Center for Assessment Summer Internship 2021

Evaluating the Efficacy of Multiple Imputation Methods for Missing Educational Assessment and Growth Data

Allie Cooperman

June 4, 2021

The Pandemic's Impact on Academic Achievement

- Researchers and policymakers are starting to examine the impact of the COVID-19 pandemic on student learning.
- "Learning loss" analyses will require new and innovative methods for evaluating educational assessment data (e.g., Ho, 2021).
- An overarching question for these analyses concerns the extent to which we can appropriately compare 2021 test scores to those from before the pandemic.

The Missing Data Problem

- One potential roadblock to generating valid skip-year comparisons is anticipated missingness in the 2021 data.
- Factors like differential rates of participation and “opt-out” testing can introduce non-ignorable missingness patterns.
- Can we create “adjusted” test scores for 2021 that allow researchers and policymakers to adequately understand students’ learning trajectories?

Multiple Imputation

Multiple imputation (MI) uses information from the observed data to generate model parameter estimates through three steps:

- **Imputation**: A prediction model generates a set of plausible values for the missing observations, resulting in M imputed data sets.
- **Analysis**: The analysis (e.g., regression, student growth percentiles) is conducted on each of the M data sets.
- **Pooling**: Parameter estimates are constructed by pooling across the M analyses.

In the context of learning loss analyses, researchers may implement MI to estimate mean scale score or student growth percentile (SGP) values.

Enders, 2010; Fox & Weisberg, 2018; van Buuren, 2018

Simulation Overview: Procedure

- Observations were amputed from a simulated data set (available in the *SGPdata* R package; Betebenner et al., 2021).
- Missingness types:
  - Missing completely at random (MCAR)
  - Missing at random based on status and demographics (MAR Demog)
  - Missing at random based on status and growth (MAR Growth)
- Either 30%, 50%, or 70% of the data were simulated as missing.
- MI was then used to generate "adjusted" mean scale scores and student growth percentiles.

Simulation Overview: MI Methods

Six MI methods were examined using the *mice* R package (van Buuren & Groothuis-Oudshoorn, 2011):

- Cross-sectional multi-level modeling with the *pan* package (L2PAN; Zhao & Schafer, 2018)
- Cross-sectional multi-level modeling with the *lmer* function (L2LMER; Bates et al., 2015)
- Longitudinal multi-level modeling with *pan* (L2PAN_LONG)
- Longitudinal multi-level modeling with *lmer* (L2LMER_LONG)
- Quantile regression (RQ)
- Predictive mean matching (PMM)

These methods were also compared to when no imputation was implemented (i.e., "Observed").

Simulation Overview: Evaluation

Percent Bias

- Calculated as $\left|\frac{\text{Raw Bias}}{\text{True Value}}\right| \times 100$
- Ideally less than 5% (Miri et al., 2020; Qi et al., 2010)

Confidence Interval (CI) Coverage Rate

- Calculated as the proportion of times that the simplified CI (Vink & van Buuren, 2014) contains the true value
- Ideally as close to $1 - \alpha$ as possible (Demirtas, 2004; Qi et al., 2010)

Simplified F_1 Statistic

- Tests the null hypothesis that the true and imputed values are equivalent
- The *p*-value should ideally be greater than α (van Buuren, 2018)

MI Method Comparison: Scale Scores

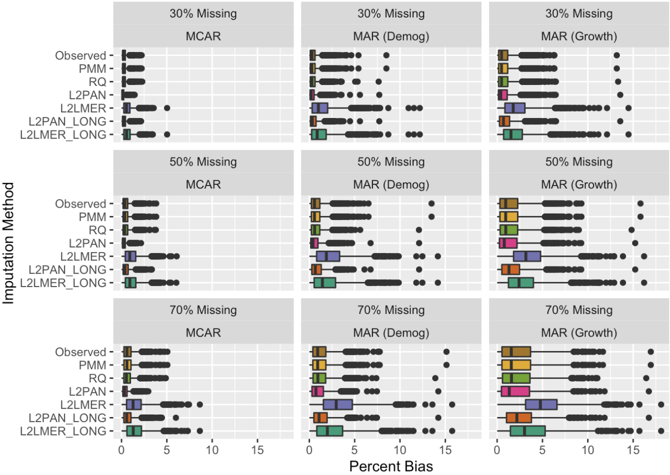

Figure 1: Scale score percent bias by imputation method, missingness percentage, and missingness type

MI Method Comparison: Scale Scores

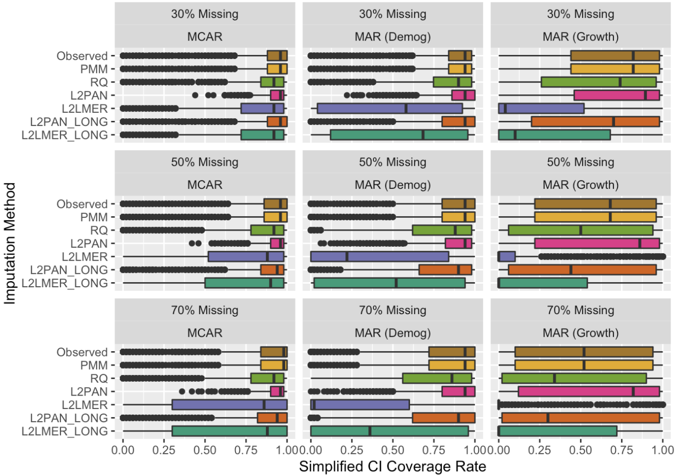

Figure 2: Scale score coverage rate by imputation method, missingness percentage, and missingness type

MI Method Comparison: Scale Scores

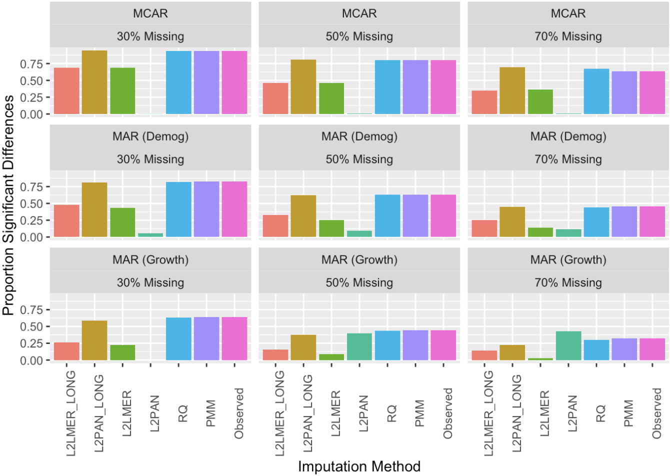

Figure 3: Proportion of times that the imputed scale score differed from the true value based on the simplified F1 statistic

MI Method Comparison: SGPs

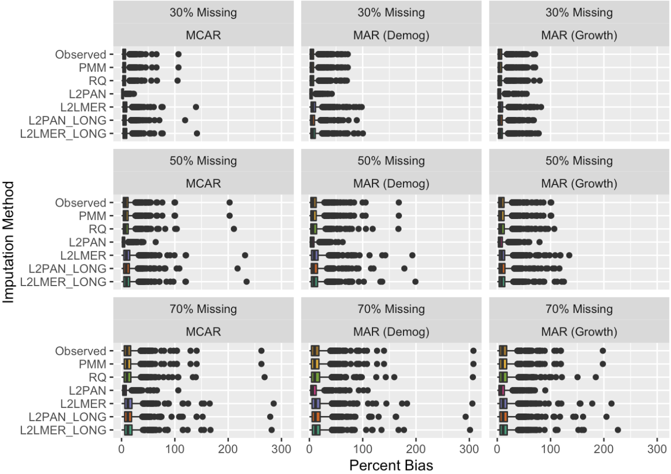

Figure 4: SGP percent bias by imputation method, missingness percentage, and missingness type

MI Method Comparison: SGPs

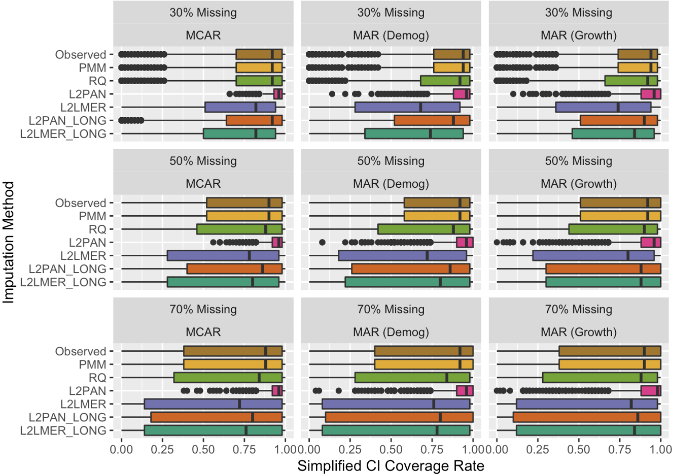

Figure 5: SGP coverage rate by imputation method, missingness percentage, and missingness type

MI Method Comparison: SGPs

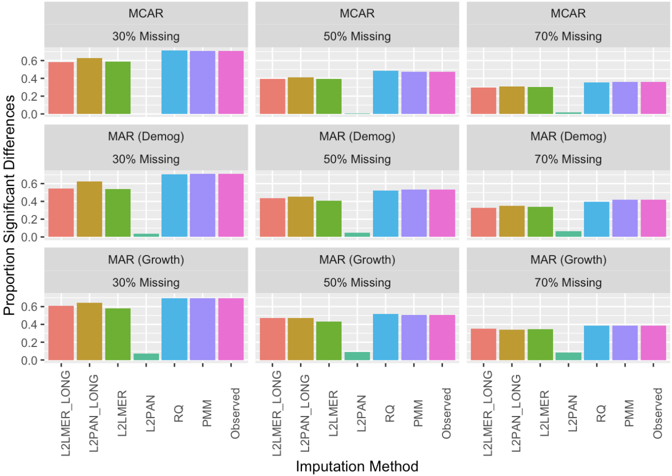

Figure 6: Proportion of times that the imputed SGP differed from the true value based on the simplified F1 statistic

MI Method Comparison: Basic Regression Models

Table 1: Linear fixed-effect regression models for absolute scale score or SGP bias; coefficients with $p < 0.01$ are bolded.

	Scale Scores	SGPs
Grade/Content Area Size	**-0.01 (0.00)**	**-0.02 (0.00)**
50% Missing	**2.76 (0.25)**	**1.57 (0.09)**
70% Missing	**5.86 (0.60)**	**3.13 (0.15)**
MAR with Demographics	**3.60 (0.42)**	**0.55 (0.07)**
MAR with Growth	**8.27 (1.26)**	**0.59 (0.08)**
L2LMER_LONG	1.12 (0.84)	**2.58 (0.28)**
L2PAN_LONG	**-3.36 (0.53)**	**2.31 (0.28)**
L2LMER	**2.96 (0.31)**	**2.64 (0.26)**
L2PAN	**-4.62(0.85)**	0.10 (0.10)
RQ	**-3.92 (0.66)**	**1.85 (0.26)**
PMM	**-3.91 (0.69)**	**1.75 (0.26)**
R^2	0.36	0.18
Within R^2	0.34	0.17

MI Method Comparison: Summary

- The percent bias tends to be lower when imputing scale scores as compared to SGPs.
- MI efficacy declines with smaller grade/content area size, higher missingness percentages, and when data are missing at random based on status and growth.
- L2PAN demonstrates the relative best performance among the examined MI methods, as evidenced by
  - relatively smaller percent bias;
  - higher coverage rates; and
  - fewer statistically significant F_1 statistics.

Characteristics Influencing L2PAN Efficacy: Scale Scores

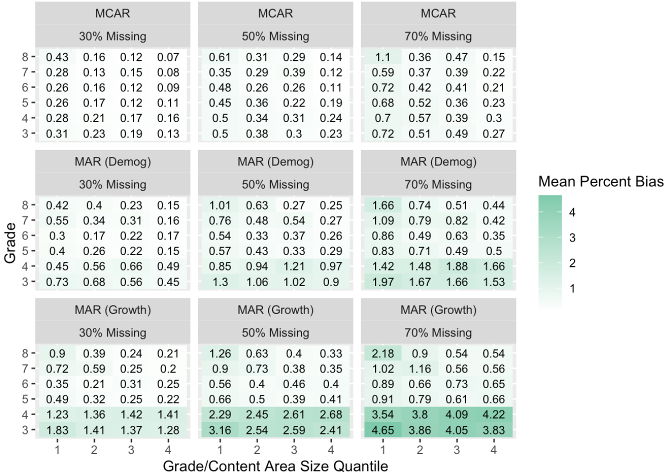

Figure 7: Average scale score percent bias by grade/content area size quantile, grade, and missingness characteristics

Characteristics Influencing L2PAN Efficacy: Scale Scores

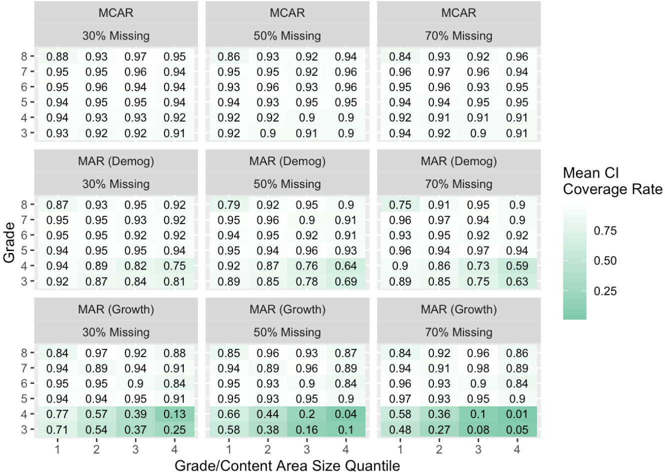

Figure 8: Average scale score coverage rate by grade/content area size quantile, grade, and missingness characteristics

Characteristics Influencing L2PAN Efficacy: SGPs

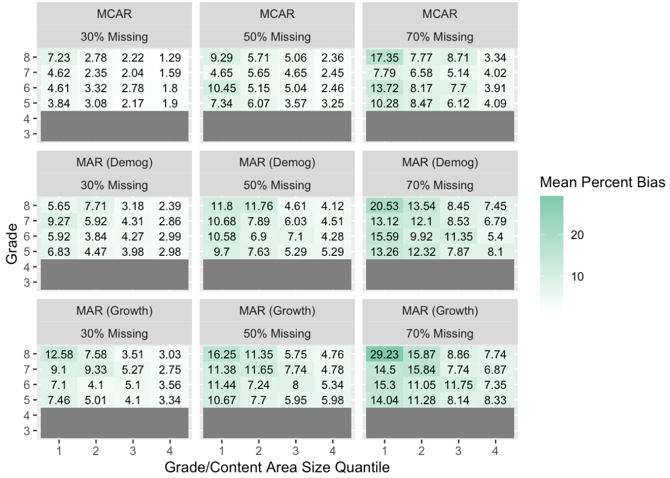

Figure 9: Average SGP percent bias by grade/content area size quantile, grade, and missingness characteristics

Characteristics Influencing L2PAN Efficacy: SGPs

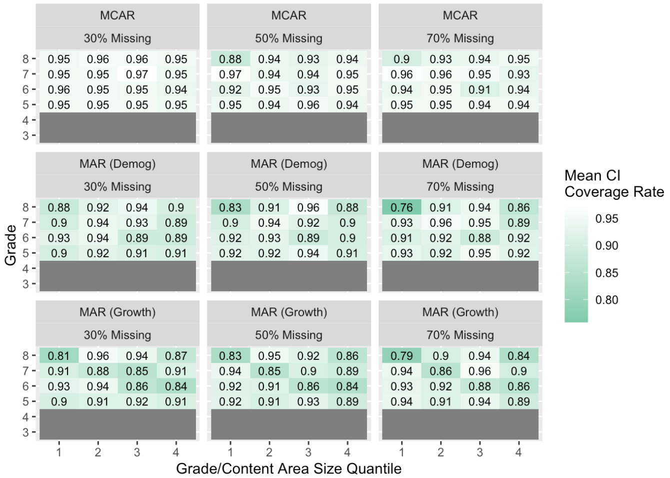

Figure 10: Average SGP coverage rate by grade/content area size quantile, grade, and missingness characteristics

Summary

- Based on percent bias, CI coverage rates, and the simplified F_1 statistic, cross-sectional L2PAN method generally outperforms the other MI methods when imputing mean scale scores and SGPs.

- MI with L2PAN tends to perform worse among cases of smaller grade/content area sizes, when higher percentages of data are missing, and when data are missing based on status and growth.

- Patterns of MI efficacy differ based on whether the scale scores or SGPs are being imputed.

Recommendations

- It is important that researchers and policymakers examine their missingness patterns *prior* to imputation.
- MI should be used with great caution when more than 50% of the data are missing (and note that missingness rates may differ among schools).
- Individualized analyses should include diagnostic checks to examine the MI performance with a particular set of data (for a review, see Nguyen et al., 2017; Stuart et al., 2009)

Future Directions

- Fit a series of more complex generalized linear models to better understand the relationships among the simulation design factors and MI efficacy.
- Replicate analyses using simulated data that incorporates a "Covid effect."
- Consider ways to improve MI for lower grades (if improvement is even possible).
- Explore the possibility of propensity score weighting for drawing appropriate comparisons between 2019 and 2021.

References

- Bates, D., Maechler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software, 67*(1), 1-48. https://doi.org/10.18637/jss.v067.i01.
- Betebenner, D. W., Van Iwaarden, A. R., & Domingue, B. (2021). SGPdata: Exemplar data sets for student growth percentile (SGP) analyses. R package version 25.1-0.0. https://centerforassessment.github.io/SGPdata/
- Demirtas, H. (2004). Simulation driven inferences for multiply imputed longitudinal datasets. *Statistica neerlandica, 58*(4), 466-482. https://doi.org/10.1111/j.1467-9574.2004.00271.x
- Enders, C. K. (2010). *Applied missing data analysis*. New York, NY: The Guilford Press.
- Fox, J. & Weisberg, S. (2018). *Multiple imputation of missing data. Appendix of An R companion to applied regression, third edition*. Thousand Oaks, CA: Sage Publications, Inc.

References

- Ho, A. (2021, Feburary 26). *Three test-score metrics that all states should report in the COVID-19-affected spring of 2021.* Harvard Graduate School of Education.
  https://scholar.harvard.edu/files/andrewho/files/threemetrics.pdf
- Miri, H. H., Hassanzadeh, J., Khaniki, S. H., Akrami, R., & Sirjani, E. (2020). Accuracy of five multiple imputation methods in estimating prevalence of Type 2 diabetes based on STEPS surveys. *Journal of Epidemiology and Global Health, 10*(1), 36-41. https://doi.org/10.2991/jegh.k.191207.001
- Nguyen, C. D., Carlin, J. B., & Lee, K. J. (2017). Model checking in multiple imputation: An overview and case study. *Emerging Themes in Epidemiology, 14*(8). https://doi.org/10.1186/s12982-017-0062-6
- Qi, L., Wang, Y.-F., & He, Y. (2010). A comparison of multiple imputation and fully augmented weighted estimators for Cox regression with missing covariates. *Statistics in Medicine, 29*(25), 2592-2604.
  https://doi.org/10.1002/sim.4016

References

- Stuart, E. A., Azur, M., Frangakis, C., & Leaf, P. (2009). Multiple imputation with large data sets: A case study of the Children's Mental Health Initiative. *American Journal of Epidemiology, 169*(9), 1133–1139. https://doi.org/10.1093/aje/kwp026
- van Buuren, S. (2018). *Flexible imputation of missing data*. CRC Press. https://stefvanbuuren.name/fimd/
- van Buuren, S., & Groothuis-Oudshoorn, K. (2011). mice: Multivariate imputation by chained equations in R. *Journal of Statistical Software, 45*(3), 1-67. https://www.jstatsoft.org/v45/i03/
- Vink, G., & van Buuren, S. (2014). Pooling multiple imputations when the sample happens to be the population. arXiv Pre-Print 1409.8542.
- Zhao, J. H., & Schafer, J. L. (2018). pan: Multiple imputation for multivariate panel or clustered data. R package version 1.6.