Proceedings of the 90th Annual International Meeting of the Psychometric Society, Minneapolis, United States (2025).

Assessing Metacognitive Calibration in Biological Model Evaluation using Rasch Measurement

M.G. Behrendt [*†], J.M. Wheeler [‡], J.T. Dauer [¶], and C.A.C Clark [‡]
†Department of Psychology, Baker University, Baldwin City, 66006, KS, USA
‡Department of Educational Psychology, University of Nebraska—Lincoln, City, 68588, NE, USA
¶School of Natural Resources, University of Nebraska—Lincoln, City, 68583, NE, USA
*Corresponding author. Email: mbehrendt@bakeru.edu

Abstract
Metacognitive calibration, the alignment between confidence and accuracy, is essential for self-regulated learning, yet remains challenging to measure because observed performance reflects both individual metacognitive skill and task characteristics. This study developed and evaluated the Biology Metacognitive Scale (BMS), a Rasch-based measure of confident accuracy, a specific facet of metacognitive calibration defined as instances where students both know the correct answer and feel confident in their knowledge. Fifty undergraduate life science students evaluated 36 biology models for conceptual errors during fMRI scanning, indicating both accuracy and confidence for each judgment. Rasch measurement was used to disentangle person ability from item difficulty, placing both on a common latent scale. The final 15-item BMS demonstrated adequate psychometric properties: unidimensionality (29.4% variance explained), acceptable internal consistency (α = .73), and appropriate item fit statistics (infit/outfit: 0.80–1.36). External validity was supported by strong correlations with expert ratings of biological reasoning one year later ($r = .54$, $p < .001$) and task accuracy ($r = .88$, $p < .001$), while showing discriminant validity from general cognitive ability ($r = .18$, ns) and course grades ($r = .20$, ns). The Wright map revealed meaningful variation in item difficulty, demonstrating how task characteristics influence metacognitive performance. This preliminary work contributes to ongoing methodological development in metacognition measurement and provides a psychometric foundation for future research examining how confident accuracy relates to cognitive processes and instructional interventions in STEM education contexts.

Keywords: metacognitive calibration, Rasch analysis, biological model evaluation

1. Introduction to Metacognitive Calibration

Metacognitive calibration, the alignment between a learner's confidence and accuracy, plays a central role in how individuals monitor and regulate performance (Flavell, 1979; Tobias & Everson, 2009; Molenbergh et al., 2016). Measuring metacognitive calibration, however, is challenging because judgments are influenced by both latent metacognitive skills and task characteristics, including item difficulty and cognitive demand (Fleming & Lau, 2014; Dunlosky & Metcalfe, 2008; Händel et al., 2020). When these influences are not separated, observed performance may reflect artifacts of the task rather than true individual differences in metacognitive calibration. Traditional behavioral approaches to measuring metacognition, such as confidence bias and signal detection-based measures, are often confounded by performance level or response bias (Fleming & Lau, 2014; Fleming & Daw, 2017).

In educational contexts, a specific pattern of metacognitive performance holds practical significance: *confident accuracy*, defined as instances in which students both know the correct answer and feel confident in their knowledge. While this operationalization focuses on one facet of metacognitive functioning and does not distinguish among different types of miscalibration (e.g., accurate but uncertain vs. confidently incorrect), it

provides a clear construct relevant to self-regulated learning. Students who consistently demonstrate confident accuracy possess both content mastery and metacognitive awareness of that mastery, positioning them to engage effectively in self-directed learning.

1.1 Purpose of Study

Recent reviews highlight the ongoing need for methodological development in this area, as no single existing metric fully addresses the complexities of metacognition measurement (Rahnev, 2025). As such, this study uses Rasch measurement, which offers a principled approach to disentangle person ability from item difficulty (Rasch, 1960; Wright, 1977; Engelhard, 2013). By placing persons and items on the same latent scale, Rasch models enable invariant comparisons and provide interpretable parameter estimates essential for measuring constructs like metacognitive calibration.

While the current study emphasizes psychometric evaluation, this measurement work represents one component of a broader research agenda investigating metacognitive calibration using complementary behavioral, neuroimaging, and qualitative approaches. The purpose of this study is to develop and evaluate the Biology Metacognitive Scale (BMS), a Rasch-based measure of students' confident accuracy during biological model evaluation. The development of the BMS is a preliminary contribution to the ongoing development of psychometrically sound tools for assessing metacognition in STEM education contexts. As such, this study provides a measurement framework for future research examining how confident accuracy relates to cognitive processes and instructional interventions.

2. Methods

The procedures described here reflect the measurement-focused component of a larger multi-method study examining metacognitive calibration through neuroimaging, behavioral, and qualitative analyses (Clark et al., 2025; Behrendt et al., 2024; Dauer et al., 2024). All data for the present study were collected during functional MRI scanning session, which constrained sample size due to the cost and logistical demands of neuroimaging research.

2.1 Participants

This study includes fifty undergraduate life science students recruited from multiple sections of an introductory biology course (M_{age} = 19.62, SD = .90). Most were first-year students (70%), and the sample reflected the demographic composition of the students who take that course. While course instructors regularly used biological models in their teaching, students had not previously encountered the specific biology models used in the study. Participants received course-relevant instruction before testing and completed the study for $50 compensation.

2.2 Procedures and Biology Model Assessment Task

During an MRI session, students completed a biology model assessment in which they evaluated whether the visual models contained conceptual errors. Models were presented as diagrams, flow charts, or schematic images depicting biological processes across evolution, ecology, genetics, and physiology (see Appendix 1 for examples). The assessment consisted of three runs of 12 trials each (36 total biological models were evaluated). On each trial, participants viewed a biology model for 10 seconds and made two sequential judgments. First, they indicated whether the model contained an error (yes/no). Second, they rated their confidence in their judgment (confident/not confident).

Following the MRI session (brain findings published separately), participants completed the Kaufman Brief Intelligence Test, Second Edition (KBIT-2; Kaufman & Kaufman, 2004) to assess general cognitive ability. Approximately one year later, most participants (N = 44) returned for a follow-up session in which they received $100 compensation and subject-matter experts rated their biological reasoning on a subset of six biology model assessment tasks using a standardized rubric. Both the KBIT-2 and expert ratings served as external validity criteria, allowing examination of the BMS's relationship with general cognitive ability and domain-specific reasoning.

2.3 Biology Metacognitive Scale (BMS)

The BMS consisted of 15 items selected from an original pool of 36 through an iterative process balancing psychometric quality and content coverage. Items were removed for two primary reasons. First, several items showed ceiling effects, with nearly all participants responding correctly. Second, preliminary Rasch analyses identified items with poor fit statistics (e.g., infit and outfit). Although the items on the final scale were substantially reduced, the final set of items maintained the original balance of error-containing and error-free biological models and preserved a similar balanced representation across content areas (evolution, ecology, genetics, and physiology).

2.4 Dichotomous Rasch Model

The Rasch model estimates the probability that a participant will give an accurate *and* confident response. The model treats calibrated responding as a function of a person's ability to align confidence with accuracy (θ_j) and the difficulty of the item (δ_i). The probability of person j producing a calibrated response on item i is:

$$P_{ij} = P\left(Y_{ij} = 1 \middle| \theta_j\right) = \frac{\exp\left(\theta_j - \delta_i\right)}{1 + \exp\left(\theta_j - \delta_i\right)}$$

Students with higher θ_j values demonstrate greater confident accuracy more consistently across biology items, reflecting stronger metacognitive ability to both acquire knowledge and appropriately monitor their confidence. Students with lower θ_j values are less likely to achieve confident accuracy, which may result from knowledge gaps, uncertainty despite correct understanding, or misplaced confidence in incorrect responses. Higher δ_i values indicate items that are more challenging for students to evaluate accurately and confidently, while lower δ_i values indicate items that are easier to answer correctly and confidently.

2.5 Data Analysis

Rasch model fit was evaluated using variance explained by the Rasch model, item separation reliability, infit and outfit mean-square values, and a principal components analysis of residuals. Principal components analysis of residuals was used to assess unidimensionality by examining whether any secondary dimensions explained substantial variance beyond the primary Rasch dimension. Additionally, to complement the Rasch analysis, we conducted exploratory factor analysis (EFA) to examine the scale's dimensionality and calculated Cronbach's alpha to assess internal consistency reliability. All Rasch analyses were conducted using the eRm package (Mair & Hatzinger, 2007), while the PCA, EFA, and reliability analyses were conducted using the psych package (Revell, 2017) in R version 4.2.1 (R Core Team, 2022).

3. Results and Evaluation of Psychometric Quality

3.1 Unidimensionality

Several sources of evidence suggested that the BMS functions as a unidimensional scale. The Rasch model converged after 11 iterations and explained 29.4% of the variance in responses, which exceeds the 20% threshold commonly used as evidence of adequate unidimensionality for practical measurement purposes (Reckase, 1979). Principal components analysis of the standardized residual correlations revealed a first contrast eigenvalue of 1.86, which is below the critical value of 2.0 (Linacre, 2016; Wind & Hua, 2022), indicating that residual patterns primarily reflect random variation rather than a substantial secondary dimension. Exploratory factor analysis showed a one-factor structure (Figure 1), with the scree plot displaying an elbow after the first component along with the second eigenvalue less than 1.

3.2 Internal Validity Evidence

Table 1 presents the item parameter estimates for the Rasch model along with their infit and outfit mean square statistics. All items demonstrated appropriate fit, with infit mean square values ranging from 0.796 to 1.229 and outfit mean square values ranging from 0.669 to 1.362, falling within the recommended 0.5 to 1.5 range (Engelhard, 2013). Item difficulties spanned approximately three logits, ranging from -1.147 to 1.588. This

indicates that items provided adequate coverage across ability levels, though the distribution suggests that additional difficult items could improve precision for high-ability students.

The Wright map (Figure 2) illustrated good alignment between the distribution of person abilities and item difficulties along the latent trait continuum. Person ability estimates ranged from -2.89 to 2.92 logits, with most students clustered in the middle range of the scale, indicating that the items were appropriately targeted to the sample's ability distribution.

To complement the Rasch analysis, internal consistency reliability and item-total correlations were evaluated. Cronbach's alpha was acceptable (α = 0.73), indicating adequate coherence among items. Item-total correlations ranged from .27 to .63, with all items contributing positively to the overall scale.

3.3 External Validity Evidence

Criterion-related validity was examined through correlations with four external measures. BMS scores showed strong positive relationships with expert ratings of students' biological reasoning one year later ($r = .54$, $p < .001$) and with accuracy on the biology model assessment task ($r = .88$, $p < .001$). Expert ratings were obtained from three researchers with extensive familiarity with the larger metacognition project, who had been working on this research program for multiple years. These researchers independently evaluated participants' responses to 4 of the 15 BMS items using a standardized 5-point rubric. Inter-rater agreement was substantial, with quadratic weighted kappa (QWK) coefficients ranging from .641 to .699, which indicates that raters applied the rubric consistently. In contrast, correlations between BMS scores and general cognitive ability (KBIT-2, $r = .18$) and final course grades ($r = .20$) were weak and not statistically significant. This pattern of correlations provides preliminary evidence for convergent validity (stronger relationships with conceptually related measures of domain-specific reasoning and performance) and discriminant validity (weaker relationships with general cognitive ability and overall academic achievement), suggesting that the BMS may capture aspects of metacognitive calibration distinct from general intelligence or accumulated knowledge.

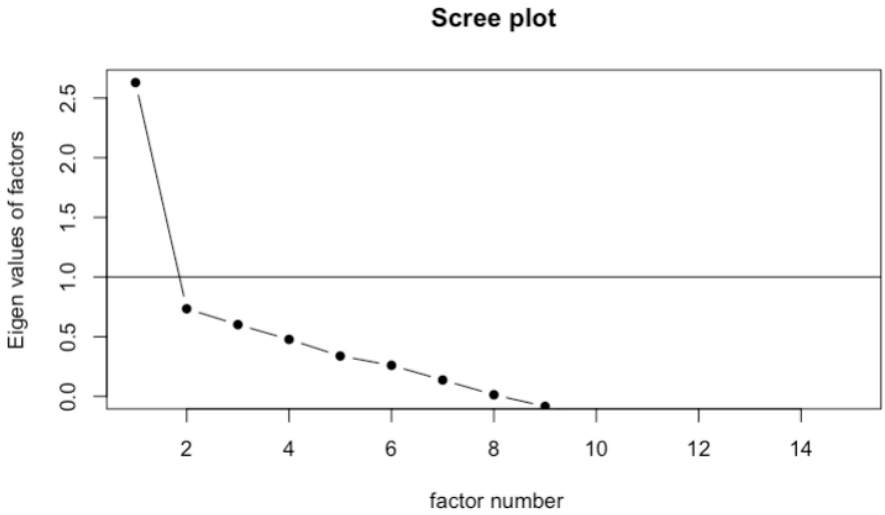
4. Discussion

This study developed the Biology Metacognitive Scale (BMS) as a preliminary measure of one facet of metacognitive calibration: *confident accuracy* during biological model evaluation. Given ongoing challenges in measuring metacognition (e.g., Rahnev, 2025), this work contributes to an area requiring continued methodological development. The application of Rasch measurement allowed for the separation of person ability from item difficulty, providing a framework for examining how confidence and accuracy interact during complex tasks. The Wright map revealed meaningful variation in item difficulty, highlighting how task characteristics influence metacognitive performance, consistent with findings that metacognitive calibration varies with task complexity (Händel et al., 2020).

Although this study provided a variety of validity and reliability evidence and demonstrated how Rasch measurement can be useful for the measurement of metacognition, there were a few limitations worth noting. The sample size (N = 50) was determined by the neuroimaging component of the larger study and may limit generalizability. Binary scoring simplified interpretation but does not distinguish among different types of metacognitive miscalibration, such as appropriate uncertainty versus misplaced confidence. The scale focuses on one aspect of metacognitive functioning within biological model evaluation, and generalizability to other STEM domains remains unknown.

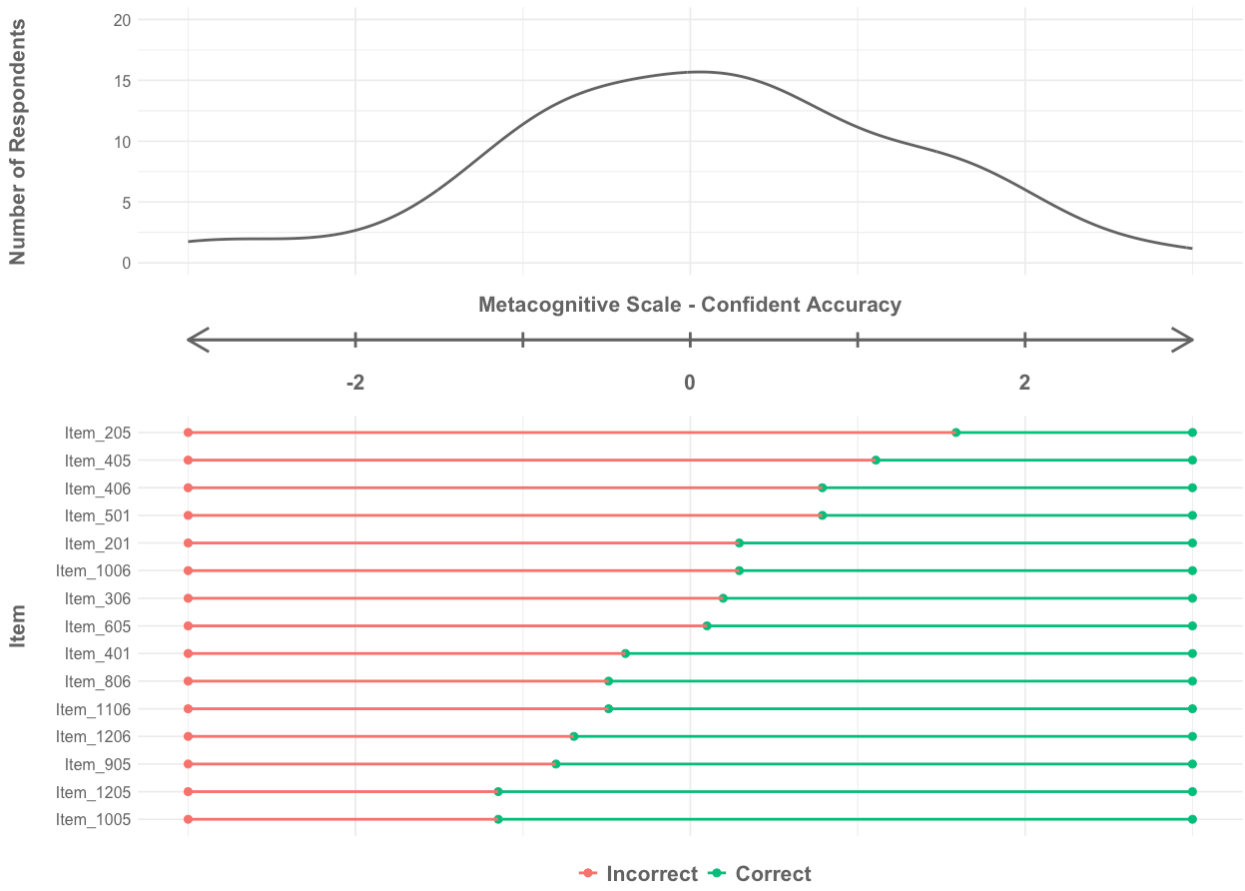
Future work should refine the item pool, expand the difficulty range to better capture high-ability students, and explore polytomous scoring to capture more nuanced patterns of metacognition (Fleming & Lau, 2014). Broader sampling and extension to other STEM domains would strengthen generalizability. Despite these limitations, this preliminary work suggests that Rasch measurement may offer useful approaches for examining metacognitive constructs in STEM education contexts.

Figure 1. *Scree Plot from Exploratory Factor Analysis of the Biology Metacognitive Scale*



Note. The scree plot shows eigenvalues from the exploratory factor analysis. A clear elbow appears after the first component, and the second eigenvalue is below 1.0, supporting a one-factor solution consistent with the scale's intended unidimensional structure.

Figure 2. *Wright Map of Person Ability and Item Difficulty for the Biology Metacognitive Scale*



Note. The Wright Map shows the distribution of respondent abilities (top) and item difficulties (bottom) on a common logit scale. Red lines indicate the ability range where the Rasch model predicts probability of a confident accuracy response as p(Y=1) < .50, while green lines indicate where p(Y=1) > .50. The transition point represents each item's difficulty parameter (δ). The distribution demonstrates adequate alignment between person abilities and item difficulties, supporting the scale's targeting.

Table 1. Item fit statistic from Rasch analysis

Item	Infit MNSQ	Infit ZSTD	Outfit MNSQ	Outfit ZSTD	Item Location
Item_205	0.839	-0.828	0.891	-0.144	1.588
Item_405	1.066	0.459	1.164	0.608	1.108
Item_406	0.887	-0.771	0.827	-0.62	0.789
Item_501	1.140	0.987	1.105	0.478	0.789
Item_201	0.904	-0.731	0.823	-0.805	0.293
Item_1006	1.229	1.701	1.231	1.073	0.293
Item_306	0.865	-1.068	0.790	-1.004	0.196
Item_605	0.978	-0.135	0.902	-0.417	0.100
Item_401	0.801	-1.573	0.691	-1.465	-0.387
Item_806	1.138	1.016	1.362	1.462	-0.488
Item_1106	1.121	0.904	1.229	0.986	-0.488
Item_1206	0.913	-0.572	0.778	-0.842	-0.695
Item_905	0.896	-0.665	0.747	-0.925	-0.803
Item_1205	1.150	0.875	1.094	0.385	-1.147
Item_1005	0.796	-1.195	0.669	-1.026	-1.147

Note. Infit and outfit are mean square fit statistics indicating item conformity to the Rasch model. Values between 0.5 and 1.5 are productive for measurement; values below 0.5 are less productive but not degrading; values between 1.5 and 2.0 are unproductive but not degrading; and values above 2.0 distort the measurement system (Engelhard, 2013). All BMS items fell within the productive range.

Acknowledgement

The authors thank the students whose participation made this research possible.

Funding Statement This research was supported by a grant from the National Science Foundation (Award No. 2000549).

Competing Interests None.

References

Behrendt, M.G., Clark, C.A.C, Elliott, M., & Dauer, J.T. (2024). Relation of life sciences students' metacognitive monitoring to neural activity during biology error detection. *npj Science of Learning, 9*(1), 1—12.

Boone, W. J. (2016). Rasch analysis for instrument development: Why, when, and how?. *CBE—Life Sciences Education, 15*(4), 1—7.

Clark, C.A.C., Behrendt, M.G., Long, T.M., & Dauer, J.T. (2025). Neural responses to errors in models and their relevance for long-term knowledge retention among undergraduate life sciences students. *Trends in Neuroscience and Education, 39,* 100253.

Dauer, J.T., Behrendt, M.G., Elliott, M., Gettings, B., Long, T.M. & Clark, C.A.C. (2024). Individual variation in undergraduate student metacognitive monitoring and error detection during biology model evaluation. *Frontiers in Education, 9,* 1356626.

De Boer, H., Donker, A. S., Kostons, D. D., & Van der Werf, G. P. (2018). Long-term effects of metacognitive strategy instruction on student academic performance: A meta-analysis. *Educational Research Review, 24,* 98—115.

Dunlosky, J., & Metcalfe, J. (2008). *Metacognition*. Thousand Oaks, CA: Sage Publications.

Engelhard Jr, G. (2013). *Invariant measurement: Using Rasch models in the social, behavioral, and health sciences.* Routledge.

Flavell, J. H. (1979). Metacognition and cognitive monitoring: A new area of cognitive-developmental inquiry. *American Psychologist, 34*(10), 906—911.

Fleming, S. M., & Daw, N. D. (2017). Self-evaluation of decision-making: A general Bayesian framework for metacognitive computation. *Psychological review, 124*(1), 91—114.

Fleming, S. M., & Lau, H. C. (2014). How to measure metacognition. *Frontiers in Human Neuroscience, 8,* 1—9.

Händel, M., de Bruin, A. B., & Dresel, M. (2020). Individual differences in local and global metacognitive judgments. *Metacognition and Learning, 15*, 51—75.

Kaufman, A. S. & Kaufman, N. L. (2004). *Kaufman Brief Intelligence Test*, Second Edition. Bloomington, MN: Pearson, Inc.

Linacre, J. M. (2016). A user's guide to Winsteps Raschmodel computer programs: program manual 3.92. 0. *Mesa-Press, Chicago.*

Mair, P., & Hatzinger, R. (2007). Extended Rasch modeling: The eRm package for the application of IRT models in R. *Journal of Statistical Software*, *20*, 1-20.

Molenberghs, P., Trautwein, F. M., Böckler, A., Singer, T., & Kanske, P. (2016). Neural correlates of metacognitive ability and of feeling confident: A large-scale fMRI study. *Social cognitive and affective neuroscience, 11*(12), 1942—1951.

R Core Team. (2022). R: A language and environment for statistical computing. R Foundation for Statistical Computing. https://www.R-project.org/

Rahnev, D. (2025). A comprehensive assessment of current methods for measuring metacognition. *Nature Communications, 16*(1), 701.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests.* Copenhagen: Danish Institute for Educational Research.

Reckase, M. D. (1979). Unifactor latent trait models applied to multifactor tests: Results and implications. *Journal of educational statistics, 4*(3), 207-230.

Revelle, W. R. (2017). psych: Procedures for personality and psychological research.

Tobias, S., & Everson, H. T. (2009). The importance of knowing what you know: A knowledge monitoring framework for studying metacognition in education. In *Handbook of Metacognition in Education* (pp. 107—127). Routledge.

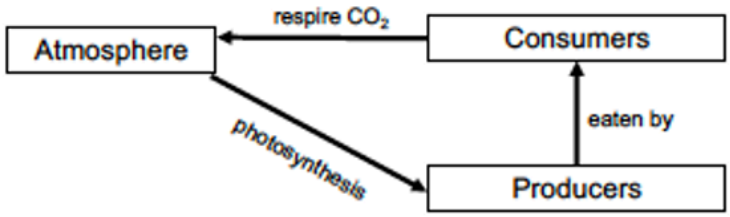
Veenman, M. V. (2017). Assessing metacognitive deficiencies and effectively instructing metacognitive skills. *Teachers College Record, 119*(13), 1—20.

Wind, S., & Hua, C. (2022). *Rasch measurement theory analysis in R*. Chapman and Hall/CRC.

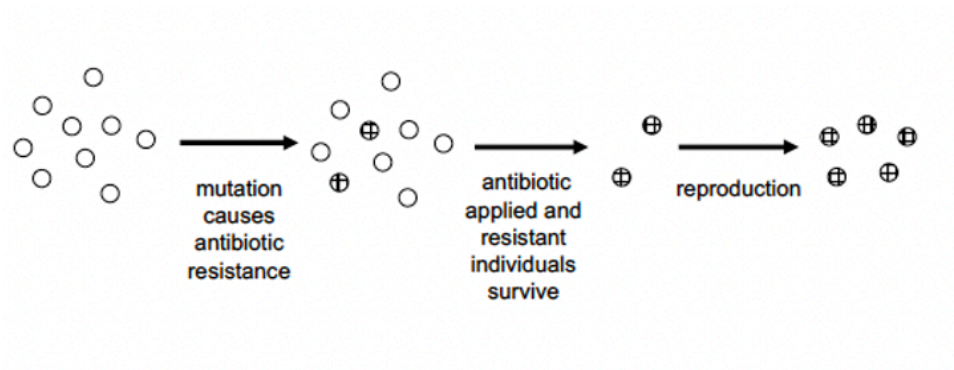
Wright, B. D. (1977). Solving measurement problems with the Rasch model. *Journal of educational measurement,* 97-116.

Appendix 1. Example Biology Models used in the BMS

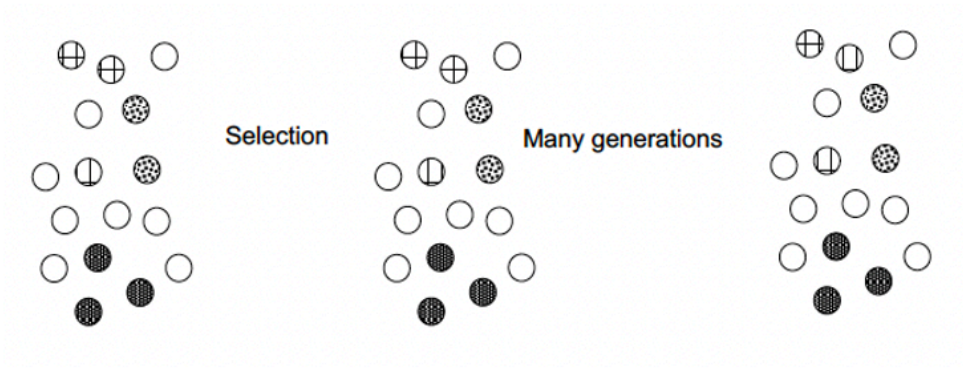
Item 205



Item 401



Item 1005



Example biology models shown to undergraduates during the model evaluation task. Item 205 is an error-containing model of the carbon cycle, Item 401 is an error-free model illustrating how antibiotic resistance emerges within a population, and Item 1005 is an error-containing model of how natural selection affects diversity within a population.