Designing, Validating and Piloting a Genomics and Bioinformatics Assessment

Chad Campbell, Teaching and Learning, The Ohio State University
Ross H. Nehm, Teaching and Learning, The Ohio State University
Brian Morton, Biological Sciences Barnard College, Columbia University

Abstract. Over the past decade, hundreds of studies have introduced genomics and bioinformatics (GB) curricula and laboratory activities at the undergraduate level. While these publications have facilitated the teaching and learning of cutting-edge content, one key aspect of evidence-based practice has been left behind: the development of assessment tools capable of generating valid and reliable inferences about student learning. Content validity is a core facet of construct validity, and must be used to guide instrument and item development. Based on previous work which reported on the correspondence of content validity evidence gathered from independent sources, our current work details: (1) the process of item development using this evidence and (2) the results from a pilot administration of the assessment. By including only the subtopics that were shown to have robust support across our content validity sources, 22 GB subtopics were established for inclusion in our assessment. An expert panel subsequently developed, evaluated, and revised two multiple-choice items to align with each subtopic, producing a final item pool of 44 items. These items were piloted with student samples of varying content exposure levels. We report on Classical Test Theory (CTT) and Rasch analyses of individual items and overall instrument quality.

Designing, Validating and Piloting a Genomics and Bioinformatics Assessment

Genomics and bioinformatics (GB) are new biological disciplines emblematic of the revolutionary changes occurring in the life sciences. These fields—part of what has been called the "New Biology" (National Research Council, 2009)--have not only brought forth novel approaches for investigating life, but have also re-conceptualized the information, skills, and performances that life science students need to know to reason about and contribute to the life sciences. Changes in the disciplinary structure of the biological sciences also necessitate the creation of new assessment tools capable of measuring whether students are acquiring the new knowledge, skills and performances emblematic of these new disciplines.

In order to build new assessment tools for the "New Biology", content validity must first be established. However, in a review of the GB education literature by Campbell and Nehm (2011) it was shown that there has been little discussion about the boundaries of these content domains, consensus standards, or agreed-upon core ideas (*sensu* NRC, 2012). A review of genomics and bioinformatics education (GBE) research (Campbell and Nehm, in press) also revealed that in approximately 100 publications claiming affective, cognitive or procedural learning gains, only 7% provided any supporting validity or reliability evidence, calling into question the robustness of efficacy claims. This less than encouraging finding highlights the need for the creation of assessments in GBE that adhere to the standards set forth by the educational research community (e.g., AERA, APA and NCME, 1999).

Our previous research used an expert survey and textbook analyses to establish content validity evidence. In short 29 subtopics were proposed for belonging on a GB assessment. These subtopics were evaluated by the experts and searched for within the textbooks in order to provide

support for their inclusion on our assessment. Our results indicated that 24 of the 29 subtopics had robust support.

Building on past work, we (1) designed assessment items that aligned with our content validity evidence; and (2) piloted the assessment on student sample populations with various levels of GB content exposure.

Research questions

The overarching goals of our study were to develop items for the new assessment using our content validity evidence, and evaluate the quality of the assessment and its constituent items relative to a theoretical framework for assessment validation. We investigated four specific research questions: (1) Can relevant multiple-choice items be developed that align with consensus topics? (2) To what degree do the designed items validly and reliably measure student GB knowledge? (3) In what ways do the designed items differentiate (or fail to differentiate) between undergraduate students with different levels of genomics and bioinformatics content exposure? And (4) To what extent does the sources of validity and reliability evidence inform us about our assessment's construct validity? We used a variety of methods to answer each of these research questions.

Methods

Assessment design, administration and analysis. Based on our previous content validity work, items were developed to align with each of the sub-domains and to appropriately cover the overall GB domain. Although there are drawbacks to multiple-choice (MC) formats (reviewed in Nehm and Schonfeld, 2008), they are well suited to assessing broad content

domains in a small amount of time (Brennan, 2006; Downing, 2006; Nitko and Brookhart, 2006). The MC item development and evaluation methods are described below.

Item development and evaluation. The first phase of our work was to develop a full set of MC GB items at the "knowledge" and "comprehension" levels (in alignment with Bloom's taxonomy; Bloom et al., 1956). (Note that a parallel suite of items from higher-order levels will form the second phase of our project). A panel consisting of three experts in the biological sciences (one Ph.D. in Integrative Biology, one Ph.D. in Genetics and one M.S. in Molecular Genetics) developed two multiple-choice items to align with each subtopic. These items were developed to: (1) match the subtopic descriptions used in the survey and textbook analyses; (2) be scientifically accurate; and (3) be at an appropriate level for assessing undergraduate GB knowledge. Items were developed by locating and modifying existing items in GBE research publications and textbooks on genetics and genomics (Brown, 2007; Campbell and Heyer, 2007; Griffiths, et. al., 2008; Hartl, 2011; Hartl and Jones, 2009; Hartwell, 2008; Higgs and Attwood, 2005; Gibson and Muse, 2009; Pierce, 2012; Pevsner, 2009; Primrose and Twyman, 2004; Snustad and Simmons, 2009), or were developed *de novo* when items with proper alignment to the subtopics could not be located.

To empirically evaluate if the items had been developed in accordance with the specifications outlined above, additional experts with doctoral degrees and expertise in genetics, genomics, bioinformatics and computational biology were recruited. Four experts with Ph.D.s (Biology, Molecular Biology, Bioengineering and Developmental Physiology) evaluated each assessment item for its accordance with our item guidelines noted above. The items were modified based on these evaluations in one of two ways: (1) by rewriting the stem or responses to make them clearer or more scientifically accurate and (2) by discarding the item and replacing

it with one which was more appropriate for an undergraduate level of knowledge, that aligned more accurately with the content topic, or did not contain information relevant to other items on the assessment. Modifications were also made to the draft items throughout the item creation and evaluation process to ensure that they: (1) complied with the multiple-choice item writing guidelines described in the *Handbook of Test Development* (Downing, 2006) and (2) were grammatically correct and linguistically clear. An English educator independent of the project (faculty member with a Ph.D. in English Education) performed the latter aspects of item evaluation.

Assessment formats and administration. To prepare the assessment for the online pilot test, items were first separated into two different forms (A and B, with several overlapping items) to minimize instrument length. In order to determine which items should be present on each form, two experts in the biological sciences ranked each item's difficulty from 1 to 10. A weighted Kappa (κ_w=0.98) was calculated (in place of the more common Cohen's kappa) because of the ordinal rankings. Based on these rankings, 16 of the least difficult items were chosen to be present on both assessment forms. The remaining items were then divided so that: (1) all 22 sub-topics were present on each form and (2) the average difficulty of the items was approximately the same on each form. Instrument form (A, B) difficulty averages were calculated by summing the difficulty rankings for all of the items on each assessment form (for both raters, individually), and then dividing by 30. The two instrument forms had randomized item order (and assessment form was randomly assigned to participants). Additionally, students were asked to self-report: gender; if English was their first language; and English reading and writing skills.

The assessment was administered to three different populations of students at a large Midwestern university: those with low, medium and high GB content exposure. The assessment was administered using web-based SurveyMonkey (Professional Version) software. Students were asked to complete the survey on their own time. Voluntary response rates were high (> 80%) and 535 students provided consent and completed the entire survey (Table 1). Assessment Form A was administered to 254 students and Form B was administered to 281 students. The "low" content exposure level consisted of students from an introductory biology course taken in the freshman year of college; a total of 135 students took the assessment in this content exposure level (61 responding to Form A and 74 responding to Form B). Within this exposure level, 83 participants were female, 50 were male, and 2 did not report their gender. The "medium" content exposure level consisted of students from an introductory genetics course taken by all biology-related majors sometime between their sophomore and junior years of college. A total of 351 students responded in this exposure level (170 receiving Form A and 181 receiving Form B). Within this exposure level, 185 participants were female, 165 were male, and one participant did not report gender. The "high" content exposure level consisted of graduate and advanced undergraduate students; a total of 49 students took the assessment with 23 responding to Form A and 27 responding to Form B. Within this exposure level, 22 participants were female and 27 were male.

From the total student sample, 51(10%) students indicated that English was not their first language; however, all students indicated that their English reading and writing skills were "Good", "Very Good" or "Excellent". These data suggest that all students sampled had adequate English skills and were able to read and understand the assessment items.

Table 1
Sample information; n=535

Content Exposure Level	Assessment Form		Gender	
	A n = 254	B n = 281	Female n = 290	Male n = 242
Low n = 135	61	74	83	50
Medium n = 321	170	181	185	165
High n = 49	23	26	22	27

Assessment and item analyses. We empirically evaluated the quality of the instrument and its constituent items by analyzing: (1) Rasch fit statistics for the two assessment forms; (2) internal consistency/reliability; (3) internal structure/uni-dimensionality; (4) item fit statistics/performance; and (5) performance relative to extrinsic variables (e.g., content exposure level and gender). The CTT and Rasch analyses used to perform these evaluations are discussed below.

Rasch fit statistics for the two assessment forms. Rasch analysis was also used to analyze our assessment data because recent work in psychometrics has challenged the methodological appropriateness (and validity inferences) derived from Classical Test Theory (CTT) raw measures (for an overview of this perspective, see Bond and Fox, 2001). Unlike CTT analyses, Rasch analyses are capable of constructing linear measures from numerical raw data. To analyze the dataset using Rasch, we used WINSTEPS v3.68.2 to: (1) remove any persons with negative point bi-serial correlations (as this is an indication of problematic response patterns); (2) identify items that had outfit mean square values outside of the acceptable range (between 0.7-1.3 for MC data; Wright and Linacre, 1994); (3) delete person responses (on the items found in the previous step) that had Z standard values above two; (4) repeat steps two and three until all items

demonstrate appropriate model fit; and (5) examine whether the modifications made in steps two and three (above) detracted from overall measurement validity.

Both forms of the assessment must be shown to fit the Rasch model prior to being combined; consequently, forms A and B were analyzed independently following the methodologies described above. Once both forms demonstrated robust fit to a Rasch model, a common item cross-plot was then used to compare the two assessment forms to determine if the data could be analyzed together. This approach allows one Rasch model to predict item and person measure scores for all 44 items and 535 persons.

Internal consistency of student responses. The response reliability (for both forms of the assessment) was analyzed using Kuder Richardson 20 (KR-20). This is a measure of an assessment's internal consistency. Acceptable values are generally considered to exceed 0.7 (Doran, 1980). Rasch person and item reliabilities were also calculated using WINSTEPS. Acceptable values for person reliability separating between two or three groups is > 0.8 while acceptable values for item reliability are > 0.9 (Linacre, 2012).

Internal structure. The internal structure of an assessment focuses on the number of dimensions or latent factors captured by the items. In general, an assessment should measure one construct or dimension at a time to ensure the accuracy and clarity of inferences that can be made about that measure (Brennan, 2006). If an assessment contains multiple dimensions but only one measurement statistic, interpretations or inferences about that measurement will be unclear. For example, a well-performing student could have performed well on one dimension but poorly on another, or instead, could have performed moderately on both dimensions. Unless the assessments are uni-dimensional (or have separate measures for each dimension) it is not generally possible to meaningfully quantify performance.

We used Rasch-residual-based principal component analysis (PCAr) to determine the dimensionality of the assessment. This methodology is not comparable to traditional PCA or factor analysis methods where identifying underlying latent factors occurs. PCAr, in contrast, is an attempt to explain variance by analyzing residual contrasts. These residuals are the unexplained data remaining after removal of the data pertaining to the primary factor. In the Rasch model much of the unexplained variance is expected to be due to random fluctuations, therefore if a contrast is found in the residuals it must be compared to the "noise" level. For an assessment to be considered multi-dimensional, the eigenvalue for the first contrast should be > 2.0 and explain more than 5.0% of the overall variance. There are many exceptions to this rule, however, and interested readers are encouraged to consult the WINSTEPS manual (Linacre, 2012, and online at www.rasch.org). (Note that traditional factor analysis and principal component analysis of the raw response data was not possible because of the two forms of the assessment).

Item performance. Individual item quality was empirically evaluated using: (1) item difficulty (P), (2) item discrimination (DI), (3) item distractor response rates, (4) Rasch fit statistics, and (5) person-item alignment on a Wright map (produced from the combined Rasch analysis). Item difficulty (P) is the measure of how difficult an item is, and was calculated by dividing the number of correct responses on an item by the total number of responses; acceptable P values are between 0.3 and 0.9 (30% and 90%; Doran, 1980). The item discrimination index (DI) was used to determine how well an item differentiated between low- and high-performers on this assessment. DI was calculated by dividing the student population into thirds (based on their total test scores) and subsequently subtracting the number of correct responses in the lower

third from the number of correct responses in the upper third, and finally dividing by the number of students within each third; acceptable DI values are above 0.3 (Doran, 1980).

Item distractor response rates (that is, the number of individuals choosing an incorrect answer option) were analyzed by dividing the number of times a distractor was selected by the total number of responses. Acceptable distractor percentage values should be greater than 5% (values < 5% indicate non-functioning distractors (Haladyna and Downing, 1993). Items with a difficulty index of 90% or higher were removed from the distractor analysis.

Rasch fit statistics were also used to analyze assessment quality. Items with outfit or infit mean square values outside the range of 0.7-1.3, and Z standard values > 2.0, were identified as misfitting with the Rasch model. Items that display misfit need to be reviewed to indicate how and why the item does not fit the data. Those items that display misfit *outfit* mean square values are indicative of high-ability students answering the item incorrectly, or low-ability students answering the item correctly. Items with misfit *infit* mean square values are indicative of unexpected responses from persons with the same ability as the item difficulty (Linacre, 2012). These statistics were analyzed in order to identify items in need of improvement.

Wright Maps visualize item difficulty measures and person ability measures derived from a Rasch analysis on a linear scale with the easiest items, and least able persons, at the bottom of the scale, and the most difficult items and most able persons at the top of the scale. A Wright Map allows one to compare how well the sample population aligns with the assessment items. Persons at the same position on the scale as an item are modeled to have a 50% chance of answering that item correctly. Items higher on the scale than the person have less than a 50% chance of being answered correctly, while those items lower on the scale than the person have a greater than 50% chance of being answered correctly. The distributions of items (on the right

side of the map) and persons (along the left side of the map) should overlap, with few gaps between persons or items. The presence of a gap indicates that either there are no items to assess students at that ability level, or conversely there were no students at the ability level in the sample for which the item discriminates best. Thus, the Wright Map is a holistic visual snapshot of instrument performance relative to the sample of participants.

Performance of students between classification groups. Three comparison groups exposed to differing amounts of GB knowledge were used to compare performance on the assessment. To compare total *assessment* scores between the three samples, total raw scores were used in a three-way ANOVA followed by a Fisher LSD multiple comparison test. This same methodology was also used to compare person-measure scores derived from Rasch analysis. To compare student performance at the *item* level, a test of independent proportions was used to compare the low and high exposure sample scores. Items were predicted to have statistically significant ($p=0.05$) response proportions given different correct response rates. Given the large number of comparisons, a Bonferroni correction was applied to our critical value.

Results

Item and instrument design and analysis. Two MC assessment items were developed to align with each of the subtopics delineated in previous work. Subsequently, these items were evaluated in field tests.

Item development and evaluation. During the process of item development, the subtopics "Comparative Genomics: Applications" and "Applications" were found to require cognitive processes ("skills") in addition to GB knowledge (i.e., data analysis and graph interpretation). Because these skills were: (1) not part of the intended construct (see above), (2)

could introduce construct-irrelevant measurement variance; and (3) were skills peripheral to successful performance on other items in other subtopics, they were removed from the assessment and will be included in a parallel GB assessment focusing on these GB skills.

Two items were created for each of the 22 subtopics (n=44) and evaluated by a panel of experts to determine if they: (1) matched the subtopic descriptions used in the survey and textbook analyses; (2) were scientifically accurate; and (3) were at an appropriate level for assessing undergraduate GB knowledge. The experts (n=4) unanimously agreed that 42 (95%) of the items matched the category descriptions. The remaining two items had two experts who were unsure if the item matched the category description, and two experts who agreed that the item matched the description. The experts also unanimously agreed that 41 (93%) of the items were scientifically accurate. The remaining three items had two experts indicate that they were unsure whether the item was scientifically accurate, and two others who indicated that the item was scientifically accurate. The experts unanimously agreed that 38 (86%) of the items were appropriately aligned with undergraduate knowledge. Five of the remaining items had two experts who were unsure if the item was appropriate for an undergraduate assessment and two who agreed that the item was appropriate. The final remaining item: had two experts agree that it was appropriate; one who was unsure if it was appropriate; and one who thought the item was too difficult for undergraduates. Overall, for the majority of items the experts unanimously agreed that they: (1) matched their category description; (2) were scientifically accurate; and (3) were appropriate for an undergraduate assessment. For the remaining items that did not receive unanimous support, the experts who did not mark the item in the "Agree" category indicated that they were "Unsure" as the item topic was outside of their expertise. Only one item (Item 29)

received a ranking of "Disagree" by one expert in terms of whether it was at an appropriate for undergraduates. (Note: All items are available from the authors upon request).

GB assessment analysis. We empirically evaluated the quality of the instrument and its constituent items by analyzing: (1) Rasch fit statistics for the two assessment forms; (2) internal consistency/reliability; (3) internal structure/uni-dimensionality; (4) item fit statistics/performance; and (5) performance relative to extrinsic variables (e.g., content exposure level and gender).

Rasch fit statistics for the two assessment forms. Each assessment form was analyzed separately for goodness of fit to the Rasch model prior to combining the data. Persons with negative point bi-serial correlations were removed from both assessments. Specifically, four individuals (1.57%) were removed from Form A and four individuals (1.42%) were removed from Form B. Item fit statistics were then analyzed and one item ("Item 16", MNSQ=1.68) contained an outfit mean square value outside of the acceptable range on assessment Form A and three items ("Item 10", MNSQ=1.53; "Item 16", MNSQ=1.7; and "Item 43", MNSQ=1.52) contained outfit mean square values outside the acceptable range on Form B. Item responses for the above items were analyzed and responses with Z standard values > 3 were removed; six responses (2.17%) were removed from "Item 10", fourteen responses (2.65%) were removed from "Item 16", six (2.4%) were removed from Form A, and eight (2.89%) were removed from Form B, and seven responses (2.53%) removed from "Item 43". A total of 27 responses (0.17%) were removed from the analysis, which resulted in outfit mean square values being within acceptable limits (i.e., fit the Rasch model).

A common item-difficulty measure cross-plot was used to determine if the two forms of the assessment could be combined into one assessment analysis. A regression value at or near 1.0

is considered appropriate for joining assessment forms without modification. The R^2 value for our dataset was 0.987, indicating robust agreement.

Internal consistency/reliability. KR-20 was used to calculate the reliability for CTT data from both forms of the assessment. Form A had a value of 0.69 and Form B had a value of 0.70. Both forms are within the standard criterion value for acceptable reliability. Rasch person and item reliabilities were also calculated using WINSTEPS. The person reliability score was 0.75, which is slightly below the criterion value (r_{person}=0.80). The item reliability score was 0.98, which is above the criterion value of 0.9. Thus, both CTT and Rasch analyses produced acceptable reliability values for the instrument.

Internal structure/dimensionality. The PCAr data for item dimensionality indicated that 28.2% of the variance was explained by our measures (11.6% for persons and 16.6% for items) leaving a remaining 71.8% of the variance as unexplained. While an unexplained variance of less than 50% is optimal, our high amount of unexplained variance is not unexpected due to the high number of individuals sampled in the medium content exposure group (n = 349, 66%). This unbalanced sampling will lead to a smaller standard deviations and therefore smaller amounts of variance explained. Our first contrast eigenvalue was 2.4; 3.9% of the total variance and 5.4% of the unexplained variance. While our eigenvalue was greater than the criterion value of 2.0, the total variance explained was less than 5%, providing little support for construct multi-dimensionality.

Item fit statistics/performance. For the sixteen items common to both forms of the assessment, item difficulty (P) values were calculated using the entire sample population (n=535). Difficulty values for the remaining 28 items were calculated using only the population unique to each assessment form (Form A, n=254; Form B, n=281). The P values varied between

12% and 94%; thirty-three of the items had acceptable difficulty values (between 30% and 90%), ten of the remaining items had difficulty values less than 30% (Items 8, 10, 13, 16, 22, 23, 30, 36, 38 and 43) and one item had a difficulty value greater than 90% (Item 6; See Table 3 column P for exact values). Overall, the average item difficulty index value was 46%, which falls within the criterion range.

As above, item discrimination (DI) values were calculated using only the sample populations which took each form of the assessment. DI values ranged between -0.13 and 0.7 (Table 3: columns DI A and DI B). For the 16 items that were in common to both assessments, ten of the items had acceptable DI values on both assessments (Items 1, 2, 3, 5, 7, 15, 19, 24, 25, and 27), four items had DI values below 0.3 on both assessments (Items 6, 16, 26 and 43) and two items (Items 17 and 41) had mixed results on the assessment forms. For the remaining fourteen items on Form A, ten of the items had discrimination values above 0.3 (Items 4, 9, 14, 21, 28, 29, 31, 34, 36 and 40) and four had values below 0.3 (items 8, 11, 20 and 38). For the remaining fourteen items on Form B, five of the items (Items 18, 32, 37, 39 and 42) had discrimination values above 0.3 and nine items had discrimination values below 0.3 (Items 10, 12, 13, 22, 23, 30, 33, 35 and 44). Overall, the average discrimination values for the two forms of the assessment were above the criterion 0.3 value (DI A=0.32 and DI B=0.32 respectively).

Item distracter response rates were analyzed to determine if any of the item distracters were non-functioning. Eight items were found to contain distracters with response percentages below 5% (Table 4; bold). The remaining 36 items did not contain any non-functioning distracters.

Table 3

Classical Test Theory Item Statistics

Item #	Item Name	P n=535	DI A n=254	DI B n=281
1	Bioinformatics 1	65%	0.63	0.59
2	Bioinformatics 2	46%	0.58	0.48
3	Comparative genomics 1	48%	0.55	0.57
4	Comparative genomics 2	59%	0.38	
5	Definition of a genome 1	58%	0.45	0.52
6	Definition of a genome 2	**94%**	**0.12**	**0.09**
7	Expression 1	63%	0.70	0.53
8	Expression 2	**28%**	**0.06**	
9	Expression: Analysis 1	52%	0.49	
10	Expression: Analysis 2	**20%**		**0.03**
11	Genome annotation by pattern 1	31%	**0.24**	
12	Genome annotation by pattern 2	31%		**0.01**
13	Genome annotation by similarity 1	**25%**		**0.23**
14	Genome annotation by similarity 2	40%	0.31	
15	Genome content and organization 1	70%	0.60	0.48
16	Genome content and organization 2	**20%**	**-0.05**	**-0.13**
17	Genome evolution 1	44%	0.32	**0.23**
18	Genome evolution 2	59%		0.61
19	Genome mapping 1	39%	0.46	0.57
20	Genome mapping 2	42%	**0.19**	
21	Genome sequencing: Concepts of 1	62%	0.37	
22	Genome sequencing: Concepts of 2	**27%**		**0.15**
23	Genome sequencing: Techniques 1	**29%**		**-0.08**
24	Genome sequencing: Techniques 2	70%	0.48	0.62
25	Genome structure 1	76%	0.52	0.57
26	Genome structure 2	90%	**0.19**	**0.17**
27	Integration of genomics with biology 1	42%	0.57	0.45
28	Integration of genomics with biology 2	59%	0.49	
29	Metagenomics 1	44%	0.54	
30	Metagenomics 2	**24%**		**0.20**
31	Regulation: Epigenetics 1	57%	0.45	
32	Regulation: Epigenetics 2	36%		0.44
33	Regulation: RNA 1	42%		**0.11**
34	Regulation: RNA 2	51%	0.46	
35	Regulation: Transcription factors 1	53%		**0.18**
36	Regulation: Transcription factors 2	**26%**	0.43	
37	Sequence databases 1	46%		0.55
38	Sequence databases 2	**16%**	**0.18**	
39	Variation: Linkage disequilibrium and haplotypes 1	41%		0.42
40	Variation: Linkage disequilibrium and haplotypes 2	62%	0.48	
41	Variation: Nucleotide 1	58%	**0.20**	0.39
42	Variation: Nucleotide 2	51%		0.53
43	Variation: Structural 1	**12%**	**0.05**	**-0.05**
44	Variation: Structural 2	30%		**0.04**
	Mean	46%	.32	.32

P = Item Difficulty, DI A = Item Discrimination Index for assessment Form A, DI B = Item Discrimination Index for assessment Form B

Table 4
Items with distracter response percentages below 5%

Item #	Item Name	P	Distractor 1	Distractor 2	Distractor 3
1	Bioinformatics 1	65%	7.48%	**4.30%**	23.55%
4	Comparative genomics 2	59%	18.11%	18.90%	**4.33%**
5	Definition of a genome 1	58%	15.70%	26.17%	**0.56%**
19	Genome evolution 2	39%	10.47%	47.66%	**3.36%**
20	Genome mapping 1	42%	21.26%	**3.94%**	33.07%
21	Genome mapping 2	62%	**4.72%**	27.17%	6.30%
25	Genome sequencing: Techniques 2	76%	**4.30%**	7.66%	11.96%
27	Genome structure 2	42%	**1.68%**	27.48%	29.16%

Rasch fit statistics were used to analyze the composite data from Forms A and B. Five items were found to have mean square outfit values above the accepted range of 0.7-1.3: Items 8, 10, 16, 23 and 44 (Table 9). All of these items also had outfit Z standards values above 2.0, indicating that they do not fit the Rasch model (Table 5). All items had acceptable mean square infit values, negating the need to examine their Z values. In sum, 39 of the 44 items fit the Rasch model quite well.

A Wright map was generated to analyze person ability and item difficulty distributions. Figure 1 displays the Wright map from the combined data set. Overall, the distributions of persons and items are well matched; however, there appears to be an indication that the item difficulties are slightly higher than the students' ability levels. Some items are too easy for the sample population (Items 6 and 26) and others are too difficult (Items 10, 16, 38 and 43). Twenty-one of the subtopics have two items that are at different difficulty levels (as was intended by design); however, one subtopic (Genome annotation by pattern) has both items at the same approximate difficulty level.

Table 5

GB Assessment fit statistics from the combined Rasch analysis

Item #	Measure	Model SE	Infit (MNSQ)	Infit (ZSTD)	Outfit (MNSQ)	Outfit (ZSTD)
1	-0.9	0.1	0.84	-4.42	0.77	-4.21
2	0	0.09	0.95	-1.73	0.94	-1.43
3	-0.09	0.09	0.93	-2.41	0.92	-2.01
4	-0.56	0.14	1.01	0.29	0.99	-0.08
5	-0.57	0.1	0.96	-1.21	0.95	-1.04
6	-3.46	0.21	0.97	-0.1	0.96	-0.06
7	-0.83	0.1	0.84	-4.47	0.78	-4.23
8	1.08	0.16	1.15	1.89	**1.4**	**3.08**
9	-0.24	0.14	0.97	-0.76	0.94	-0.86
10	1.5	0.16	1.12	1.26	**1.35**	**2.21**
11	0.78	0.15	1.14	2.07	1.24	2.36
12	0.73	0.14	1.21	3.3	1.28	2.9
13	1.06	0.15	1.01	0.2	1.12	1.04
14	0.33	0.14	1.06	1.16	1.13	1.7
15	-1.14	0.1	0.9	-2.33	0.83	-2.67
16	1.69	0.13	1.14	1.73	**1.34**	**2.69**
17	0.09	0.09	1.09	2.63	1.09	2.07
18	-0.67	0.13	0.88	-2.53	0.84	-2.63
19	0.34	0.1	0.96	-1.06	0.93	-1.41
20	0.27	0.14	1.17	3.24	1.25	3.26
21	-0.68	0.14	1.02	0.45	1	0.04
22	0.95	0.14	1.05	0.77	1.15	1.36
23	0.83	0.15	1.23	3.26	**1.36**	**3.37**
24	-1.12	0.14	0.84	-2.67	0.75	-2.75
25	-1.58	0.11	0.84	-2.88	0.74	-3.07
26	-2.87	0.16	1.01	0.14	0.92	-0.38
27	0.22	0.1	0.89	-3.32	0.9	-2.21
28	-0.58	0.14	0.96	-0.87	0.96	-0.5
29	0.1	0.14	0.94	-1.29	0.93	-1.03
30	1.06	0.15	1.06	0.84	1.3	2.45
31	-0.49	0.14	0.96	-0.76	0.95	-0.69
32	0.47	0.13	0.96	-0.83	0.92	-1.09
33	0.16	0.13	1.18	3.96	1.28	4.15
34	-0.18	0.14	0.99	-0.2	0.99	-0.2
35	-0.38	0.13	1.15	3.4	1.19	3.27
36	1.1	0.16	0.87	-1.68	0.94	-0.48
37	-0.05	0.13	0.95	-1.29	0.95	-0.93
38	1.85	0.19	0.99	-0.04	1.06	0.36
39	0.21	0.13	0.96	-0.84	0.97	-0.43
40	-0.74	0.14	0.93	-1.27	0.98	-0.23
41	-0.55	0.1	1.08	2.26	1.08	1.69
42	-0.25	0.13	0.95	-1.14	0.97	-0.5
43	2.34	0.15	1.03	0.27	1.26	1.44
44	0.77	0.14	1.2	3.07	**1.34**	**3.32**
Mean	0.00	0.13	1.01	-0.09	1.04	0.17
SD	1.11	0.03	0.11	2.14	0.18	2.17

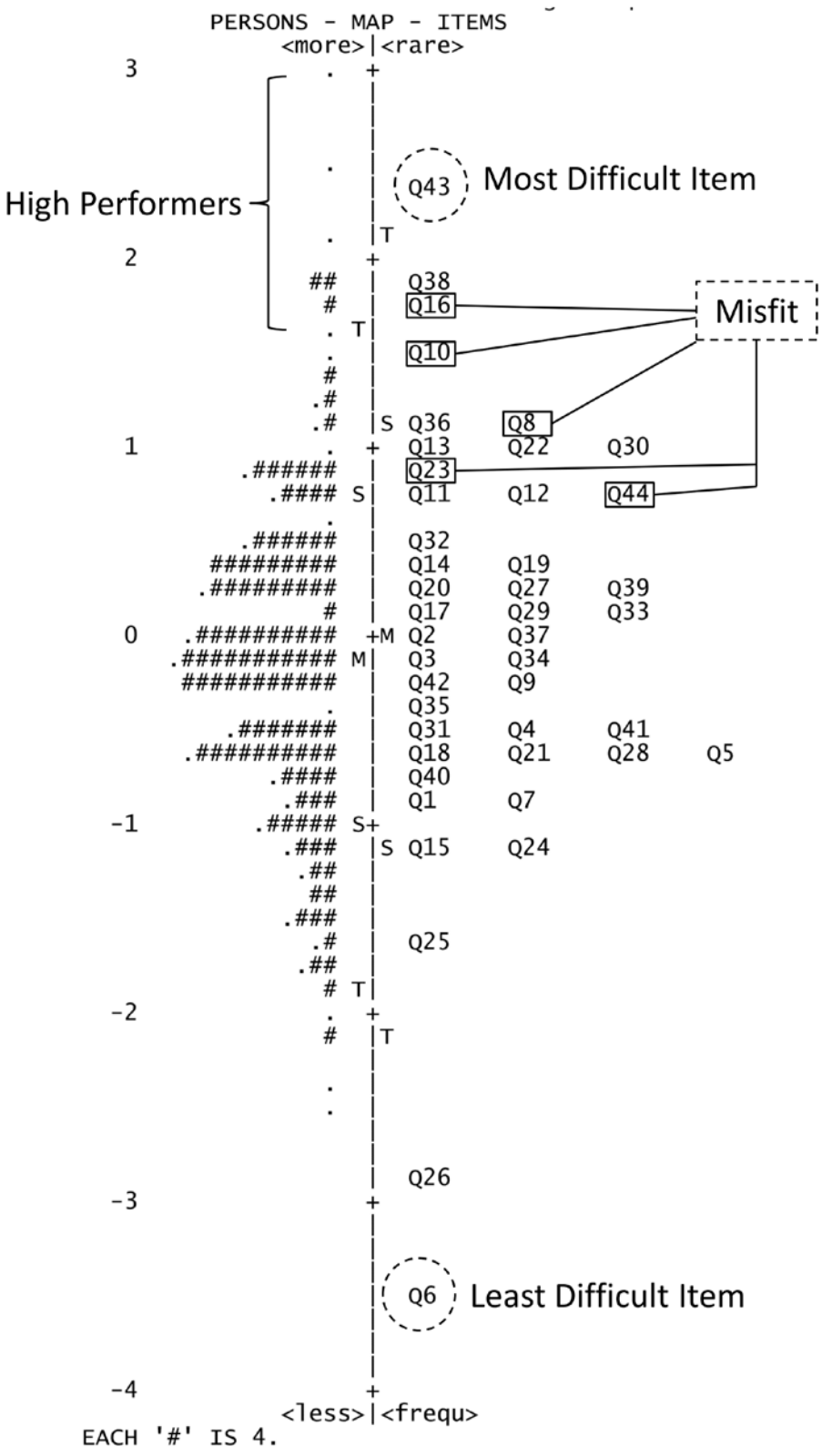

Figure 1. Wright map of person abilities and item difficulties on a Logit scale.

Performance relative to extrinsic variables. Average total scores were calculated for each of our three exposure groups. The mean total score was 14.58 (n=535). When the different assessment forms were compared, the mean raw total score was 14.93 (n=254) for Form A and 14.26 (n=281) for Form B. When the different content exposure levels were compared (Figure 2), the low exposure raw score mean was 10.36 (n=135), the medium content exposure level was 15.44 (n=351) and the high content exposure level was 20.02 (n=49). When student gender was compared, the average total score was 13.99 (n=290) for females and 15.3 (n=242) for males. To determine if these students' mean scores were significantly different a three-way ANOVA (assessment form x content exposure level x gender) was performed. (N.B. The three individuals that did not report gender were removed; remaining n=532). Based on a residual plot, skewness and kurtosis statistics and Levene's homogeneity of variance test (p=0.450), the assumptions were satisfied for performing an ANOVA.

The ANOVA revealed significant effects at the p=0.05 level for assessment form, content exposure level, gender, and interaction effect form*exposure (F_{form}=19.9, df=1, 532, p. <0.001; $F_{exposure}$=130.5, df=2, 532, p. <0.001; F_{gender}=4.38, df=1, 532, p=0.037; $F_{form*exposure}$=12.63, df=, 532, p. <0.001; Appendix, Table 1). A post-hoc Fisher LSD (Appendix, Table 2) was performed to determine which comparison(s) of the content exposure effect were significant. The analysis indicated that all comparisons (low to medium, low to high, medium to high) were significantly different at p. <0.001. The profile plot of the interaction effect form*exposure indicated that only the high content exposure level (that is, the small sample of advanced undergraduates and graduate students) showed significant differences between assessment forms. While all of the above interactions were statistically significant, the effect size statistics ($\eta_p^2{}_{form}$=0.037; $\eta_p^2{}_{exposure}$=0.334; $\eta_p^2{}_{gender}$=0.008; $\eta_p^2{}_{form*exposure}$=0.046) indicated that only the content exposure

effect (i.e., GB content) had practical meaning, as values below 0.2 are small effects, those above 0.2 and below 0.4 are medium effects and those above 0.4 are strong effects (Cohen, 1988).

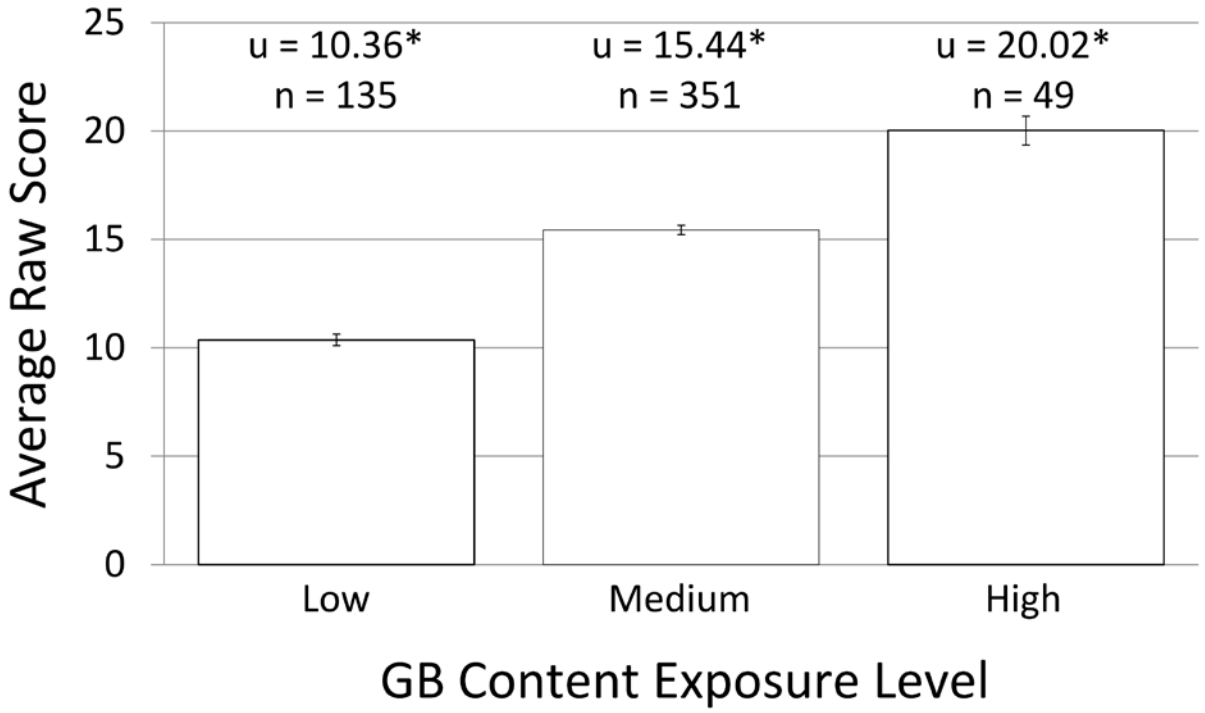

Figure 2. Average student raw scores at various content exposure levels (* indicates significance at p=0.05)

For the Rasch analysis, the average person-measure score was -0.168 (n=527; eight individuals were removed from the data set based on their negative point bi-serial values in the individual assessment form Rasch analysis). When the different assessment forms were compared, the mean person-measure score was -0.104 (n=250) for Form A and -0.225 (n=277) for Form B. When different content exposure levels were analyzed (Figure 3), the low exposure level mean person score was -0.946 (n=129), the mean medium exposure score was -0.031 (n=349) and the mean high exposure level score was 0.909 (n=49). When the mean person-measure score for gender was analyzed, females averaged -0.285 (n=286) and males averaged -0.022 (n=238).

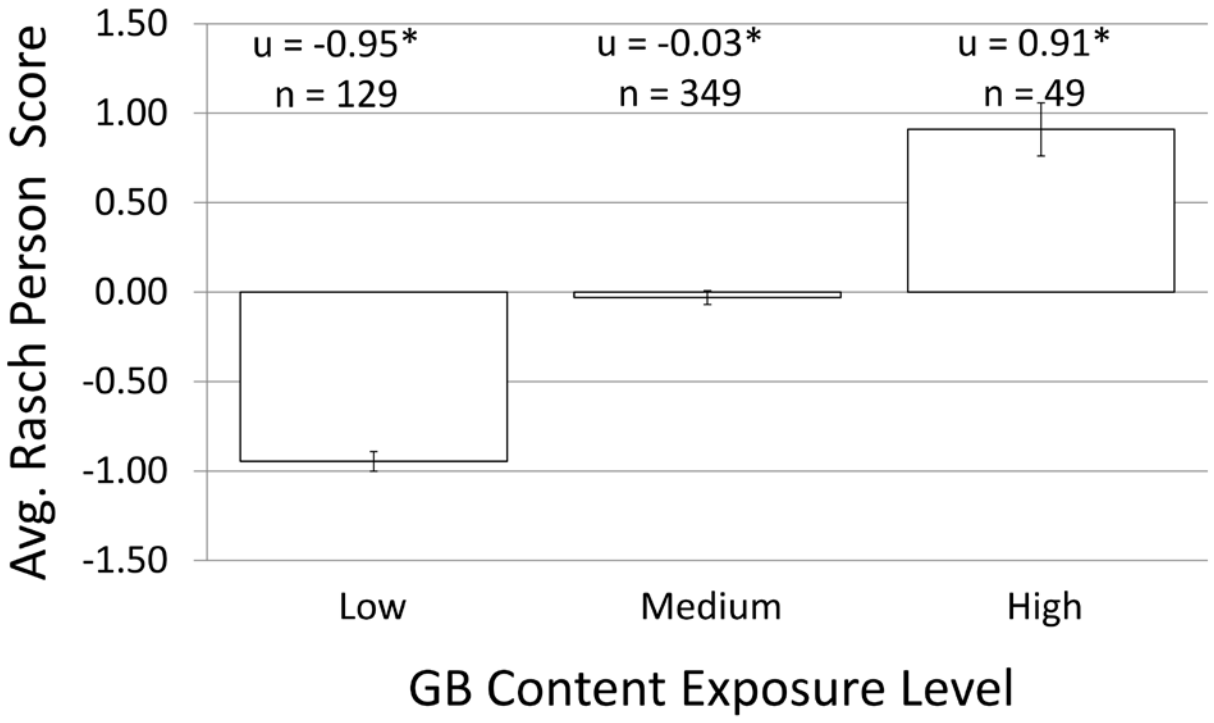

Figure 3. Average student Rasch person-measure scores at various content exposure levels (* indicates significance at p=0.05)

To determine if the differences in the mean person-measure scores were statistically significant, a three-way ANOVA and post-hoc Fisher LSD test were performed. Based on a residual plot, skewness and kurtosis statistics, and Levene's homogeneity of variance test (p = 0.856), the assumptions were satisfied for performing an ANOVA. As found with the CTT data, the ANOVA at the p=0.05 level had significant effects for form, exposure, gender and the interaction effect form*exposure (F_{form}=18.5, df=1, 524, p. <0.001; $F_{exposure}$=121.7, df=2, 524, p. <0.001; F_{gender}=3.1, df=1, 524, p.=0.014; $F_{form*exposure}$=12.4, df=2, 524, p. <0.001; Appendix, Table 3). A post-hoc Fisher LSD test (Appendix, Table 4) indicated that all comparisons of mean person-measure scores for each content exposure level were significant at p. <0.001. The profile plot of the interaction effect form*exposure indicated that only the high content exposure level showed significant differences between assessment forms. While all of the above interactions

were statistically significant, the effect size statistics ($\eta_p^2{}_{form}$=0.035; $\eta_p^2{}_{exposure}$=0.322; $\eta_p^2{}_{gender}$=0.012; $\eta_p^2{}_{form*exposure}$=0.046) indicated that only the exposure effect had practical value.

To compare student performance at the *item* level, a test of two independent proportions was performed using the raw response data from students in the low and high content exposure levels. Z scores were compared to a critical value (Z=3.04) found by using a Bonferroni corrected critical value for a one-tailed test at the p=0.05 level. Eighteen items (Items 1, 2, 3, 5, 7, 15, 18, 19, 24, 25, 27, 31, 34, 36, 38, 39, 41) were found to have statistically significant differences in the proportion of students that answered the item correctly between our low and high content exposure levels (Appendix, Table 5 bold values). In addition, negative Z-score values were found for four items (Items 10, 12, 16, 44; Appendix, Table 5, values in italics) indicating that the students in the low content exposure level performed *better* than those in the high content exposure level.

Summary of GB assessment analysis. The data from our two assessment forms was joined into one Rasch model because both were shown to independently fit a Rasch model. Furthermore, our assessment was shown to have reliable student responses and little evidence to indicate a multi-dimensional structure. Twenty-seven of our 44 items (Table 6) require at least some degree of revision. Ten items need major modifications or replacement; five of those items (Items 8, 10, 16, 23 and 44) had unacceptable values from both our CTT and IRT analyses, and the remaining five items (Items 6, 12, 22, 33 and 43) had very low DI values but acceptable IRT fit statistics. Ten additional items warranted item review at a minimum; three items (Items 13, 30 and 38) had mildly low DI values and very high difficulty values; five items (Items 11, 17, 26, 35 and 41) had only have mildly low DI values, one item (Item 36) had only a high difficulty value, and the remaining item (Item 20) had both low DI values and at least one distracter with a

low response percentage. The last seven items (Items 1, 4, 5, 19, 21, 25, and 27) warranted only distracter reviews.

Table 6
Summary of items in need of revision

Item #	Item Name	P	DI	Rasch fit	Distractor response
1	Bioinformatics 1				Too Low
4	Comparative genomics 2				Too Low
5	Definition of a genome 1				Too Low
6	Definition of a genome 2	Too High	Too Low		
8	Expression 2	Too Low	Too Low	Too High	
10	Expression: Analysis 2	Too Low	Too Low	Too High	
11	Genome annotation by pattern 1		Too Low		
12	Genome annotation by pattern 2		Too Low		
13	Genome annotation by similarity 1	Too Low	Too Low		
16	Genome content and organization 2	Too Low	Too Low	Too High	
17	Genome evolution 1		Too Low		
19	Genome mapping 1				Too Low
20	Genome mapping 2		Too Low		Too Low
21	Genome sequencing: Concepts of 1				Too Low
22	Genome sequencing: Concepts of 2	Too Low	Too Low		
23	Genome sequencing: Techniques 1	Too Low	Too Low	Too High	
25	Genome structure 1				Too Low
26	Genome structure 2		Too Low		
27	Integration of genomics with biology 1				Too Low
30	Metagenomics 2	Too Low	Too Low		
33	Regulation: RNA 1		Too Low		
35	Regulation: Transcription factors 1		Too Low		
36	Regulation: Transcription factors 2	Too Low			
38	Sequence databases 2	Too Low	Too Low		
41	Variation: Nucleotide 1		Too Low		
43	Variation: Structural 1	Too Low	Too Low		
44	Variation: Structural 2		Too Low	Too High	

Our ANOVA and Fisher LSD values from both the student raw scores and Rasch measure scores indicated that student means were significantly different between the exposure groups, gender, and the interaction between forms and exposure. However, the only meaningful statistical significance (moderate effect sizes) was level of content exposure. Additionally, our analyses indicated that 18 items had the ability to differentiate between students at the low and high content exposure levels.

Discussion

A prior literature review (Campbell and Nehm, in press) revealed that there is a lack of consensus in the field about what genomics and bioinformatics (GB) knowledge, skills, and dispositions should be the focus of undergraduate life science education. This finding, coupled with the fact that the GB education (GBE) literature has yet to produce assessments with robust validity or reliability evidence necessary for measuring the efficacy of instructional innovations, motivated our work on the development and evaluation of a new knowledge instrument grounded in content validity evidence. Our study described the first steps that have been taken to gather content validity evidence using the following research questions as guides: (1) What topics do GB experts consider to be central to the domain? (2) To what extent does GB textbook content coverage align with expert judgments? (3) Can relevant multiple-choice items be developed that align with consensus topics? (4) To what degree do the designed items validly and reliably measure student knowledge? (5) In what ways do the designed items differentiate (or fail to differentiate) between undergraduate students with different levels of genomics and bioinformatics content exposure? And (6) To what extent does the validity and reliability evidence contribute to establishing the assessment's construct validity?

What topics do GB experts consider to be central to the domain? We determined the topics that GB experts considered to be central to the domain through the use of an expert survey. Twenty-five GB subtopics were shown to have adequate support based on the results of our CTT and IRT analyses. The four subtopics which did not have adequate expert support for inclusion on our assessment ("Metabolomics", "Proteomics: Models", "Proteomics: Protein-protein interactions" and "Systems biology") all belonged to what we thought would be sub-

disciplines within GB, however based on our results it appears as though Metabolomics, Proteomics and Systems biology may be viewed as their own independent fields by the experts.

To what extent does GB textbook content coverage align with expert judgments?

We determined whether GB textbook content coverage aligned with expert judgments through the analysis of genomics textbooks; 28 subtopics were shown to have adequate support. This result is not surprising, as textbooks were consulted during the process of initial subtopic generation. The finding that one subtopic did not have adequate support ("Phenotype from genotype") was surprising as there is considerable research dedicated to the maintenance of gene expression within a cell leading to a phenotype. It is unclear why textbooks do not contain content relevant this subtopic. One potential possibility is that this subtopic does not belong in the GB domain, but instead is more related to the domain of developmental biology.

When our textbook data were compared to the expert survey data, 24 subtopics were found to have adequate support across both methodologies, providing support for the claim that they should be considered as part of the GB construct. It was interesting to note that while the experts did not support the inclusion of our "Proteomics" subtopics, almost every textbook contained them. This is perhaps an artifact of the textbook authors designing their content in a more logical or education friendly manner; in other words, a textbook is not confined to only contain information relevant to one construct or discipline. This finding highlights the need for multiple sources of content validity evidence so that they can be compared to one another.

Can relevant multiple-choice items be developed that align with consensus topics?

Multiple-choice (MC) items were developed that aligned with the consensus topics. Twenty-two supported subtopics (after removal of two subtopics found to be problematic) were used to design items for an undergraduate assessment with the intention of being able to differentiate

between students at various different levels of exposure to GB content knowledge. Four content experts examined these items and indicated that they: (1) matched the subtopic descriptions used in the survey and textbook analyses; (2) were scientifically accurate; and (3) were at an appropriate level for assessing undergraduate GB knowledge.

To what degree do the designed items validly and reliably measure student GB knowledge? We determined the degree to which the designed items validly and reliably measured student GB knowledge by analyzing student response reliabilities, the uni-dimensionality of the assessment, and item performance statistics using both CTT and IRT methodologies. Student responses were found to be reliable through the use of KR-20, Rasch item, and Rasch person reliability statistics.

PCAr was used to examine the uni-dimensionality of the assessment. While this analysis indicated that our assessment had little supporting evidence for multi-dimensionality, it must be noted that the construct of GB knowledge is very broad. Furthermore, less than 50% of the score variance was explained by our measures, indicating that there is a possibility that there are many dimensions on our assessment each explaining only a small portion of the variance that may be below the "noise" level in our analysis. These dimensions may account for a larger percentage of the variance if our item pool was larger. It will be of great importance to analyze the assessments' uni-dimensionality with a larger and more diverse population in order to further the argument against assessment multi-dimensionality.

Our CTT and IRT analyses helped to further establish support for item validity, and indicated that seventeen of the items required no modification, ten items required major modifications, nine items needed some item review, and the remaining eight items required at least one distractor to be modified. As this is the first iteration of this assessment, it is not

surprising that many modifications were required. It is of interest that both “Item 43” and “Item 44” are included in the items that need major modification as they both belong to the subtopic “Variation: Structural”; it is not clear if this subtopic is particularly hard to assess or if the items were poorly designed. Furthermore, it is also of interest that almost twice as many items had poor discrimination values on assessment Form B than on Form A. Although this could be a result of chance, there may be some undiscovered variable that accounts for this result, or perhaps this is a drawback of using CTT statistics to evaluate our assessment. Finally, it is also of interest to note that one of the items that required major modification (due to poor discrimination) was the easiest item on the assessment (P = 94%). Further analysis revealed that if 6% of the incorrect responses were present in the lower-scoring students, the discrimination value could only be as high as 0.12 (which is less than the criterion value). This indicates that some of item discrimination scores need to be reevaluated based on their difficulty scores.

In what ways do the designed items differentiate (or fail to differentiate) between undergraduate students with different levels of genomics and bioinformatics content exposure? We determined that our assessment differentiated between undergraduate students with different levels of genomics and bioinformatics content exposure based on our comparisons of student raw scores and Rasch person-measures. Both raw score and Rasch person-measure ANOVAs indicated that differences in mean scores between assessment forms, content exposure level, and gender were statistically significant. However, only the statistical differences between the content exposure levels had any practical value (based on effect size calculations). It was likely that the statistically significant differences between form and gender were an artifact of sample size. Also of interest was the interaction effect between form and exposure level. Our profile plot indicated that the significant difference was present only at the high content exposure

level, suggesting that while the easier items on the two assessment forms were comparable, those which discriminate at a higher difficulty may not have been equally distributed between the assessment forms.

Both our CTT and Rasch analysis indicated statistical and practical significance between the scores of students at various content exposure levels. It is interesting to note that the same conclusions would have been made based on the results of both analyses. This finding adds to the robustness of instrument quality. Furthermore, while many researchers are beginning to abandon the use of raw score counts (due to more modern definitions of measurement), perhaps these results indicate that the counting methodology may still hold value as a quick and non computer-based approach to estimate instrument quality.

To what extent do sources of validity and reliability evidence inform us about our assessment's construct validity? Finally, we were able to determine the degree to which our validity and reliability evidence informed us about the instrument's construct validity. Through our research, we have established the following validity evidence; (1) content validity through our expert survey analyses, textbook analyses, expert item evaluation, CTT analyses, and IRT analyses; (2) internal structure evidence through the use of PCAr; and (3) generalization evidence through our ANOVA and LSD analyses of different populations with different expertise levels. We have also established reliability evidence by examining the internal consistency using KR-20, and item and person reliability analyses. While the body of evidence that we have gathered is substantial, we have yet to establish substantive validity evidence or consequential validity evidence. Such work is planned for both of these areas in order to expand the evidence for construct validity.

Moving forward. The results discussed above are major steps forward in the development of a robust GB assessment with validity and reliability evidence necessary for evidence-based measurement of GB knowledge and evidence-based educational reform. We plan to move our research forward though the modification of assessment items and further validation of our assessment. Only through the completion of these steps can we begin to move the teaching and learning within GB in a scientific and evidence-based direction.

Acknowledgments

We thank several colleagues including Dr. Lynn Caporale, Dr. Harald Vaessin, Dr. Brady Bernard, Dr. Brad Goodner, Dr. Gregory Booton, Dr. Simon Queenbourough, Dr. Iris Meier, and Dr. Guo-Liang Wang for helping make our assessment design and pilot as constructive as possible. Portions of this work were funded by a NSF CCLI program grant 0837397. Any opinions, findings, and conclusions or recommendations expressed in this publication are those of the authors and do not necessarily reflect the view of the National Science Foundation. The methodology for the administration of our assessment was submitted to the Ohio State University IRB and found to exempt from review (Protocol Number 2012E0509).

References

AERA, APA, & NCME (1999). Standards for Educational and Psychological Testing. Washington, D.C.: Authors.

Bloom BS, Englehart MD, Furst EJ, hill WH, Krathwohl DR (1956). Taxonomy of Educational Objectives: The Classification of Educational Goals, Handbook I: Cognitive domain. White Plains, NY: Lonman.

Bond TG, Fox CM (2001). Applying the rasch model: Fundamental measurement in the human sciences. New Jersey, Lawrence Earlbaum.

Brennan RL (2006). Educational Measurement (Fourth ed.). Rowman and Littlefield Publishers.

Brown TA (2007). Genomes 3 (3rd ed.). New York, NY: Garland Science Publishing.

Campbell MA, Heyer LJ (2007). Discovering genomics, proteomics and bioinformatics (2nd ed.). San Francisco, CA: Pearson Education.

Campbell C, Nehm RH (2011). Assessing the educational efficacy of genomics and bioinformatics curricula and labs. Paper presented at Society for the Advancement of Biology Education Research (SABER) conference. Minneapolis, Minnesota. July 29-30.

Campbell C, Nehm RH (2012). Building new assessments for the "New Biology": Establishing content validity for a genomics and bioinformatics test. Paper presented at National Association for Research in Science Teaching (NARST) annual conference, Indianapolis, IN, March 25-March 28.

Campbell C, Nehm RH (in press). A Critical analysis of assessment quality in genomics and bioinformatics education research. CBE Life Sciences Education.

Cohen J (1988). Statistical power analysis for the behavioral sciences (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.
Doran R (1980). Basic measurement and evaluation of science instruction. Washington D.C.: National Science Teachers Association.
Downing SM (2006). Selected-response Item Formats in test development. Handbook of Test Development 287-301. New Jersey: Lawrence Erlbaum Associates Inc.
Griffiths AJF, Wessler SR, Lewontin RC, Carroll SB (2008). Introduction to genetic analysis (9th ed.). New York, NY: W.H. Freeman and Company.
Haladyna TM, Downing SM (1993). How many options is enough for a multiple-choice test item? Educational and Psychological Measurement *53*, 999-1010.
Hartl DL (2011). Essential genetics: A Genomics perspective (5th ed.). Sudbury, MA: Jones and Bartlett Publishers.
Hartl DL, Jones EW (2009). Genetics: analysis of genes and genomes (7th ed.). Sudbury, MA: Jones and Bartlett Publishers.
Hartwell LH, Hood L, Goldberg ML, Reynolds AE, Silver LM, Veres RC (2008). Genetics: From genes to genomes (3rd ed.). New York, NY: McGraw-Hill Companies.
Higgs PG, Attwood TK (2005). Bioinformatics and molecular evolution. Malden, MA: Blackwell Publishing.
Gibson G, Muse SV (2009). A Primer of genome science (3rd ed.). Sunderland, MA: Sinauer Associates, Inc. Publishers.
Linacre JM (2012). A user's guide to WINSTEPS, MINISTEP Rasch-model computer programs v3.75.0. Retrieved from www.winsteps.com.
Messick S (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. Am Psychol *50*(9), 741-749.
Messick S (1999). Assessment in higher education: Issues of access, quality, student development, and public policy: A Festschrift in honor of Warren W. Willingham . New Jersey: Lawrence Erlbaum Associates Inc.
National Research Council (2009). A new biology for the 21st century: Ensuring the United States leads the coming biology revolution. Washington, DC: The National Academies Press.
Nehm RH, Schonfeld IS (2008). Measuring knowledge of natural selection: A comparison of the CINS, an open-response instrument, and an oral interview. JRST *45*(10), 1131–1160.
Nitko AJ, Brookhart SM (2007). Educational Assessment of Students (5th ed.). New Jersey: Pearson Prentice Hall.
Pierce BA (2012). Genetics: A Conceptual approach (4th ed.). New York, NY: W.H. Freeman and Company.
Pevsner J (2009). Bioinformatics and functional genomics (2nd ed.). Hoboken, NJ: Wiley and Sons, Inc.
Primrose SB, Twyman RM (2004). Principles of genome analysis and genomics (3rd ed.). Malden, MA: Blackwell Publishing.
Salzberger T (2010). Does the Rasch model convert an ordinal scale into an interval scale? Rasch Measurement Transactions *24*(2), 1273-1275. Retrieved from http://www.rasch.org/rmt/rmt242a.htm
Snustad DP, Simmons MJ (2009). Principles of genetics (5th ed.). Hoboken, NJ: Wiley and Sons, Inc.

Wright BD, Linacre JM (1994). Reasonable mean-square fit values. Rasch Measurement Transactions *8*(3), 370.

Appendix

Table 1

ANOVA of student raw total scores

Source	Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared	Observed Power[b]
Form	281.834	1	281.834	19.908	.000	.037	.994
Exposure	3693.839	2	1846.919	130.465	.000	.334	1.000
Gender	61.951	1	61.951	4.376	.037	.008	.551
Form * Exposure	357.537	2	178.768	12.628	.000	.046	.997
Form * Gender	.887	1	.887	.063	.802	.000	.057
Exposure * Gender	7.058	2	3.529	.249	.779	.001	.089
Form * Exposure * Gender	9.088	2	4.544	.321	.726	.001	.101
Error	7361.357	520	14.156				
Total	125109.000	532					

a. R Squared = .384 (Adjusted R Squared = .371)

b. Computed using alpha = .05

Table 2

Fisher LSD post-hoc for student raw total scores

	(I) Exposure	(J) Exposure	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval Lower Bound	Upper Bound
LSD	Low	Medium	-5.07*	.383	.000	-5.82	-4.31
		High	-9.66*	.629	.000	-10.89	-8.42
	Medium	Low	5.07*	.383	.000	4.31	5.82
		High	-4.59*	.574	.000	-5.72	-3.46
	High	Low	9.66*	.629	.000	8.42	10.89
		Medium	4.59*	.574	.000	3.46	5.82

Based on observed means.

The error term is Mean Square(Error) = 14.156.

*. The mean difference is significant at the 0.05 level.

Table 3

ANOVA of student person measure scores

Source	Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared	Observed Power[b]
Form	9.599	1	9.599	18.468	.000	.035	.990
Exposure	126.528	2	63.264	121.721	.000	.322	1.000
Gender	3.137	1	3.137	6.035	.014	.012	.689
Form * Exposure	12.936	2	6.468	12.445	.000	.046	.996
Form * Gender	.000	1	.000	.000	.988	.000	.050
Exposure * Gender	.605	2	.302	.582	.559	.002	.147
Form * Exposure * Gender	.774	2	.387	.744	.476	.003	.176
Error	266.110	512	.520				
Total	440.809	524					

a. R Squared = .376 (Adjusted R Squared = .363)

b. Computed using alpha = 0.05

Table 4

Fisher LSD post-hoc for student person measure scores

	(I) Exposure	(J) Exposure	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval Lower Bound	95% Confidence Interval Upper Bound
LSD	Low	Medium	-.9149*	.07474	.000	-1.0617	-.7681
		High	-1.8562*	.12124	.000	-2.0944	-1.6180
	Medium	Low	.9149*	.07474	.000	.7681	1.0617
		High	-.9413*	.11000	.000	-1.1574	-.7252
	High	Low	1.8562*	.12124	.000	1.6180	2.0944
		Medium	.9413*	.11000	.000	.7252	1.1574

Based on observed means.

The error term is Mean Square(Error) = .520.

*. The mean difference is significant at the 0.05 level.

Table 5

Z Score from a test of two independent proportions. Critical Value = 3.04

Item #	Item Name	Z Score
1	Bioinformatics 1	**8.17**
2	Bioinformatics 2	**7.15**
3	Comparative genomics 1	**4.86**
4	Comparative genomics 2	2.07
5	Definition of a genome 1	**4.41**
6	Definition of a genome 2	1.22
7	Expression 1	**7.08**
8	Expression 2	3.01
9	Expression: Analysis 1	3.00
10	Expression: Analysis 2	*-0.32*
11	Genome annotation by pattern 1	1.32
12	Genome annotation by pattern 2	*-1.70*
13	Genome annotation by similarity 1	2.64
14	Genome annotation by similarity 2	1.27
15	Genome content and organization 1	**5.31**
16	Genome content and organization 2	*-1.97*
17	Genome evolution 1	1.55
18	Genome evolution 2	**3.24**
19	Genome mapping 1	**7.73**
20	Genome mapping 2	2.84
21	Genome sequencing: Concepts of 1	**3.12**
22	Genome sequencing: Concepts of 2	0.85
23	Genome sequencing: Techniques 1	0.17
24	Genome sequencing: Techniques 2	**9.97**
25	Genome structure 1	**8.26**
26	Genome structure 2	2.17
27	Integration of genomics with biology 1	**7.10**
28	Integration of genomics with biology 2	1.87
29	Metagenomics 1	2.74
30	Metagenomics 2	1.63
31	Regulation: Epigenetics 1	**3.76**
32	Regulation: Epigenetics 2	1.61
33	Regulation: RNA 1	0.67
34	Regulation: RNA 2	**3.92**
35	Regulation: Transcription factors 1	0.66
36	Regulation: Transcription factors 2	**6.29**
37	Sequence databases 1	2.97
38	Sequence databases 2	**3.97**
39	Variation: Linkage disequilibrium and haplotypes 1	**3.90**
40	Variation: Linkage disequilibrium and haplotypes 2	2.25
41	Variation: Nucleotide 1	**5.16**

42	Variation: Nucleotide 2	1.60
43	Variation: Structural 1	0.25
44	Variation: Structural 2	*-2.17*

Bold = significant difference between low and high scores
Italic = Negative Z-score; low group scored better than high group