Applying Item Response Theory to Examine Extreme Survey Response Style

Xiaolin Wang

Amy K. Ribera

Robert M. Gonyea

Center for Postsecondary Education

Indiana University, Bloomington

Authors' Note

This paper is a work in progress. Please do not cite without the permission of the authors. Paper presented at the annual meeting of the American Education Research Association, San Antonio Texas, April 27, 2017. Direct correspondence concerning this paper to Xiaolin Wang (xw41@indiana.edu) located at 1900 E. 10th Street, Eigenmann Hall Suite 419, Bloomington, IN 47406-7512.

Applying Item Response Theory to Examine Extreme Survey Response Style

Abstract

Response style effect is a well-known survey limitation. By applying a generalized item response theory (IRT) model to the Global Perspective Inventory data from the 2014 National Survey of Student Engagement (NSSE), this study provides estimates of college students' extreme response style (ERS) tendency. Furthermore, findings reveal significant group differences in ERS tendency by two student characteristics—first-generation status and major choice (STEM vs non-STEM).

Keywords: Extreme response style, survey methodology, higher education

Applying Item Response Theory to Examine Extreme Survey Response Style

Messick (1989) defines validity as an integrative evaluative judgment of the degree to which empirical evidence and theoretical rationales support the appropriateness of inferences and actions based on scores or other modes of assessment. One of the many methods to build a case for the validity is by examining the stylistic tendencies of survey respondents also known as *response styles* (Cronbach, 1946). According to Adcock and Collier (2001), response style effects could lead to bias in analysis by inflating or deflating respondents’ survey scores and potentially threaten measurement validity. To further investigate this issue in higher education assessment, we analyze the responses of 22,450 senior college students who participated to the National Survey of Student Engagement in 2014. We isolated our analysis to 71 four-year institutions that opted to administer the module item set, Global Learning Inventory. This study examines the extreme response styles of these students in order to reveal group differences by individual factors such as gender, race, first-generation status, sexual orientation, disability status, major choice, and enrollment status. Furthermore, we argue the advantages of applying an item response theory framework to examine ERS as opposed to the traditional classical test theory framework.

Relevant Literature

Survey respondents have known to follow many types of response styles such as the tendency to unconditionally agree with the items and the tendency to select middle points on a Likert scale (Baumgartner & Steenkamp, 2001). Among the different response styles, the tendency to select endpoints on a Likert scale is known as *extreme response style* (ERS; Greenleaf, 1992). For example, on a seven-point Likert scale ranging from “strongly disagree” to “strongly agree,” ERS refers to the tendency to select the endpoints, “strongly disagree” or “7=strongly agree,” as opposed to the middle points ranging from “disagree” to “agree”. Response styles such as ERS could contaminate group comparison results and lead to incorrect conclusions. With the existence of ERS, group score differences are mixtures of both true response difference and response style difference, which threatens survey validity since it is measuring more than the construct of interest. Since responses from surveys are commonly used for group comparisons, an understanding of whether ERS exists and if it potentially contaminates group comparison results would be very meaningful.

ERS and Demographic

To further explore threats of validity on survey data, a line of research has begun to investigate ERS differences across demographic groups (Bachman & O’Mally, 1984; Chen, Lee, & Stevenson, 1995; Hui & Triandis, 1989; Gilman et al., 2008). In the past researchers have simply counted the number or the percent of extreme responses on a survey as a quantification of ERS. For example, based on the number of extreme responses throughout the survey, Bachman and O’Malley (1984) found that African-American adolescents were more likely to select endpoints than the white adolescents. Hui and Triandis (1989) used percentage of extreme responses throughout surveys as operational quantifications of ERS tendency and compared the ERS tendencies between Hispanic and non-Hispanic groups utilizing questionnaire data collected by the U.S. Navy recruit stations. The study results showed that the Hispanics exhibited a stronger ERS tendency when the items were on a 5-point Likert scale. Chen, Lee, and Stevenson (1995) also used the percentage of extreme responses as a measure of ERS tendency.

They studied the effects of age, education level, gender, and household income by fitting a regression model to survey data collected from a large sample of U.S. adults serving on a customer panel. The model fitting results suggested that age, education level, and household income were significantly related with ERS tendency while gender was not. Gilman et al. (2008) compared the percent of extreme responses on the Adolescent Multidimensional Life Satisfaction Report scales across four nations: US, Ireland, China, and South Korea. Results showed that there were significant differences in ERS tendencies across nations on all scales. Specifically, Irish respondents generally showed the highest ERS tendencies while Korean respondents showed the lowest ERS tendencies. Gilman et al. further conducted MANOVA analysis exploring gender and gender-by-nation effects for ERS. Results showed that none of the two factors significantly affect ERS.

Yet, these studies applied the classical test theory (CTT) framework and used either number of extreme responses or percent of extreme responses as quantifications of ERS, which has its limitations. According to Podsakoff et al. (2003), ERS is the product of the interaction between characteristics of the respondent and survey items. Respondents have different tendencies to select endpoints and items elicit ERS to different degrees. The number or percentage of extreme responses based on the CTT framework does not separate respondent and item effects. As a result, researchers have turned to modern psychometric techniques such as item response theory (IRT) in order to study response styles (Bolt & Johnson, 2009; Bolt & Newton, 2011; Jin & Wang, 2014; Johnson, 2003).

Advantage of Item Response Theory

IRT is a set of probability model that describe the relationship between the item response with item and person characteristics (Embreston & Reise, 2013). One advantage of IRT over CTT is that IRT separates item and person parameters so that both individuals' ERS tendencies and items' ERS elicitation degrees could be quantified. Several IRT models have been proposed for response style research (Bolt & Johnson, 2009; Bolt & Newton, 2011; Jin & Wang, 2014; Johnson, 2003). One of the models was a generalized IRT-ERS model developed by Jin and Wang (2014). The generalized IRT-ERS model is able to provide both purified estimates of construct of interest and respondents' ERS tendencies. Modified based on the partial credit model (Masters, 1982), the generalized IRT-ERS model is written as:

$$\log(\frac{P_{nij}}{P_{ni(j-1)}}) = \mathrm{a}_\mathrm{i}\ [\theta_\mathrm{n} - (\delta_\mathrm{i} + \omega_\mathrm{n}\,\tau_\mathrm{ij})],$$

where P_{nij} is the probability of the *n*-th respondent selecting category *j* on item *i*, a_i is the discrimination parameter of item *I* which indicates how well the item differentiate respondents of high level of the construct of interest and respondents of low level of the construct of interest, θ_n is the estimate of the construct of interest for the *n*-th respondent, δ_i is the overall location parameter of item *i*, τ_{ij} is the relative location of selecting category *j* on item *i*, and ω_n is a weight parameter of the *n*-th respondent on the relative difficulties. A greater ω_n implies smaller log ratio between adjacent score categories, which makes it harder to select endpoints (i.e., smaller ERS tendency). θ_n, δ_i, and τ_{ij} are all assumed to follow normal distributions. a_i is assumed to follow a log normal distribution and ω_n is assumed to follow a log-normal distribution with mean zero and variance σ_ω^2.

Purpose

In the current study, the generalized IRT-ERS model will fit to data from a national higher education survey in order to examine the respondents' ERS tendencies. Further, ERS tendencies will be compared across groups for eight demographic factors: gender, enrollment status, international status, first-generation status, race, majoring in STEM , sexual orientation, and disability status. Since most of the ERS comparisons in previous studies were between cultures, and only a few studies were within the realm of education, not to mention in higher education surveys, our study will greatly contribute to response style research in higher education.

Methods

Data

In this study, the data used consisted of the responses from 22,450 college seniors who participated in the 2014 administration of the National Survey of Student Engagement (NSSE). The analysis was limited to a set of 21 items measuring the cognitive and social elements of students' global perspective (GPI). The items asked about their experiences with global learning and views on intercultural understanding. An example of the GPI items is "I understand the reasons and causes of conflict among nations of different cultures." All of the GPI items are on a 5-point Likert scale ranging from "1 = Strongly Disagree" to "5 = Strongly Agree" showing how respondents agree with the statements. Table 1 shows the demographic composition of the respondents receiving the GPI items.

Analyses

Model Fit. To estimate parameters in complex models which are not accommodated in available software, Markov chain Monte Carlo (MCMC) estimation method is a popular and dependable alternative, which is gaining increased popularity in educational research (Curtis, 2010; Fox, 2010). MCMC method is a process of generating random samples from theoretical multivariate distributions of parameters. It is gaining popularity partly due to the availability of multiple software and packages to implement the process and the straightforwardness and ease of the implementation. MCMC could be implemented in *WinBUGS*, *JAGS*, and *rjags*, etc. based on the Bayesian posterior distributions of parameters (i.e., product of prior beliefs of the parameter distributions and likelihood function) derived automatically by the software and packages.

We fitted the complex generalized IRT-ERS model described above to the students' responses to the GPI items. The model fitting was accomplished by the *rjags* R package in R software (R Core Team, 2016). Since MCMC estimation is very time-consuming, we randomly sampled data from 3,000 cases. The prior distributions for the parameters in the model were specified as: $N(0, 1)$ for θ_n, log-normal (0, 1) for a_i, $N(0, 1)$ for δ_i and τ_ij, and Gamma (1, .1) for σ_ω^2. ω_n is the parameter of our interests. The MCMC estimation was set to include 3 chains and 15,000 iterations with 5,000 burn-in iterations.

Next, we checked the MCMC convergence by examining the trace plots of the parameters. Because of the large number of parameters, we only show trace plots of several example parameters in Figure 1. The trace plots are quite stable, suggesting convergence of the chains and stability of parameter estimation. In addition, the Gelman-Rubin's diagnostic plots were produced (but not displayed here) and the shrink factors for a large number of selected parameters were close to 1, which again suggests the chain convergence (Sinharay, 2003).

After the model fitting, the means of the parameter values across the latter 10,000 MCMC iterations were adopted as the parameter estimates, which are called the expected a priori (EAP) estimates. The ω estimates, which are of our research interest, for several selected example response patterns are shown in Table 2. As shown in Table 2, generally, response patterns consisting of larger number of endpoints (i.e., 1 or 5) are related with smaller ω estimates (i.e., higher ERS tendency), as expected.

Group Comparison. The next step was to apply either a t-test or an ANOVA to compare the estimated ω across demographic groups. The eight demographic factors considered in this study included:

1. gender (0 = female, 1 = male),
2. international status (0 = no, 1 = yes),
3. full-time enrollment (0 = no, 1 = yes),
4. having a disability (0 = no, 1 = yes),
5. majoring in a STEM field (0 = no, 1 = yes),
6. first generation status (0 = no, 1 = yes),
7. race (1 = Asian, Native Hawaiian, or other Pacific Islander, 2 = Black or African American, 3 = Hispanic or Latino, 4 = White, 5 = American Indian, Alaska Native, Other, Multiracial, 6 = I prefer not to respond), and
8. sexual orientation (1 = heterosexual, 2 = gay/lesbian, 3 = bisexual, 4 = questioning or unsure, 6 = I prefer not to respond).

Among the eight demographic factors, sex orientation was recoded so that the category of “Another sexual orientation” was treated as missing. The logic was that, the frequency of this category was low and it included a mixture of all other sexual orientations. We conducted t-tests to compare mean ω estimates between groups for gender, international status, full-time enrollment, having a disability, majoring in a STEM field, and first generation status and conducted ANOVA analyses for race and sexual orientation.

Results

Table 2 shows the ERS group mean comparison results for the eight demographic factors. Results in Table 2 suggest significant mean ω differences between STEM major and first-generation status (p<.05). Specifically, STEM major students had significantly ($t_{(1245)}$ = 2.18, *p* = .03) smaller mean ω (1.73) than non-STEM major students (2.05), which implied a greater ERS tendency of STEM major students than non-STEM students. Similarly, non-first-generation students had significantly ($t_{(1452)}$ = 3.623, *p* = .00) smaller mean ω (1.73) than first-generation major students (2.31), which implied a greater ERS tendency of non-first-generation students than the first-generation students.

While there was no significant difference in ERS tendency for the other demographic factors, the group comparison results could provide a general picture of the ERS tendency in each group. Specifically, female respondents had a slightly lower mean ω (1.89) than male respondents (1.91); non-international students had a lower mean ω (1.96) than international students; full-time students had a lower mean ω (1.88) than non-full-time students; students with disabilities had a lower mean ω (1.71) than students without disabilities; Black or African American students had the lowest mean ω (1.63) among the race groups while “I prefer not to respond” and Hispanic or Latino had the highest mean ω’s (2.51 and 2.30 respectively); and gay/lesbian students had the lowest mean ω (1.29) while “I prefer not to respond” and “bisexual” had the highest mean ω’s (3.18 and 2.22 respectively). The comparisons in the mean ω’s suggested that female, non-international students, full-time students, students with disabilities, Black or African, and gay/lesbian students had higher ERS tendencies than their comparison groups.

Discussion

Extreme response style effects threaten self-report survey validity. Previous literature (Bachman & O’Mally, 1984; Chen, Lee, & Stevenson, 1995; Hui & Triandis, 1989; Gilman et al., 2008), which adopted the classical test theory framework and quantified extreme response style by the number/percent of extreme responses, were not able to separate item and person effects. Thus, we applied a generalized IRT-ERS model to explore the extreme response style effects in a higher education assessment tool and investigate whether the extreme response style tendencies differ across demographic groups.

Based on the t-test and ANOVA analysis results about the mean ERS tendency group difference for eight demographic factors (i.e., gender, enrollment status, international status, first generation status, race, STEM major, sexual orientation, and disability status), we found significant (p<.05) ERS tendency difference for two demographic factors: STEM major and first-generation status. Specifically, STEM students and non-first generation students were more likely to select endpoints.

While this study reveals some interesting findings, we noted several limitations in our study. First, the sample sizes were unbalanced for different demographic groups. For example, the percent of international students in the sample was only 5.7% and the rest 94.3% were American students. It is unknown to what extent the unbalanced sample sizes affect our group comparison results. Further, the present study only examines one set of items on the survey (i.e., GPI). Although it is generally assumed that extreme response styles are not related with item contents, future studies may analyze data from other content areas to make sure it is unrelated with students' ERS tendencies. Additionally, only one IRT model was applied in this study; whether all models yield the same results needs to be checked in later studies. Moreover, the study results indicate ERS tendency difference across demographic groups. Future research investigating reasons (e.g., personality characteristics) why ERS differs across groups is warranted. Another future study could compare respondents' ERS tendencies for more demographic factors. Also, researchers might utilize the respondents' purified estimates of the construct of interest (i.e., purified GPI score in this study) and examine how analysis based on the purified scores differs from those based on the raw scores.

Conclusion

In closing, we found ERS tendencies could vary significantly between demographic groups. We recommend when evaluating the validity of scores, researchers should consider assessing survey response style effects through an IRT lens. It is our recommendation that researchers should examine if different demographic groups show different ERS tendencies in order to better interpret the data. The issue of ERS is especially important when the survey results are utilized for policy and high-stakes decision making. If ERS differences are ignored, conclusions based on survey results could be misleading.

References

Adcock, R. (2001, September). Measurement validity: A shared standard for qualitative and quantitative research. In *American Political Science Association* (Vol. 95, No. 03, pp. 529-546). Cambridge University Press.

Bachman, J. G., & O'Malley, P. M. (1986). Self-concepts, self-esteem, and educational experiences: The frog pond revisited (again). *Journal of Personality and Social Psychology*, *50*(1), 35.

Bolt, D. M., & Johnson, T. R. (2009). Addressing score bias and differential item functioning due to individual differences in response style. *Applied Psychological Measurement*.

Bolt, D. M., & Newton, J. R. (2011). Multiscale measurement of extreme response style. *Educational and Psychological Measurement*, *71*(5), 814-833.

Cronbach, L. J. (1946). Response sets and test validity. *Educational and psychological measurement, 6*(4), 475-494.

Chen, C., Lee, S. Y., & Stevenson, H. W. (1995). Response style and cross-cultural comparisons of rating scales among East Asian and North American students. *Psychological Science*, 170-175.

Embretson, S. E., & Reise, S. P. (2013). *Item response theory*. Psychology Press.

Greenleaf, E. A. (1992). Measuring extreme response style. *Public Opinion Quarterly*, *56*(3), 328-351.

Hui, C. H., & Triandis, H. C. (1989). Effects of culture and response format on extreme response style. *Journal of Cross-Cultural Psychology*, *20*(3), 296-309.

Jin, K. Y., & Wang, W. C. (2014). Generalized IRT models for extreme response style. *Educational and Psychological Measurement*, *74*(1), 116-138.

Johnson, T. R. (2003). On the use of heterogeneous thresholds ordinal regression models to account for individual differences in response style. *Psychometrika*, *68*(4), 563-583.

Messick, S. (1989). Validity. In R.L. Linn (Ed.), *Educational measurements*, (3rd ed., pp. 13-103). New York: Macmillian.

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika*, *47*(2), 149-174.

Team, R. C. (2013). R: A language and environment for statistical computing.

Table 1

Sample Demographic Composition

Variable	Category	N	%
Gender	Female	14,277	63.6
	Male	8,173	36.4
	Total	22,450	100
International	No	16,761	94.3
	Yes	1,009	5.7
	Total	17,770	100.0
First Generation	No	10,517	58.7
	Yes	7,413	41.3
	Total	17,930	100.0
Full-time	No	3,402	15.2
	Yes	19,048	84.8
	Total	22,450	100.0
Having a	No	15,847	88.7
	Yes	1,479	8.3
	I prefer not to respond	547	3.1
	Total	17,873	100.0
STEM Major	No	13,610	76.0
	Yes	4,307	24.0
	Total	17,917	100.0
Race	Asian, Native Hawaiian, or Other Pacific Islander	1,307	7.3
	Black or African American	1220	6.8
	Hispanic or Latino	850	4.7
	White	12,381	69.1
	American Indian, Alaska	1,347	7.5
	I prefer not to respond	818	4.6
	Total	17,923	100.0
Sexual Orientation	Heterosexual	8,650	87.5
	Gay	163	1.6
	Lesbian	92	.9
	Bisexual	224	2.3
	Another sexual orientation	72	.7
	Questioning or unsure	83	.8
	I prefer not to respond	605	6.1
	Total	9,889	100.0

Table 2

Example Response Patterns and Parameter Estimates

Response Pattern	# of endpoints	*ω*	theta
32522 55555 52545 225155	13	.26	.77
45511 31422 24432 334112	7	.53	-1.46
32425 43443 42444 445245	3	1.18	.42
34512 44444 42444 324244	2	1.68	.76
24423 44444 42444 424244	0	2.12	1.11

Note. Lower ω means greater mean ERS tendency.

Figure 1

Trace Plots and Density Plots for Selected Parameters

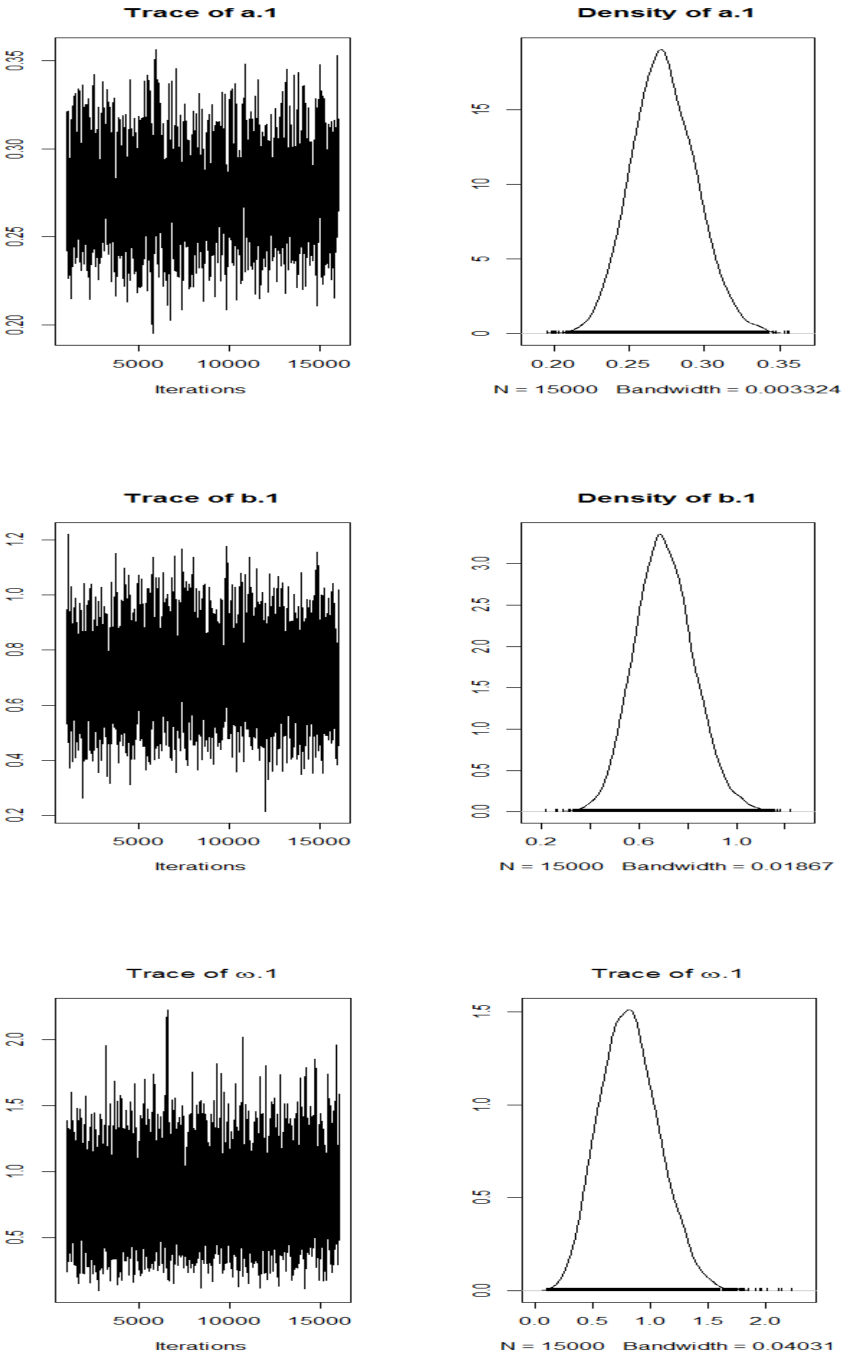

Figure 1 (continue)

Trace Plots and Density Plots for Selected Parameters

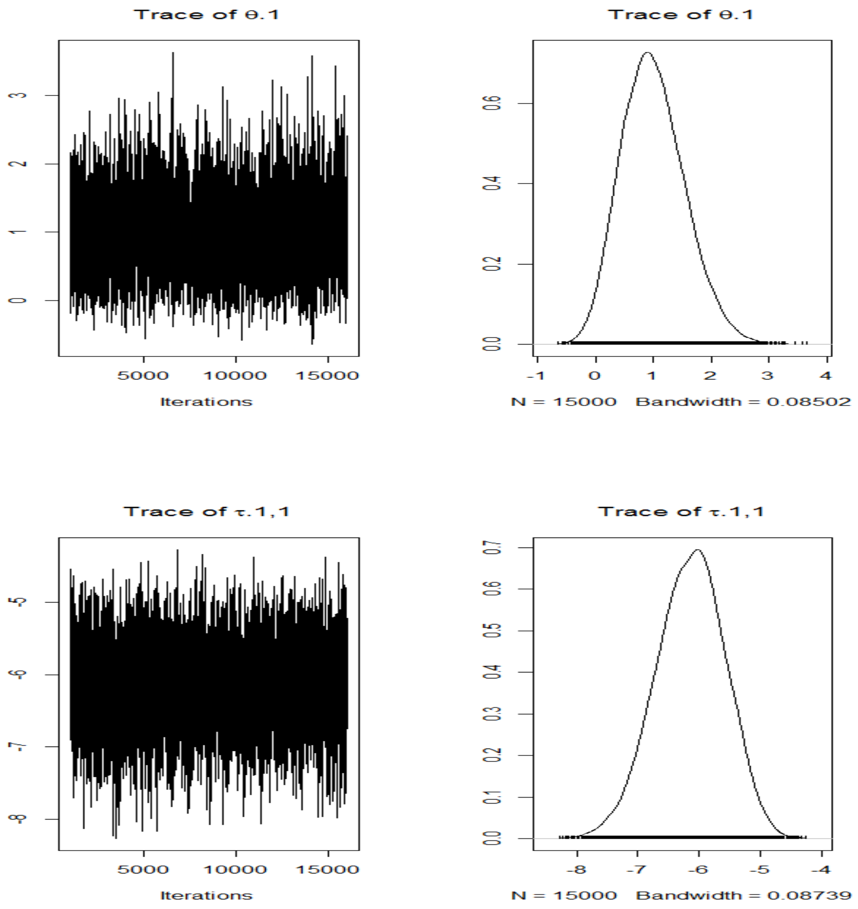

Table 3

ERS Group Mean Comparison Results

Variable	Category	Mean ω	t/F	p
Gender	Female	1.89	.16	.87
	Male	1.91		
International Status	No	1.96	.93	.36
	Yes	2.34		
Full-time Enrollment	No	1.97	.54	.59
	Yes	1.88		
First Generation Status	No	1.74	3.62	***.00***
	Yes	2.31		
Having a Disability	No	1.99	1.12	.26
	Yes	1.71		
STEM Major	No	2.05	2.18	***.03***
	Yes	1.73		
Race	Asian, Native Hawaiian, or Other Pacific Islander	2.12	1.23	.29
	Black or African	1.63		
	Hispanic or Latino	2.30		
	White	1.96		
	American Indian, Alaska Native, Other, Multiracial	1.72		
	I prefer not to respond	2.51		
Sexual Orientation	Heterosexual	2.01	2.29	.06
	Gay/Lesbian	1.29		
	Bisexual	2.22		
	Questioning or unsure	1.84		
	Prefer not to respond	3.18		

Note. Lower ω means greater mean ERS tendency.