Task Development

Task template. We developed 4 MBR tasks using the task structure depicted in Figure 2. Each task begins with a description of a real-world phenomenon related to natural selection or genetics. After the phenomenon or scenario is described, a set of data is provided, and the respondent is asked to analyze data and propose an initial model that would explain those data (inductive reasoning). Next, respondents are presented with an alternative model and additional data. They consider the alternative model and data and evaluate whether these new data support the model (abductive reasoning). Finally, respondents are provided with data that contradict the alternative model and asked whether these data support the model. They are then prompted to write a revised model that fits all the data presented in the task (hypothetico-deductive reasoning).

NGSS alignment. While the focus of the assessments was model-based reasoning using biology core ideas (LS3.A & LS4.B), students were required to use aspects of all three NGSS dimensions when completing the tasks. For example, students had to analyze graphical data to identify patterns that could be used as evidence to support a model-based explanation.

Task contexts. Two of the tasks were aligned to assess understanding of natural selection using the contexts of pocket rocket mice and lactose tolerance. The other two tasks assess understanding of genetics with one using Tay-Sachs as a context targeting both Mendelian and molecular genetics and the other using polydactyl cats as a context targeting only Mendelien genetics.

Pilot Test

Four pilot test forms were constructed, each being made up of 2 MBR tasks and 2 sets of content-focused (CF) items. The CF items were selected from the Conceptual Inventory of Natural Selection (CINS) (Anderson, et al., 2002) and the Genetics Literacy Assessment Instrument (GLAI) (Bowling, et al., 2008).

Because COVID-19 made it impossible to pilot test with high school students, we utilized Amazon’s Mechanical Turk (MTurk) to collect data from 234 adults with at least a high school diploma from the U.S. Although this population is not an exact match to the intended student population, it provides us with data that can be used to inform item revisions and evaluate scoring rubrics. The population is a closer match to the teacher population and had the potential to give us more correct responses at the upper end of the distribution, which might be difficult to obtain in a high school sample.

Results and Discussion

The MTurk data fit the partial credit Rasch model well with person and item reliabilities of .80 and .97, respectively. The average person measure was small and positive (.32), indicating that the items were well matched to the MTurk workers’ ability. As shown in the Wright map in Figure 3, the multiple-choice items within the tasks that ask respondents to predict or identify patterns (shaded yellow) were easier than the constructed-response items (shaded gray, green and blue). The items that asked respondents to evaluate whether data support a proposed model (shaded gray) were easier than the items that required respondents to write a model-based explanation (shaded green and blue).

A principal component analysis of the Rasch residuals showed that the first component of the correlation matrix was 2.5 indicating the MBR tasks and CF items were predominantly unidimensional, although the construct may be broad. Cross-plots tell the person measures based only on the MBR tasks against the person measures based only on the CF items and show a small but statistically significant correlation between performance on the CF items and MBR tasks (p = 0.59, p < .001 for genetics; p = 0.42, p < .001 for natural selection). This suggests that the MBR tasks measure a different aspect of the construct than the CF items, which is to be expected given that the MBR tasks require both content knowledge of biology and model-based reasoning, while the CF items focus solely on content knowledge.

The results suggest that instruments made up of content-focused items and MBR tasks measure the intended outcome and will be appropriate for use with the teacher population. We are currently pilot testing the assessments with a small sample of high school students to ensure that they perform as expected for the targeted student population. We will also be conducting a second MTurk pilot test of these assessments along with the assessments to measure self-efficacy and tolerance to ambiguity. The MTurk respondents will be narrowed to 18-year-old, high school graduates to better mimic the target student population.

References