Tornabene *et al. Evo Edu Outreach (2018) 11:6*
https://doi.org/10.1186/s12052-018-0082-x

Evolution: Education and Outreach

RESEARCH ARTICLE

Open Access

Testing validity inferences for Genetic Drift Inventory scores using Rasch modeling and item order analyses

Robyn E. Tornabene[1*], Erik Lavington[2] and Ross H. Nehm[1]

Abstract

Background: Concept inventories (CIs) are commonly used tools for assessing student understanding of scientific and naive ideas, yet the body of empirical evidence supporting the inferences drawn from CI scores is often limited in scope and remains deeply rooted in Classical Test Theory. The Genetic Drift Inventory (GeDI) is a relatively new CI designed for use in diagnosing undergraduate students' conceptual understanding of genetic drift. This study seeks to expand the sources of evidence examining validity and reliability inferences produced by GeDI scores. Specifically, our research focused on: (1) GeDI instrument and item properties as revealed by Rasch modeling, (2) item order effects on response patterns, and (3) generalization to a new geographic sample.

Methods: A sample of 336 advanced undergraduate biology majors completed four equivalent versions of the GeDI. Rasch analysis was used to examine instrument dimensionality, item fit properties, person and item reliability, and alignment of item difficulty with person ability. To investigate whether the presentation order of GeDI item suites influenced overall student performance, scores were compared from randomly assigned, equivalent test versions varying in item-suite presentation order. Scores from this sample were also compared with scores from similar but geographically distinct samples to examine generalizability of score patterns.

Results: Rasch analysis indicated that the GeDI was unidimensional, with good fit to the Rasch model. Items had high reliability and were well matched to the ability of the sample. Person reliability was low. Rotating the GeDI's item suites had no significant impact on scores, suggesting each suite functioned independently. Scores from our new sample from the NE United States were comparable to those from other geographic regions and provide evidence in support of score generalizability. Overall, most instrument features were robust. Suggestions for improvement include: (1) incorporation of additional items to differentiate high-ability persons and improve person reliability, and (2) re-examination of items with redundant or low difficulty levels.

Conclusions: Rasch analyses of the GEDI instrument and item order effects expand the range and quality of evidence in support of validity claims and illustrate changes that are likely to improve the quality of this (and other) evolution education instruments.

Keywords: Concept inventories, Assessment, Validity evidence, Psychometrics, Genetic drift, Evolution

*Correspondence: robyn.tornabene@stonybrook.edu
[1] Science Education Program, Institute for STEM Education, Stony Brook University, 092 Life Sciences Building, Stony Brook, NY 11794-5233, USA
Full list of author information is available at the end of the article

Introduction

The accurate measurement of student understanding is an essential feature of educational practice because it provides evidence-based insights into students' conceptual ecologies, guides learning progression development, and permits empirical evaluation of the efficacy of alternative educational interventions (National Research Council 2001). A diverse array of assessment tools and types have been developed for evolution educators (Table 1). They range from static, multiple-choice formats (e.g., Price et al. 2014) to open-ended questions whose answers can be scored by computers (e.g., Moharreri et al. 2014). Available assessment tools cover many different evolutionary concepts, including natural selection, evo-devo, genetic drift, and macro-evolution. These assessments vary significantly in the types of information that they can reveal about student understanding, in the situations in which they are most appropriately implemented, and in the robustness of the inferences that they are able to support (American Association for the Advancement of Science (AAAS) 2011; American Educational Research Association, American Psychological Association, and National Council on Measurement in Education (AERA, APA, NCME) 2014; Nehm and Schonfeld 2008).

Concept inventories (CIs) are a type of research-based educational assessment designed to rapidly reveal (through easy administration and scoring) students' preferences for normative (i.e., scientifically accurate) or non-normative (e.g., preconceptions, misconceptions) facets of core ideas (e.g., natural selection, genetic drift) (Nehm and Haertig 2012, p. 56–57). Although CIs have become indispensable tools for assessing undergraduate students' conceptual understandings of many core ideas in the sciences (e.g., force and motion, chemical bonding), few have been carefully evaluated in terms of (1) the forms of validity outlined in the *Standards for Educational and Psychological Testing* (American Educational Research Association, American Psychological Association, and National Council on Measurement in Education (AERA, APA, NCME) 2014), (2) item order effects and associated response biases (Federer et al. 2015, 2016), or (3) item properties using ratio-scaled data (generated by Rasch or Item Response Theory [IRT] analyses; Boone, Staver and Yale 2014). Consequently, validity evidence—that is, evidence that the measures derived from CIs accurately

Table 1 Evolution education instruments measuring knowledge of evolutionary processes: potential to elicit scientific and naive ideas about adaptive and non-adaptive evolution

Instrument	Format[a] and target population	Conceptions measured[b]			
		NS-S	NS-N	GD-S	GD-N
Bishop and Anderson's diagnostic instrument (Bishop and Anderson 1990)	Combination MC and OR[c]: undergraduates (introductory biology non-majors)	Intended	Intended	Possible[d]	Possible[d]
Concept Inventory of Natural Selection (CINS) (Anderson, Fisher and Norman 2002)	20 MC: undergraduates	Intended	Intended		
Assessing Contextual Reasoning about Natural Selection (ACORNS) (Nehm et al. 2012)	Flexible number OR: undergraduates	Intended	Intended	Possible[d]	Possible[d]
Conceptual Assessment of Natural Selection (CANS) (Kalinowski et al. 2016)	24 MC: undergraduates (introductory biology majors)	Intended	Intended	[e]	
Daphne Assessment for Natural Selection (DANS) (Furtak et al. 2014)	26 MC: high school	Intended	Intended		
Genetic Drift Inventory (GeDI) (Price et al. 2014)	22 TF: undergraduates (upper-division biology majors)		Intended	Intended	Intended
Evo-devo concept inventory (Perez et al. 2013)	11 MC: undergraduates	Intended	Intended		
Measure of understanding of macroevolution (MUM) (Nadelson and Southerland 2009)	27 MC and 1 OR: undergraduate	OR: possible[f]	OR: possible[f]	OR: possible[f]	OR: possible[f]

N.B. Readers interested in genetics-focused concept inventories that contain individual items dealing with non-adaptive change may wish to consult Price et al. (2014)

[a] *MC*Multiple choice, *OR* Open response, *TF* True–false

[b] *NS-S* natural selection-scientific ideas, *NS-N* natural selection-naive ideas, GD-S genetic drift-scientific ideas, *GD-N* genetic drift-naive ideas; "Intended" indicates that the instrument intentionally targeted ideas of this type

[c] Bishop and Anderson's instrument includes 2 OR, 3 MC with OR explanation, and 1 question about belief in evolution

[d] Open response format affords the possibility of capturing reasoning about genetic drift, although, in line with instrument's intent, scoring guide focuses on natural selection

[e] Includes one question (item 20) asking whether chance plays a role in whether a cactus will produce a seedling

[f] MC items address macroevolution. OR item asks student to explain how two species might have arisen from one. Authors state that item does not address speciation by means beyond natural selection, though they include a student response mentioning genetic drift

reflect the construct of interest–remains limited. Given the centrality of accurate measurement to evidence-based educational practices, evolution education research must include the study of instrument quality. Such studies help to support instructional decisions firmly rooted in high-quality evidence.

Given the paucity of work on evolution education instrument quality (Nehm et al. 2010), our study examines the psychometric properties of a relatively new evolution education instrument known as the Genetic Drift Inventory (GeDI). As the only assessment instrument focusing on non-adaptive evolutionary mechanisms, the GeDI fills a crucial gap in available evolution education instruments and holds potential to offer insights into a much neglected area of student thinking about evolution. To date, validity evidence for the GeDI remains limited to Classical Test Theory frameworks (Price et al. 2014), despite the availability of more robust IRT and Rasch approaches (Boone et al. 2014). In order to build a larger body of validity evidence in support of evolution education assessments in general, and to empirically examine the strengths and weaknesses of the inferences that may be drawn from GeDI scores in particular, our study explores three research questions: (1) How well does the GeDI function when studied within the context of the Rasch model? (2) Does the presentation order of instrument scenarios (and associated item suites) impact measures of student understanding? And (3) Does the GeDI measure student knowledge in a manner that is generalizable across geographic regions of the United States (e.g., Northeast, Southeast, and Midwest) when administered to students of similar academic backgrounds? Prior to discussing our psychometric approach, we begin with a brief review of the position of genetic drift within evolution education, continue with an overview of Classical Test Theory and Item Response Theory frameworks for instrument evaluation, and end with a summary of GeDI instrument properties and prior validation work relative to these frameworks.

Genetic drift and evolution education

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

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

The Genetic Drift Inventory

The Genetic Drift Inventory (known as the GeDI; Price et al. 2014) is a 22-item CI designed to measure advanced undergraduate biology majors' understanding of four key concepts and six alternative conceptions (or "misconceptions") of genetic drift. To date, it is the only concept inventory to focus on non-adaptive evolutionary processes (Table 1). The GeDI features four scenarios, each followed by one to three question stems containing a number of associated agree-disagree statements (i.e., items; see Table 2 for details). The 22 items target an individual key concept (15 items) or a misconception (7 items). Misconceptions targeted by the GeDI are limited to those expected to be harbored by upper division majors whose knowledge of genetic drift is developing but often conflated with other evolutionary mechanisms (see Price et al. 2014 for more information on misconception delineation by expertise levels). For scoring, GeDI authors recommend that all items are given equal weight (e.g., 17/22 = 77%). To compensate for the high guessing rate for dichotomous questions, GeDI developers

Table 2 GeDI scenarios and associated items

Scenario	Items	Example
1	1–8	Small subpopulation of land snails colonize a new island
2	9–11	Dung beetles geographically isolated by canals
3	12–18	Biologist randomly selects fruit flies to breed in captive populations
4	19–22	Nearsighted island population of humans before and after a devastating storm

recommended: comparing raw scores before and after instruction, using higher than usual raw score cut-points to define success, or consideration of only the percentage correct above 50% (the score that could potentially be obtained by guessing alone) (Price et al. 2014). All of these scoring recommendations are grounded in Classical Test Theory (see below).

Instrument evaluation using Item Response Theory

The frameworks for developing and evaluating assessment instruments have changed substantially over the past few decades, and faculty at all educational levels need to be familiar with these changes in order to understand the strengths and weaknesses of the measures that are derived from evolution education instruments (American Educational Research Association, American Psychological Association, and National Council on Measurement in Education (AERA, APA, NCME) 2014). Classical Test Theory (CTT) and Item Response Theory (IRT) are two conceptual and empirical frameworks commonly used for analyzing and evaluating measurement instruments.

IRT and Rasch frameworks address many inherent limitations of CTT (Bond and Fox 2007; Nehm and Schonfeld 2008; Boone et al. 2014). A broad advantage is the existence of diverse IRT and Rasch models suitable for different types of data (unidimensional, multidimensional, dichotomous, polytomous, large and small data sets, and multi-matrix sampling), permitting analyses to be more closely matched to study type. IRT/Rasch analyses report ratio-scaled scores for both persons and items on the same Logit scale, facilitating more accurate score inferences and comparisons between persons, items, or persons and items. A variety of fit statistics are also calculated to allow more thorough evaluation of item, person, and instrument function. Some of the salient features of IRT/Rasch that are relevant to our analysis of the GeDI instrument are reviewed below. Readers interested in a more technical treatment of these frameworks are encouraged to consult Bond and Fox (2007), Boone et al. (2014), and de Ayala (2009). Overall, IRT and Rasch frameworks afford a robust psychometric evaluation and inferential potential for educational measurement instruments. Our analysis of the GeDI provides an example of how Rasch analysis can offer greater insights into the measurement capabilities and limitations of measurement instruments.

Item Response Theory in educational measurement

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

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

Ratio-scale logit scores for persons and items

The vast majority of evolution education instruments have been developed and evaluated using CTT as a guiding framework. IRT/Rasch frameworks address many inherent limitations of CTT (Bond and Fox 2007; Nehm

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

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

Instrument dimensionality

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

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

Wright maps

A display of all person and item measures for a unidimensional construct on a shared logit scale (commonly known as a Wright map, e.g., Fig. 1) is another powerful and unique application of Rasch. This side-by-side comparison enables researchers to examine the alignment of test items to test taker "ability" and to identify possible measurement gaps (i.e. difficulty/ability ranges in which items are lacking). Items are represented by their respective number on the right side of the scale, while persons are represented by "X's" on the left side of the scale. Given the probabilistic nature of the analysis, each person has a 50% chance of correctly answering an item with an equivalent measure. In a well-designed instrument, question difficulty should be aligned with test-taker ability, with items present that are able to differentiate among

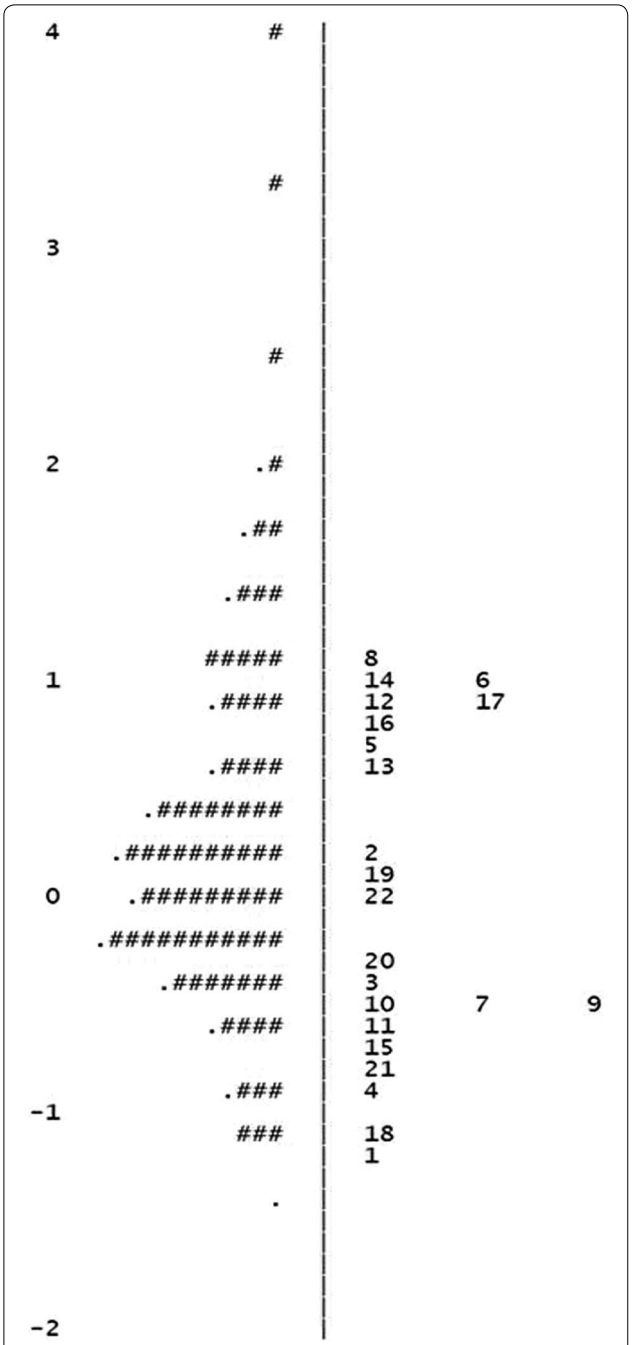

Fig. 1 Wright map derived from Rasch analysis of GeDI responses. The distribution of persons (on the left) and GeDI items (on the right) are illustrated on the same log interval (−2 to 4) scale. Each # = 4 persons

learners at all ability levels. Thus, instrument evaluation using a Wright map includes examining the match of the "spread" of test takers to the "spread" of test items. Items that are too easy appear below the lowest test takers (having been estimated to be correctly answered by everyone, these items likely add little value to the measures), while items that are too difficult appear above the highest test takers (these items may be too challenging for the sample). If all of the persons are plotted above the highest item or below the lowest item, then the items lack alignment with ability level. Multiple items aligned at the same difficulty levels on the Wright map, and testing the same concept or misconception, add little to measurement and are candidates for elimination. Large clusters of persons at the same ability level indicate locations where additional items could be added to better separate their abilities. Overall, the Wright map is a useful visual tool for examining instrument properties and person-item relationships.

Item and person fit

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

In Rasch measurement, fit is expressed as weighted ("infit") or unweighted ("outfit") values for the mean square parameter (MNSQ), and calculation of fit is based on a Chi square test of how well the empirical data fit the Rasch model (Bond and Fox 2007, p. 238). For a standard multiple choice assessment, MNSQ values above 1.3 are considered to be "underfitting", indicating that the response pattern for that item is erratic. Values below 0.7 are considered to be "overfitting", indicating that the response pattern is overly predictable. Both overfit and underfit suggest that the item is not functioning properly (i.e., eliciting information consistent with test-taker ability). Cut off values of 0.7 and 1.3 are used for the MNSQ parameter to ensure an adequate match between the empirical data and the statistical model (Boone et al. 2014; Bond and Fox 2007). Z-Standard (ZSTD) scores are transformed *t* test statistics that report the probability of MNSQ scores occurring by chance when the data fit the Rasch model (Linacre 2017). Ideal ZSTD scores range

from 0 to 2. However, as sample size increases, accumulation of random responses tends to elevate ZSTD scores (Smith et al. 2008). For this reason, and because ZSTD statistics are based on MNSQ statistics, ZSTD values are considered secondary to MNSQ scores. Depending upon measurement goals and sample sizes, ZSTD scores may be ignored if MNSQ values are acceptable (Linacre 2017). With multiple indicators of fit that correspond to different causes of misfit as well as parameters to report the probability of fit statistics, Rasch and IRT provide a much more detailed characterization of item fit properties compared to CTT.

Item and person reliability

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

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

Missing data

A key benefit of IRT and Rasch modeling is the ability to readily accommodate "missing" data. Because person estimates are based on the probability a person will correctly respond to a given item of a particular difficulty, failure to answer a few items among many others whose difficulty is known does not significantly impact person estimates; the model is able to predict how a person would likely have answered a skipped question based on responses to items of similar difficulty. Similarly, because item measures are estimated based on the probability that a person of a determined ability will select a correct answer for that item, item estimates are not impacted by the absence of a few individuals' responses from among many responses of known ability. These properties ensure that Rasch person scores are item-independent and item scores are sample-independent, characteristics which afford researchers the widespread benefit of being able to confidently utilize partially completed student response sets. Accommodation of missing data is also essential for computer adaptive testing (Bond and Fox 2007) and multi-matrix studies in which participants are assigned only a subset of items from the total collection of questions (cf. Sirotnik and Wellington 1977; e.g., Schmiemann et al. 2017). Such designs allow testing of a wider variety of items while minimizing participant test fatigue. In sum, Rasch and IRT hold considerable potential for expanding the body of empirical evidence on instrument quality, yet remain broadly underutilized in science education measurement.

Item order effects on student performance

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

Generalizability of instrument scores

Evidence for generalization validity is important to substantiate claims that an instrument measures a trait in the same manner across different populations and administration contexts. Instruments are designed to measure a specific construct under specific circumstances, such as a particular educational level (e.g., undergraduate biology majors, elementary students) under certain

administration conditions (e.g., unproctored computerized testing, timed paper-and-pencil tests), and for particular purposes (e.g., formative evaluation of instructional interventions, employment screening). Explicit delineation of such contexts and evidence to support validity and reliability of inferences generated under these circumstances should accompany instruments (American Educational Research Association, American Psychological Association, and National Council on Measurement in Education (AERA, APA, NCME) 2014). Under alternative administration contexts (e.g., sample populations, testing conditions), items are subject to differing interpretations or stress factors which may bias responses. For instance, a question may be beyond the comprehension level of a group, may be scrutinized more stringently by those with greater subject expertise, or may contain terms whose meaning differs according to the cultural or regional background of a sample. Accordingly, biased item responses compromise the validity of inferences about the latent trait (American Educational Research Association, American Psychological Association, and National Council on Measurement in Education (AERA, APA, NCME) 2014). When an instrument is used in a new context, evidence is needed to support the validity and reliability of inferences generated in the new context.

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

Evidence used to support instrument quality

The GeDI has only been evaluated using Classical Test Theory methods despite many known limitations of using raw data to interpret item and instrument properties (as discussed above; Boone et al. 2014). A summary of the forms of evidence used to support validity inferences for the GeDI are shown in Table 3. The present study expands upon prior validity and reliability work by (1) employing Rasch Modeling, which produces more accurate ratio-scaled scores and can contribute evidence to examine dimensionality, construct validity, internal structure validity, item properties, and reliability, (2)

Table 3 Summary of validity and reliability evidence for the GeDI

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

[a] Based on Campbell and Nehm (2013); Messick (1995); Nitko and Brookhart (2010)

examining item order effects, and (3) studying a participant population from a new geographic region of the country (Table 3).

Methods

Item order

The GeDI features four scenarios, each followed by one to three question stems containing a number of associated agree-disagree statements (i.e., items; see Table 2). The GeDI's scenarios differ in situational features (cf. Table 2) and difficulty, two factors which have been shown to demonstrate item-order effects in prior studies (reviewed in Federer et al. 2015). In order to determine whether the sequence of scenarios and related items within the GeDI instrument impacted student performance, four complete forms of the GeDI were generated, which differed only in the presentation sequence of scenarios. A four-by-four Latin square design was used to rotate scenario sequence among the test forms (see Table 4). Each of the four scenarios (and related items) constituted a block in the square; the original order of the scenarios and items (Price et al. 2014) was used to seed the Latin square, and the original order of the items within a block was maintained throughout all forms (see Table 4).

Sample and administration

The GeDI forms (Table 4) were administered online using a learning management system in the spring semester of an upper division (300-level) genetics class at a large, Northeastern Doctoral-granting university. This course was chosen because it aligns with the target population for GeDI use and is among the course types used in the development and initial validation studies of the GeDI (Price et al. 2014). Students were randomly assigned to one of four experimental groups, each of which had access to only one of the four forms of the assessment (Table 4). Extra credit was offered as an incentive for participation. Students were allowed one attempt to complete the activity, with a generous 60-min time limit to allow a maximum of 2–3 min per response. While we did not collect data on exact completion time, it was less than 1 h. Random student identification numbers were assigned to anonymize response data. The assessment was open for a period of 1 week beginning in the 10th week of the semester, prior to which no instruction relating to genetic drift had occurred. Of the 480 students enrolled, 336 (70%) completed the assessment in the following distribution: $n_{form\ 1} = 91$, $n_{form\ 2} = 78$, $n_{form\ 3} = 80$ and $n_{form\ 4} = 87$.

Table 4 Design of GeDI forms

	Position 1	Position 2	Position 3	Position 4
GeDI form 1 sequence	Scenario 1	Scenario 2	Scenario 3	Scenario 4
GeDI form 2 sequence	Scenario 2	Scenario 3	Scenario 4	Scenario 1
GeDI form 3 sequence	Scenario 3	Scenario 4	Scenario 1	Scenario 2
GeDI form 4 sequence	Scenario 4	Scenario 1	Scenario 2	Scenario 3

The presentation sequence of the four scenarios (and accompanying item suites that comprise the GeDI) was rotated among four equivalent test forms using to a Latin square design. A Latin square is an array of *n* rows and *n* columns, with each row and each column containing units 1 through *n* exactly once. By rotating experimental treatments in this manner, the researcher can generate data to determine whether an adjacent treatment (or, in this case, item suite placement) influences overall performance. For a description of scenarios and associated items see Table 2

Data analysis

In order to empirically evaluate the validity and reliability inferences derived from GeDI scores, Rasch modeling was performed using WINSTEPS v 3.68.2. Dimensionality was examined via a principal components analysis (PCA) of Rasch residuals. The overall fit of items and persons to the unidimensional Rasch model were examined by infit and outfit mean square (MNSQ) values and Z standard (ZSTD) values. A Wright map was generated to visualize item difficulty relative to test-taker ability, and individual item fit values were considered. Item reliability was calculated to determine whether responses were varied enough to confirm the item difficulty structure, person reliability was calculated to determine whether the items differentiated among achievement levels sufficiently. To determine if item order impacted test performance, a one-way ANOVA was performed on Rasch-scaled scores for the four GeDI forms. Finally, total scores and item difficulty ranks were compared across administrations in order to examine score generalizability.

Results

Dimensionality

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

Overall model fit

Overall item fit and person fit values are measures of how well a dataset fits the Rasch model. Values outside of the acceptable range (MNSQ = 0.7–1.3 and Z standard values < 2.0) suggest that test takers were responding in a manner that was either overly predictable, very erratic, or inconsistent with their ability. Excellent overall person fit values (infit MNSQ = 1.0, ZSTD = 0.0; outfit MNSQ = 0. 99, ZSTD = 0.0) and item fit values (infit MNSQ = 1.0, ZSTD = 0.0; outfit MNSQ = 0. 99, ZSTD = − 0.1) were apparent (Table 5) and indicated that the participant sample responses fit the Rasch model very well.

Individual item fit

Rasch infit and outfit MNSQ and ZSTD values were used to assess how well individual GeDI items aligned with the student population and with the Rasch model. Infit and outfit MNSQ values for all test items were within acceptable ranges (Table 5) and are thus functioning to elicit responses consistent with test-taker ability. Five items (2, 9, 16, 19, and 22) had infit and/or outfit ZSTD values outside of the acceptable range. According to Linacre (2017), misfitting ZSTD scores are very sensitive to sample size, and may be disregarded when samples are large (over 300 observations) and MNSQ scores are acceptable. This is because ZSTD values reflect how perfectly data fit the Rasch model rather than how usefully data fit the model, and in large samples (over 300 observations), the accumulation of rare individual atypical responses can inflate ZSTD scores without having a bearing on the usefulness of the data.

Table 5 GeDI Rasch fit properties

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

Italics values refer to higher than expected values

Wright map

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

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

Reliability

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

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

Item order effects

Raw score group means for all four forms of the GeDI were very similar, ranging between 12.02 and 12.20 (SD 3.30–3.98) out of a possible 22 (Table 6). A one-way ANOVA confirmed that there was no statistically significant difference in Rasch-scaled scores for each of the four GeDI forms ($F[3332] = 0.038$, $p = 0.990$). This result indicates that the order of scenarios did not impact overall performance. Comparisons of mean item measures for the first, second, third, or fourth rotation position showed no apparent differences in item difficulty when controlling for the number of statistical tests (Fig. 2). Detailed information on item measures for all items and rotation positions is available in Additional file 1.

Comparisons with other undergraduate participant samples

Given that evolution acceptance, religion, and demographic variables differ across the United States, it is important to determine if instrument properties generalize. GeDI scores from our sample of undergraduates from the Northeastern United States were nicely aligned with the scores obtained by Price et al. (2014) from similar courses from other regions of the country (Table 6). In particular, no significant difference was found between raw scores from the 300-level genetics class in our sample ($M = 12.35$, $SD = 3.59$) and those of 300-level genetics classes in the Midwest ($M = 11.94$, $SD = 3.35$; $t(475) = 0.481$, $p = 0.631$) or the Southeast ($M = 12.35$, $SD = 3.29$; $t(652) = 0.890$, $p = 0.374$). Similarity in scores across institutions indicates that the GeDI is functioning to elicit similar responses in comparable populations across the country and may suggest generalizability of score inferences (cf. Messick 1995). It should be noted that no Rasch-scaled scores are available from prior GeDI administrations so comparisons are limited to raw scores. Raw score similarity also provides limited evidence that Rasch-based validity measures obtained with our population may generalize to the GeDI as a whole, though this should be confirmed in future studies.

To examine whether individual items functioned similarly across administrations, item difficulty rank from our sample was compared to CTT-based item difficulty (*P*) rank from Price et al. 2014 (Table 7). Overall, most items maintained a similar or only slightly shifted difficulty order, though a few notable differences in item difficulty across administrations were found. Among items targeting key concepts, the hardest and easiest items maintained the same difficulty position and mid-level items showed only minor rearrangement. Item 3 (relating to a loss of variation associated with genetic drift), initially ranked as an easy item by Price et al. (2014),

Table 6 Comparison of performance on GeDI by form, course, and region

Course, region (number tested)	Mean of items correct (SD)
300-level genetics, Northeast (N = 336)[a]	12.11 (3.59)
Form 1 (n = 81)[a]	12.02 (3.30)
Form 2 (n = 78)[a]	12.15 (3.61)
Form 3 (n = 80)[a]	12.20 (3.98)
Form 4 (n = 87)[a]	12.09 (3.54)
300-level genetics, Southeast (N = 318)[b]	12.35 (3.29)
300-level genetics, Midwest (N = 141)[b]	11.94 (3.35)
300-level cell biology, Northwest (N = 51)[b]	13.35 (3.64)
300-level evolution, Northwest (N = 91)[b]	14.47 (3.78)
400-level evolution, Midwest (N = 60)[b]	16.66 (3.44)

All institutions were Doctoral-granting. Maximum number of correct items is 22

Raw scores were used for comparison as Rasch-scaled data were not available from prior studies

[a] Denotes present study

[b] Denotes data from Price et al. 2014

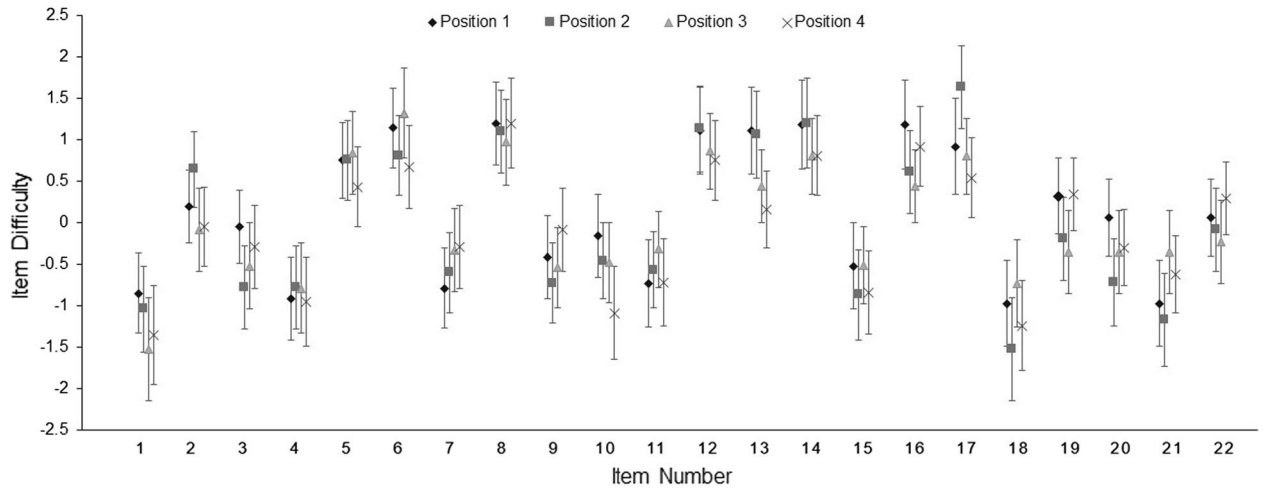

Fig. 2 Position effects on item difficulty. GeDI items appear on the X axis and mean item measures for each GeDI item are plotted on the Y axis. Symbols denote scenario position (1, 2, 3, or 4) in a counterbalanced rotation sequence of scenarios and accompanying item suites (see "Methods" section for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

Table 7 GeDI item difficulty rank in initial and present administrations

Items addressing key concepts		Items addressing misconceptions	
Price et al. (2014)	Present study	Price et al. (2014)	Present study
16	16	8	8
13	13	6	6
10	3	12	14
4	10	17	12
15	15	5	17
3	4	14	5
1	1	2	2
		11	19
		18	22
		20	20
		9	9
		7	7
		19	11
		22	21
		21	18

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

ranked among the more difficult key concept items in our administration. Among items targeting misconceptions, many items maintained a similar difficulty ranking, but items 14, 19 and 22 (all of which addressed "genetic drift is random mutation") were notably more difficult in the present administration. Items 11 and 18 (addressing "genetic drift is gene flow") were notably easier in the present administration. Overall, most of the GeDI questions maintained the general difficulty hierarchy across diverse samples, which may be used as evidence in support of generalization validity. Difficulty rank differences in items 3, 11, 14, 18, 19 and 22 should be investigated further.

Discussion

How well does the GeDI function when studied within the context of the Rasch model?

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

Rasch analysis is useful to help a test developer to improve test quality because it can provide information on how items function individually and as a whole. While the GeDI overall functioned very well within the

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

Does the presentation order of instrument scenarios (and associated item suites) impact measures of student understanding?

The GeDI features four scenarios differing in taxonomic context and item difficulty, two factors which have been associated with item position effects in studies with other instruments (cf. Federer et al. 2015). Rearranging the order of GeDI scenarios and associated item suites had no significant impact on test scores, thus each scenario is functioning independently to assess student knowledge and does not appear to be impacting responses to subsequent items. Almost no other concept inventories in biology education have been tested for order effects.

Does the GeDI measure student knowledge in a manner that is generalizable across geographic regions of the United States?

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

Genetic drift, natural selection, and their interrelationships

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

in genetic drift often confuse drift with scientific or naive ideas about natural selection or other evolutionary events such as bottlenecks/population boundaries, random mutation, migration/gene flow, or speciation in general (Andrews et al. 2012; Beggrow and Nehm 2012). In contrast, students with more advanced ideas about drift tend to place inaccurate constraints on the situations under which drift occurs (Andrews et al. 2012).

The development of the GeDI to target stage 2 (mid-level) misconceptions about genetic drift is an important addition to the body of evolution measurement tools because it both gauges understanding of a previously neglected evolutionary mechanism and holds potential to capture some simultaneous reasoning about natural selection (as it relates to drift). Given the incoherence of naive student thought about evolution (inappropriately both conflating and failing to recognize simultaneous adaptive and non-adaptive processes), attention toward developing instruments that can simultaneously capture thought on adaptive and nonadaptive mechanisms is warranted. Few instruments are capable of simultaneously eliciting thought about natural selection and genetic drift, and none were designed with the intent to robustly measure knowledge of both processes (Table 1). More fully capturing the array of student thought about diverse evolutionary mechanisms, including how thoughts on diverse mechanisms intersect, will better equip educators to develop appropriate instructional strategies and develop curricula.

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

Limitations and further research

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

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

Conclusions

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

evolution instruments that were validated using CTT methods.

Additional file

Additional file 1: Table S1. Mean GeDI item measures (measure) and standard error (SE) by vignette rotation position.

Authors' contributions

EL designed the data collection procedure and collected the data. RT performed all analyses, interpreted the data, and wrote most of the manuscript. RHN designed the study, assisted with the analyses and interpretation, and helped edit the manuscript. All authors read and approved the final manuscript.

Author details

[1] Science Education Program, Institute for STEM Education, Stony Brook University, 092 Life Sciences Building, Stony Brook, NY 11794-5233, USA. [2] Department of Ecology, Evolution, and Natural Resources, Rutgers University, 14 College Farm Road, New Brunswick, NJ 08901, USA.

Acknowledgements

RT thanks the National Association for Research in Science Teaching (NARST) for a NARST Scholarship for Classroom Teachers and Informal Educators that facilitated work on this project. RHN thanks the National Science foundation TUES-1322872 for partial support. Any opinions, findings, conclusions or recommendations expressed in this publication are those of the authors and do not necessarily reflect the views of the National Science Foundation. The authors also thank Gena Sbeglia for comments on the manuscript.

Competing interests

The authors declare they have no competing interests.

Availability of data

Please contact the first author for access to data.

Publisher's Note

Received: 22 March 2018 Accepted: 9 July 2018
Published online: 17 July 2018

References

American Association for the Advancement of Science (AAAS). Vision and change in undergraduate biology education. Washington, DC; 2011. http://visionandchange.org/ Accessed 20 February 2018.

American Educational Research Association, American Psychological Association, and National Council on Measurement in Education (AERA, APA, NCME). The standards for educational and psychological testing. Washington, DC: American Educational Research Association, American Psychological Association, and National Council on Measurement in Education (AERA, APA, NCME); 2014.

Anderson DL, Fisher KM, Norman GJ. Development and evaluation of the conceptual inventory of natural selection. J Res Sci Teach. 2002;39(10):952–78.

Andrews TM, Price RM, Mead LS, McElhinny TL, Thanukos A, Perez KE, Lemons PP. Biology undergraduates' misconceptions about genetic drift. CBE Life Sci Educ. 2012;11(3):248–59.

Andrich D, Marais I, Humphry S. Using a theorem by Andersen and the dichotomous Rasch model to assess the presence of random guessing in multiple choice items. J Educ Behav Stat. 2012;37(3):417–42.

Beggrow EP, Nehm RH. Students' mental models of evolutionary causation: natural selection and genetic drift. Evol Educ Outreach. 2012;5(3):429–44.

Bishop BA, Anderson CW. Student conceptions of natural selection and its role in evolution. J Res Sci Teach. 1990;27(5):415–27.

Bond TG, Fox CM. Applying the Rasch model: fundamental measurement in the human sciences. 2nd ed. Mahwah: Lawrence Erlbaum Associates; 2007.

Boone WJ, Staver JR, Yale MS. Rasch analysis in the human sciences. Dordrecht: Springer; 2014.

Campbell CE, Nehm RH. A critical analysis of assessment quality in genomics and bioinformatics education research. CBE Life Sci Educ. 2013;12(3):530–41.

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

de Ayala RJ. The theory and practice of item response theory. New York: The Guilford Press; 2009.

Federer MR, Nehm RH, Opfer JE, Pearl D. Using a constructed-response instrument to explore the effects of item position and item features on the assessment of students' written scientific explanations. Res Sci Educ. 2015;45(4):527–53.

Federer MR, Nehm RH, Pearl DK. Examining gender differences in written assessment tasks in biology: a case study of evolutionary explanations. CBE Life Sci Educ. 2016;15(1):ar2.

Furtak E, Morrison D, Kroog H. Investigating the link between learning progressions and classroom assessment. Sci Educ. 2014;98(4):640–73.

Gershon R. Guessing and measurement. Rasch Meas Trans. 1992;6(2):209–10.

International Baccalaureate Organization. Diploma programme biology guide. Cardiff: International Baccalaureate Organization; 2014.

Kalinowski ST, Leonard MJ, Taper ML. Development and validation of the conceptual assessment of natural selection (CANS). CBE Life Sci Educ. 2016;15(4):ar64.

Leary LF, Dorans NJ. Implications for altering the context in which test items appear: a historical perspective on an immediate concern. Rev Educ Res. 1985;55(3):387–413.

Linacre M. Sample size and item calibration stability. Rasch Meas Trans. 1994;7(4):328.

Linacre, M. A users guide to winsteps/ministep Rasch model computer programs. Program Manual 4.0.0. 2017. http://www.winsteps.com/a/Winsteps-ManualPDF.zip. Accessed 10 Feb 2018.

MacNicol K. Effects of varying order of item difficulty in an unspeeded verbal test. Unpublished manuscript, Educational Testing Service. Princeton; 1956.

Masel J. Rethinking Hardy–Weinberg and genetic drift in undergraduate biology. BioEssays. 2012;34(8):701–10.

Messick S. Validity of psychological assessment. Am Psychol. 1995;50:741–9.

Miller K, Levine J. Miller and Levine biology. Upper Saddle River (NJ): Pearson Prentice Hall; 2017.

Moharreri K, Ha M, Nehm RH. EvoGrader: an online formative assessment tool for automatically evaluating written evolutionary explanations. Evol Educ Outreach. 2014;7(1):15.

Mollenkopf WG. An experimental study of the effects on item-analysis data of changing item placement and test time limit. Psychometrika. 1950;15(3):291–315.

Nadelson LS, Southerland SA. Development and preliminary evaluation of the measure of understanding of macroevolution: introducing the MUM. J Exp Educ. 2009;78(2):151–90.

National Research Council. Knowing what students know: the science and design of educational assessment. Washington, DC: National Academies Press; 2001.

Nehm RH. Evolution (Chapter 14). In: Kampourakis K, Reiss M, editors. Teaching biology in schools. Routledge: New York; 2018.

Nehm RH, Beggrow EP, Opfer JE, Ha M. Reasoning about natural selection: diagnosing contextual competency using the ACORNS instrument. Am Biol Teach. 2012;74(2):92–8.

Nehm RH, Ha M. Item feature effects in evolution assessment. J Res Sci Teach. 2011;48(3):237–56.

Nehm RH, Haertig H. Human vs. computer diagnosis of students' natural selection knowledge: testing the efficacy of text analytic software. J Sci Educ Technol. 2012;21(1):56–73.

Nehm RH, Schonfeld IS. Measuring knowledge of natural selection: a comparison of the CINS, an open-response instrument, and an oral interview. J Res Sci Teach. 2008;45(10):1131–60.
Nehm RH, Schonfeld IS. The future of natural selection knowledge measurement: a reply to Anderson et al. (2010). J Res Sci Teach. 2010;47(3):358–62.
Neumann I, Neumann K, Nehm R. Evaluating instrument quality in science education: Rasch-based analyses of a nature of science test. Int J Sci Educ. 2011;33(10):1373–405.
NGSS Lead States. Next generation science standards: for states, by states. Washington, DC: The National Academies Press; 2013.
Nitko A, Brookhart S. Educational assessment of students, 6th edn. New York: Pearson; 2010.
Nowicki S. HMH biology 2017. Boston: Houghton Mifflin Harcourt Publishing Company; 2017.
Opfer JE, Nehm RH, Ha M. Cognitive foundations for science assessment design: knowing what students know about evolution. J Res Sci Teach. 2012;49(6):744–77.
Perez KE, Hiatt A, Davis GK, Trujillo C, French DP, Terry M, Price RM. The EvoDevoCI: a concept inventory for gauging students' understanding of evolutionary developmental biology. CBE Life Sci Educ. 2013;12(4):665–75.
Price RM, Andrews TM, McElhinny TL, Mead LS, Abraham JK, Thanukos A, Perez KE. The Genetic Drift Inventory: a tool for measuring what undergraduates have mastered about genetic drift. CBE Life Sci Educ. 2014;13(1):65–75.
Price RM, Perez KE. Beyond the adaptationist legacy: updating our teaching to include a diversity of evolutionary mechanisms. Am Biol Teach. 2016;78(2):101–8.
Price RM, Pope DS, Abraham JK, Maruca S, Meir E. Observing populations and testing predictions about genetic drift in a computer simulation improves college students' conceptual understanding. Evol Educ Outreach. 2016;9(1):8.
Raîche G. Critical eigenvalue sizes (variances) in standardized residual principal components analysis. Rasch Meas Trans. 2005;19(1):1012.
Romine WL, Schaffer DL, Barrow L. Development and application of a novel Rasch-based methodology for evaluating multi-tiered assessment instruments: validation and utilization of an undergraduate diagnostic test of the water cycle. Int J Sci Educ. 2015;37(16):2740–68.
Schmiemann P, Nehm RH, Tornabene RE. Assessment of genetics understanding: under what conditions do situational features have an impact on measures? Sci Educ. 2017;26(10):1161–91.
Sirotnik K, Wellington R. Incidence sampling: an integrated theory for matrix sampling. J Educ Meas. 1977;14(4):343–99.
Smith AB, Rush R, Fallowfield LJ, Velikova G, Sharpe M. Rasch fit statistics and sample size considerations for polytomous data. BMC Med Res Methodol. 2008;8(1):33.
Stony Brook University. Undergraduate course bulletin. 2017. http://sb.cc.stonybrook.edu/bulletin/current/courses/index.pdf. Accessed 12 Dec 2017.
Urry LA, Cain ML, Wasserman SA, Minorsky PV, Reece JB. Campbell biology. 11th ed. Boston: Pearson; 2017.
Wright BD, Stone M. Best test design: Rasch measurement. Chicago: MESA Press; 1979.