Supplemental Material

CBE—Life Sciences Education

Limeri *et al*.

Supplemental Materials for:
Undergraduate Lay Theories of Abilities: Mindset, universality, and brilliance beliefs uniquely predict undergraduate educational outcomes

Lisa B. Limeri*, Nathan T. Carter, Franchesca Lyra, Joel Martin, Halle Mastronardo, Jay Patel, Erin L. Dolan

Corresponding author: Lisa B. Limeri, llimeri@ttu.edu

Table of Contents

Section 1: Phase 1 methods and results Page S2
Section 2: Phase 3 methods and results Pages S3-6
Section 3: CFA residual covariances and item deletion (Phase 3) Page S7
Section 4: Measurement invariance methods and results (Phase 3) Pages S8-10
Section 5: Different Item Functioning methods and results (Phase 3) Pages S11-12
Section 6. Item Response Theory analysis methods and results (Phase 3) Pages S13-18
Section 7. Survey items (Phase 4) Pages S19-24
Section 8. Correlation table (Phase 4) Page S25
References Pages S26-27

Section 1: Phase 1 methods

Semi-structured interview methods.

We used a screening survey to select interview participants so that we could ensure our sample represented the diversity of undergraduates in the United States. The screening survey asked students about their racial and ethnic background, gender identity, and major(s). To recruit a diverse sample, we advertised the study at institutions that are diverse in terms of their research activity, student population (including Hispanic-Serving Institutions and Historically Black Colleges and Universities), and institution type (including community colleges and both public and private institutions). We asked instructors of introductory biology, chemistry, math, and physics courses at 14 institutions to distribute study information and a link to the screening survey to their students. We received 337 responses to our screening survey. We selected participants who were diverse in their personal characteristics and identities. When possible, we invited one student from each class from each institution to interview. If the volunteer did not respond to our invitation, we instead selected another volunteer from the same class. We prioritized selecting selected students with gender identities and racial/ethnic identities that were uncommon and/or not already represented in our sample to maximize the diversity of perspectives in our sample. Further, at minority-serving institutions, we prioritized the minority population being served (e.g., we invited Black students from Historically Black Universities and Hispanic students from Hispanic-Serving Institutions). Ultimately, we invited 84 volunteers to interview. Of these, 39 did not respond to our invitation and we conducted 45 interviews.

We drafted the initial interview questions based on candidate terms we identified from our prior work that we hypothesized undergraduates might interpret consistently: critical thinking, problem-solving, and ability to learn. For "intelligence" and each of these terms, we asked students how they defined the term, whether they think it is possible to improve that ability, and whether that ability is relevant to their academic performance. We also asked students if there were any other cognitive factors that influenced their academic performance to identify new candidate terms. We iteratively revised the interview questions throughout the project to include new candidate terms that arose and to remove candidate terms that we decided were not being interpreted consistently enough to continue considering.

Three analysists (LBL, HM, & JP) read and coded interviews separately, then met as a group to discuss to consensus and refine the codebook. Analyses occurred in real-time while data was being collected, and results from analyses were used to modify the interview protocol. As analyses revealed that a term was ambiguous, it was removed from the interview protocol, and when new, candidate terms were identified, they were added to the protocol for testing.

Cognitive interview methods.

We recruited participants using the same screening survey we used to select participants for the semi-structured interviews.

Section 2: Phase 3 methods and results

Data collection methods

We recruited participants by asking instructors of introductory biology, chemistry, physics, and math courses information about the study and a link to the survey to their students. We initially targeted 20 institutions that we selected to represent a range of institutional characteristics and student populations. We chose institutions based on national representation of undergraduate enrollment from the National Center for Education Statistics. For example, according to NCES statistics from 2018, 29% of undergraduates in the United States are enrolled at 2-yr institutions, so 5 of the 20 (25%) focal institutions were 2-yr institutions. We capped participation from each institution to 100 respondents to ensure that our sample would reflect a diversity of undergraduate institutions in the United States and not be inundated with responses from students from a single institution. Respondents were compensated with a $10 gift card. We began data collection during fall 2020, but did not reach our desired sample size. Over the winter we compared the representation of student groups in our sample to national statistics from NCSES and refined our recruiting strategy to focus on student groups that were underrepresented in our sample compared to national representation of undergraduates. We collected the rest of our sample during the spring 2021 semester. The vast majority of participants (93%; 1,114/1,194) came from the 20 focal institutions where we recruited, but some participants shared study information through their networks, so we also received responses from an additional 80 students from 48 other institutions.

Supplemental Table 1. Institutional information about the 68 institutions attended by respondents in Phase 2. Student Population acronyms: PWI = Primarily White Institution; HBCU = Historically Black Colleges and Universities; HSI = Hispanic Serving Institution; AANAPISI = Asian American and Native American Pacific Islander-Serving Institution

Number of respondents	Institutional ownership	Carnegie Classification	Student population
1	Private	Associates	PWI
7	Public	Masters	PWI
1	Private	Masters	PWI
1	Private	Baccalaureate	PWI
1	Private	Baccalaureate	Women's College
1	Public	Masters	PWI
1	Public	Masters	HSI & AANAPISI
1	Public	Masters	PWI
1	Private	Very High Research Activity	PWI
2	Private	Baccalaureate	HBCU
1	Public	Masters	PWI
83	Public	Masters	PWI
43	Public	Associates	AANAPISI
1	Private	Very High Research Activity	PWI

91	Public	Very High Research Activity	HSI
87	Public	Associates	PWI
31	Public	Baccalaureate	PWI
12	Public	Masters	HBCU
10	Public	Associates	PWI
100	Public	Very High Research Activity	PWI
2	Public	Very High Research Activity	AANAPISI
94	Private	Baccalaureate	PWI
35	Private	Masters	PWI
10	Public	High Research Activity	HBCU
81	Public	Masters	PWI
1	Private	Doctoral	PWI
2	Public	Associates	AANAPISI
100	Public	Doctoral	PWI
1	Public	High Research Activity	HBCU
1	Private	Very High Research Activity	PWI
33	Public	High Research Activity	HBCU
10	Public	Baccalaureate	PWI
2	Public	High Research Activity	PWI
79	Private	Associates	PWI
1	Public	Very High Research Activity	PWI
23	Public	Associates	HSI
1	Public	Very High Research Activity	PWI
1	Private	Masters	PWI
1	Public	Very High Research Activity	PWI
1	Private	High Research Activity	PWI
3	Public	Associates	HSI
5	Public	Masters	AANAPISI
1	Public	Masters	HBCU
9	Private	Baccalaureate	HBCU & Women's College
1	Public	Masters	HSI
1	Public	Associates	PWI
1	Public	Doctoral	PWI
1	Private	Baccalaureate	PWI
2	Public	Very High Research Activity	PWI
1	Public	Very High Research Activity	HSI
1	Private	High Research Activity	PWI
1	Public	Very High Research Activity	PWI
1	Public	Very High Research Activity	PWI

96	Public	Very High Research Activity	PWI
12	Public	Doctoral	AANAPISI
5	Public	Very High Research Activity	AANAPISI
1	Public	High Research Activity	AANAPISI
1	Public	Very High Research Activity	PWI
1	Private	Very High Research Activity	PWI
1	Public	Very High Research Activity	PWI
1	Private	Doctoral	AANAPISI
85	Public	Very High Research Activity	HSI
2	Public	Doctoral	PWI
1	Private	Very High Research Activity	PWI
1	Public	Very High Research Activity	PWI
1	Public	Very High Research Activity	PWI
2	Private	Very High Research Activity	PWI
1	Private	High Research Activity	PWI

Data quality

We received 1,522 complete survey responses. The survey included two directed response questions to screen out respondents who were paying insufficient attention and selecting random responses rather than effortfully reading and responding to items. These two questions directed respondents to select a specific response. We removed all responses that failed to select the directed response to either item (n = 266). In addition, we manually inspected the responses and further removed responses where it was clear that individuals had responded twice (e.g., entered the same name and two different emails, one institutional and one personal, n = 27) and two bouts of “spam” responses to gain the financial compensation being offered (n = 35). [We detected and removed two bouts of “spam” responses based on abnormal patterns of responses, including selecting the first institution from the drop-down list, which was a decoy entry that is not a real institution, suspicious email addresses and names, numerous responses from the same IP address, and an unusually large number of responses received within minutes of each other in the middle of the night.] After these screening procedures, we had a final sample of 1,194 participants.

Methods: testing alternative confirmatory factor models methods and results

We compared the fit of models based on both relative metrics of model fit (CFI and TLI) as well as absolute metrics of model fit (RMSEA and SRMR). We also report Chi-Square following convention, but note that with large sample sizes, chi-square tests have a very high type II error rate (i.e., they will be significant even for well-fitting models).

It is notable that while the absolute fit indices (RMSEA and SRMR) suggest that the 5-factor with higher-order factors model is an acceptable fit to the data, the relative fit indices (CFI & TLI) do not reach the commonly-used cut-off for acceptable model fit (>0.9; table 2). However, it is noted that CFI and TLI can be misleading metrics when the null model has unusually good model-data fit. To investigate this possibility, we calculated the fit of the null model using the nullRMSEA() command available in the semTools package (Jorgensen et al., 2021). The RMSEA of the null model is 0.160, which is very close to the recommended cut-off of 0.158. For this reason, we are not concerned about the low CFI and TLI values.

Section 3: CFA residual covariances and item deletion (Phase 3)

We examined the residual covariances (standardized to be on the scale of correlations). There were two five item pairs with residual covariances greater than |.20|:

1. Growth_02 and Fixed_07 (.305)

“I can vastly improve my ability to think creatively”
“Even if I try my best I could never become extremely creative”

2. Fixed_01 and Fixed_02 (.309)

“My intellectual ability will remain about the same over time."
"My ability to think creatively will stay about the same throughout college."

3. Universal_01 and Universal_02 (.246)

"Some people will always be less effective at learning than those who have a natural talent for it."
"People with a natural talent will achieve greater success in STEM than others."

4. Nonuniversal_08 and Nonuniversal_11 (.243)

"Some people are just naturally better at analyzing information than others."
"Some people will always be able to learn better than others."

5. Growth_06 and Fixed_07 (.222)

"It's possible that I could become as creative as highly successful STEM professionals one day."
“Even if I try my best I could never become extremely creative”

Growth_02 and Fixed_07 had in common the notion of creativity; Fixed_07 also showed unique overlap with Growth_06 and therefore we eliminated Fixed_07 only. Fixed_01 and Fixed_02 were both about changing over time; we believed the more general wording was better and therefore eliminated Fixed_02 which was only about the duration of college. Nonuniversal_08 and _11 had in common that it is referential to others and about a cognitive process. We eliminated only Nonuniversal_8 due to the brevity of _11. None of these items were selected for inclusion in the short form.

After removing these items, we re-fit the CFA which improved the model fit: RMSEA = .055, t-size adjusted RMSEA = .057, SRMSR = .063, and CFI & TLI were both .89.

Section 4: Measurement invariance methods and results (Phase 3)

Measurement invariance analysis methods. We conducted measurement invariance analyses to ensure that the survey items function equivalently across groups. Measurement invariance tests for equivalence using a factor analytic framework. Following best practices (Vandenberg & Lance, 2000), we used an omnibus approach to test for measurement invariance. We estimated nested sets of multiple-groups CFA model. In the constrained model, all parameters are set equal across groups. In the free model, all parameters were freely estimated across both groups. If the free model is a better fit to the data than the constrained model, that is an indication that there are differences of some kind across groups. Further testing would be necessary to investigate the nature of the differences. However, if the constrained model fits equally well as or better than the free model, we can conclude that there are no differences across groups, and measurement invariance can be confirmed.

We examine a variety of demographic characteristics: gender, race/ethnicity, disability status, generation in college, English as first language, institution type (e.g., research intensive or teaching intensive), institution population (e.g., Historically Black College or University or Primarily White Institution), and discipline enrolled (e.g., biology, chemistry, math, or physics). Demographic identities are complex and diverse. However, for many identities, we received too few respondents to enable estimation of models to test for measurement invariance or DIF. For some comparisons, we aggregated or excluded participants with low-frequency identities so that models could be estimated. We recognize that students with different identities have distinct sets of experiences and perspectives and thus aggregation and exclusion limit our ability to draw inferences about these groups.

Gender was dichotomized because there were too few respondents who selected an identity other than “man” or “woman” to include other gender identities in analyses. Participants who selected a non-binary or other identity were excluded from these analyses. For racial/ethnic identity, we examined two different dichotomous comparisons. First, we compared students who identified as “white” (they may have also selected other racial/ethnic identities) to students who did not select “white” as a group they identify with. Second, we compared students who identified as Under-Represented Minority (URM; as defined by the NSF, includes students who identified as Black or African American, Hispanic or Latin(x), Native American, and/or Native Hawaiian or Pacific Islander) to students who did not identify with a URM group. For disability status, we aggregated responses of students who indicated they had any type of disability or impairment to compare against students who indicated no disabilities or impairments. First generation was defined as having no parents/guardians who completed a 4-year degree. We aggregated institution types into three larger categories for comparison: research intensive (RI: High Research Activity and Very High Research Activity), Four-year teaching intensive (Doctoral, Masters, and Baccalaureate institutions), and two-year institutions (community colleges). We compared institution population by comparing students enrolled in Primarily-White Institutions to those enrolled in Minority-Serving Institutions. We compared responses of students enrolled in the four main types of introductory courses we recruited at: biology, chemistry, physics, and math. Many students were cross-enrolled, so we estimated these differences using four sets of dummy codes (i.e., comparing students enrolled in a biology course to students not enrolled in a biology course, etc.).

Measurement invariance results. The model-data fit metrics for measurement invariance tests are presented in supplemental table 2. We interpreted fit statistics following guidelines from Cheung and Rensvold (2002): ΔCFI values = .01 or ΔRMSEA values = -.015 indicate worse model fit. We also examine Bayesian Information Criteria (BIC), which is specifically designed for model comparison. For each of these comparisons, the constrained model was an equivalent or better fit than the free model (with the exception of generation in college, which could not be estimated because the covariance matrix was not positive-definite). Thus, we failed to detect evidence of measurement invariance across any of the groups tested.

Supplemental Table 2. *Measurement invariance results (Phase 3)*
Note: In the free model for Generation in college, the covariance matrix was not positive definite. Thus, this model cannot be interpreted. RI = Research Intensive institutions (High Research Activity and Very High Research Activity Carnegie Classifications); 4yr = Four-year teaching intensive institutions (Doctoral, Masters, and Baccalaureate Carnegie classifications); CC = Community Colleges (two-year institutions)

Variable	Model	CFI	TLI	RMSEA	RMSEA 90% CI	SRMR	Chi^2	df	BIC
Gender	Constrained	0.849	0.851	0.062	0.060-0.063	0.073	6857	2493	139221
	Free	0.853	0.846	0.063	0.061-0.064	0.065	6566	2336	139962
Race: white	Constrained	0.846	0.849	0.062	0.060-0.064	0.076	6958	2493	140261
	Free	0.853	0.846	0.063	0.061-0.065	0.068	6619	2336	140936
Race: URM	Constrained	0.836	0.838	0.063	0.062-0.065	0.088	7096	2493	140261
	Free	0.851	0.844	0.062	0.061-0.064	0.068	6542	2336	140635
Disability	Constrained	0.847	0.849	0.063	0.060-0.064	0.082	6950	2493	139675
	Free	0.852	0.845	0.063	0.061-0.065	0.069	6647	2336	140414
Generation	Constrained	0.848	0.850	0.061	0.060-0.063	0.073	6853	2493	140740
	Free*	0.853	0.846	0.062	0.061-0.064	0.066	6573	2336	141461
Language	Constrained	0.848	0.851	0.062	0.060-0.064	0.073	7113	2493	142870
	Free	0.851	0.844	0.063	0.062-0.065	0.069	6894	2336	143687
Institution type: CC	Constrained	0.844	0.846	0.063	0.061-0.065	0.081	7230	2493	143069
	Free	0.848	0.841	0.064	0.062-0.066	0.069	6952	2336	143821
Institution type: 4-yr	Constrained	0.849	0.852	0.062	0.060-0.063	0.075	6997	2493	143069
	Free	0.852	0.845	0.063	0.061-0.065	0.069	6766	2336	143888
Institution type: RI	Constrained	0.849	0.852	0.062	0.060-0.063	0.079	6967	2493	143069
	Free	0.852	0.845	0.063	0.061-0.065	0.069	6730	2336	143873
Institution population	Constrained	0.839	0.842	0.063	0.061-0.065	0.086	7126	2493	143069
	Free	0.851	0.844	0.063	0.061-0.064	0.070	6680	2336	143565
Discipline: biology	Constrained	0.852	0.855	0.061	0.059-0.063	0.073	6884	2493	143069
	Free	0.853	0.846	0.063	0.061-0.065	0.069	6699	2336	143931

Discipline: chemistry	Constrained	0.851	0.853	0.061	0.060-0.063	0.076	6939	2493	143069
	Free	0.852	0.845	0.063	0.061-0.065	0.070	6751	2336	143934
Discipline: physics	Constrained	0.848	0.850	0.062	0.060-0.064	0.074	7106	2493	143069
	Free	0.851	0.843	0.063	0.063-0.065	0.069	6890	2336	143872
Discipline: math	Constrained	0.852	0.854	0.061	0.059-0.063	0.075	6900	2493	143069
	Free	0.851	0.843	0.063	0.062-0.065	0.069	6890	2336	143872

Section 5: Different Item Functioning methods and results (Phase 3)

Differential Item Functioning (DIF) methods. It is so critical that the measure functions equivalently across groups that we also searched for any issues by conducting differential item functioning (DIF) analyses, which are situated within an item response theory framework. DIF analyses examine whether groups differ in their response characteristic curves. In other words, significant DIF suggests that individuals in different groups who are at the same level of the latent trait have different probabilities of selecting a given response. DIF is useful for measurement development because it is highly sensitive (i.e., it can detect differences even with very small effect sizes) and works at the item level. Thus, these analyses can pinpoint particular items that are nonequivalent. We conducted these analyses using the multiplegroups() and DIF() functions in the mirt package (Chalmers, 2012).

we conduct DIF analyses using unidimensional IRT models due to the computational complexity, long run times, and estimation issues that arise in IRT models with several factors. Thus, we conducted DIF analyses separately for each of the five factors. For each factor, we first identified anchor items (items that are known to function equivalently across groups) by estimating a model in which everything is free except the latent means and variances for groups are set to 0 and 1, respectively (using the multiplegroups() function). We then examine whether any items show potential for DIF using the DIF() function. Any items with non-significant X2 values are identified as anchor items. We used a conservative critical value cut-off of 0.01 because DIF is particularly sensitive to type I (false positive) error and we are conducting a large number of tests. Selecting a cut-off of 0.01 is less conservative than a Bonferroni correction would be, but we chose to err on the side of detecting all issues with items rather than failing to detect potentially serious problems with item functioning. Once anchor items have been identified, we then fit another model that constrains these items across groups. We then investigate whether any remaining items are now non-significant and if so, fit a new model adding these items to the constrained parameters list. If any items remain that display significant DIF, this process is repeated one final time. If any items still have significant DIF in that third model, then they are confirmed to have issues with DIF. We used DIF analysis to test for differential item functioning across all of the same demographic groups we tested with measurement invariance. The exception is that DIF analyses comparing URM to non-URM students could not be calculated for growth items due to missing response patterns, (i.e., insufficient sample size). This is because DIF analyses are highly sensitive but have very high sample size requirements.

DIF analysis results. DIF analyses flagged potential issues with 7 items. One fixed item (10) had DIF with respect to English as a first language. Two fixed items (2 & 11) had DIF with respect to disability status. One growth item (8) had DIF for students enrolled in physics compared to students not enrolled in physics. One non-universal item (10) had DIF with respect to English as a first language. One universal item (9) had DIF for students enrolled in math compared to those not in math. One Brilliance item (6) had DIF for students who attend community colleges compared to all other institution types.

Effect sizes of all DIF detected were very small (cohen's d < 0.01). Nonetheless, these metrics were taken into consideration when making decisions about the recommended short form of the measure. We ultimately retained two of the items in the short form that had displayed DIF because they had excellent other psychometric properties, were important for conceptual coverage, and the effect sizes of DIF were near zero: brilliance 6 (cohen's d = 0.007) and universal 9 (cohen's d = 0.0008).

Section 6. Item Response Theory analysis methods and results

Testing assumptions. We fit Graded Response Models (GRMs) to each first-order factor of our factor model: growth, fixed, universal, non-universal, and brilliance. The Graded Response model assumes that each model must be unidimensional. We reasoned that the good model-data fit of the CFA with each factor separated suggested that each factor is unidimensional. We further evaluated this assumption by examining model-data fit for each of the five GRM models. We examined model-data fit using both item-level fit (S-X^2) as well as overall model fit (M2). S-X^2 values, along with degrees of freedom (df) and p-values are presented in Supplemental Table 3 below. To interpret significance of mis-fit, we used Bonferroni p-value correction to reduce the risk of Type I error. Thus, the critical value for each lay theory model is as follows: growth items 0.05 / 10 = 0.005; fixed items 0.05 / 11 = 0.005; brilliance items 0.05 / 6 = 0.008; non-universal items 0.05 / 12 = 0.004, and universal items 0.05 / 11 = 0.005. Overall model fit was estimated using the M2() function in the mirt package.

The overall model-fit metrics SRMR, TLI, and CFI indicated acceptable or good fit for all 5 models. RMSEA was unacceptable for all 5 models, but RMSEA is known to become unreliable when df are low, as they are in these models (Kenny et al., 2015). None of the growth, fixed, or non-universal items indicated mis-fit, indicating support that each of these meet the unidimensionality assumption. Only 2 items showed mis-fit for the universal model. Neither of these items were selected for the recommended short form of the measure as a precaution. All of the brilliance items showed significant mis-fit. However, item-fit metrics are relative, meaning that if all items in a model fit well, item fit metrics can be poor (Orlando & Thissen, 2000). To investigate this possibility, we examined the item fit graphs for brilliance items, presented below in Supplemental Figure 1. Inspecting the graphs of item-data fit suggest that all items fit model expectations well. The high discrimination (alpha) parameters for the brilliance items, strong global model-data fit, and visual inspection of item fit graphs each support that brilliance items fit the data well.

Supplemental Table 3. Full set of items in the ULTrA measure with item-fit metrics and alpha (α) and beta (β) parameter values from Graded Response Models fit for each construct. Items selected for the short form of the measure are indicated with bolding and an asterisk preceding the item. Note that Fixed_02, Fixed_07, and Nonuniversal_8 were deleted from the final model due to substantial residual covariances.

Item ID	Item text	S-X^2	df	p	α	β1	β2	β3	β4
Growth 1	***If I try, I can become as effective at learning as STEM experts.**	51.5	52	0.49	3.11	-2.09	-1.30	-0.90	0.28
Growth 2	I can vastly improve my ability to think creatively.	46.2	61	0.92	1.75	-3.01	-1.56	-1.06	0.35
Growth 3	***I can become as good at analyzing information as highly successful STEM professionals if I try hard enough.**	31.9	43	0.89	4.20	-2.00	-1.21	-0.91	0.10
Growth 4	***If I want to, I can become as effective at applying knowledge**	53.0	44	0.17	4.38	-1.93	-1.21	-0.87	0.20

	as STEM experts.								
Growth 5	***I could improve my intellectual abilities to the same level as successful STEM professionals.**	52.8	40	0.08	4.77	-1.86	-1.09	-0.75	0.22
Growth 6	It’s possible that I could become as creative as highly successful STEM professionals one day.	63.2	54	0.18	2.59	-2.25	-1.29	-0.97	0.22
Growth 7	I can improve how well I can learn complex concepts in STEM.	90.3	60	0.01	2.05	-3.23	-2.10	-1.65	-0.11
Growth 8	I can improve my intellectual abilities to a large extent.	64.1	59	0.30	2.03	-2.84	-1.70	-1.23	0.08
Growth 9	***I can become excellent at applying knowledge to solve challenging problems.**	60.4	56	0.32	2.22	-3.05	-2.04	-1.63	-0.09
Growth 10	I can greatly improve how well I analyze information.	59.1	61	0.55	1.93	-3.30	-2.37	-1.82	-0.13
Fixed 1	My intellectual ability will remain about the same over time.	103.9	96	0.27	1.29	-0.72	1.06	1.33	2.95
Fixed 2	My ability to think creatively will stay about the same throughout college.	106.6	101	0.33	1.34	-0.91	0.95	1.26	2.30
Fixed 3	Even if I try to improve, there will be limits to how effectively I can analyze information.	86.2	96	0.75	1.50	-1.25	-0.05	0.26	1.97
Fixed 4	***My ability to apply knowledge will change very little over time.**	109.9	102	0.28	1.29	-0.64	1.13	1.48	2.51
Fixed 5	I will always learn at about the same pace that I do now.	95.1	97	0.54	1.59	-1.05	0.34	0.74	2.05
Fixed 6	***I will never be able to reach the highest level of intellectual ability.**	114.5	104	0.23	1.16	-1.04	0.22	0.66	1.95
Fixed 7	Even if I try my best, I could never become extremely creative.	99.3	101	0.53	1.41	-0.63	0.56	0.92	2.24
Fixed 8	***At the end of college, my ability to analyze information will be at about the same level that it is now.**	75.5	70	0.31	2.12	-0.02	1.43	1.74	2.60
Fixed 9	***It would be very difficult for me to improve how well I can apply knowledge.**	96.5	80	0.10	2.02	-0.51	0.88	1.31	2.55
Fixed 10	***How well I learn is something that I cannot change very much.**	87.9	88	0.51	2.07	-0.51	0.65	1.06	2.08
Fixed 11	How effectively I learn is relatively constant over most of my life.	119.9	91	0.02	1.70	-0.82	0.43	0.77	2.26

Brilliance 1	People have to be naturally brilliant to reach the top of a STEM field.	87.4	44	0	2.75	-0.72	0.21	0.52	1.60
Brilliance 2	***Excelling in STEM requires natural talent.**	71.4	35	0	3.38	-0.80	0.14	0.49	1.69
Brilliance 3	***People who are highly successful in STEM have a natural talent for it.**	120.7	38	0	2.73	-1.17	-0.28	0.18	1.51
Brilliance 4	***Becoming a top student in STEM requires an innate talent that just can't be taught.**	129.0	36	0	3.75	-0.50	0.44	0.84	1.76
Brilliance 5	***People have to be naturally talented to excel in challenging STEM courses.**	82.0	38	0	3.56	-0.57	0.42	0.77	1.83
Brilliance 6	***Being a highly successful STEM professional requires natural talent that just can't be taught.**	75.5	34	0	4.26	-0.42	0.48	0.85	1.70
Non-universal 1	***Some people will always be less effective at learning than those who have a natural talent for it.**	111.7	101	0.22	1.83	-1.66	-0.62	-0.08	1.48
Non-universal 2	People with a natural talent will achieve greater success in STEM than others.	116.1	103	0.18	1.84	-1.32	-0.21	0.38	1.71
Non-universal 3	***Only some people have the intellectual ability to become a successful STEM professional.**	102.1	95	0.29	2.17	-0.68	0.27	0.73	1.82
Non-universal 4	***Only people with a natural talent can become good enough at applying knowledge to solve the most difficult problems.**	100.5	89	0.19	2.04	-0.51	0.71	1.15	2.30
Non-universal 5	***Even if they try, some people could never become as effective at analyzing information as their peers.**	109.7	95	0.14	2.24	-0.99	-0.07	0.27	1.50
Non-universal 6	Some people will always be able to think more creatively than others because they are naturally creative.	117.5	98	0.09	1.86	-1.88	-0.96	-0.56	0.88
Non-universal 7	Only some people can become great at applying knowledge to solve challenging problems.	108.4	98	0.22	1.82	-0.88	0.41	0.83	2.25
Non-universal 8	Some people are just naturally better at analyzing information than others.	119.6	101	0.10	1.43	-2.53	-1.48	-1.19	0.97
Non-universal 9	Only some people are capable of becoming very creative.	88.1	105	0.88	1.72	-1.06	0.12	0.59	1.96
Non-universal 10	Not everyone has the intellectual ability to earn a STEM degree.	127.7	106	0.07	1.73	-1.00	-0.07	0.35	1.51

Non-universal 11	Some people will always be able to learn better than others.	107.9	103	0.35	1.58	-2.30	-1.31	-0.91	0.84
Non-universal 12	***Only people with a natural talent can become excellent at analyzing information.**	115.0	93	0.06	1.75	-0.53	0.87	1.46	2.67
Universal 1	Everyone has the potential to become as good at analyzing information as STEM experts.	80.3	76	0.35	2.80	-1.95	-1.12	-0.76	0.35
Universal 2	Everyone has the potential to become as creative as successful STEM professionals.	98.4	72	0.02	2.77	-2.09	-1.09	-0.77	0.38
Universal 3	***Everyone has the intellectual ability to become a successful STEM professional if they want to.**	68.0	62	0.28	3.64	-1.81	-0.93	-0.63	0.34
Universal 4	***Anyone who tries could become as good at applying knowledge as STEM experts.**	101.2	70	0.01	3.02	-2.01	-1.10	-0.76	0.30
Universal 5	Everyone has the intellectual ability to have a highly successful career in STEM if they work hard.	112.5	73	0.00	2.88	-2.08	-1.19	-0.84	0.13
Universal 6	Anyone who wants to could be able to think creatively as well as highly successful STEM students.	108.2	71	0.00	2.65	-2.15	-1.03	-0.63	0.53
Universal 7	Everyone has the intellectual ability to succeed in a STEM career if they want to.	80.5	65	0.09	3.30	-1.90	-0.94	-0.65	0.38
Universal 8	***With enough hard work, anyone could become as good at analyzing information as highly successful STEM professionals.**	67.6	60	0.23	3.85	-2.06	-1.08	-0.81	0.16
Universal 9	***With enough motivation, anyone can become as good at applying knowledge as high achieving STEM students.**	86.0	64	0.04	3.10	-2.28	-1.33	-1.01	0.12
Universal 10	Everyone has the intellectual ability to earn a STEM degree if they work hard.	85.8	76	0.21	2.79	-2.16	-1.23	-0.87	0.11
Universal 11	***Anyone could become as effective at learning as highly successful STEM students.**	82.7	61	0.03	3.67	-1.95	-1.07	-0.71	0.29

Supplemental Table 4. Global model-fit metrics from Graded Response Models fit for each construct.

	M2	**df**	**p**	**RMSEA (95% CI)**	**SRMSR**	**TLI**	**CFI**
Growth	775.374	35	0	0.136 (0.127 - 0.144)	0.077	0.942	0.955
Fixed	808.945	44	0	0.121 (0.114 - 0.129)	0.086	0.882	0.905
Brilliance	262.599	9	0	0.154 (0.139 - 0.171)	0.0526	0.947	0.968

Non-Universal	1029.914	54	0	0.124 (0.118 - 0.131)	0.075	0.916	0.931
Universal	1023.712	44	0	0.138 (0.131 - 0.146)	0.073	0.951	0.960

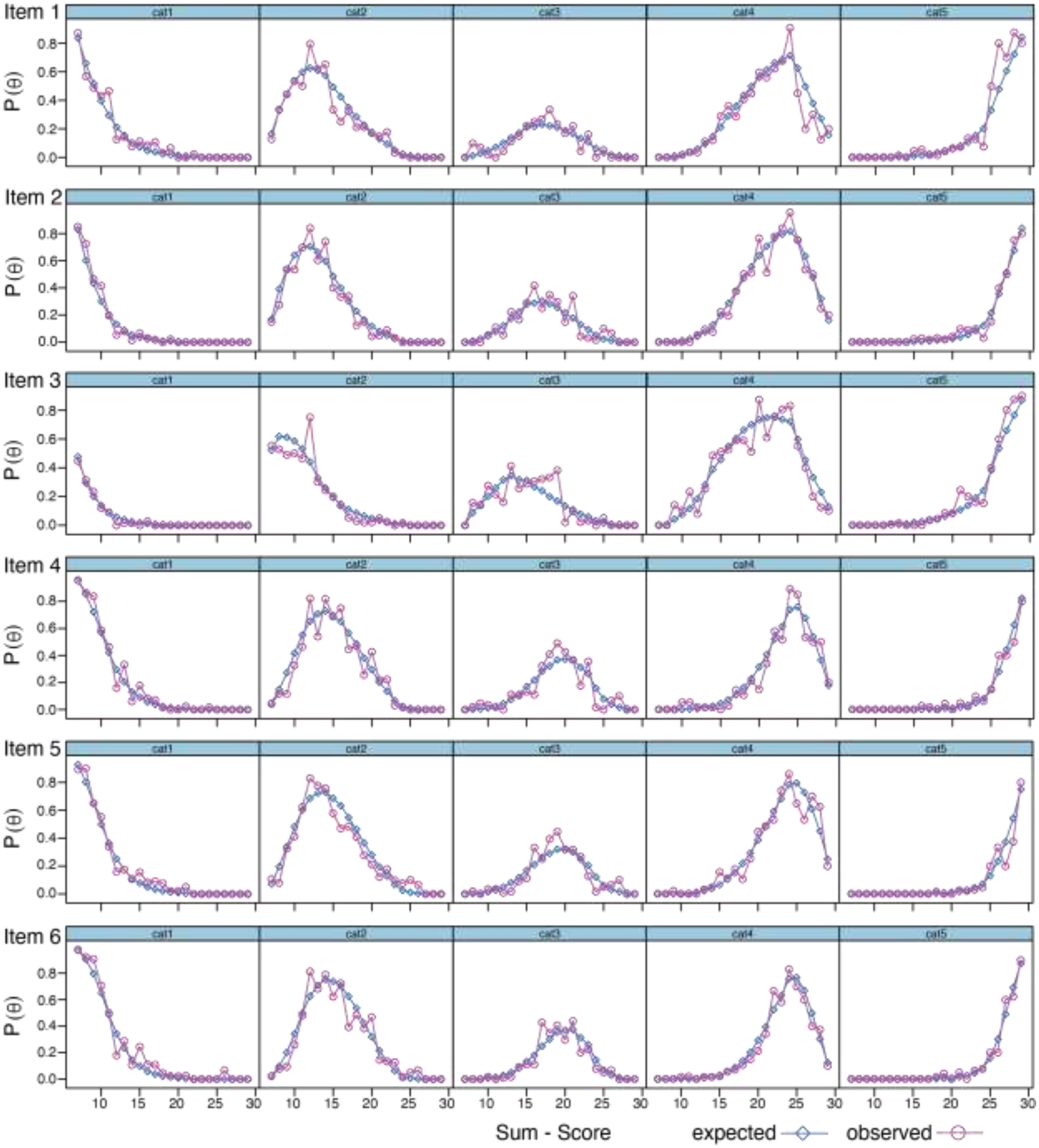
A second assumption of the Graded Response Model is local independence, which is the assumption that associations between items are fully explained by the latent variable. We assessed local dependence by examining standardized residual correlations between items. The assumption of local dependence would be met if we observed no substantial residual correlations. Scholars have recommended using absolute values of 0.2 or 0.3 as cut-offs (Chen & Thissen, 1997; Christensen et al., 2017). Residual correlations for each of the five graded response models are presented in Supplemental Table 5 below.

Supplemental Table 5. Descriptive statistics of standardized residual correlations for each Graded Response Model.

Model	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
Growth	-0.135	-0.103	-0.09	-0.013	0.111	0.17
Fixed	-0.135	-0.102	-0.086	-0.025	0.093	0.187
Brilliance	-0.22	-0.171	-0.165	-0.079	0.006	0.215
Non-Universal	-0.196	-0.137	-0.115	-0.036	0.131	0.213
Universal	-0.198	-0.164	-0.148	-0.082	-0.118	0.3

The highest absolute value residual correlation in any model is 0.3. Thus, our models do not show strong local dependence.

Interpreting IRT model parameters. We then interpreted the alpha and beta parameters of each item, also presented in Supplemental table 3. The alpha parameter, called the discrimination parameter, represents how well the item can discriminate respondents at different levels of the latent trait. Higher alpha values are ideal. The beta parameter, called the location parameter, describes the location in the latent trait at which the item best discriminates among respondents (i.e., where "information" is highest). A set of items that cover a range of beta values is ideal. Each item has a beta parameter for each border between response options. Since there are 5 response options (strongly disagree, somewhat disagree, neither agree nor disagree, somewhat agree, strongly agree), there are 4 beta parameters presenting the 4 borders between 5 options. For example, the first beta parameter represents the location at which a participant becomes more likely to select "somewhat disagree" than "strongly disagree" for each item. Items that were selected to be retained in the short version of the measure are indicated with bolding and an asterisk.



Supplemental Figure 1. Figures showing the alignment of model-expected and observed response curves for the 6 items in the Graded Response Model of the Brilliance items.

Section 7. Survey items (Phase 4)

Undergraduate Lay Theories of Abilities (ULTrA) short form, developed in the present study

- 25 items measuring 5 dimensions
- Prompt: Please indicate the extent to which you agree or disagree with the following statements. There are no correct answers, we want to understand how **you** think about these ideas. Note that STEM stands for Science, Technology, Engineering, and Mathematics. STEM professionals are individuals in a career in a STEM field, such as scientists, engineers, medical doctors, and other healthcare professionals.
- Response scale: 1 = Strongly disagree; 2 = Somewhat disagree; 3 = Neither agree nor disagree; 4 = Somewhat agree; 5 = Strongly agree; Prefer not to respond

Items

Fixed Belief

1. At the end of college, my ability to analyze information will be at about the same level that it is now.
2. How well I learn is something that I cannot change very much.
3. My ability to apply knowledge will change very little over time.
4. I will never be able to reach the highest level of intellectual ability.
5. It would be very difficult for me to improve how well I can apply knowledge.

Growth Belief

6. I can become as good at analyzing information as highly successful STEM professionals if I try hard enough.
7. If I want to, I can become as effective at applying knowledge as STEM experts.
8. I can become excellent at applying knowledge to solve challenging problems.
9. If I try, I can become as effective at learning as STEM experts.
10. I could improve my intellectual abilities to the same level as successful STEM professionals.

Non-Universal Belief

11. Even if they try, some people could never become as effective at analyzing information as their peers.
12. Only people with a natural talent can become good enough at applying knowledge to solve the most difficult problems.
13. Only people with a natural talent can become excellent at analyzing information.
14. Some people will always be less effective at learning than those who have a natural talent for it.
15. Only some people have the intellectual ability to become a successful STEM professional.

Universal Belief

16. With enough hard work, anyone could become as good at analyzing information as highly successful STEM professionals.
17. Anyone who tries could become as good at applying knowledge as STEM experts.
18. Anyone could become as effective at learning as highly successful STEM students.
19. Everyone has the intellectual ability to become a successful STEM professional if they want to.

20. With enough motivation, anyone can become as good at applying knowledge as high achieving STEM students.

Brilliance Belief

21. Excelling in STEM requires natural talent.
22. People who are highly successful in STEM have a natural talent for it.
23. Becoming a top student in STEM requires an innate talent that just can’t be taught.
24. People have to be naturally talented to excel in challenging STEM courses.
25. Being a highly successful STEM professional requires natural talent that just can’t be taught.

<u>Goal orientation</u> Achievement Goal Questionnaire-Revised (Elliot & Murayama, 2008)

- 6 items measuring 2 dimensions
- Prompt: Please indicate the extent to which you agree or disagree with the following statements. There are no correct answers, we want to understand how **you** think about these ideas.

Evidence of validity

We measured achievement goal orientation using the Achievement Goal Questionnaire-Revised (AGQ-R; Elliot & Murayama, 2008). We selected this measure because prior studies have collected strong validity evidence for its utility measuring the academic achievement goals of undergraduates (Cook et al., 2017; de Castella & Byrne, 2015; Elliot & Murayama, 2008; Yan & Wang, 2021). In their revision of the original measure, Elliot and Murayama (2008) presented multiple sources of validity evidence. The re-worded many items to better align with achievement goal theory, constituting validity evidence related to content. They also presented validity evidence related to internal structure by conducting a confirmatory factor analysis on responses collected from 229 undergraduates that achieved excellent model-data fit. Finally, they present evidence related to relations to other variables by demonstrating that achievement goals are related to fear of failure, motivation, and exam performance in expected ways, but are not so high as to raise concern about conceptual overlap. Further, other studies using this measure have replicated the same factor structure and also presented evidence of relations to other variables (Cook et al., 2017; de Castella & Byrne, 2015; Yan & Wang, 2021). The AGQ-R has items measuring all four types of goal orientations. However, we only included items measuring mastery-approach and performance-avoid goals because we had strong theoretical and empirical basis for theorizing about how they should relate to mindset beliefs. We excluded the other two dimensions because we could not make strong a priori theoretical predictions.

Items

Mastery approach

1. My aim is to completely master the material presented in this class.
2. I am striving to understand the content of this course as thoroughly as possible.
3. My goal is to learn as much as possible.

Performance avoid

4. My aim is to avoid doing worse than other students.

5. I am striving to avoid performing worse than others.
6. My goal is to avoid performing poorly compared to others.

Belonging (Hoffman et al., 2002)

- 26 items measuring 5 dimensions

Evidence of validity

We measure sense of belonging using a scale developed by Hoffman and colleagues (2002) because there is strong evidence of validity for its utility measuring sense of belonging for undergraduates. Hoffman and colleagues rigorously developed the instrument through an iterative process that involved collecting validity evidence, refining the measure, and collecting more evidence. They present strong validity evidence related to content, response process, and internal structure for using the resulting instrument to measure undergraduates' sense of belonging (Hoffman et al, 2002). The measure contains 26 items measuring five dimensions: perceived peer support, perceived faculty support, perceived classroom comfort, perceived isolation, and empathic faculty understanding.

Items

Perceived peer support

1. I could contact another student from class if I had a question about an assignment.
2. Other students are helpful in reminding me when assignments are due or when tests are approaching.
3. If I miss class, I know students who I could get the notes from.
4. I have developed personal relationships with other students in class.
5. I have met with classmates outside of class to study for an exam.
6. I discuss events which happen outside of class with my classmates.
7. I invite people I know from class to do things socially.
8. I have discussed personal matters with students who I met in class.

Perceived faculty support

9. I feel comfortable seeking help from a teacher before or after class.
10. I feel comfortable asking a teacher for help if I do not understand course-related material.
11. If I had a reason, I would feel comfortable seeking help from a faculty member outside of class time (i.e., during office hours, etc.).
12. I feel comfortable talking about a problem with faculty.
13. I feel comfortable socializing with a faculty member outside of class.
14. I feel comfortable asking a teacher for help with a personal problem.

Perceived classroom comfort

15. Speaking in class is easy because I feel comfortable.
16. I feel comfortable volunteering ideas or opinions in class.
17. I feel comfortable contributing to class discussions.
18. I feel comfortable asking a question in class.

Perceived isolation

19. It is difficult to meet other students in class.

20. No one in my classes knows anything personal about me.
21. I rarely talk to other students in my classes.
22. I know very few people in my classes.

Empathetic faculty understanding

23. I feel that a faculty member would take the time to talk to me if I needed help.
24. I feel that a faculty member would be sympathetic if I was upset.
25. I feel that a faculty member would be sensitive to my difficulties if I shared them.
26. I feel that a faculty member really tried to understand my problem when I talked about it.

<u>Self-handicapping</u> (Midgley et al., 2000)

- 6 items measuring one dimension
- Response scale: Items are anchored at 1 = "Not at all true," 3 = "Somewhat true," and 5 = "Very true."

Evidence of validity

We measured self-handicapping using the 6-item sub-scale of the Patterns of Adaptive Learning Scale (PALS; Midgley et al., 2000). The PALS has been developed and refined over time by a group of researchers, grounded in goal orientation theory. The revision published in 2000 includes evidence of validity based on multiple sources, including content, internal structure, and relations to other variables, as well as evidence of reliability, with a middle school student population (Midgley et al., 2000). Subsequent studies have used the self-handicapping sub-scale with undergraduates and collected evidence of validity based on internal structure as well as evidence of reliability (e.g., Yu & McLellan, 2020).

Items

1. Some students fool around the night before a test. Then if they don't do well, they can say that is the reason. How true is this of you?
2. Some students purposely get involved in lots of activities. Then if they don't do well on their class work, they can say it is because they were involved with other things. How true is this of you?
3. Some students look for reasons to keep them from studying (not feeling well, having to help their parents, taking care of a brother or sister, etc.). Then if they don't do well on their class work, they can say this is the reason. How true is this of you?
4. Some students let their friends keep them from paying attention in class or from doing their homework. Then if they don't do well, they can say their friends kept them from working. How true is this of you?
5. Some students purposely don't try hard in class. Then if they don't do well, they can say it is because they didn't try. How true is this of you?
6. Some students put off doing their class work until the last minute. Then if they don't do well on their work, they can say that is the reason. How true is this of you?

<u>Evaluative concerns:</u> Adapted from Wout, Steele, & Murphy, 2010

- 5 items measuring 1 dimension

- Response scale: 1=Not at all to 6=Extremely

Evidence of validity
We measured evaluative concerns using six items adapted from Wout, Steele, & Murphy (2010) by Muenks et al., (2020). A limitation of this work is that there is evidence of reliability, but relatively little evidence of validity available for this measure. Evidence of validity based on content is supported by the current researchers' judgment that the items relate to undergraduates' evaluative concerns in the classroom and that the items were written by experts in this field.

Items:

1. In class, how much did you worry that you might have said the wrong thing?
2. In class, how much did you worry that you might have made a mistake in front of your professor?
3. In class, how much did you worry that the professor might have underestimated your intelligence?
4. In class, how much did you worry that your professor might have thought that you were a slow learner?
5. In class, how much did you worry that your professor might not believe in your abilities to do well in this class?

Intent to persist: Adapted from (Estrada et al., 2011):

- 3 items measuring 1 dimension
- Prompt: Please rate how likely you are to pursue…
- Response Scale: 1 = not at all likely; 2 = somewhat likely; 3 = unsure; 4 = somewhat likely; 5 = very likely; prefer not to respond

Evidence of validity
We chose to measure intent to persist as an indicator of actual persistence based on evidence that intent to persist is a strong predictor of actual persistence (Estrada et al., 2011). We measure intent to persist in science using three items adapted from Estrada and colleagues (2011), who report strong evidence that responses to these items correlate well with longitudinal behavioral indicators persistence.

Items:

1. A career in science
2. A career in research
3. Graduate education in science

Demographic questions
Which of the following accurately describes your family's education?

- Continuing generation: At least one of my parent(s)/guardian(s) has earned a 4-year college degree.

- First generation: None of my parent(s)/guardian(s) has earned a 4-year college degree.
- Prefer not to respond.

With which race(s) and ethnicity/ies do you identify? Select all that apply:

- African American or Black
- East Asian (e.g., China and Japan)
- South Asian (e.g., the Indian sub-continent)
- Southeast Asian (e.g., Vietnam)
- Latinx or Hispanic
- Middle Eastern or North African
- Native American or Alaskan Native
- Native Hawaiian or Pacific Islander
- White
- Other: ____
- Prefer not to respond

What gender do you identify as?

- Man
- Woman
- Non-binary
- Not listed above: _____
- Prefer not to respond

Section 8: Correlation table

Supplemental Table 6. Correlation Table for variables measured in Phase 4. * p < 0.05; ** p < 0.01; *** p < 0.001

	Mean	St. Dev.	1	2	3	4	5	6	7	8	9	10	11	12
1. Growth mindset	4.30	0.75												
2. Fixed mindset	2.03	0.71	-0.35***											
3. Brilliance belief	2.46	0.95	-0.24***	0.28***										
4. Universal belief	4.03	0.87	0.38***	-0.19***	-0.42***									
5. Non-universal belief	2.38	0.89	-0.24***	0.34***	0.67***	-0.53***								
6. Sense of belonging	3.44	0.71	0.23***	-0.28***	-0.11***	0.18***	-0.16***							
7. Mastery-approach goals	4.51	0.63	0.27***	-0.23***	-0.12***	0.22***	-0.18***	0.25***						
8. Performance-avoid goals	3.77	1.16	-0.05	0.11***	0.13***	0.02	0.12***	-0.03	0.12***					
9. Self-handicapping	2.15	0.93	-0.15***	0.23***	0.12***	-0.03	0.15***	-0.08**	-0.12***	0.12***				
10. Evaluative concerns	2.76	1.1	-0.08**	0.20***	0.13***	0.02	0.08**	-0.27***	0.02	0.19***	0.21***			
11. Intent to persist	3.66	1.08	0.28***	-0.13***	-0.08**	0.08**	-0.08**	0.09**	0.17***	-0.05	-0.06*	0.08**		
12. Course grade	3.09	1.01	0.15***	-0.09**	-0.01	-0.06*	0.02	0.06*	0.09**	-0.03	-0.18***	-0.08**	0.14***	
13. Overall GPA	3.32	0.7	0.16**	-0.21***	-0.06	-0.03	-0.06	0.24***	0.15*	0.04	-0.20**	-0.07	0.01	0.77***

References

Chalmers, R. P. (2012). mirt: A multidimensional item response theory package for the R environment. *Journal of Statistical Software*, *48*(6), 1–29. https://doi.org/10.18637/jss.v048.i06

Chen, W.-H., & Thissen, D. (1997). Local Dependence Indexes for Item Pairs Using Item Response Theory. *Journal of Educational and Behavioral Statistics*, *22*(3), 265–289. https://doi.org/10.3102/10769986022003265

Cheung, G. W., & Rensvold, R. B. (2002). Evaluating goodness-of-fit indexes for testing measurement invariance. *Structural Equation Modeling*, *9*(2), 233–255.

Christensen, K. B., Makransky, G., & Horton, M. (2017). Critical Values for Yen's Q3: Identification of Local Dependence in the Rasch Model Using Residual Correlations. *Applied Psychological Measurement*, *41*(3), 178–194. https://doi.org/10.1177/0146621616677520

Estrada, M., Woodcock, A., Hernandez, P. R., & Schultz, P. W. (2011). Toward a model of social influence that explains minority student integration into the scientific community. *Journal of Educational Psychology*, *103*(1), 206–222. https://doi.org/10.1037/a0020743

Hoffman, M., Richmond, J., Morrow, J., & Salomone, K. (2002). Investigating "sense of belonging" in first-year college students. *Journal of College Student Retention: Research, Theory & Practice*, *4*(3), 227–256. https://doi.org/10.2190/DRYC-CXQ9-JQ8V-HT4V

Jorgensen, T. D., Pornprasertmanit, S., Schoemann, A. M., & Rosseel, Y. (2021). *semTools: Useful tools for structural equation modeling* (0.5-5) [Computer software]. https://CRAN.R-project.org/package=semTools

Kenny, D. A., Kaniskan, B., & McCoach, D. B. (2015). The performance of RMSEA in models with small degrees of freedom. *Sociological Methods & Research*, *44*(3), 486–507. https://doi.org/10.1177/0049124114543236

Muenks, K., Canning, E. A., LaCosse, J., Green, D. J., Zirkel, S., Garcia, J. A., & Murphy, M. C. (2020). Does my professor think my ability can change? Students' perceptions of their STEM professors' mindset beliefs predict their psychological vulnerability, engagement, and performance in class. *Journal of Experimental Psychology: General*. https://doi.org/10.1037/xge0000763

Orlando, M., & Thissen, D. (2000). Likelihood-based item-fit indices for dichotomous item response theory models. *Applied Psychological Measurement*, *24*(1), 50–64.

Vandenberg, R. J., & Lance, C. E. (2000). A Review and Synthesis of the Measurement Invariance Literature: Suggestions, Practices, and Recommendations for Organizational Research. *Organizational Research Methods*, *3*(1), 4–70. https://doi.org/10.1177/109442810031002