Equity Implications of Open-Source and Subscription-Based IRT Software: A Comparison of Scored Performance

Abstract

Equitable technology in education describes the access of technology-based resources to learners, regardless of cultural, ethnic, or financial status. In higher education, access to accurate and reliable technology is still a hurdle for individuals due to cost and access to adjacent resources such as mentors and computers. In order to inform communities on the efficacy of and access to technology for IRT analysis, this study examines the differences of applicable software that are either open-source or subscription-based. Item and person traits and standard errors of a shortened Raven's Progressive Matrices test were estimated through CTT and IRT parameterization strategies with a three-parameter logistic model (3PL) and a nominal response model (NRM) to showcase differences amongst available technologies.

Keywords: equity, social justice, IRT software, open-source, performance measures

Purpose

Limited resources are a common equity issue for learners, researchers, and practitioners. To overcome the inequity problem, it is important to recognize and open doors of access and opportunity for everyone by redistributing resources and services (Lalas et al., 2019, pp.42-43). Consequently, testing the efficacy of open-source statistical software would help not only researchers but also practitioners to collect data, develop research questions, implement interventions, and apply findings (Coburn et al., 2013). Studies showed how important it is to involve marginalized people to feel empowered to pursue their educational degree or to incorporate them into their future learning, personal interest, and/or career pathways. For example, Bevan et al. (2017) and Ryoo et al. (2019) highlighted the problems of equity and

access for minorities in computer science fields. Furthermore, extending access to communities outside of the U.S. is another a critical ethnographic element to consider when it comes to utilization of software. The diversity in environments and interactions in global settings impacts the access and resources towards software use (Resta et al., 2018).

Psychometric methods, namely those whose goal is to classify an underlying latent trait, have direct consequences on the tested population. Kane (2008) discusses the consequences of measurement theory on test takers. Particularly in cases of critical decision making, a software's ability to produce valid estimates in measurement ultimately affects the direction of test development and scoring practices. Moreover, equitable technology means marginalized populations have access to an equal caliber of measurement tools without restriction due to limited personal, school, or community resources. To combat these limitations, open-source resources need to meet expectations and quality of subscription-based software. Little is known about whether open-source software, particularly in the field of educational measurement, can hold its own against cost-driven, subscription-based software.

A simple, globally supported intelligence test was chosen to examine such differences. Raven's Standard Progressive Matrices (SPM) is one of the longest standing tests of intelligence, conceived by John C. Raven in 1936. Due to the nature of its design, it is used as a measure of fluid reasoning through non-verbal abstract reasoning. The SPM was first standardized in 1938 on a sample of 1,407 children in Ipswich, England (Raven, 1941). Over the following decades, the test gained a great breadth of international norms before it was introduced in the United States. Furthermore, differing versions emerged, such as the Colored Progressive Matrices (CPM), Advanced Progressive Matrices (APM), and Standard Progressive Matrices Plus (SPM-

Plus) (Raven, 1941). The test properties have stood the test of time, as the stimuli and responses are independent of language, reading, writing, and advanced (computational) math skills.

Research Questions

1. Does subscription-based software outperform open-source resources, and what does this result mean for the equitability of use for learners, practitioners, and researchers?
2. Multiple software and corresponding packages inevitably differ for calculating CTT and IRT estimates, but is this difference noticeable when all settings are fixed?
3. How should psychometricians decide on which program to calibrate models and generate theta estimates?

Methods

Polytomous SPM response data was obtained from Mendeley Data. Responses were then converted into a dichotomous format. The sample consisted of 499 undergraduate students from a French business school who were native French speakers.

Six software and corresponding packages were used to fit the two types of SPM response data (Table A). To analyze the SPM response data, three subscription-based software (STATA, MPlus, IRTPRO) and three open-source software (SAS OnDemand for Academics, RStudio, jMetrik) were selected. Estimates for item and person traits were produced through CTT and IRT parameterization strategies and calibration methods involving marginal maximum likelihood (MML) scoring and Bayesian expected a posteriori (EAP) estimation. Holding MML and EAP constant in all models, theta value estimates were examined to compare 3PL and NRM estimates. The differences among software and considerations for equitable access to technology are teased apart using performance estimates (i.e., unbiased RMSE, mean biased error, mean absolute error; Table B and C) and user experience impressions.

Preliminary Findings

Comparing software and packages is challenging given the variety of argument settings and function defaults that can be chosen by authors. One major takeaway is that the openness of the program matters. Our results revealed that open-source software appeared to have limitations in terms of usability and flexibility, however, even subscription-based software revealed limitations in the versatility of coding complex models. While R outperformed all models with the 3PL model fit for ability estimates, it failed to do so for the nominal model where subscription-based software like IRTPRO and Stata exceeded.

References

Birnbaum, A. (1968). *Some latent trait models and their use in inferring an examinee's ability.* In F.M. Lord & M.R. Novick, Statistical theories of mental test scores (pp. 392-479). Reading, MA: Addison-Wesley.

Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. Applied *Psychological Measurement, 16,* 159-176.

Muraki, E. (1997). *A generalized partial credit model.* In W. van der Linden & R. K. Hambleton (Eds.), Handbook of modern item response theory (pp. 153-164). New York: Springer.

Samejima, F. (1969). *Estimation of latent ability using a response pattern of graded scores.* Psychometric Monograph, No. 17, 34, Part 2.

Thissen, D. (1982). Marginal maximum likelihood estimation for the one-parameter logistic model. *Psychometrika, 47,* 201-214.

Bock, R. D. (1972). Estimating item parameters and latent ability when responses are scored in two or more nominal categories. *Psychometrika, 37*, 29-51.

Cai, L., Thissen, D., & du Toit, S.H.C. (2017). *IRTPRO 4.2 for Windows* [Computer software]. Skokie, IL: Scientific Software International, Inc.

Chalmers, R. P. (2012). mirt: A Multidimensional Item Response Theory Package for the R Environment. *Journal of Statistical Software*, 48.

Chon, K.H., Lee, W. & Ansley, T.N. (2007). *Assessing IRT model-data fit for mixed format tests.* (CASMA Report No. 26). Iowa City, Iowa.

Kane, M. T. (2010). *Errors of measurement, theory, and public policy.* Princeton, NJ: Educational Testing Service.

Matlock Cole, K., & Paek, I. (2017). PROC IRT: A SAS Procedure for Item Response Theory. *Applied Psychological Measurement, 41*(4), 311–320. https://doi.org/10.1177/0146621616685062

Meyer, J. P. (2014). *Applied measurement with jMetrik.* Routledge.

Meyer, J. P. (2018). jMetrik, version 4.1.1.

Muthen, L. K. & Muthen, B. O. (2017). *Mplus User's Guide.* Eight Edition. Los Angeles, CA: Muthen & Muthen.

Myszkowski, N., & Storme, M. (2018). *A snapshot of g? Binary and Polytomous Item-Response Theory investigations of the Last Series of the Standard Progressive Matrices (SPM-LS)*(VI)[Mendeley Data]. https://doi.org/10.17632/h3yhs5gy3w.1

Oscar Perpinan Lamigueiro (2018). tdr: Target Diagram. R package version 0.13. https://CRAN.R-project.org/package=tdr

Oscar Perpinan Lamigueiro (2018). tdr: Target Diagram. R package version 0.13. https://CRAN.R-project.org/package=tdr

Perpinan-Lamigueiro O. (2018). tdr: Target Diagram. R package version 0.13.

Preston, K., Reise, S., Cai, Li., Hays, R. D. (2011). Using the nominal response model to evaluate *Psychometrika, 75*(3), 454–473. http://dx.doi.org/10.1007/s11336-010-9163-7

R Core Team (2013). *R: A language and environment for statistical computing.* R Foundation for Statistical Computing, Vienna, Austria. http://www.R-project.org/

R. Philip Chalmers (2012). mirt: A Multidimensional Item Response Theory Package for the R Environment. *Journal of Statistical Software, 48*(6), 1-29. doi:10.18637/jss.v048.i06

Raven, J.C. (1941) Standardization of Progressive Matrices, 1938. *British Journal of Medical Psychology, 19*, 137-150. https://doi.org/10.1111/j.2044-8341.1941.tb00316.x

Bevan, B., Ryoo, J., & Shea, M. (2017). What if? Building creative cultures for STEM making and learning. *Afterschool Matters, 25*, 1–8. https://eric.ed.gov/?id=EJ1138042

Birnbaum, A. (1968). Using the nominal response model to evaluate response category discrimination in the PROMIS emotional distress item pools. *Educational and Psychological Measurement, 71*(3), 523-550. https://doi.org/10.1177/0013164410382250

Lalas, J. W., Charest, B., Strikwerda, H., & Ordaz, M. (2019). *Nurturing hope, sense of belonging and engagement through equity.* In K. Scorgie & C. Forlin (Eds.), Promoting social inclusion: Co-creating environments that foster equity and belonging international perspectives on inclusive education (pp. 41–52). Emerald Publishing. https://doi.org/10.1108/S1479-363620190000013004

Penuel, W. R., Coburn, C. E., & Gallagher, D. J. (2013). Negotiating problems of practice in research–practice design partnerships. *Teachers College Record, 115*(14), 237–255. https://doi.org/10.1177/016146811311501404

Resta, P., Laferrière, T., McLaughlin, R., & Kouraogo, A., et al. (2018). Issues and challenges related to digital equity: An overview. In J. Voogt (Ed.), *Second handbook of information technology in primary and secondary education, Springer international handbooks of education* (pp. 987–1004). Cham, Switzerland: Springer International Publishing AG.

Ryoo, J. J., Margolis, J., Estrada, C., Tanksley, T. C., Guest-Johnson, D., & Mendoza, S. (2019). *Student voices: Equity, identity, and agency in CS class- rooms.* In 2019 Research on Equity and Sustained Participation in Engineering, Computing, and Technology (RESPECT). https://doi.org/10.1109/ RESPECT46404.2019.8985947

SAS Institute Inc. (2018). SAS Enterprise Edition 3.8, Cary, NC.

StataCorp. 2021. Stata Statistical Software: Release 17. College Station, TX: StataCorp LLC.

Suh, Y., & Bolt, D. M. (2010). Nested logit models for multiple-choice item response data. *Psychometrika, 75*(3), 454–473. https://doi.org/10.1007/s11336-010-9163-7

Thissen, D., Cai, L., & Bock, R.D. (2010). *The nominal item response model.* In M. Nering & R. Ostini (Eds.), Handbook of polytomous item response theory models: Developments and applications.

Wickham, H. (2016). *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York. ISBN 978-3-319-24277-4, https://ggplot2.tidyverse.org.

Yilmaz, H. B. (2019). A comparison of IRT model combinations for assessing fit in a mixed format elementary school science test. *International Electronic Journal of Elementary Education, 11*(5), 539-545. https://doi.org/10.26822/iejee.2019553350

Appendix

Table A

Program Versions and Corresponding Model Fit Used.

Program (Package)	Version	Date Updated	Models Fitted
R (mirt)	4.1.3	3/3/2022	Nominal/3PL/3PLNRM
SAS (IRT)	3.8 (Enterprise Edition)	9/1/2020	Nominal/3PL
IRTPro	5.1.0.20	5/2021	Nominal/3PL
Stata (IRT)	17.8 BE	4/20/2021	Nominal/3PL
jMetrik	4.1.1	2/23/2018	3PL
MPlus	8.3	4/30/2019	Nominal/3PL

Table B

Measures for 3PL Unbiased RMSE, Mean Biased Error, Mean Absolute Error across Software.

	θ						SE					
Program	Median	Mean	SD	MBE	MAE	RMSE	Median	Mean	SD	MBE	MAE	RMSE
R	0.032008	0.146	0.908	NA	NA	NA	0.357	0.384	0.075	NA	NA	NA
SAS	0.032122	0.000	0.935	-0.015	0.047	0.065	NA	NA	NA	NA	NA	NA
IRTPro	0.019000	0.000	0.933	-0.014	0.036	0.055	0.387	0.414	0.064	0.030	0.030	0.040
Stata	-0.009010	0.000	0.915	-0.015	0.040	0.055	0.368	0.401	0.074	0.017	0.022	0.032
jMetrik	-0.094440	-0.074	0.937	-0.088	0.100	0.111	0.377	0.407	0.076	0.023	0.025	0.028
MPlus	0.037000	-0.002	0.909	-0.016	0.093	0.130	0.390	0.413	0.068	0.030	0.040	0.051

Table C

Measures for Nominal Unbiased RMSE, Mean Biased Error, Mean Absolute Error across Software.

	θ						SE					
Program	Median	Mean	SD	MBE	MAE	RMSE	Median	Mean	SD	MBE	MAE	RMSE
R	0.236	0.010	0.823	NA	NA	NA	0.590	0.543	0.172	NA	NA	NA
SAS	0.043	0.000	0.904	-0.010	1.376	1.672	NA	NA	NA	NA	NA	NA
IRTPro	-0.153	-0.217	0.839	-0.227	1.333	1.611	0.374	0.415	0.121	-0.128	0.142	0.170
Stata	0.043	0.000	0.904	-0.010	1.376	1.672	0.387	0.409	0.130	-0.134	0.138	0.165
MPlus	0.045	0.003	0.905	-0.007	1.377	1.673	0.388	0.409	0.130	-0.134	0.137	0.164