earned grades ranging from A to D (9 "A's," 8 "B's," and 4 "C's" or below). We also interviewed two students who had taken the CU nonmajors genetics course (which addresses all of the learning goals in Table 2 except goal 6 and goal 8) and 10 students who had completed one introductory biology course in the Ecology and Evolutionary Biology Department (this course includes a brief section on Mendelian inheritance and population genetics, but it does not explicitly cover any other concepts addressed in the GCA). In total, we interviewed 11 males and 22 females.

We conducted one-on-one interviews with these students, asking them to think aloud as they worked through the questions. In the first five interviews, we gave questions without correct answers or distracters and asked students to provide answers in their own words. In the subsequent interviews, students selected a multiple-choice answer to each question and then explained to the interviewer why they thought their choice was correct and the other answers incorrect. All the authors reviewed the interview transcripts and together revised the GCA to include student-generated correct and incorrect answers in simple language with minimal jargon.

At least 10 students were interviewed on the final versions of all 25 GCA questions. For each question, at least five students chose the right answer using correct reasoning. However, for 11 questions, some students chose the right answer for incomplete or incorrect reasons. For these 11 questions, additional students were interviewed, resulting in an average total of 26 student interviews per question. For each question on the final version of the GCA, at least 86% of these students who chose the right answer did so using correct reasoning.

Of the 78 total distracters on the GCA, 43 were chosen by at least 15% of the students on the pretest. We obtained an average of five student interview explanations for each of the 43 distracters, with the exception of choice a on question 14 (see Supplemental Material 1, example B), which no student chose. We used the multiple reasons for incorrect choices to identify the commonly held misconceptions that lead students to select particular distracters.

During this process, we discovered that interviewing students at different achievement levels was essential for assessment development. In general, students who earned "A's" in a genetics course were able to select the correct answers. More importantly, their explanations of why answers were correct helped us determine whether students were picking the correct answers for the right reasons. Students who earned "B's" and "C's" sometimes revealed persistent misconceptions, and their responses helped us write better distracters for multiple-choice questions. Finally, students who received "D's" often based their answer choices on noncontent clues to the correct answer. For example, one such student chose the correct answer to a question by eliminating other choices that included the words "only" or "never." When students selected the correct answer to a question for a nonscientific reason, we revised the question so the answer could not easily be guessed using such strategies.

Faculty Reviews

To determine whether other faculty who teach genetics would see the GCA as valuable for assessing their students' conceptual understanding of genetics, we presented question ideas and examples to faculty groups at CU. We also asked 10 Ph.D. geneticists (experts) at other institutions to take the GCA online, respond to three queries about each question, and offer suggestions for improvement. A summary of their responses is presented in Table 3. The expert suggestions were primarily to reword a few of the questions to increase precision and eliminate possible ambiguities. Although student interview data indicated no difficulties in interpreting the questions that elicited expert comments, we will incorporate some of these suggestions into future updated versions of the GCA to maximize its perceived usefulness to faculty.

Table 3. Summary of expert responses to three queries about the 25 GCA questions

Subject of query	Agreement of experts >90%, >80%, >70%: No. of questions		
The question tests achievement of the specified learning goal	21	3	1
The information given in this question is scientifically accurate	25	0	0
The question is written clearly and precisely	22	3	0

Large-scale Administration and Analysis

To validate our assessment for use in a variety of instructional situations, increase our student sample size, and decrease any possible effect of having developers of the GCA involved in teaching the assessed students, we arranged to have the GCA given to students in five genetics courses at three different institutions during the 2007–2008 academic year. In the MCDB Department at CU, we administered the GCA in both nonmajors and majors genetics courses. The fall semester nonmajors human genetics course (n = 61 students) has no prerequisites and is taken primarily by nonscience majors. For the majors genetics course, which is taught by different instructors in the fall (n = 107) and spring (n = 321) semesters, the prerequisite is a one-semester course, Introduction to Molecular and Cellular Biology, which has no overlapping learning goals with Genetics except basic understanding of the Central Dogma (DNA → RNA → protein). Genetics is required of all MCDB majors and is also taken by majors in Integrative Physiology and Psychology, as well as by premedical students in various other majors. The GCA also was given in two majors genetics courses at other institutions: a small liberal arts college (n = 30 students) and a large private research university (n = 88 students) during the fall 2007 semester. Students at these two institutions typically take genetics as sophomores, and at both institutions, the prerequisite is a two-semester introductory course that spans molecular to ecosystem biology. Both these introductory courses include units on the Central Dogma and basic Mendelian Genetics.

In all five courses, students took the GCA pretest as a 30-min survey on paper during the first day of class. Course credit was awarded for taking the pretest, but it was not graded. At the end of the course, students were given the identically worded posttest, imbedded as a small proportion of the final exam (the first 25 questions). These questions were graded along with the rest of the exam; students were not aware in advance about inclusion of the GCA questions.

In total, 607 students took both the pre- and the posttest. The actual number of students in each course was higher than reported here, but only students who took both pre- and posttests were included in our analysis. The mean pretest scores, posttest scores, and normalized learning gains ($[100 \times (\text{post} - \text{pre})/(100 - \text{pre})]$; Hake, 1998) were calculated for all five courses. In addition, we evaluated how well pretest, posttest, and normalized learning gain scores correlated with average exam scores in the spring 2008 CU majors course, a large course (n = 321) that included several elements not typically found in a traditional lecture course (e.g., an average of

Table 4. Mean pretest, posttest, and learning gain scores for students, TAs/LAs, and genetics experts

	n[a]	Mean pretest (± SE), %	Mean posttest (± SE), %	Mean learning gain (± SE), %
Students	607	40.5 (± 0.6)	74.0 (± 0.7)	56.7 (± 1.0)
TAs/LAs	18	76.9 (± 3.7)	87.8 (± 3.8)	40.0 (± 12.1)
Genetics experts	10	NA[b]	93.0 (± 5.2)	NA

[a] Number of people who took the GCA. Students were enrolled in either majors or nonmajors genetics courses at three different institutions. TAs and LAs were graduate and undergraduate students, respectively, at CU. Genetics experts from several institutions (see text) who took the GCA are included for comparison.

[b] NA, not applicable.

five clicker questions per class with peer discussion, weekly peer-led study groups that encouraged student-student interaction, and a "help room" staffed by course instructors and teaching assistants [TAs] for >30 h/wk to facilitate problem solving). Because the five courses in which the GCA was administered involved multiple uncontrolled variables (e.g., different institutions, student populations, instructors, and teaching approaches), we did not attempt to compare students' overall course performances, and almost all our analyses are based on pooled data.

Statistical Characterization of Assessment

Assessment instruments are commonly evaluated by statistical tests for several attributes, including item difficulty, item discrimination, and reliability (Adams *et al.*, 2006; Ding *et al.*, 2006). The item difficulty index (P) for a question is calculated as the total number of correct responses (N_1) divided by the total number of responses (N); thus, P is the fraction of correct answers. The item discrimination index (D) measures how well each question distinguishes between students whose total pre- or posttest scores identify them as generally strong or weak. To calculate D for individual questions on either the pre- or the posttest, we divided the 607 students into top, middle, and bottom groups based on their total scores for that test (Morrow *et al.*, 2005) and used the following formula: $D = (N_H - N_L)/(N/3)$, where N_H is number of correct responses by the top 33% of students, N_L is number of correct responses in the bottom 33%, and N is total number of student responses (Doran, 1980).

To evaluate reliability, we compared pretest responses from the CU fall 2007 (n = 107) and spring 2008 (n = 321) MCDB majors courses. Assuming that there was little variation in these two large student populations within the same school year, we chose to use the test-retest method to calculate an r value called the coefficient of stability (Crocker and Algina, 1986). We used the test-retest method rather than an internal measure of reliability such as Cronbach's α, which does not measure the stability of test scores over time (Crocker and Algina, 1986). To determine whether the spread of incorrect and correct answer choices was similar in the two semesters, we also compared the percentages of students who answered each choice for every question using chi-square analysis.

Institutional Review Board (IRB) Protocols

We received approval for administration of pre- and posttests in CU classes (exempt status, protocol 0108.9) and for student interviews (expedited status, protocol 0603.08) from the CU IRB. Data obtained from other institutions contained no student identifiers.

RESULTS

The mean pretest scores, posttest scores, and normalized learning gains for the 607 students from five institutions are listed in Table 4. Ninety-six percent of the students had positive normalized learning gains. At CU, the pre- and posttests also were taken by the graduate TAs and a group of undergraduates who served in the course as learning assistants (LAs), helping to staff the genetics "help room" and leading study groups. On both the pre- and posttests the TAs and LAs, all of whom had had previous genetics course work, performed significantly better (analysis of variance [ANOVA], Tukey's post hoc test, $p < 0.05$) than students in the courses. Genetics experts from a variety of institutions (n = 10) had a mean score that was significantly higher than student pre- and posttest mean scores and TA/LA pretest mean scores (ANOVA, Tukey post hoc test, $p < 0.05$) but not significantly higher than the TA/LA posttest mean score (Table 4).

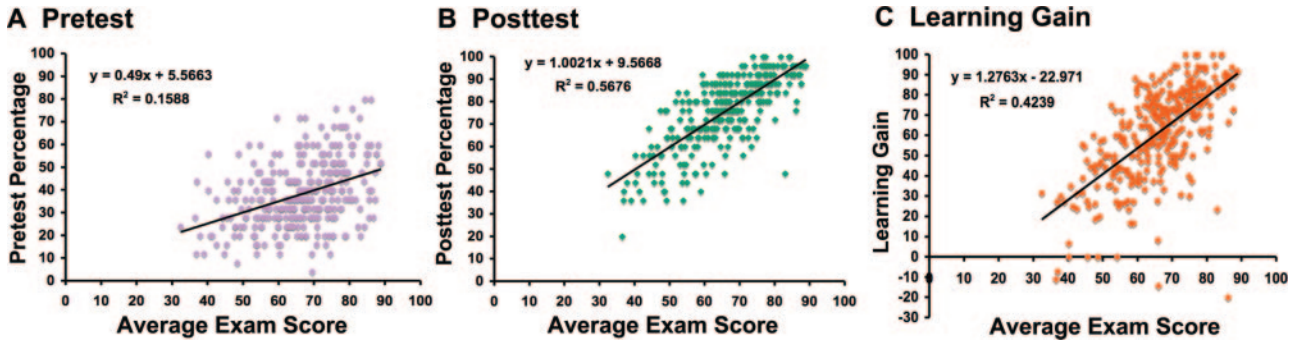

Figure 1. Correlation of pretest score versus average exam score (A), posttest score versus average exam score (B), and learning gain versus average exam score (C). The equation of the trend line and R^2 value are displayed on each graph.

Next, we determined how well pretest, posttest, and learning gain scores correlated with course exam scores for the spring 2008 CU majors course. Because the posttest questions were included on the final exam, we recalculated the average exam score for each student so that it no longer included their score on the 25 GCA questions. Correlations with average exam scores were higher for posttest scores and learning gains than for pretest scores (Figure 1).

Descriptive Statistics

As described above, the item difficulty index (P) for a question is equal to the fraction of students who answered it correctly; that is, P values are high for easy questions and low for difficult questions. Figure 2 shows the pre- and posttest P values for each question on the GCA.

The item D measures how well each question distinguishes "strong" (top 33% of the total scores on the GCA) from "weak" (bottom 33%) students. The higher the D value, the better the question discriminates between strong and weak students (Doran, 1980). Figure 3 shows the D values calculated for each question on both the pretest and the posttest.

Reliability

We measured reliability of the GCA by comparing pretest scores from the CU fall 2007 and spring 2008 majors genetics courses. Students in both courses must take the same prerequisite one-semester introductory course, which is typically their only previous college biology experience. Using the test-retest method as described above, we calculated the mean coefficient of stability for the pretest in these two semesters to be 0.93. The closer the coefficient of stability is to 1, the greater the reliability of the assessment. Although there is no minimum standard for coefficient of stability measurements, values of 0.80–0.90 were reported for commercially available tests (Crocker and Algina, 1986).

We also compared the range of all correct and incorrect answer choices for the fall 2007 and spring 2008 pretests by using chi-square analysis. This analysis helped us to determine how much students from these two semesters differed in their preferences for particular distracters. Only four questions: 8 ($p = 0.029$), 10 ($p = 0.004$), 22 ($p = 0.039$), and 24 ($p = 0.001$) showed a significantly ($p < 0.05$) different spread of answers in the fall and spring pretests.

DISCUSSION

Summary of Results

We developed an assessment instrument in simple language, named the GCA, for gauging student understanding of basic concepts in genetics at the undergraduate level. We validated the GCA through student interviews and expert responses, and we showed that it is reliable when administered in two different courses at the same level (coefficient of stability 0.93). Chi-square analysis revealed that only four questions exhibited a significantly different ($p < 0.05$) spread of distracter choices on pretests in the fall 2007 and spring 2008 semesters for the CU majors genetics course, and interviews indicated that interpretation of these distracters was consistent. The exam scores that students earned in the CU spring 2008 majors genetics course were not well correlated to GCA pretest scores, but were well correlated to both posttest scores and normalized learning gains (Figure 1). We saw similar trends in the other courses that administered the GCA (data not shown). The stronger correlation of posttest scores and normalized learning gains with exam scores suggests that this instrument measures knowledge that is gained during a genetics course and valued by course instructors.

Uses for the GCA

When administered as a pre- and posttest to measure normalized learning gains (Hake, 1998), the GCA can judge student learning in a variety of ways. Like the Force Concept Inventory in physics (Hestenes, 1992) and other similar assessments (see *Introduction*), the GCA can be

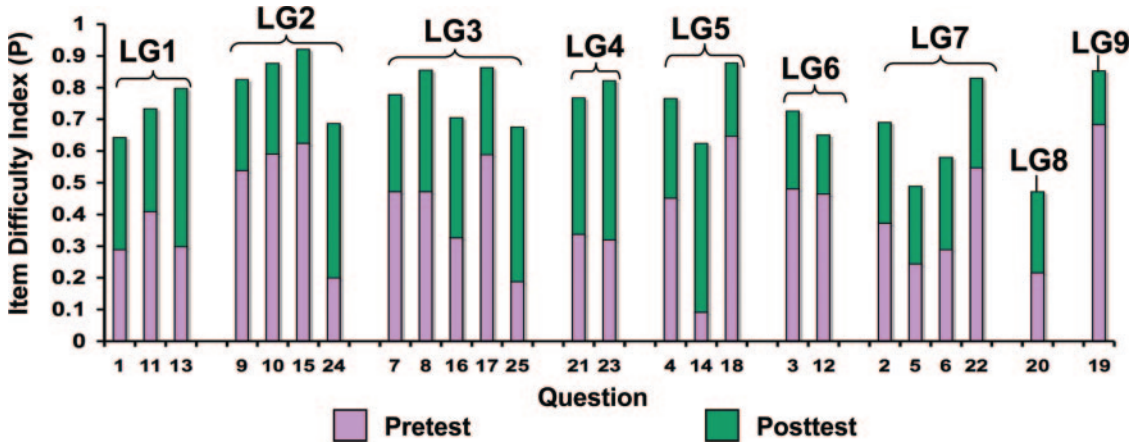

Figure 2. P values for each question on the GCA pretest and posttest. P values represent percentages of correct answers; therefore, lower values indicate more difficult questions. Results are based on combined responses from 607 students in five different genetics courses. Different colored bars show the increase in correct answer percentages between pretest and posttest for each question, indicating the extent of student learning on the corresponding concepts. Questions are grouped according to learning goal (see Table 2).

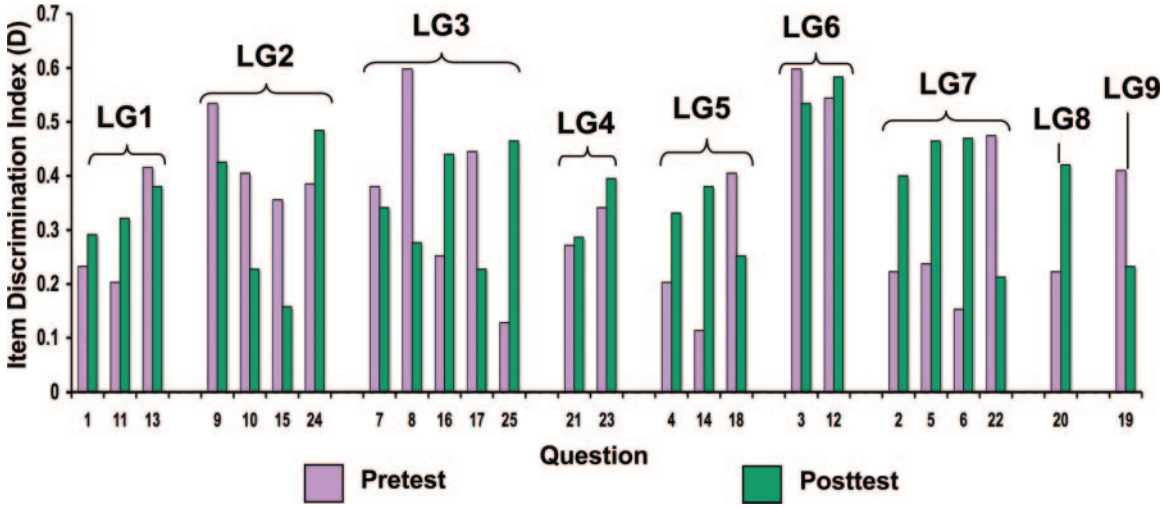

Figure 3. D values for questions on the GCA pretest and posttest. Results were calculated (see *Methods*) from the same data set as in Figure 2. Questions that have higher D values more effectively discriminate between students whose overall test scores identify them as strong or weak students (Doran, 1980). Questions that show high D values on the pretest (only strong students answered them correctly) and low D values on the posttest (most students answered correctly) correspond to concepts on which most students gained understanding during the course. Questions with high D values on both the pre- and posttests correspond to concepts that primarily only the stronger students understood at the end of the course. Questions are grouped according to learning goal (see Table 2).

used to compare student learning outcomes achieved with different modes of instruction (e.g., traditional lecture, lecture with in-class questions and peer discussion, workshop classes built around group problem solving). The mean normalized learning gain for all students provides an overall measure of course effectiveness (e.g., Hake, 1998), whereas the distribution of learning gains achieved by individual students gives a more nuanced picture of instructional impact on the student population (e.g., Knight and Wood, 2005).

For finer-grained evaluation of instructional approaches within a course, the GCA can test the achievement of specific learning goals, assuming that particular questions are assessing achievement of specific goals (Table 2). For example, the gain in understanding of different concepts is evident from the distribution of P values for each question on the assessment (Figure 2), where a higher P value corresponds to a higher percentage of students answering that question correctly. Questions with low P values on the pretest and high P values on the posttest represent concepts for which student understanding improved during the course.

The D values (Figure 3) also can be used to evaluate student learning for different concepts. For example, a high positive D value for a question on the pretest indicates that strong incoming students, as judged by their total assessment score on the pretest, are more likely to understand the corresponding concept than are weak students. If a question still has a high positive D value on the posttest, we can conclude that in spite of instruction, only the stronger students understand the corresponding concept. For example, questions 3 and 12, which we classified under learning goal 6 ("Describe the processes that can affect the frequency of phenotypes in a population over time"), consistently showed high positive D values on both the pre- and posttests (Figure 3). These results suggest that the mechanism by which new alleles arise in a population remains a difficult concept for all but the stronger students.

The extensive student interviews we conducted provided us further insight into some commonly held student misconceptions that are addressed in questions 3 and 12. For question 3, a student who earned a grade of "B+" in majors genetics said that the origin of mutations is not random because "that is too easy, it is just like saying luck is the right answer." When asked about the appearance of a new allele in an isolated population, another student who earned a grade of "A−" in majors genetics said that the environment rather than mutations is probably responsible because "It seems more like an outside factor rather than an inside factor would be important" and "it is so rare to have a mutation, so that is not likely to do anything." High posttest D values for questions 3 and 12 on the GCA and comments during student interviews suggest that many students retain misconceptions about core evolution concepts even after taking a genetics course. In future genetics courses at CU, we will try to dispel these misconceptions through additional instruction coupled with new formative assessments, such as clicker and homework questions.

In contrast to the above-mentioned example, the GCA also can identify concepts on which the majority of students gained understanding. Questions addressing such concepts showed P values that were low on the pretest and high on the posttest. In addition, such questions had D values that were high on the pretest but low on the posttest, indicating that although primarily only the stronger students answered these questions correctly on the pretest, both strong and weak students answered them correctly at the end of the course. Questions 8 and 17, which address learning goal 3 ("Describe the mechanisms by which an organism's genome

is passed on to the next generation") show these characteristics (Figures 2 and 3).

Interviews on these questions (8 and 17) revealed that even students who did not perform well overall in their genetics course (awarded a grade of "C" or below) were able to provide correct answers and adequate explanations on this topic. For example, a "C−" student who correctly answered question 8, which asks about whether a somatic skin cell mutation can be inherited (see Supplemental Material 1), said "Since the mutation is in a single skin cell, it will not occur in his gametes and he will not pass it on to the next generation." Question 17 asks students to consider a diploid germ cell with two pairs of chromosomes whose members carry different alleles (Pp on chromosome pair 1 and Qq on chromosome pair 2) and predict all the possible allele combinations in sperm produced from this cell. A student who received a grade of "D" in the course selected the correct answer in the interview and stated that after normal meiosis "you must have one copy of each chromosome." High pretest and low posttest item D values along with high posttest P values, similar to those observed for questions 8 and 17, can inform instructors that their course was successful in helping both strong and weak students learn the corresponding concepts.

A comparison of P and D values also can suggest that some ways of teaching a concept are more effective than others. For example, for a question on mitochondrial inheritance (question 13), students from two courses (designated courses 1 and 2) had statistically equivalent mean pretest P values (ANOVA, Tukey post hoc test, $p > 0.05$) and similar mean pretest D values (Figure 4). Because D values have no sampling distribution (Crocker and Algina, 1986), it is not possible to determine whether their differences between courses 1 and 2 are significant. However, on the posttest, course 1 had a significantly higher mean P value (ANOVA, Tukey's post hoc test $p < 0.05$) and a lower mean D value than course 2, indicating that a higher fraction of students in course 1 understood the concept at the end of the course. When we compared how this topic was taught and assessed in the two courses (based on classroom observations), we found a clear difference. Instructor 1 discussed mitochondrial inheritance in lecture, formatively assessed understanding in class using a clicker question with peer instruction (Mazur, 1997), and asked about this concept on homework problems as well as on two different exams. Instructor 2 lectured on mitochondrial inheritance but only asked students about this concept on exams. These differences and the resulting performance on this question indicate that the GCA can be useful in evaluating instructional approaches and materials.

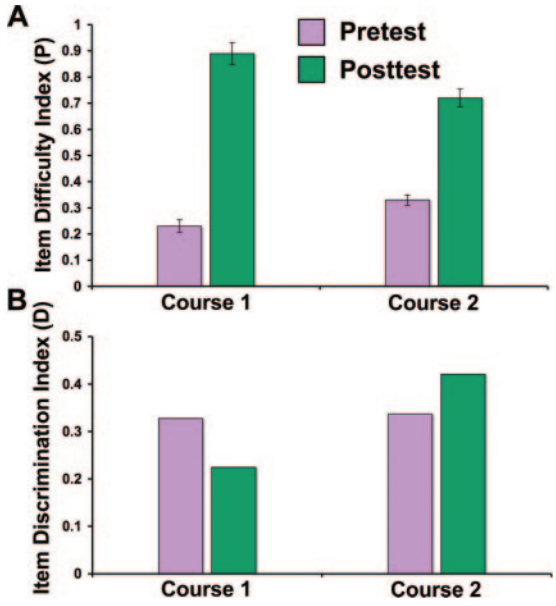

Figure 4. Item difficulty (A) and item discrimination (B) index values for question 13 in two majors genetics courses taught by different instructors. Students in courses 1 and 2 showed similar P and D index values on the pretest, but on the posttest, students in course 1 had a significantly ($p < 0.05$) higher P value and a lower D value compared with students in course 2. These results suggest that the instruction in course 1 was more effective in promoting student learning gains for the concept addressed in this question (see legends to Figures 2 and 3).

Statistical Criteria for Utility

Some of the statistical criteria used by psychometricians to gauge the usefulness of standardized tests must be viewed differently for assessments such as the GCA and other concept inventories. For standardized assessments such as the Scholastic Aptitude Test (SAT), it is considered desirable for D values to be ≥0.3 for all questions (Doran, 1980). Another commonly calculated statistical parameter, the point biserial coefficient (r_{pbs}), also indicates how well performance on each question correlates with performance on the test as a whole, and values ≥0.2 for all questions are considered desirable (Kline, 1986). However, these guidelines are useful only for design of assessments such as the SAT, where all questions are testing essentially the same set of skills at one point in time. In contrast, the questions on the GCA are specifically intended to assess understanding of many different concepts, both before and after instruction, so that instructors can evaluate success in achieving a variety of different learning goals. For example, GCA questions 8 and 17 mentioned above had D values above the cut-off value of 0.3 for the pretest but below this value for the posttest, which would disqualify them from inclusion on the assessment if we were to apply accepted psychometric standards. However, as demonstrated above, it is precisely this difference between D values on the pre- and posttests that can provide useful information about student learning of the corresponding concepts as well as identifying areas where improved instruction may be needed. Because the standard interpretation of point biserial coefficient cut-offs is similarly inapplicable to assessments such as the GCA, we have not reported r_{pbs} values for our questions.

Conclusion, Future Directions, and Accessibility of Materials

The GCA is a validated, reliable conceptual assessment instrument that can be used by genetics instructors to measure

students' learning gains as well as to identify strengths and weaknesses in teaching approaches for specific concepts. Due to the concern that circulation of the GCA among students might decrease its value to instructors, we have not published the full set of questions in this article. However, we hope that appropriate use of the GCA will become widespread, and we will supply the complete set of questions, with answers, on request. Interested instructors should contact M.K.S.

We welcome comments and suggestions from GCA users, and we will continue to revise the GCA as we obtain feedback. We will review all suggestions and validate any significant revisions by student interviews before they are incorporated into a new version of the GCA. At least eight institutions have agreed to test updated versions of the GCA during the 2008–2009 academic year. Analysis of the combined data from these and additional future tests will be posted online at www.colorado.edu/sei/departments/mcdb_assessment.htm.

ACKNOWLEDGMENTS

We thank Carl Wieman, Kathy Perkins, and Wendy Adams of the SEI for intellectual support throughout this project, guidance in statistical analysis of our results, and helpful comments on the manuscript. Sylvia Fromherz and Mark Winey provided ideas for the original questions on the pilot version of the assessment. Special thanks to Christy Fillman, Sylvia Fromherz, Ken Krauter, Robyn Puffenbarger, Ronda Rolfes, Tin Tin Su, and Mark Winey for administering the assessment in courses and sharing data. We also thank our GCA expert reviewers for valuable comments on the assessment questions. Finally, we are grateful to the Science Education Initiative of CU for full support of M.K.S. and partial support of J.K.K. during this project.

REFERENCES

Adams, W. K., Perkins, K. K., Podolefsky, N. S., Dubson, M., Finkelstein, N. D., and Wieman, C. E. (2006). New instrument for measuring student beliefs about physics and learning physics: the Colorado learning attitudes about science survey. Phys. Rev.-PER *2*, 010101.

Anderson, D. L., Fisher, K. M., and Norman, G. J. (2002). Development and evaluation of the conceptual inventory of natural selection. J. Res. Sci. Teach. *39*, 952–978.

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

Chabay, R., and Sherwood, B. (1997). Qualitative understanding and retention. AAPT Announcer *27*, S12.

Chattopadhyay, A. (2005). Understanding of genetic information in higher secondary students in northeast India and the implications for genetics education. Cell Biol. Educ. *4*, 97–104.

Crocker, L., and Algina, J. (1986). Introduction to Classical and Modern Test Theory, New York: Holt, Rinehart, and Winston.

Crouch, C. H., and Mazur, E. (2001). Peer Instruction: ten years of experience and results. Am. J. Phys. *69*, 970–977.

Ding, L., Chabay, R., Sherwood, B., and Beichner, R. (2006). Evaluating an electricity and magnetism assessment tool: brief electricity and magnetism assessment. Phys. Rev.-PER *2*, 010105.

Doran, R. (1980). Basic Measurement and Evaluation of Science Instruction, Washington, DC: National Science Teachers Association.

Freeman, S., O'Connor, E., Parks, J. W., Cunningham, M., Hurley, D., Haak, D., Dirks, C., and Wenderoth, M. P. (2007). Prescribed active learning increases performance in introductory biology. CBE Life Sci. Educ. *6*, 132–139.

Garvin-Doxas, K., Klymkowsky, M., and Elrod, S. (2007). Building, using, and maximizing the impact of concept inventories in the biological sciences: report on a National Science Foundation sponsored conference on the construction of concept inventories in the biological sciences. CBE Life Sci. Educ. *6*, 277–282.

Hake, R. R. (1998). Interactive-engagement versus traditional methods: a six-thousand-student survey of mechanics test data for introductory physics courses. Am. J. Phys. *66*, 64–74.

Hake, R. R. (2002). Lessons from the physics education reform effort. Conserv. Ecol. *5*, 28

Hestenes, D. (1992). Force concept inventory. Phys. Teach. *30*, 141–158.

Hufnagel, B. (2002). Development of the astronomy diagnostic test. Astron. Educ. Rev. *1*, 47–51.

Kline, P. (1986). A Handbook of Test Construction: Introduction to Psychometric Design, London, United Kingdom: Methuen.

Klymkowsky, M. W., Garvin-Doxas, K., and Zeilik, M. (2003). Bioliteracy and teaching efficacy: what biologists can learn from physicists. Cell Biol. Educ. *2*, 155–161.

Knight, J. K., and Wood, W. B. (2005). Teaching more by lecturing less. Cell Biol. Educ. *4*, 298–310.

Lewis, J., Leach, J., and Wood-Robinson, C. (2000a). All in the genes? Young people's understanding of the nature of genes. J. Biol. Educ. *34*, 74–79.

Lewis, J., Leach, J., and Wood-Robinson, C. (2000b). Chromosomes: the missing link—young people's understanding of mitosis, meiosis, and fertilisation. J. Biol. Educ. *34*, 189–199.

Lewis, J., Leach, J., and Wood-Robinson, C. (2000c). Genes, chromosomes, cell division and inheritance—do students see any relationship? Int. J. Sci. Educ. *22*, 177–195.

Lewis, J., Leach, J., and Wood-Robinson, C. (2000d). What's in a cell? Young people's understanding of the genetic relationship between cells within an individual. J. Biol. Educ. *34*, 129–132.

Marbach-Ad, G. (2001). Attempting to break the code in student comprehension of genetic concepts. J. Biol. Educ. *35*, 183–189.

Marbach-Ad, G., and Stavy, R. (2000). Students' cellular and molecular explanations of genetic phenomena. J. Biol. Educ. *34*, 200–205.

Mazur, E. (1997). Peer Instruction: A User's Manual, Upper Saddle River, NJ: Prentice Hall.

Morrow, J., Jackson, A., Disch, J., and Mood, D. (2005). Measurement and Evaluation in Human Performance, Champaign, IL: Human Kinetics.

Orcajo, T. I., and Aznar, M. M. (2005). Solving problems in genetics II: conceptual restructuring. Int. J. Sci. Educ. *27*, 1495–1519.

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

Thornton, R. K., and Sokoloff, D. R. (1998). Assessing student learning of Newton's laws: the force and motion conceptual evaluation and the evaluation of active learning laboratory and lecture curricula. Am. J. Phys. *66*, 338–352.

Tsui, C. Y., and Treagust, D. (2004). Conceptual change in learning genetics: an ontological perspective. Res. Sci. Tech. Educ. *22*, 185–202.

Udovic, D., Morris, D., Dickman, A., Postlethwait, J., and Wetherwax, P. (2002). Workshop biology: demonstrating the effectiveness of active learning in an introductory biology course. BioScience *52*, 272–281.

Venville, G. J., and Treagust, D. F. (1998). Exploring conceptual change in genetics using a multidimensional interpretive framework. J. Res. Sci. Teach. *35*, 1031–1055.

Wood-Robinson, C., Lewis, J., and Leach, J. (2000). Young people's understanding of the nature of genetic information in the cells of an organism. J. Biol. Educ. *35*, 29–36.

Zohar, A., and Nemet, F. (2002). Fostering students' knowledge and argumentation skills through dilemmas in human genetics. J. Res. Sci. Teach. *39*, 35–62.