Investigating Sense of Belonging of Graduate Students Using Explanatory Item Response Models

Carlos Chavez, Tai Do, Michael C. Rodriguez
University of Minnesota

Minnesota Youth Development Research Group
www.mnydrg.com

April 13, 2023

Paper presented at the annual meeting of the
National Council on Measurement in Education, Chicago, IL.

Citation:

Chavez, C., Do, T., & Rodriguez. M.C. (2023, April 13). *Investigating sense of belonging of graduate students using explanatory item response models* [Paper presentation]. National Council on Measurement in Education Annual Meeting, Chicago, IL. https://conservancy.umn.edu/handle/11299/194886

Investigating Sense of Belonging of Graduate Students Using Explanatory Item Response Models

Introduction

Sense of belonging and school climate are posited to be important psychosocial factors that facilitate graduate student success (Fong et al., 2021; Johnson & Strayhorn, 2022). Sense of belonging refers to the perception that an individual is a valued member of and a part of the campus community (Hurtado & Carter, 1997). Within the field of higher education, sense of belonging encompasses perceived support and connectedness on campus, the experience of mattering, feeling cared about, accepted, respected, valued, and the perception that one is important to their campus community (Strayhorn, 2019). For graduate students, perceiving a sense of belonging has been associated with graduate student balancing of life demands, investment in graduate school, relationships with peers, and relationships with faculty (Blakewood Pascale, 2018).

Given the increasing number of graduate student enrollees in higher education within the past four decades (Redden, 2021), it is now more important than ever that researchers, administrators, and practitioners (i.e., stakeholders) understand what contributes to graduate student sense of belonging. For example, the quality of advisor-advisee relationships can play a significant role in facilitating graduate student sense of belonging, and as a result, their success in graduate school (Holloway-Friesen, 2021). It is particularly important that the measures developed to assess noncognitive abilities, such as sense of belonging, have evidence to support their interpretation and use.

The purpose of this study was to address this challenge by using explanatory item response models (EIRM). EIRM can provide insight as to the person-characteristics and item-characteristics that may explain the likelihood of selecting an item response. An advantage of EIRM is being able to incorporate random effects, for instance, the potential variation between research universities. The ability to explain variation at multiple levels of analysis (i.e., items responses within students within institutions) can provide strong empirical evidence for the interpretation and use of a measure.

Investigating Graduate Student Experiences

Although researchers have increasingly published on graduate student success in recent years (Blakewood Pascale, 2018; Trent et al., 2021), research on the graduate student experience remains infrequent and sparse compared to undergraduate student success (Shepard et al., 2022). In an effort to address these issues, the SERU Consortium developed the Graduate Student Experience in the Research University (GradSERU) survey. The GradSERU purpose is to provide decisionmakers with a glimpse of graduate student experiences, the impact of various institutional support systems, and the aspirations and career goals that students are pursuing and achieving.

The GradSERU is administered to multiple institutions in North America, and as a result, may have potential notable variations in graduate student experiences across the different universities (and particularly across departments and fields within an institution). Embedded within the survey are several items that could contribute to a measure sense of belonging. Providing evidence to support the interpretation and use of a sense of belonging measure for graduate students is challenging given the potential variation in graduate student experiences across institutions and program areas.

Validity Evidence

At the core of any educational and psychological measure is the evidence that supports how the derived score is used and for what purpose. The extent to which theory and evidence support the interpretation and use of test scores is more generally referred to as validity (AERA, 2014). In this paper we were guided by Kane's (2013) approach to validity argumentation, that is, a validity argument includes the evaluation of evidence for the interpretation and use of test scores. Although validity may be ever-changing, it is important to assess and disseminate existing evidence that supports the interpretation and use of test scores.

Perhaps the most common use of test scores is at the individual level, that is, to make decisions about an individual given their performance through content standards (expectations of what test takers should know or can do) or a norm-referenced standard. Although this is typical for most achievement-based measures (e.g., math knowledge or verbal reasoning tests), there is little precedence for social and emotional competency measures to be used for decisions at the individual level. For instance, if a program wishes to use scores of sense of belonging to determine a graduate student's standing in their graduate program, there is little theoretical and practical use for such scores. It is crucial in establishing evidence to support interpretation and use, that one considers the stakeholders involved in each step of the measure development process.

Assessment stakeholders are important considerations for any validity argument. In this instance, given the purpose and needs of the GradSERU, key stakeholders include university institutions, their graduate programs and faculty, and most importantly, their graduate students. The influence that community, institutional policies and environment have on an individual's sense of belonging cannot be ignored (Garcia, 2020). Thus, there is a need to gather evidence which supports the use of aggregate scores or group-scores. In this case, since there is a lot of unique variability across institutions and even more variability within their graduate programs, we need to account for this variability in the models.

Statistical Paradigms

In this paper, we utilized quantitative methodologies to investigate the quality of validity evidence for a sense of belonging measure with the existing items of the GradSERU. These methodologies arose from statistical and quantitative work in education research. We explored statistical paradigms that were robust to certain assumptions (e.g., independence, linearity, etc.). First, we established how hierarchical linear models (also known as multilevel or mixed-effects models) account for nested sampling designs. Then, we reviewed how item response theory has incorporated similar designs and approaches from generalized linear models to explain item responses in survey data.

Hierarchical Linear Models

A typical sampling approach in education research is to sample schools and then sample classrooms or students within schools. The resulting data may exhibit a dependency in the data between the nested sampling structure and the variables of interest. The hierarchical linear model (HLM) framework was developed to tackle this methodological issue (Raudenbush & Bryk, 2002). In this framework we can partition the variance due to nested sampling. For instance, in a typical HLM model, we can partition the variance due to individuals within institutions, and variation due to institutions. Furthermore, in an HLM we can model multiple units of measurement that may be attributed to person characteristics (e.g., math achievement scores) or school characteristics (e.g., district median income).

However, one significant trait of this paradigm is the flexibility in the model to incorporate fixed and random effects in the model. The random effects are typically the group-level specific distribution that allows individuals or institutions to vary on the outcome or allows for individual (or group) deviations from the fixed effects (McCormick et al., 2021; Raudenbush & Bryk, 2002; Sulis & Toland, 2017). The flexibility of the HLM paradigm also lends itself to other statistical frameworks such as item response theory (IRT).

Explanatory Item Response Theory

The development of IRT provided a quantitative paradigm that accounts for test taker performance and the extent to which their performance is associated with the latent trait (de Ayala, 2009; Hambleton & Jones, 2005). IRT models are quite flexible in their application to real test data; however, in practice only a few are used. When an item is dichotomously scored the typical IRT models include the 1-, 2-, and 3-parameter logistic models. For polytomously scored items (e.g., partial credit, rating scales, etc.), the typical IRT models include the Master's partial credit, graded response, and rating scale models. Across all IRT models, the general goal is to describe the probability for an examinee with a given ability level, θ, to correctly respond to an item, given a set of item parameters such as item difficulty, discrimination, and lower asymptote. At face-value, these models allow us to estimate person propensities using the item properties (de Boeck & Wilson, 2011; Wilson et al., 2008). This approach can be extended to include person properties and item properties for measuring abilities, skills, attitudes, etc. This approach to IRT is known as the explanatory approach and includes EIRM.

Similar to the HLM paradigm, EIRMs (and IRT models in general) can be thought of as mixed-effects models and are also referred to as generalized linear or nonlinear mixed models. We can think of test data as repeated observations varying within individuals. Furthermore, these models incorporate random components which help define the distribution function of the outcome variable—identical to the random variance components in HLM. The flexibility to incorporate random variance components also allows us to extend beyond the individual level and into classroom and schools for example (Fox, 2005; Sulis & Toland, 2017).

With this in mind, we can include person and item properties, and any relevant interactions between them. To this extent, Wilson and De Boeck (2011) proposed that EIRMs can be doubly descriptive, explanatory or doubly explanatory. A doubly descriptive EIRM is simply a regular IRT model with no item or person predictors. The inclusion of item (e.g., content domain) or person properties (e.g., race or ethnicity) attempts to explain possible variation in the item responses and is known as an explanatory model, and the inclusion of both item and person properties is known as a doubly explanatory model. In education applications, EIRMs have not caught on substantially and they have seen limited use in measures of social and emotional competencies. In one such case, the researchers used a partial credit EIRM to investigate whether the rating scale item responses differed due to item reference (self versus other) for American Indian and Latino high school students, compared to White students (Rodriguez et al., 2018).

Summary and Research Questions

In higher education, the term sense of belonging refers to a student's perception that they are a welcomed and valued member of the campus community. Student experiences at the undergraduate level have frequently been studied; however, there is a lack of research that focuses on graduate student experiences. For graduate students, feeling welcomed and valued in their program or at their institution is a significant factor that impacts student retention, degree completion, and student well-being. Furthermore, variation between students and institutions may explain responses to sociocultural measure such as sense of belonging. Existing methodology can be used to account for such differences and characteristics. The results from such analysis provide rigorous psychometric evidence to evaluate and support the potential uses of such scores. Thus, our research questions were:

- *What is the extent to which variation in sense of belonging item responses is a function of individuals and their institutions?*
- *Is there an interaction effect between major area of study and the estimated category thresholds?*

Method

The data came from the 2021 administration of the GradSERU. The survey provided a way for graduate students to voice their academic, social, and personal experiences across campus. An intended use of the data was to improve graduate programs and curricula, improve student services and policies, and promote positive program and institutional climates. Fifteen R1 institutions across the U.S. participated in the GradSERU survey, including 25,442 students, where 87.3% of participants completed the survey and 90% completed at least half of the survey.

Across participants, 31% of respondents were first year students, 27% were second year, 15% were third year, 11% were fourth year, 9% were fifth year, and 7% were in their sixth year or later. Only students pursuing masters (35% of total) or doctoral (65%) research degrees were invited to participate. The average age for students was 29 years (*SD* = 7.1).

Data Analysis

For the data analysis we employed an extension of the latent regression Rasch model (LRRM) by including an additional random effect due to institutions. We used the *eirm* package in R (Bulut, 2021) to estimate the LRRM. The levels of analyses included item responses (level 1) nested within students (level 2) nested within institutions (level 3). The random variance components captured by the model allowed us to examine the extent to which variation in item responses was due to a design effect (i.e., we partitioned the variance explained in item responses due to variation between students and/or institutions). Pseudo-dichotomous response categories were created to estimate the response category thresholds. The model was used to estimate a difficulty parameter for each category threshold and other item properties. In order to create the pseudo-dichotomies, we used an approach developed by Stanke and Bulut (2019). Essentially, if an observed response was *Disagree*, it was scored as one, a pseudo-response for *Strongly Disagree* was scored as zero, and responses for *Agree* and *Strongly Agree* were scored as *NA*. Table 1 contains the coding scheme for polytomous responses.

Table 1
Pseudo-Indicator Coding Scheme for Category Thresholds

Threshold estimated	Category	Observed response	Score
Strongly disagree—Disagree	Strongly disagree	Disagree	0
	Disagree	Disagree	1
	Agree	Disagree	NA
	Strongly agree	Disagree	NA
Disagree—Agree	Strongly disagree	Agree	NA
	Disagree	Agree	0
	Agree	Agree	1
	Strongly agree	Agree	NA
Agree—Strongly agree	Strongly disagree	Strongly agree	NA
	Disagree	Strongly agree	NA
	Agree	Strongly agree	0
	Strongly agree	Strongly agree	1

Using this scoring scheme, the *eirm* package was used to estimate each threshold on its own based on a set of predictors.

In this case the only predictor included in the model was a categorical variable, major area of study (Major), that was coded to address the main branches of study from a student's reported graduate program. There were eight total major areas of study including: arts, humanities, communication and design business; education; health sciences; other; public safety; social and behavioral sciences and health services; and science, technology, engineering, and mathematics. We used a modification of the explanatory multi-level item response model (Sulis & Toland, 2019) by including an interaction term between Major and the pseudo-dichotomous response in our model to estimate the effect that Major had on the threshold locations.

Analytical Model

Given that the Sense of Belonging items are polytomously scored, we relied on polytomous IRT models for interpretation of the item parameters. Beginning with the partial credit model in Equation 1:

$$P\left(\tau_h \middle| \theta_j, \delta_{hi}\right) = \frac{\exp\left[\sum_{h=0}^{\tau_h}(\theta_j - \delta_{hi})\right]}{1 + \sum_{h=0}^{\tau_h} \exp\left[\sum_{n=0}^{\tau_h}(\theta_j - \delta_{hi})\right]} \qquad (1)$$

where we estimated the log-odds of selecting response category *h* over *h* – 1 on item *i* for person *j*. For easier notational expression, we used exp[z] in place of e^z. The latent trait of person *j* was notated by θ_j; δ_{hi} was the estimated threshold *h* relative to *h* – 1 location for item *i*. We interpreted the estimated item parameter threshold as the location on the scale where there was an equal likelihood of selecting a response category over the previous category.

Although our analysis was based on the interpretation and conceptualization of the PCM, due to computational limitations, we resorted to a simpler ML-EIRM that estimated each threshold as a binary outcome rather than a multinomial outcome. The explanatory multilevel partial credit model from our analysis is shown in Equation 2:

$$\mathrm{P}(\tau_h = 1|\theta, \alpha_i) = \frac{\exp[\alpha_i \mathrm{x}_i + \theta_{0jk} + \theta_{00k}]}{1 + \exp[\alpha_i \mathrm{x}_i + \theta_{0jk} + \theta_{00k}]} \tag{2}$$

where $P(\tau_h = 1)$ was the equal probability of endorsing one category over another in item i for student j in the k^{th} institution as a function of random effects due to student and institution (θ) and item difficulty (α_i).

$$\mathrm{P}(\tau_h = 1|\theta, \alpha_i, \beta_j) = \frac{exp[\alpha_i \mathrm{x}_i + (\sum_{j=1}^{J} \beta_j \mathrm{W}_{ijk} + \theta_{0jk}) + \theta_{00k}]}{1 + exp\left[\alpha_i \mathrm{x}_i + \left(\sum_{j=1}^{J} \beta_j \mathrm{W}_{ijk} + \theta_{0jk}\right) + \theta_{00k}\right]} \tag{3}$$

We then added a fixed effect for major area of study, $\beta_{ijk} W_{ijk}$, to persons, as seen in Equation 3. Since the item intercept in the PCM was fixed at 1, $\alpha_i \mathrm{x}_i$ simplified to α_i. The variability in response selection was partitioned into variance due to items within students (level 1), variance between students (level 2), and variance between institutions (level 3). The responses were linked to an overall item intercept, α_i, a random intercept for students ($\theta_{0jk} \sim N\left[0, \sigma^2_{\theta_j}\right]$), and a random effect for institution ($\theta_{00k} \sim N[0, \sigma^2_{\theta_k}]$). To identify the model we fixed the variance of the random term for students to one. The full model including variables is presented in Equation 4:

$$P(\tau_h = 1|\theta, \alpha_i, \beta_j) = \frac{exp[(\alpha_{000}) + (\sum_{j=1}^{J} \beta_{\mathrm{0jk}} \mathrm{Major}_{0jk} + \theta_{0jk}) + \theta_{00k} + u]}{1 + exp\left[(\alpha_{000} + \tau_{100}) + \left(\sum_{j=1}^{J} \beta_{\mathrm{0jk}} \mathrm{Major}_{0jk} + \theta_{0jk}\right) + \theta_{00k} + u\right]} \tag{4}$$

Model Assumptions

The three model assumptions for a MLIRT model were similar to those of regular IRT models (Sulis & Toland, 2017). First, we assumed that the items were measuring a single latent trait at multiple levels (unidimensionality). Second, we assumed that the estimated latent trait score derived from the responses of students was the result of the latent trait measured and not affected by responses to other items (local independence). Third, we assumed that the data fit the functional form of the model (model specification).

A common approach to assessing the first assumption is to use confirmatory factor analysis to confirm that the items are measuring a single common factor. We conducted CFA using Mplus 8 and found that the 12 items we selected showed modest to adequate fit (RMSEA = .158; CFI = .953; TLI = .942; Brown, 2014). The results of our CFA provided enough evidence to suggest that the 12 items were measuring a single latent trait. The factor loadings for the 12 items ranged from .60 to .91.

Sulis and Toland (2017) suggested using the Q3 index as a tool to assess the conditional independence assumption. The Q3 is a correlation between the residuals for a pair of items. In this case the residual was the difference between the observed response and the predicted response. In this formulation of the model, the observed response of the model was [0,1] instead of [0,1,2,3] since we created pseudo-dichotomies for the response categories.

Model fit was assessed using information criteria and log-likelihood: AIC, BIC, log-likelihood, and deviance. Lower values indicate better model fit. For nested models, we conducted a likelihood ratio test to determine whether the more complex model better fit the data than the simpler model.

Results

We analyzed graduate student responses using a combination of HLM and EIRM to investigate the extent to which variation in item responses was due to differences between students and institutions. Further analysis was done to examine the functioning of ordinal response options given a student's reported major area of study. As a first step we fitted the partial credit explanatory item response model with no predictors ($EIRM_0$) to the data and then included an interaction term between the response category and major ($EIRM_1$). Then, to account for differences between institutions and accommodate the multilevel nature of the data, we fitted another partial credit EIRM to the data and included a random effect for institution ($ML\text{-}EIRM_0$) and then included the interaction term between response category and major ($ML\text{-}EIRM_1$).

Given the complexity of the analytical models used in our analysis, we evaluated model fit to determine if the complexity in our models was warranted. Table 2 contains the model fit statistics for each model. Lower values indicate better model fit. We determined that our most complex model, $ML\text{-}EIRM_1$, had the best fit across all fit indices.

Table 2
Model Fit Statistics

Model	AIC	BIC	logLik	Deviance
$EIRM_0$	400444	400487	-200218	400436
$EIRM_1$	399732	400006	-199841	399682
$ML\text{-}EIRM_0$	400127	400181	-200058	400117
$ML\text{-}EIRM_1$	399465	399749	-199707	399413

Research Question 1

Our first research question concerned the extent to which variation in item responses was due to variation between institutions. We estimated the intraclass correlation (ICC) for each level of analysis as an estimate of the extent to which variation in the outcome variable, in this case item responses, was a function of a higher order variance such as persons or institutions. The ICC for item responses was .12 (accounting for 12% of the variance), the ICC for persons or students was .81, and the ICC for institutions was .01.

Descriptive statistics for student and institution thetas are presented in Table 3.

Table 3
Descriptive Statistics for Person and Institution Thetas

Model	Theta	*M*	*SD*	Min	Max	*n*
$EIRM_0$	θ_j	-0.04	1.80	-5.39	3.46	21,189
$EIRM_1$	θ_j	-0.04	1.78	-5.62	3.59	21,189
$ML\text{-}EIRM_0$	θ_j	0.00	1.78	-5.58	3.91	21,189
	θ_k	-0.05	0.23	-0.64	0.22	15
$ML\text{-}EIRM_1$	θ_j	0.00	1.77	-5.64	4.00	21,189
	θ_k	-0.05	0.22	-0.62	0.17	15

Note. θ_j is for students and θ_k is for institutions. Units are in logits.

The person level thetas, θ_j, in both EIRM models were close to zero, although they were slightly negatively shifted from zero. We expected the average person theta to be zero, as we observed for the person thetas in the ML-EIRM. The institution level thetas, θ_k, were also negatively shifted from zero. We observed much smaller standard deviations in the institution level thetas, typical for group mean scores; the range of institution thetas was much smaller (-0.62 to 0.17 logits). The within institution (i.e. student) thetas had more variation (*SD* = 1.77 logits) and a much larger range (-5.64 to 4.00 logits). Table 4 contains the descriptive statistics for person thetas by major.

Table 4
Descriptive Statistics for Person Thetas by Major Area of Study

Major	*M*	*SD*	Min	Max	*N*
STEM	0.00	1.78	-5.6	3.99	10,943
Business	-0.03	1.83	-5.63	3.43	1,287
AHCD	0.00	1.73	-5.17	3.83	2,193
SBSHS	0.01	1.69	-5.09	4.00	2,795
Public Safety	0.00	1.73	-5.12	3.89	556
Health Sciences	-0.01	1.81	-5.58	3.75	1,402
Education	-0.01	1.78	-5.64	3.58	1,795
Other	-0.01	1.75	-5.26	3.23	218

Note. AHCD = Arts, Humanities, Communication and Design; SBSHS = Social and Behavioral Sciences and Human Services.

Correlations between person theta scores across the four models are presented in Table 5. The correlation between institution theta scores for the multilevel models was *r* = .99. It appeared that model choice had little to no effect on rank ordering of persons.

Table 5
Correlations Between Institution Theta and Person Theta

	EIRM_0	EIRM_1	ML-EIRM_0
EIRM_1	.99		
ML-EIRM_0	.98	.99	
ML-EIRM_1	.99	.98	.99

Research Question 2

Table 6 contains the results from the four models. We relied on the full model, ML-EIRM$_1$, to guide our results.

Table 6

Threshold Estimates for the Polytomous (EIRM$_0$), the Polytomous EIRM with Person Predictors (EIRM$_1$), the Multi-Level Polytomous EIRM (ML-EIRM$_0$) and the ML-EIRM with Person Predictors (ML-EIRM$_1$) Models

	EIRM$_0$	ML-EIRM$_0$	EIRM$_1$	ML-EIRM$_1$			
	α_i	α_i	α_i	α_i	$\alpha_i - \tau_i$	Sig	DIF
Strongly disagree—Disagree (τ_1)	-3.37	-3.41	-3.41	-3.44			
Disagree—Agree (τ_2)	-2.72	-2.76	-2.76	-2.78			
Agree—Strongly agree (τ_3)	0.95	0.91	1.12	1.10			
STEM τ_1			-3.45	-3.49	-0.05		
STEM τ_2			-2.79	-2.83	-0.05	*	
STEM τ_3			1.05	1.05	-0.05	*	
Business τ_1			-3.58	-3.52	0.09		
Business τ_2			-3.16	-3.19	-0.40	*	
Business τ_3			0.44	0.42	-0.68	*	C
AHCD τ_1			-2.97	-3.02	-0.46		
AHCD τ_2			-2.82	-2.85	-0.06		
AHCD τ_3			0.45	0.43	-0.67	*	C
SBSHS τ_1			-3.00	-3.07	-0.42		
SBSHS τ_2			-2.66	-2.69	0.09		
SBSHS τ_3			0.72	0.69	-0.41	*	
Public Safety τ_1			-3.07	-2.99	-0.50		
Public Safety τ_2			-3.06	-3.09	-0.30	*	
Public Safety τ_3			0.50	0.47	-0.63	*	
Health Sciences τ_1			-3.47	-3.60	0.11		
Health Sciences τ_2			-2.84	-2.87	-0.09		
Health Sciences τ_3			0.92	0.91	-0.19	*	
Education τ_1			-3.66	-3.67	0.19		
Education τ_2			-2.60	-2.62	0.16	*	
Education τ_3			0.77	0.76	-0.34	*	

Note. AHCD = Arts, Humanities, Communication and Design; SBSHS = Social and Behavioral Sciences and Human Services. α_i is the threshold location. The DIF (differential item functioning) column indicates C-level DIF.

* $p < .01$ for the interaction terms.

The model parameter estimates were presented in logits and represented item difficulty (α_i) or location for each threshold. The first three rows of the table represent the estimated threshold parameters for each sequential pair of categories: *Strongly Disagree* to *Disagree* (τ_1), *Disagree* to *Agree* (τ_2), and *Agree* to *Strongly Agree* (τ_3). Consistent across the four models, we observed that the distance between the first, second, and third thresholds positively increased in difficulty for all 12 Sense of Belonging items. For the full model (ML-EIRM$_1$), the first threshold was the average first threshold across items for all the Major groups and this indicated the level of Sense of Belonging for students to have a 50% chance to select *Disagree* or *Strongly Disagree* (τ_1 = -3.44). The categories were ordered, since each subsequent threshold was higher than the prior value, indicating the need for a higher level of Sense of Belonging to have a higher likelihood of selecting the next higher response option.

However, the interpretation of the main effects was perhaps not tenable due to significant interaction effects in the full model. Because of the significant interaction effects, the item threshold locations may have depended on major area of study. For the full ML-EIRM$_1$ model, we used STEM as the reference group. The estimate for STEM τ_1 in Table 6 (α_i = -3.49) represented the first threshold location for this group.

For the interactions between Major and the first threshold, none of them were statitstically significant ($p < .01$). For the interactions between Major and the second threshold, half of the effects were not significant at $p < .01$. For these groups, the item difficulty location may have been dependent on their Major. Interestingly, almost all of the interaction effects between Major and and the third threshold were significant; the only interaction effect that was not significant at $p < .01$ was for students that were coded as Other for their major.

Our second research question was about the impact that Major may have on item difficulty and threshold location. We used the ETS guidelines for interpreting DIF where a difference ±0.64 logits between the reference and focal group was flagged as C-level DIF, a level requiring close examination of the item. The reference group for Major was STEM. In Table 6, we examined the interaction term of the category threshold and Major and considered this as an indicator for detecting overall DIF within the EIRM framework. In the full model, we only observed significant interaction terms for the third threshold and Business and ACHD, with values greater than 0.64, both negative, suggesting that for these majors it required less Sense of Belonging to have an equal likelihood of endorsing *Agree* and *Strongly Agree*. And these two C-level DIF cases were just slightly over the 0.64 threshold.

Discussion

Sense of belonging is important for graduate student success in higher education. Sense of belonging is also an understudied construct in higher education, and extant measures of sense of belonging have sparse psychometric support. In this paper we used a 12-item measure from an existing survey as a measure of sense of belonging. That is, although the survey was not developed specifically for measuring sense of belonging, we found theoretical and empirical support that these items captured what we theorized to be sense of belonging.

Thus, we investigated the impact of nested data on item responses for a Sense of Belonging measure. It is meaningful to account for the nested structure of the data to address potential biases in the measurement qualities of Sense of Belong. The development of HLM provides the tools necessary for modeling nested data, such that EIRMs can be seen as extensions of a mixed-effects logistic regression model, or as it is sometimes referred to as a

Generalized Linear (or Non-Linear) Mixed-Effects Model. Through extending EIRMs, we were able to answer our first research question regarding variation in item responses due to person and institutions.

We observed that the average category thresholds aligned with the ordinal structure of the item responses (i.e., lower theta values were associated with selecting either *Strongly Disagree* or *Disagree*). Furthermore, the variation in item responses due to differences between institutions was very small (ICC = ~ 1.0 to 1.5%). This result is not surprising as we would expect that differences between persons to account for most of the variation in social and emotional competency based items.

Our second research question was about the extent to which a person's major area of study may impact the location of overall category thresholds. For this we used a person variable, major area of study, to examine invariance of threshold locations. A two-way interaction effect between threshold location and major identified that less than half of the estimated interaction effects were significant, and only two had practical significance. The descriptive statistics for person theta scores by Major did not meaningfully differ from each other and all had comparable standard deviations. Of those interaction effects, two of them showed logit differences between the reference group (STEM) and focal group greater than 0.64. Furthermore, the order of thresholds for the Public Safety group was not aligned with the rating scale response options, indicating disorder. The second threshold was located slightly lower on the logit scale than the first threshold when we should anticipate the order of thresholds to be in logit order; however, we acknowledge that this group is the smallest Major and the differences in disordered thresholds were very small (negligible). For the most part, item functioning does not depend on student's Major. Additionally, including Major as a covariate in our model reduced the ICC for institutions by 0.5% and reduced the ICC for persons by 0.6%.

Limitations

Our findings, although consistent with what we would expect, are not without limitations. First, the models we estimated only account for the overall (item average) category threshold as we did not estimate the overall difficulty location for each item. To do this we would have to include an interaction term between the pseudo-dichotomized polycategory response variable and each item. This would be computationally taxing with little success for model convergence. However, future researchers could benefit from such analysis and we encourage methodological development in this arena. Second, we are slightly limited in our approach to estimating polytomous items. In this paper we created pseudo-dichotomies for each category threshold, similar to the Rasch model or 1PL IRT model. There is further evidence needed to compare model estimations and efficiency between the pseudo-dichotomy approach presented in this paper and a fully polytomous approach (e.g., specifying a multinomial distribution rather than a binomial distribution as a link function). Finally, we are limited in this approach to DIF. Typically, DIF analysis is a pairwise comparison between the focal and reference groups. In the ML-EIRM$_1$ model, we only used STEM as the reference group, which limits the information regarding DIF for all possible comparisons among groups. That is, we would expect the reference group to change with each comparison in order to fully capture the logit difference across category thresholds. One solution to this, albeit time consuming, is to change the reference group and estimate models separately.

Takeaway

Despite technical and computational limitations, the information provided from this study can be meaningful in large-scale data assessment to make decisions regarding institutional policies and graduate programs. The variation between institutions was not meaningful (ICC = ~1%). Although it appears that institutions have very similar levels of Sense of Belong on average, the vast majority of variance lies within institutions where the measure may be more beneficial, as intended. Furthermore, since we found evidence that a student's major area of study has little to no impact on item threshold locations, the measure can be used with relative assurance of the consistency of scale meaning across Majors. Additional evidence could be gathered to support the use of this measure for comparing groups within institutions on other student characteristics (e.g., international versus domestic students, or master's and doctoral students).

There is more work to be done to explore the psychometric properties for measures of social and emotional competencies. The results of our work helps lay the foundation for future multilevel, psychometric analysis of such data.

References

American Educational Research Association, National Council on Educational Measurement, & American Psychological Association. (2014). *Standards for educational and psychological testing*. American Educational Research Association. https://www.testingstandards.net/

Andrich, D. (1978). A rating formulation for ordered response categories. *Psychometrika*, *43*(4), 561–573. https://doi.org/10.1007/BF02293814

Andrich, D. (2015). The problem with the step metaphor for polytomous models for ordinal assessments. *Educational Measurement: Issues and Practice*, *34*(2), 8–14. https://doi.org/10.1111/emip.12074

Bacci, S., & Caviezel, V. (2011). Multilevel IRT models for the university teaching evaluation. *Journal of Applied Statistics*, *38*(12), 2775–2791. https://doi.org/10.1080/02664763.2011.570316

Blakewood Pascale, A. (2018). "Co-existing lives": Understanding and facilitating graduate student sense of belonging. *Journal of Student Affairs Research and Practice, 55,* 399-411. https://doi.org/10.1080/19496591.2018.1474758

Boeck, P. D., Bakker, M., Zwitser, R., Nivard, M., Hofman, A., Tuerlinckx, F., & Partchev, I. (2011). The estimation of item response models with the lmer function from the **lme4** package in *R*. *Journal of Statistical Software*, *39*(12). https://doi.org/10.18637/jss.v039.i12

Briggs, D. C. (2008). Using explanatory item response models to analyze group differences in science achievement. *Applied Measurement in Education*, *21*(2), 89–118. https://doi.org/10.1080/08957340801926086

Brown, T. A. (2014). *Confirmatory factor analysis for applied research* (2nd ed.). Guilford.

Bryson, T. C., & Grunert Kowalske, M. (2022). Black women in STEM graduate programs: The advisor selection process and the perception of the advisor/advisee relationship. *Journal of Diversity in Higher Education, 15,* 111-123. https://doi.org/10.1037/dhe0000330

Bulut, O. (2021). *eirm: Explanatory item response modeling for dichotomous and polytomous item responses*, R package version 0.4. https://CRAN.R-project.org/package=eirm

Bulut, O., Gorgun, G., & Yildirim-Erbasli, S. N. (2021). Estimating explanatory extensions of dichotomous and polytomous Rasch models: The eirm package in R. *Psych*, *3*(3), 308–321. https://doi.org/10.3390/psych3030023

De Boeck, P., & Wilson, M. (2011). *Explanatory item response models: A generalized linear and nonlinear approach*. Springer.

Fong, C. J., Owens, S. L., Segovia, J., Hoff, M. A., & Alejandro, A. J. (2021). Indigenous cultural development and academic achievement of tribal community college students: Mediating roles of sense of belonging and support for student success. *Journal of Diversity in Higher Education.* https://doi.org/10.1037/dhe0000370

Fox, J.-P. (2005). Multilevel IRT using dichotomous and polytomous response data. *British Journal of Mathematical and Statistical Psychology*, *58*(1), 145–172. https://doi.org/10.1348/000711005X38951

Garcia, C. E. (2020). Belonging in a predominantly White institution: The role of membership in Latina/o sororities and fraternities. *Journal of Diversity in Higher Education, 13*(2), 181-193. https://doi.org/10.1037/dhe0000126

George Mwangi, C. A., Changamire, N., & Mosselson, J. (2019). An intersectional understanding of African international graduate students' experiences in U.S. higher education. *Journal of Diversity in Higher Education, 12,* 52-64. https://doi.org/10.1037/dhe0000076

Goldberg, A. E., Kuvalanka, K., & dickey, l. (2019). Transgender graduate students' experiences in higher education: A mixed-methods exploratory study. *Journal of Diversity in Higher Education, 12,* 38-51. https://doi.org/10.1037/dhe0000074

Hambleton, R. K., & Jones, R. W. (2005). An NCME instructional module on: Comparison of classical test theory and item response theory and their applications to test development. *Educational Measurement: Issues and Practice*, *12*(3), 38–47. https://doi.org/10.1111/j.1745-3992.1993.tb00543.x

Holloway-Friesen, H. (2021). The role of mentoring on Hispanic graduate students' sense of belonging and academic self-efficacy. *Journal of Hispanic Higher Education, 20*(1), 46-58. https://doi.org/10.1177/1538192718823716

Hurtado, S., & Carter, D. F. (1997). Effects of college transition and perceptions of the campus racial climate on Latino college students' sense of belonging. *Sociology of Education, 70,* 324-345. https://doi.org/10.2307/2673270

Johnson, R. M., & Strayhorn, T. L. (2022). Examining race and racism in Black men doctoral student socialization: A critical race mixed methods analysis. *Journal of Diversity in Higher Education.* https://doi.org/10.1037/dhe0000420

Kane, M. T. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement*, *50*(1), 1–73. https://doi.org/10.1111/jedm.12000

Lewis, J. A., Mendenhall, R., Ojiemwen, A., Thomas, M., Riopelle, C., Harwood, S. A., & Huntt, M. B. (2021). Racial microaggressions and sense of belonging at a historically White university. *American Behavioral Scientist, 65,* 1049-1071. https://doi.org/10.1177/0002764219859613

Muthén, L. .K., & Muthén, B. O. (2012). *Mplus* (Version 7) [Software program]. https://www.statmodel.com/

Raudenbush, S. W., & Bryk, S. (2002). *Hierarchical linear models: Applications and data analysis methods* (2nd ed.). SAGE.

Redden, E. (2021, October 17). U.S. graduate enrollments grew in 2020. *Inside Higher Ed*. https://www.insidehighered.com/news/2021/10/18/graduate-enrollment-grew-2020-despite-pandemic

Rodriguez, M. C. (2018, July 11). A psychometric perspective on SEL assessment [Blog post]. *Measuring SEL: Using data to inspire practice*. CASEL Measuring SEL Network. http://measuringsel.casel.org/psychometric-perspective-sel-assessment/

Rodriguez, M. C., Bulut, O., Vue, K., & Cabrera, J. (2018). Response processes in noncognitive measures: Validity evidence from explanatory item response modeling. University of Minnesota Digital Conservancy. https://hdl.handle.net/11299/195226

Rodriguez, S. L., Perez, R. J., & Schulz, J. M. (2022). How STEM lab settings influence graduate school socialization and climate for students of color. *Journal of Diversity in Higher Education, 15,* 58-72. https://doi.org/10.1037/dhe0000361

Rosseel, Y., Jorgensen, T. D., Rockwood, N., Oberski, D., Byrnes, J., Vanbrabant, L., Savalei, V., Merkle, E., Hallquist, M., Rhemtulla, M., Katsikatsou, M., Barendse, M., Scharf, F., & Du, H. (2022). *lavaan: Latent variable analysis* (Version 0.6-12). https://lavaan.ugent.be

Shepard, V., Perry, A., & Hall-Hertel, K. (2022). *Administrators in graduate and professional student services: Considering the current state of graduate and professional student affairs*. https://www.naspa.org/files/dmfile/2022-NASPAKC-Final-Rev.pdf#page=7

Stanke, L., & Bulut, O. (2019). Explanatory item response models for polytomous item responses. *International Journal of Assessment Tools in Education*, 259–278. https://doi.org/10.21449/ijate.515085

Sulis, I., & Toland, M. D. (2017). Introduction to multilevel item response theory analysis: Descriptive and explanatory models. *The Journal of Early Adolescence*, *37*(1), 85–128. https://doi.org/10.1177/0272431616642328

Trent, F., Dwiwardani, C., & Page, C. (2021). Factors impacting the retention of students of color in graduate programs: A qualitative study. *Training and Education in Professional Psychology, 15,* 219-229. https://doi.org/10.1037/tep0000319

Wilson, M., De Boeck, P., & Carstensen, C. (2008). Explanatory item response models: A brief introduction. In J. Hartig, E. Klieme, & D. Leutner (Eds.), *Assessment of Competencies in Educational Contexts* (pp. 91-120). Hogrefe & Huber Publishers.

Yang, C., Bear, G. G., & May, H. (2018). Multilevel associations between school-wide social-emotional learning approach and student engagement. *School Psychology Review*, *47*(1), 18. https://files.eric.ed.gov/fulltext/EJ1173199.pdf

Appendix

Items included in the Sense of Belonging Measure

Faculty respect students regardless of their background
Students respect other students regardless of their background
Faculty encourage expression of diverse viewpoints from their students
There are open lines of communication between students and faculty regarding student needs, concerns, and suggestions
Students are given an active role in departmental decisions that affect them
My program creates a collegial and supportive environment
There is a sense of solidarity among students
I belong in my graduate/professional program
Faculty members in my graduate/professional program are available to talk with me
Faculty members in my graduate/professional program give me positive reinforcement for my accomplishments
Faculty members in my graduate/professional program treat me fairly
Overall, the environment or climate in my program is positive and welcoming

Note. Response options for each item ranged from *Strongly Disagree, Disagree, Agree,* to *Strongly Agree*.