Supplemental Material

CBE—Life Sciences Education

Hess *et al.*

Supplemental online materials for:

Title: *Virtually the same? Evaluating the effectiveness of remote undergraduate research experiences*

Authors: Riley A. Hess, Olivia A. Erickson, Rebecca B. Cole, Jared M. Isaacs, Silvia Alvarez-Clare, Jonathan Arnold, Allison Augustus-Wallace, Joseph C. Ayoob, Alan Berkowitz, Janet Branchaw, Kevin R. Burgio, Charles H. Cannon, Ruben Michael Ceballos, C. Sarah Cohen, Hilary Coller, Jane Disney, Van A. Doze, Margaret J. Eggers, Edwin L. Ferguson, Jeffrey J. Gray, Jean T. Greenberg, Alexander Hoffmann, Danielle Jensen-Ryan, Robert M. Kao, Alex C. Keene, Johanna E. Kowalko, Steven A. Lopez, Camille Mathis, Mona Minkara, Courtney J. Murren, Mary Jo Ondrechen, Patricia Ordoñez, Anne Osano, Elizabeth Padilla-Crespo, Soubantika Palchoudhury, Hong Qin, Juan Ramírez-Lugo, Jennifer Reithel, Colin A. Shaw, Amber Smith, Rosemary Smith, Fern Tsien, Erin L. Dolan*

* Corresponding author. Email: eldolan@uga.edu

This supplement contains the following:

Measures and Assessment of Measurement Models

As noted in the main manuscript, we used several fit indices to assess how adequately our CFA models reproduced their variance-covariance matrices. First, we report a chi squared test (χ^2) for each model (Kline, 2015). Chi square is highly sensitive to misfit because it has strong assumptions, including that there is no kurtosis in the data, which is a measure of the "tailedness" of the probability distribution of a real-valued random variable (Kline, 2015). However, a significant chi square indicates misfit to some degree (Credé & Harms, 2019), and it is best practice to report it. We assessed goodness of fit using equivalence testing (Marcoulides & Yuan, 2017; Peugh & Feldon, 2020; Yuan et al., 2016). This approach has been recommended over traditional null hypothesis testing. In traditional null hypothesis testing, test statistics like RMSEA and CFI indicate that there is not enough evidence to reject the null hypothesis, but do not indicate that the data do not fit the model (Yuan et al., 2016). Equivalence testing yields adjusted, or "T-size," fit statistics, RMSEA_T and CFI_T, which allow researchers to compare the amount of misspecification in their model to a tolerable size of specification with adjusted cutoffs. To calculate RMSEA_T, CFI_T, and adjusted cutoffs, we used the R code made available in (Marcoulides & Yuan, 2017). We report adjusted cutoffs in footnotes. We supplemented evaluation of our measurement models by interpreting factor loadings and coefficient omega (Ω) values (Bandalos, 2018; Dunn et al., 2014). We report details for each scale below.

Scientific Self-Efficacy. The scientific self-efficacy scale demonstrated high internal reliability (Ω=0.91, 95% CI [0.88-0.93]). However, fit of the model was poor based on the high RMSEA value with adjusted cutoffs[1], $\chi^2(27)=124.364$ ($p<0.001$), $\text{RMSEA}_T=0.150$, $\text{CFI}_T=0.913$. To identify the source of misfit, we checked for correlated residuals using modification indices as a guide. The modification index (MI) is the chi-squared value by which model fit would improve if an additional path were added to the model (Roiger, 2020). Thus, larger values indicate larger model improvements. We began by correlating the residuals of the item pair with the highest modification index value. We continued to specify additional correlated residuals until our model fit indices reached their adjusted cutoffs. Modification indices recommended allowing the residual values of item 3 and 4 to correlate (MI=46.09), then items 8 and 9 (MI=22.503), then items 4 and 5 (MI=16.55). After adding these paths to the model, all modified fit indices reached acceptable adjusted cutoff values[2], $\chi^2(24)=60.494$ ($p<0.001$), $\text{RMSEA}_T=0.102$, $\text{CFI}_T=0.934$.

Scientific Identity. The scientific identity scale demonstrated high internal reliability (Ω=0.87, 95% CI [0.83-0.89]). However, RMSEA_T and CFI_T indicated poor model fit based on adjusted cutoffs[3], $\chi^2(14)=193.156$ ($p<0.001$), $\text{RMSEA}_T=0.271$, $\text{CFI}_T=0.522$. We examined modification indices, which recommended allowing the residual values of items 6 and 7 (MI=176.58) and items 1 and 3 (MI=27.05) to

[1] Excellent, close, fair, and mediocre model fit would have been attained by $\text{RMSEA}_T < 0.053$, 0.078, 0.105, 0.124 and $\text{CFI}_T >$ 0.957, 0.896, 0.853, 0.825, respectively.

[2] Excellent, close, fair, and mediocre model fit would have been attained by $\text{RMSEA}_T < 0.055$, 0.080, 0.107, 0.126 and $\text{CFI}_T >$ 0.957, 0.895, 0.852, 0.824, respectively.

[3] Excellent, close, fair, and mediocre model fit would have been attained by $\text{RMSEA}_T < 0.065$, 0.089, 0.116, 0.135 and $\text{CFI}_T >$ 0.951, 0.885, 0.839, 0.809, respectively.

covary. After adding these paths to the model, fit improved substantially, and RMSEA$_T$ and CFI $_T$ met the criteria for fair model fit[4] χ^2(12)=20.379 (p<0.001), RMSEA$_T$ =0.097, CFI $_T$=0.943.

Values Alignment. The values alignment scale demonstrated high internal reliability (Ω=0.81, 95% CI [0.75-0.85]). In addition, RMSEA$_T$ and CFI $_T$ indicated excellent model fit based on adjusted cutoffs[5], χ^2 (2)=0.625(p<0.01), RMSEA$_T$ =0.105, CFI $_T$=0.95. Thus, we moved forward with a one-factor model containing four items measuring values alignment.

Benefits and Costs. All scales we used to measure values, which we refer to collectively as benefits and costs, demonstrated high internal reliability: intrinsic value (Ω=0.88, 95% CI [0.85-0.91]), personal importance (Ω=0.79), 95% CI [0.73-0.83], social utility (Ω=0.74, 95% CI [0.65-0.80]), job utility (Ω=0.85, 95% CI [0.79-0.90]), life utility (Ω=0.80, 95% CI [0.74-0.84]), and costs [Ω=0.86, 95% CI [0.83-0.89])]. Given the potentially close relationships among these variables, we fit a single CFA with all six scales as separate factors. Before conducting CFAs, the costs measure was reverse-scored to match the direction of all other measures in the scale (i.e., higher rating means student perceived lower costs). Costs were not reverse-scored for any substantive analyses (i.e., higher rating means student perceived higher costs). Overall, loadings were higher than the recommended minimum value of 0.40 (Bandalos, 2018), ranging from 0.469 to 0.951. However, the RMSEA$_T$ demonstrated fair fit and CFI$_T$ indicated poor fit according to adjusted fit values[6] χ^2(174)=528.181 (p<0.001), RMSEA$_T$ =0.106, CFI $_T$=0.775. Poor CFI values are often indicative of miss-specified factor loadings (Hu & Bentler, 1999). In examining the factor loadings, we noticed that the six intrinsic value items appeared to represent two different dimensions. The first three items refer to enjoyment of research (e.g., "Research is fun for me") and the last three items are more value-oriented (e.g., "Performing well in research is important to me"). In addition, factor loadings were stronger for the first three items (0.91, 0.95, 0.87) than for the later three items (0.60, 0.57, 0.47). These results suggested that the intrinsic value factor may be better represented as two factors: enjoyment and intrinsic value. Indeed, when we split this factor in two, factor loadings for the three value-oriented items increased substantially (0.78, 0.89, 0.77), as did model fit[6] χ^2 (194)=644.326 (p<0.001), RMSEA$_T$ =0.061, CFI$_T$=0.899.

[4] Excellent, close, fair, and mediocre model fit would have been attained by RMSEA$_T$ < 0.069, 0.092, 0.12, 0.139 and CFI$_T$ > 0.951, 0.884, 0.838, 0.808, respectively.

[5] Excellent, close, fair, and mediocre model fit would have been attained by RMSEA$_T$ < 0.13, 0.148, 0.174, 0.193 and CFI$_T$ > 0.931, 0.846, 0.787, 0.751, respectively.

[6] Excellent, close, fair, and mediocre model fit would have been attained by RMSEA$_T$ < 0.033, 0.061, 0.089, 0.108 and CFI$_T$ > 0.969, 0.920, 0.884, 0.860, respectively.

Factor Loadings from Measurement Models

We present loadings from confirmatory factor analysis at Time 1 (before the URE) and Time 2 (after the URE) to gain insight into whether students are interpreting the items similarly at both timepoints. It is noteworthy that factor loadings improve from pre- to post-URE for scientific identity, values alignment, intrinsic 1, intrinsic 2, personal importance, cost, social utility, job utility, and life utility.

Table S1. Scientific Self-Efficacy Items and Factor Loadings. This measure of Scientific Self-Efficacy includes 7 published items from (Chemers et al., 2011) and (Estrada et al., 2011). Items 2 and 6 were authored based on input from the directors of the URE programs included in this study to capture the forms of scientific self-efficacy students would develop during their remote UREs. Response options were: Not confident (1), A little confident (2), Somewhat confident (3), Confident (4), Very confident (5), Extremely confident (6), and I prefer not to respond.

Please indicate how confident you are in your ability to…

Item	Content	Time 1 Factor Loadings	Time 2 Factor Loadings
1	Use technical skills (lab or field equipment, instruments, and/or bench or field techniques).	0.66	0.49
2	Use computational skills (software, algorithms, and/or quantitative techniques).	0.45	0.47
3	Generate a research question to answer.	0.87	0.89
4	Develop a hypothesis to test.	0.84	0.90
5	Figure out what data/observations to collect and how to collect them.	0.81	0.81
6	Trouble-shoot an investigation or experiment.	0.78	0.68
7	Create explanations for the results of the study.	0.81	0.76
8	Use scientific literature and/or reports to guide research.	0.68	0.72
9	Develop theories (integrate and coordinate results from multiple studies).	0.78	0.75

Table S2. Scientific Identity Items and Factor Loadings. This measure of Scientific Identity includes 7 published items from (Chemers et al., 2011) and (Estrada et al., 2011). Response options were: Strongly disagree (1), Moderately disagree (2), Slightly agree (3), Moderately agree (4), Mostly agree (5), Strongly agree (6), and I prefer not to respond (7). Response options were positively packed to avoid a ceiling effect (Brown, 2004).

Please indicate the extent to which you agree with the following statements.

Item	Content	Time 1 Factor Loadings	Time 2 Factor Loadings
1	I have a strong sense of belonging to the community of scientists.	0.52	0.67
2	I derive great personal satisfaction from working on a team of scientists.	0.59	0.75
3	I think of myself as a scientist.	0.62	0.78
4	The daily work of a scientist is appealing to me.	0.64	0.78
5	I feel like I belong in the field of science.	0.66	0.85
6	In general, being a scientist is an important part of my self-image.	0.86	0.86
7	Being a scientist is an important reflection of who I am.	0.88	0.88

Table S3. Values Alignment Items and Factor Loadings. We used a measure of values alignment borrowed from (Estrada et al., 2011). The structure of the measure was based off the Portrait Value Questionnaire (Schwartz et al., 2001). Response options were: Not like me (1), A little like me (2), Somewhat like me (3), Like me (4), Very much like me (5), Extremely like me (6), and I prefer not to respond (7). Response options were positively packed to avoid a ceiling effect (Brown, 2004).

How much is the person in the following descriptions like you?

Item	Content	Time 1 Factor Loadings	Time 2 Factor Loadings
1	A person who thinks it is valuable to conduct research that builds the world's scientific knowledge.	0.70	0.86
2	A person who feels discovering something new in the sciences is thrilling.	0.82	0.88
3	A person who thinks discussing new theories and ideas between scientists is important.	0.78	0.80
4	A person who thinks that scientific research can solve many of today's world challenges.	0.60	0.69

Table S4. Benefits and Costs Confirmatory Factor Analysis Factor Loadings. Our measure of benefits and costs is made up of seven published measures adapted from Gaspard et al. (2015). Enjoyment and Intrinsic Value were adapted from the published 6-item intrinsic value measure. Personal Importance was adapted from the 3-item personal importance measure. Utility value items, which include three factors (social, job, and life utility) was adapted from the measure of utility value. Finally, costs were adapted from the 3-item cost scale. For all measures, response options were: Strongly disagree (1), Moderately disagree (2), Slightly agree (3), Moderately agree (4), Mostly agree (5), Strongly agree (6), and I prefer not to respond (7). Response options were positively weighted to avoid a ceiling effect.

Please indicate the extent to which you agree with the following statements.

Factor	Item	Content	Time 1 Factor Loadings	Time 2 Factor Loadings
Enjoyment	1	Research is fun to me.	0.91	0.97
	2	I like doing research.	0.95	0.99
	3	I enjoy dealing with research topics.	0.87	0.92
Intrinsic Value	4	It is important to me to be good at research.	0.59	0.94
	5	Being good at research means a lot to me.	0.57	0.97
	6	Performing well in research is important to me.	0.47	0.89
Personal Importance	7	I care a lot about remembering things I learn when conducting research.	0.67	0.65
	8	I'm really keen on learning a lot about research.	0.73	0.81
	9	Research is very important to me personally.	0.84	0.90
Social Utility	10	Being well versed in research will prepare me to help my community.	0.69	0.81
	11	I can do good in the world based on my knowledge of research.	0.71	0.80
	12	If I know a lot about research, I can make a difference in the world.	0.67	0.84
Job Utility	13	Doing well in research will improve my chances of finding a job after college.	0.77	0.80
	14	The skills I develop in research will help me be successful in my career.	0.80	0.82
	15	Learning how to conduct research is worthwhile because it improves my career prospects.	0.84	0.85
Life Utility	16	Research will help me in life.	0.77	0.87
	17	I will often need research in my life.	0.81	0.85
	18	Research comes in handy in everyday life.	0.63	0.74
	19	I have to give up other activities that I like to be successful at research.	0.80	0.90
Costs	20	I have to give up a lot to do well in research.	0.95	0.93
	21	I'd have to sacrifice a lot of free time to be good at research.	0.74	0.81

Note. Enjoyment was formerly a part of the intrinsic value measure.

Table S5. Factor Correlations from the Benefits and Costs Confirmatory Factor Analysis

	Enjoyment	Intrinsic Value	Personal Importance	Social Utility	Job Utility	Life Utility	Cost
Enjoyment							
Intrinsic Value	0.592						
Personal Importance	0.813	0.786					
Social Utility	0.281	0.364	0.472				
Job Utility	0.172	0.466	0.393	0.451			
Life Utility	0.468	0.715	0.739	0.605	0.620		
Cost	-0.021	0.082	0.140	0.101	0.173	0.227	

Note. Enjoyment and personal importance were kept separate due to substantive differences in how they changed post URE.

Measurement Invariance Results

We followed the procedures outlined by (Meredith, 1993) to test our measures for invariance between Time 1 and Time 2. First, we tested the CFA models for configural invariance, which tests the hypothesis that the same general pattern of factor loadings holds across timepoints. This supports the claim that the same items are associated with the same factors across timepoints. Overall, we found most factor loadings increased slightly from Time 1 to Time 2 but maintained the same pattern of relations to their latent constructs. Thus, all measures passed the test of configural invariance (Little, 2013).

Next, we tested the measures for weak factorial invariance by constraining corresponding factor loadings to be equal across timepoints. This supports the hypothesis that the direction and strength of the relationship between the indicator and factor are the same across timepoints. If the model passed the test of *weak* factorial invariance, we tested it for *strong* factorial invariance by constraining the factor loadings and intercepts to be equal across timepoints. Strong factorial invariance supports the hypothesis that the intercept of each items' regression on the latent variable is invariant across timepoints. We evaluated invariance hypotheses by comparing the CFI and RMSEA values of the constrained models to the baseline models. To avoid sensitivity to sample size in determining factorial invariance, Chen (2007) recommends change cutoffs for factor loading invariance of ≤ -0.005 for CFI and ≥ 0.010 for RMSEA. A decrease in CFI and an increase in RMSEA indicate worsening model fit. Values alignment and benefits and costs both passed tests of weak and strong factorial invariance (Chen, 2007). Scientific self-efficacy and scientific identity pass the weak factorial tests but fail the strong factorial invariance tests. This invariance limits our interpretation of latent change between groups because it suggests that items are not operating similarly across timepoints.

Table S6. Measurement Invariance Results for Study Measures

Measure	Invariance Model	DF	AIC	BIC	CFI (ΔCFI)	ΔCFI	RMSEA	ΔRMSEA	χ^2	Δ χ^2	Δ DF	$p^{\Delta \chi^2}$
Scientific Self-Efficacy	Baseline	48	10170	10416	0.963	--	0.094	--	143.53	--	--	--
	Weak Factorial	56	10164	10377	0.962	-0.001	0.088	-0.006	153.02	9.50	8	0.302
	Strong Factorial	72	10378	10378	0.946	**-0.017**	0.099	0.005	203.43	59.90	16	0.000
Scientific Identity	Baseline	24	7974	8163	0.984	--	0.077	--	55.87	--	--	--
	Weak Factorial	30	7964	8129	0.986	0.002	0.065	-0.012	58.33	2.46	6	0.873
	Strong Factorial	36	7996	8136	0.967	**-0.017**	0.090	**0.013**	101.99	46.12	12	0.000
Values Alignment	Baseline	4	3894	3993	1.000	--	0.000	--	3.86	--	--	--
	Weak Factorial	7	3889	3975	1.000	0.000	0.000	0.000	4.77	0.91	3	0.824
	Strong Factorial	10	3890	3964	0.997	-0.003	0.033	0.003	12.41	8.54	6	0.201
Benefits and Costs	Baseline	336	19764	20452	0.944	--	0.073	--	731.92	--	--	--
	Weak Factorial	350	19762	20393	0.942	-0.002	0.072	-0.001	758.20	26.28	14	0.024
	Strong Factorial	364	19749	20323	0.942	-0.002	0.071	-0.002	773.42	41.50	28	0.048

Note. Change (Δ) values are relative to the baseline model. Change in CFI ≤ -0.005 and/or change in RMSEA ≥ 0.010 indicates a failure of the strong factorial invariance test. Criteria indicating failure to pass the invariance test are bolded.

Test for Regression to the Mean

Regression to the mean, or the tendency for extreme scores to shift closer to the mean value over time, presents a threat to the meaningfulness of significant findings for studies with only two timepoints. To address this concern, we compared the change we saw in scientific self-efficacy, scientific identity, and values alignment to simulated change data with the same parameters. The simulated dataset allows us to observe the amount of change we would expect by chance, which we compare to the amount of change we observed in our study data. Thus, highly similar datasets indicate strong regression to the mean.

We followed the data simulation procedure outlined in (Furrow, 2019). Specifically, for each of the constructs, we simulated 1,000 observations using the observed correlation between time 1 and time 2, the mean value of the measure, the number points used in the scale, and the sample size. Simulated time 1 and time 2 values were drawn from a binomial distribution that maintained the same correlations and mean values as the original dataset. Next, we created a column of change scores between the simulated time 1 and time 2 values. Likewise, we calculated changed scores for observations in our empirical dataset.

We present the simulated and actual mean change values for scientific self-efficacy, scientific identity, and values alignment in the violin plots below. Quartile 1 (Q1) includes observations with the lowest starting values of scientific self-efficacy, scientific identity, or values alignment, and Quartile 4 (Q4) represents those with the highest starting values. Note that the Y-axis on the simulated change plots is truncated compared to the Y-axis on the actual change plots. The red dots represent the mean value in each quartile.

As reported in the main manuscript, students with lower starting values of scientific self-efficacy, scientific identity, and values alignment experienced greater increases between timepoints compared to students with high starting values. Overall, the magnitude of change for all three variables in the empirical dataset was far greater than in the simulated dataset. This suggests that the change we observed empirically is greater than what we would expect if change were only due to regression to the mean.

Figure S1. Actual (left) vs. simulated (right) change in **scientific self-efficacy** based on starting quartile.

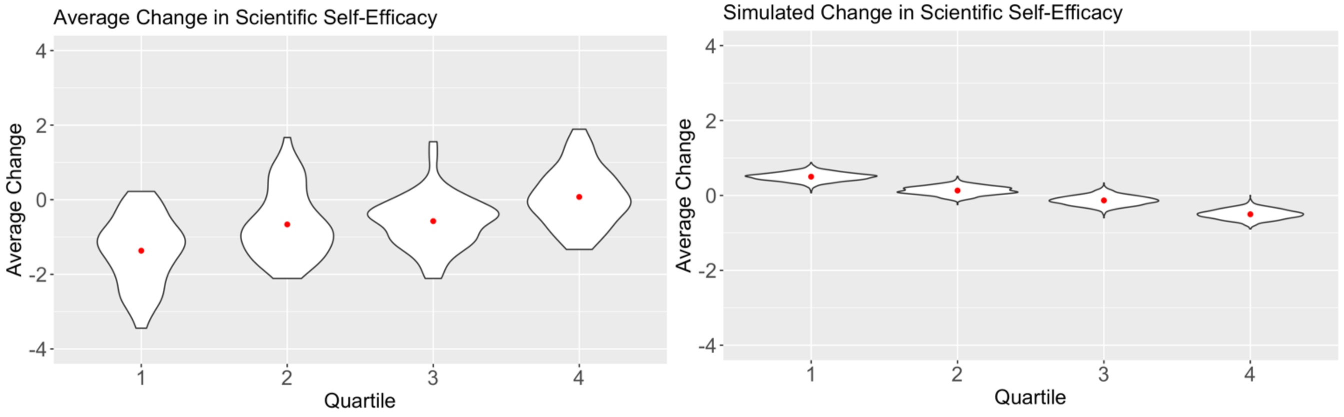


Figure S2. Actual (left) vs. simulated (right) change in **scientific identity** based on starting quartile.

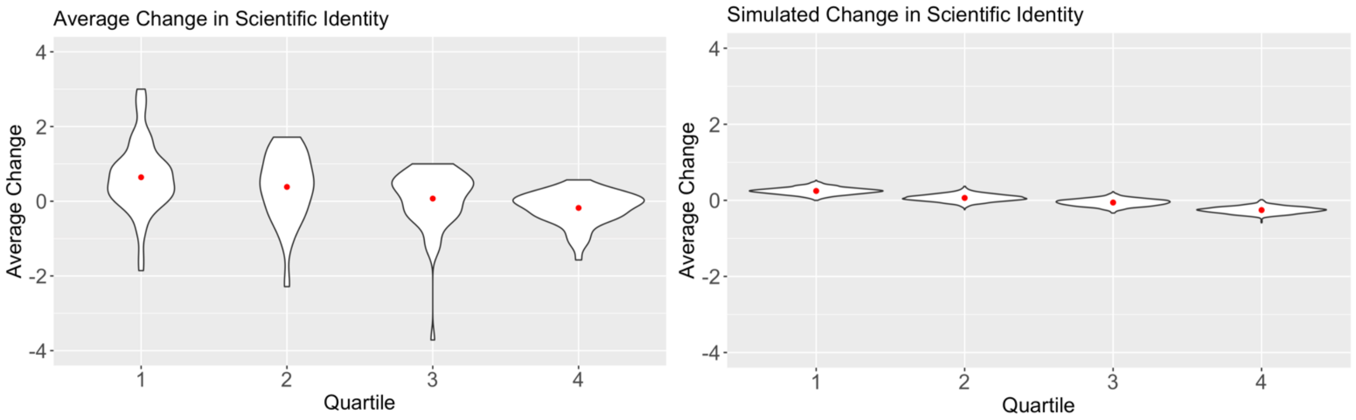
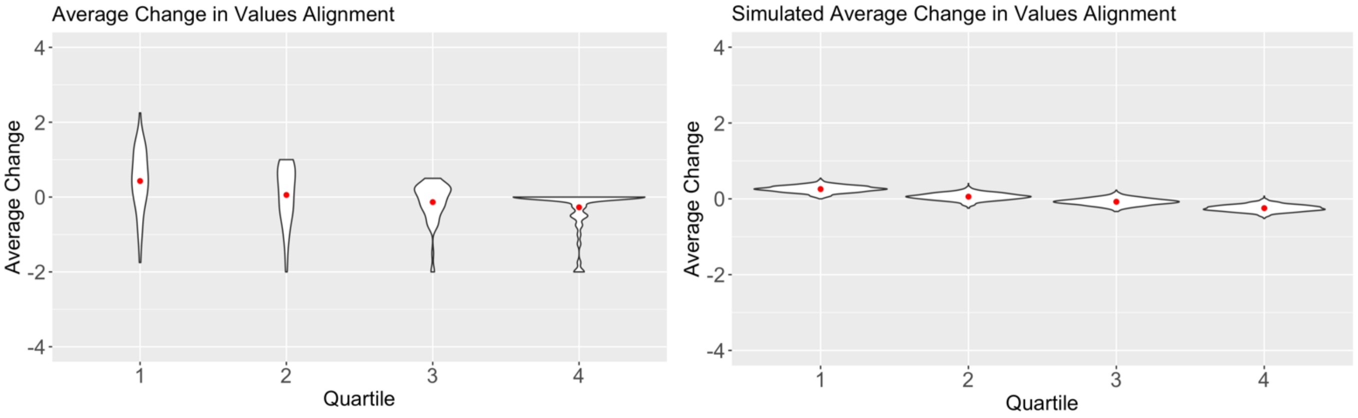


Figure S3. Actual (left) vs. simulated (right) change in **values alignment** based on starting quartile.



References

Bandalos, D. L. (2018). *Measurement Theory and Applications for the Social Sciences* (1 edition). The Guilford Press.

Brown, G. T. L. (2004). Measuring Attitude with Positively Packed Self-Report Ratings: Comparison of Agreement and Fr equency Scales. *Psychological Reports*, *94*(3), 1015–1024. https://doi.org/10.2466/pr0.94.3.1015-1024

Chemers, M. M., Zurbriggen, E. L., Syed, M., Goza, B. K., & Bearman, S. (2011). The Role of Efficacy and Identity in Science Career Commitment Among Underrepresented Minority Students. *Journal of Social Issues*, *67*(3), 469–491. https://doi.org/10.1111/j.1540-4560.2011.01710.x

Chen, F. F. (2007). Sensitivity of goodness of fit indexes to lack of measurement invariance. *Structural Equation Modeling: a Multidisciplinary Journal*, *14*(3), 464-504.

Credé, M., & Harms, P. (2019). Questionable research practices when using confirmatory factor analysis. *Journal of Managerial Psychology*, *34*(1), 18–30. https://doi.org/10.1108/JMP-06-2018-0272

Dunn, T. J., Baguley, T., & Brunsden, V. (2014). From alpha to omega: A practical solution to the pervasive problem of internal consistency estimation. *British Journal of Psychology*, *105*(3), 399–412.

Estrada, M., Woodcock, A., Hernandez, P. R., & Schultz, W. P. (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

Furrow, R. E. (2019). Regression to the Mean in Pre–Post Testing: Using Simulations and Permutations to Develop Null Expectations. *CBE—Life Sciences Education*, *18*(2), le2.

Gaspard, H., Dicke, A.-L., Flunger, B., Schreier, B., Häfner, I., Trautwein, U., & Nagengast, B. (2015). More value through greater differentiation: Gender differences in value beliefs about math. *Journal of Educational Psychology*, *107*(3), 663.

Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal*, *6*(1), 1–55. https://doi.org/10.1080/10705519909540118

Kline, R. B. (2015). *Principles and Practice of Structural Equation Modeling*. Guilford Publications.

Little, T. D. (2013). *Longitudinal structural equation modeling*. Guilford press.

Marcoulides, K. M., & Yuan, K.-H. (2017). New ways to evaluate goodness of fit: A note on using equivalence testing to assess structural equation models. *Structural Equation Modeling: A Multidisciplinary Journal*, *24*(1), 148–153.

Meredith, W. (1993). Measurement invariance, factor analysis and factorial invariance. *Psychometrika*, *58*(4), 525–543.

Peugh, J., & Feldon, D. F. (2020). "How well does your structural equation model fit your data?": Is Marcoulides and Yuan's equivalence test the answer? *CBE—Life Sciences Education*, *19*(3), es5.

Roiger, R. J. (2020). *Just Enough R!: An Interactive Approach to Machine Learning and Analytics*. CRC Press.

Schwartz, S. H., Melech, G., Lehmann, A., Burgess, S., Harris, M., & Owens, V. (2001). Extending the Cross-Cultural Validity of the Theory of Basic Human Values with a Different Method of Measurement. *Journal of Cross-Cultural Psychology*, *32*(5), 519–542. https://doi.org/10.1177/0022022101032005001

Yuan, K.-H., Chan, W., Marcoulides, G. A., & Bentler, P. M. (2016). Assessing structural equation models by equivalence testing with adjusted fit indexes. *Structural Equation Modeling: A Multidisciplinary Journal*, *23*(3), 319–330.