Psychometric Properties of PAES Math M1

# Introduction

The regular admission process to higher education in Chile is a significant annual event. It involves selecting students into programs based on their scores in various selection factors that reflect their readiness for higher education. The programs assign specific weights to these factors, especially the tests that measure students’ mathematics and reading comprehension competencies.

Since its inception in 2004, the mathematics test has played a pivotal role in the admission process for students interested in STEM, health sciences, and law programs. This is particularly true for those aspiring to study in prestigious, highly selective universities. The test’s importance has drawn the attention of researchers and stakeholders, who question its validity and argue that it may inadvertently favor certain students in the admission process.

The mathematics test has been largely criticized for the correlation between its scores and the socioeconomic characteristics of the students and their schools. Different research findings have shown that students tend to achieve higher scores in math when their families have more educational and financial resources and when they attend private schools. Furthermore, since the selection test was designed to measure knowledge of the mathematics curriculum content, students had higher chances of achieving high scores if they attended schools able to cover most of the curriculum or had access to additional mathematics instruction. This additional instruction often took the form of paid tutoring during or after secondary school, giving an advantage to students whose families could afford such resources.

DEMRE, the institution responsible for designing, administering, and analyzing the selection tests, has not ignored the criticism of the mathematics test. Over the years, they have made continuous adjustments to improve its technical properties and objectivity. While these changes have not yet fully addressed all public concerns, they demonstrate DEMRE’s responsiveness to those dissatisfied with the selection tests or the admission process.

In 2020, DEMRE announced a substantial change in the selection tests following student protests across the country. These protests were sparked by concerns about the administration of the tests and the leakage of a form used in the history selection test. From that point forward, DEMRE declared that the tests would evaluate students’ competencies in the measured subjects. This shift meant that questions would assess students’ ability to solve tasks based on knowledge acquired during secondary education.

Following DEMRE’s announcement, the well-known PSU (Prueba de Selección Universitaria or University Selection Tests) were replaced by the PDT (Prueba de Transición or Transition Tests) for the 2021 and 2022 admission processes. In 2023, the new PAES (Prueba de Acceso a la Educación Superior or Higher Education Entrance Tests) were implemented for the first time. The PAES were DEMRE’s response to calls for more equitable and fair selection tests. As part of this change, the mathematics test was divided into two separate assessments. The PAES Mathematics Competencies 1 (M1) focuses on general mathematical knowledge required for higher education studies, while the PAES Mathematics Competencies 2 (M2) is an advanced, more specific test designed for students interested in math-oriented programs.

Like its predecessors, the M1 test is mandatory for students, and its properties and results warrant careful examination. This paper analyzes the M1 test applied in the 2024 admission process from a psychometric perspective, specifically by fitting and evaluating different models within the Item Response Theory (IRT) framework to assess its effectiveness in measuring intended outcomes. The study aims to provide evidence about the psychometric qualities of the M1 test, which may help evaluate the validity of its scores and determine whether its results align with its design intentions. To this end, the research addresses two primary questions:

1. What are the characteristics of the IRT model that best fits the M1 test data?
2. How comparable are the information and efficiency of the different forms used in administering the M1 test according to various IRT models?

# Method

## Dataset

The dataset used in this study was the item response data from the **M1** test administered in the 2024 regular admission process. All the information was collected, processed, and made public by the Department of Educational Evaluation, Measurement, and Registration (DEMRE) of the University of Chile, the institution in charge of the admission process. It contains students’ correct and incorrect responses, with double marked item and omitted items treated as missing data.

## Instrument

The M1 test measures abilities (e.g., modeling and representing data) and knowledge (e.g., in numbers, algebra, and probability) in mathematics, which are considered as foundational skills for higher education. These skills are part of the national curriculum in mathematics and are supposed to be developed during primary and secondary education.

The 2024 M1 test was composed of 65 multiple-choice items, in which 60 were operational items, while the remaining 5 were pilot items and did not contribute to students score. Each item had 4 options and a single correct answer. The test had 4 equivalent forms (originally labeled as 113, 114, 115, and 116) that for convenience were referred to as forms **A**, **B**, **C**, and **D** throughout. Forms were prepared in pairs: A and C, and B and D. Each pair of forms contained the same set of items but in different order (scrabbled). Across pair of forms, 47 out of the 60 (78,5%) scored questions were common items.

## Sample

The original dataset contained the responses of 243,305 students who took the M1 test in 2024. As part of the data cleaning process, cases with no responses were removed, thus the final sample include 243,294 students. Because of the focus of this study, information about students is not provided; However, a detailed description of their relevant attributes can be found on the DEMRE website.

## Procedure

Findings to answer our research questions come from analyzing the performance of different psychometric models in estimating item and student characteristics, and from comparing M1 test forms in terms of information and relative efficacy.

## IRT Models

In order to select the most appropriate item response theory (IRT) model, we fitted and revised the widely known **1PL**, **2PL**, and **3PL** IRT models ([De Ayala 2013](#ref-de2013theory)) These models belong to a larger family of psychometric models suitable for assessing unidimensional latent traits when items are dichotomously scored (i.e., correct or incorrect answers), as is the case with the M1 test.

The three models differ in the number of parameters being estimated for each item, which are expressions of the difficulty (**a** parameter), discrimination (**b** parameter), and guessing (**c** parameter) attributable to the item -these are characteristics of the items independent of the student characteristics-. In the 1PL model, only the **b** parameter is estimated for each item, while the **a** parameter is constrained to be constant across items. The 2PL model estimates the **a** and **b** parameters for each item. The 3PL also estimates the **c** parameter for each item.

The mathematical form of the 3PL model is:

Equation 1

where:  
- is the probability of a correct response for item **i** conditional on the latent ability of student **s** ,  
- is the discrimination parameter for item ,  
- is the difficulty parameter for item ,  
- is the guessing parameter for item .

The 2PL model can be derived from Equation 1 by setting the guessing parameter () to zero, and the 1PL model can be derived from the 2PL model by setting the discrimination parameter () to a constant that is estimated from a reference item.

The discrimination () parameter indicates how effectively an item distinguishes between individuals with varying levels of ability. Items with high discrimination values provide valuable insights into differences in ability level across individuals.

The difficulty parameter () denotes the point on the ability scale where the probability of a correct response for the item equals 0.5, at least in the 1PL and 2PL models. For the 3PL model, it denotes the point where the probability of a correct response is halfway between the guessing parameter () and 1 (i.e ).

The guessing parameter () represents the probability of providing a correct response regardless of an individual’s ability. It is related to the options available for each question. Naturally, in any of the M1 test questions, students have 25% chance (1/4) of a correct answer just by chance given that the test items had four response options. However, empirical estimates have shown that () values are usually less than , where is the number of response options ([Linden 2016](#ref-van2016unidimensional)).

## Calibration

Test calibration in IRT involves obtaining the item parameters for each of the models defined above ([Baker and Kim 2004](#ref-baker2004item)). When dealing with a single test form, obtaining item parameter estimates is straightforward and can be achieved using many statistical software programs. However, when multiple test forms are considered, decisions must be made about the calibration approach to ensure item parameters across forms are on the same scale (and therefore the scores derived from them). This process is known as linking ([Kolen and Brennan 2014](#ref-KolenTestEquating)).

Depending on the data collection approach and sample characteristics, we can choose one of three commonly used IRT linking methods: separate calibration, concurrent calibration, and fixed calibration ([Lee and Lee 2018](#ref-lee2018irt)). Given that students were randomly assigned to each form and took the test at the same day, the concurrent calibration approach was used to obtain item parameter estimates across forms. This linking method combines all item response data into a single dataset for parameter estimation, which is performed in one computer run. Items that are unique to a particular form are treated as missing data in the other forms, and at the end of the process, parameters for both common items and form-specific items were on the same scale.

Item parameters calibration was performed using the marginal maximum likelihood estimation method and the expectation-maximization algorithm, as implemented in the mirt R package ([Chalmers 2012](#ref-chalmersmirt)).

## Model fit

To assess the fit of the IRT models to the observed response data, we examined both absolute and relative model fit. Absolute model fit evaluates how well a model’s predicted responses match the actual observed responses. In contrast, relative model fit involves comparing models to determine which one provides the most efficient and accurate representation of the dataset. Absolute fit focuses on the alignment between predicted and observed responses for a single model, whereas relative fit focuses on identifying the most suitable model among competing models.

Absolute model fit was assessed using the **M2** statistic ([Maydeu-Olivares and Joe 2005](#ref-maydeu2005limited)), a limited-information model fit statistic that evaluate the fit between the first and second order marginal residuals of the IRT models. Under the null hypothesis (i.e., the model is exactly true in the population,) this statistic is asymptotically chi-square distributed. However, under the alternative hypothesis (i.e., lack of fit), it has a non-central chi-square distribution when N is large ([Cai, Chung, and Lee 2023](#ref-cai2023incremental)). For this test, we conclude that the model adequately fits the data when the corresponding **p**-value was > 0.05.

Relative model fit was assessed using likelihood ratio tests (−2ΔLL). This is a statistical significance test to compare nested models, as the case is for the 1PL, 2PL and 3PL models. The −2ΔLL test consists of subtracting the maximized log-likelihood of the more complex model from that of the simpler model, multiplying such difference by -2 and comparing it to a distribution with degrees of freedom equal to the difference in parameters between models to determine its significance. For instance, after checking that the 3PL and 2PL models demonstrate adequate absolute fit, we can use the −2ΔLL test to evaluate whether constraining some item parameters in the 2PL model fits the data better than estimating additional item parameters in the 3PL model.

## Information and Standard error

Item and test information, along with standard errors, are additional metrics for evaluating the performance of an item or test in meeting assessment goals. The item/test information represents the precision with which ability is estimated across the ability continuum. The more information we have in estimating ability at certain point, let’s say between 0 and 1, the less error we have in that range of ability. Computationally, the standard error of the ability estimate is the reciprocal of the square root of the information ([Baker, Kim, et al. 2017](#ref-baker2017basics)). Thus, the point (in terms of ability) at which information reaches its maximum usually coincides with the point at which the standard error is lowest.

Test information was examined across the IRT models to ensure that the selected model is peaked around the target ability level for the M1 test.

# Results

Comparison of model results (information, factor scores, SE) Comparison of forms based on Rasch and 2PL models (information, expected true scores)

# Discussion

# Conclusion

# Things to move

The LR test for comparing the 1PL and 2PL models is more straightforward than comparing the 2PL with the 3PL model. When comparing the 1PL and 2PL models, we test the null hypothesis that all discrimination parameters are equal to the estimated discrimination value against an alternative hypothesis that at least one discrimination parameter differs from this value. This type of comparison in LR is typical and can be easily tested with the anova function in the mirt R package, as the reference distribution holds. However, when comparing the 2PL and 3PL models, the null hypothesis states that the guessing parameter is equal to zero for all items, against an alternative hypothesis that at least one guessing parameter is not equal to zero. This null hypothesis places the guessing parameters at their boundary, and the reference distribution no longer holds ([Brown, Templin, and Cohen 2015](#ref-brown2015comparing)).

Comparing the 2PL model to the 1PL model was straightforward. In order to appropriately compare the 2PL and 3PL models, we followed the alternative approach outlined by Brown, Templin, and Cohen ([2015](#ref-brown2015comparing)). This involves conducting an item-level model comparison, comparing a model where the 2PL is specified for all items to models where the 3PL is specified for one item at a time. This approach results in as many LR tests as there are items in the test, leading to a final model where some items follow a 2PL model while others follow a 3PL model.

# References

Baker, Frank B, and Seock-Ho Kim. 2004. *Item Response Theory: Parameter Estimation Techniques*. CRC press.

Baker, Frank B, Seock-Ho Kim, et al. 2017. *The Basics of Item Response Theory Using r*. Vol. 969. Springer.

Brown, Christy, Jonathan Templin, and Allan Cohen. 2015. “Comparing the Two-and Three-Parameter Logistic Models via Likelihood Ratio Tests: A Commonly Misunderstood Problem.” *Applied Psychological Measurement* 39 (5): 335–48.

Cai, Li, Seung Won Chung, and Taehun Lee. 2023. “Incremental Model Fit Assessment in the Case of Categorical Data: Tucker–Lewis Index for Item Response Theory Modeling.” *Prevention Science* 24 (3): 455–66.

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

De Ayala, Rafael Jaime. 2013. *The Theory and Practice of Item Response Theory*. Guilford Publications.

Kolen, Michael, and Robert Brennan. 2014. *Test Equating, Scaling, and Linking. Methods and Practices*. Springer New York, NY.

Lee, Won-Chan, and Guemin Lee. 2018. “IRT Linking and Equating.” *The Wiley Handbook of Psychometric Testing: A Multidisciplinary Reference on Survey, Scale and Test Development*, 639–73.

Linden, Wim J van der. 2016. “Unidimensional Logistic Response Models.” In *Handbook of Item Response Theory*, 13–30. Chapman; Hall/CRC.

Maydeu-Olivares, Albert, and Harry Joe. 2005. “Limited-and Full-Information Estimation and Goodness-of-Fit Testing in 2 n Contingency Tables: A Unified Framework.” *Journal of the American Statistical Association* 100 (471): 1009–20.