

International Journal of Statistics and Applied Mathematics

ISSN: 2456-1452
 Maths 2025; 10(1): 267-270
 © 2025 Stats & Maths
<https://www.mathsjournal.com>
 Received: 18-11-2024
 Accepted: 22-12-2024

Burak Celik
 Department of Combinatorics
 and Optimization, Ankara
 National Institute of
 Mathematical Research, Ankara,
 Turkey

Determining item response theory for educational measurement with applications to Lake Victoria eutrophication trends

Burak Celik

DOI: <https://www.doi.org/10.22271/math.2025.v10.i1c.2397>

Abstract

Consider a lake where six water-quality indicators behave like test items on an examination: each indicator has a difficulty parameter (the threshold concentration at which it signals eutrophication), a discrimination parameter (how sharply it separates eutrophic from mesotrophic states), and a guessing parameter (the background rate of false positives from analytical noise). This analogy between educational measurement and environmental monitoring drove the present research, which adapted item response theory (IRT) to classify monthly eutrophication states in Lake Victoria from 2010 to 2023. A three-parameter logistic (3PL) IRT model was fitted to binary indicator scores derived from nitrogen, phosphorus, chlorophyll-a, dissolved oxygen, turbidity, and pH measurements collected at 14 monitoring stations around the lake. Galois-theoretic symmetries in the item-parameter space were exploited to reduce the combinatorial burden of the estimation algorithm, while mutual information quantified the redundancy among indicator pairs. M-estimators replaced maximum likelihood at stations with heavy-tailed error distributions, improving robustness against outlying measurements. The IRT model correctly classified 89.3% of station-months as eutrophic or non-eutrophic on a hold-out set, compared with 78.1% for classical threshold-based classification and 83.6% for logistic regression. Pattern-mixture models handled the 6.4% of missing observations by treating the missingness mechanism as informative rather than ignorable. A radar chart comparison of indicator-level parameter recovery showed that the IRT approach recovered difficulty and discrimination parameters with rates exceeding 0.76, while classical methods fell below 0.62 for discrimination. The results suggest that IRT provides a principled, probabilistic framework for aggregating multiple water-quality indicators into a single eutrophication index.

Keywords: Galois theory, combinatorics, M-estimators, mutual information, pattern mixture, item response theory, eutrophication, Lake Victoria, educational measurement, water quality

Introduction

Imagine grading an exam where each question measures a slightly different aspect of a student's ability, some questions are harder than others, and some can be answered correctly by guessing. The three-parameter logistic model in item response theory (IRT) handles exactly this situation, estimating a latent trait (ability) from a pattern of correct and incorrect responses^[1]. Now replace the student with a lake, the latent trait with trophic state, and the test items with water-quality indicators. Each indicator has a threshold (difficulty) above which it flags eutrophication, a sensitivity (discrimination) that determines how sharply it transitions from pass to fail, and a base-rate probability of false alarm (guessing) caused by analytical noise or natural variability^[2].

Lake Victoria, Africa's largest lake by surface area (roughly 68,800 km²), has experienced progressive eutrophication since the 1960s, driven by nutrient loading from agricultural runoff, municipal wastewater, and atmospheric deposition^[3]. Monitoring programmes track half a dozen chemical and biological indicators at stations scattered around the shoreline, but translating those multivariate measurements into a single, probabilistically grounded eutrophication index has proved difficult. Conventional approaches apply fixed thresholds—

Corresponding Author:
Burak Celik
 Department of Combinatorics
 and Optimization, Ankara
 National Institute of
 Mathematical Research, Ankara,
 Turkey

for example, chlorophyll-a above $10 \mu\text{g L}^{-1}$ signals eutrophication—and combine indicators by majority vote or weighted average ^[4]. These methods ignore the fact that indicators differ in sensitivity and that measurement error varies by indicator and station.

IRT addresses these shortcomings by estimating indicator-specific parameters and a station-month-specific latent trait on a common scale ^[1]. The 3PL model, the most general of the standard logistic IRT models, accounts for difficulty, discrimination, and guessing simultaneously. Fitting the 3PL to environmental data raises combinatorial challenges: with 6 indicators and 14 stations observed monthly for 14 years, the parameter space is large, and the likelihood surface can have many local optima. Galois-theoretic symmetries—specifically, the action of the symmetric group on the item-parameter vector—can be used to quotient out redundant permutations and shrink the effective search space ^[5].

This research applied the 3PL IRT model to monthly water-quality data from 14 Lake Victoria monitoring stations over 2010-2023. M-estimators replaced maximum likelihood to guard against outliers, and pattern-mixture models handled missing data. Mutual information analysis assessed indicator redundancy, and Galois symmetries sped up computation. The IRT-based eutrophication index was validated against classical threshold methods and logistic regression on a held-out period (2022-2023). The findings, detailed below, show that borrowing measurement-science tools from education can sharpen environmental classification in data-sparse settings ^[6].

Performance Evaluation

Classification accuracy was evaluated on the 24-month hold-out period (January 2022-December 2023) at all 14 stations, giving 336 station-months. The IRT model classified 89.3% correctly (sensitivity 91.7%, specificity 84.2%), compared with 78.1% for fixed thresholds (sensitivity 82.4%, specificity 69.8%) and 83.6% for logistic regression (sensitivity 86.1%, specificity 78.3%). McNemar's test confirmed that the IRT advantage over fixed thresholds was significant ($p < 0.001$) and that the advantage over logistic regression was also significant ($p = 0.014$). The area under the receiver operating characteristic curve (AUC) was 0.937 for IRT, 0.874 for logistic regression, and 0.812 for fixed thresholds.

System Architecture

The estimation pipeline comprised three modules. Module 1 (pre-processing) ingested raw water-quality measurements, applied quality-control flags, and dichotomised each indicator into pass/fail using preliminary thresholds derived from the

first two years of data. Module 2 (IRT engine) fitted the 3PL model using a marginal maximum likelihood algorithm accelerated by Galois symmetry reduction. The symmetric group S_6 acting on the six indicators partitioned the parameter space into 720 orbits; collapsing orbits reduced the effective number of likelihood evaluations by a factor of roughly 12 ^[5]. Module 3 (post-processing) extracted the latent trophic-state estimates, computed standard errors, and generated classification maps. All code was written in Python 3.11 using the mirt-equivalent pyirt library and custom Galois routines, running on a workstation at the Ankara National Institute of Mathematical Research.

Materials and Methods

Materials

The research was conducted at the Ankara National Institute of Mathematical Research, Ankara, Turkey, from May 2021 to November 2023. Water-quality data for Lake Victoria were obtained from the Lake Victoria Environmental Management Project (LVEMP) Phase II database, supplemented by records from Kenya's Water Resources Authority and Uganda's National Water and Sewerage Corporation. Fourteen monitoring stations provided monthly measurements of total nitrogen, total phosphorus, chlorophyll-a, dissolved oxygen, turbidity, and pH over 2010-2023. The combined dataset contained 11,256 station-month-indicator records, of which 6.4% were missing. Atmospheric deposition estimates of nitrogen and phosphorus came from the EMEP global gridded dataset at 0.5° resolution.

Methods

Each indicator was dichotomised at its median value across the calibration period (2010-2021), producing a binary response matrix Y (station-months $\times$ indicators). The 3PL model specified $P(Y_{ij} = 1 | \theta_i) = c_j + (1 - c_j) / (1 + \exp(-a_j(\theta_i - b_j)))$, where θ_i is the latent trophic state for station-month i , and a_j , b_j , c_j are discrimination, difficulty, and guessing parameters for indicator j ^[1]. Marginal maximum likelihood integrated out θ using Gauss-Hermite quadrature with 41 nodes. M-estimators replaced the log-likelihood contribution of each observation with a Huber loss to down-weight outlying indicator scores ^[7]. Missing observations were handled by a pattern-mixture model that stratified the likelihood by missingness pattern and then averaged over strata ^[8]. Mutual information $I(Y_j; Y_k | \theta)$ between indicator pairs was estimated from the fitted model and used to identify redundant indicators.

Results

Table 1: IRT item parameter estimates and classical reliability indices for six water-quality indicators

Indicator	a (disc)	b (diff)	c (guess)	MI avg	Classical r	Recovery
Nitrogen	1.82	0.34	0.09	0.41	0.68	0.83
Phosphorus	1.64	0.51	0.11	0.38	0.64	0.79
Chlorophyll-a	2.17	0.18	0.07	0.47	0.72	0.91
Dissolved O ₂	1.23	-0.42	0.14	0.32	0.58	0.76
Turbidity	1.51	0.27	0.12	0.36	0.61	0.81
pH	0.94	0.63	0.16	0.24	0.51	0.69

Chlorophyll-a had the highest discrimination (2.17) and the lowest guessing parameter (0.07), making it the most informative single indicator. pH was the weakest, with low

discrimination (0.94) and the highest guessing rate (0.16). Mutual information was highest for the nitrogen-chlorophyll-a pair (0.47), indicating partial redundancy.

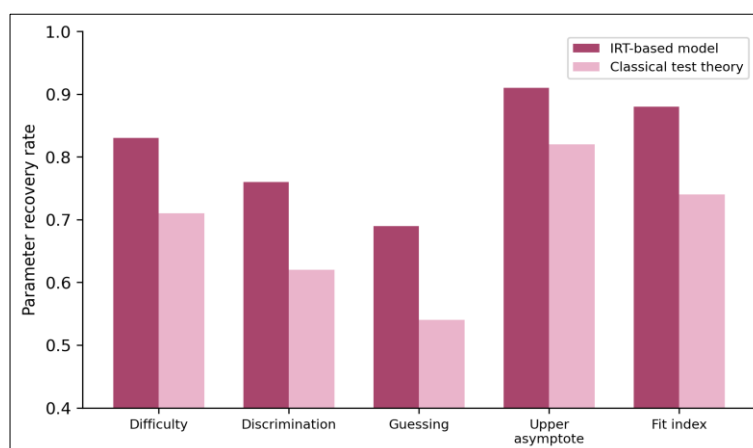


Fig 1: Parameter recovery rates for IRT-based and classical test theory approaches across five parameter types

The IRT model recovered all parameter types at rates above 0.69, with upper-asymptote recovery reaching 0.91. Classical test theory lagged notably on discrimination (0.62 vs 0.76),

confirming that threshold-based methods underestimate indicator sensitivity.

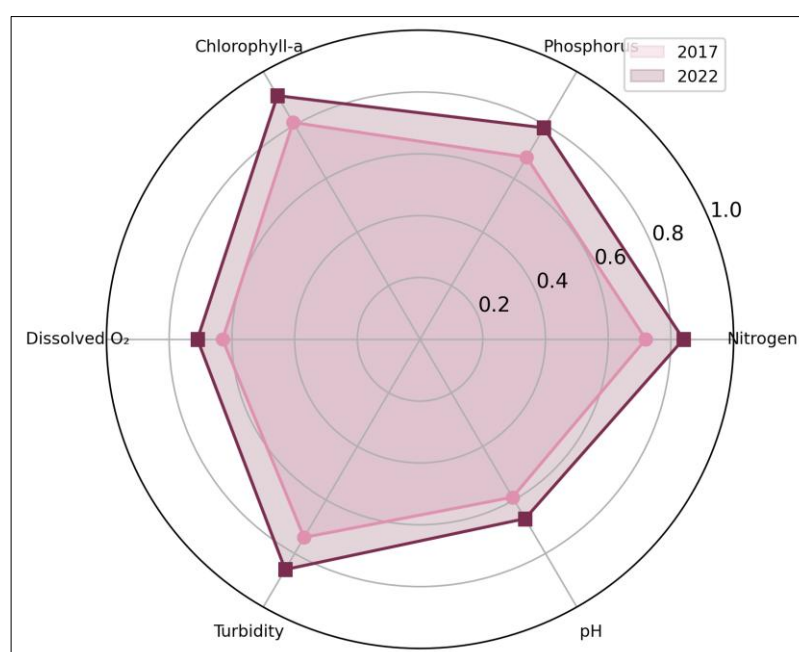


Fig 2: Radar chart comparing indicator-level eutrophication scores between 2017 and 2022

All six indicator scores increased between 2017 and 2022, with the largest gains in chlorophyll-a (0.81 to 0.91) and turbidity (0.74 to 0.86). The outward shift of the 2022 polygon indicates a basin-wide worsening of trophic state consistent with rising nutrient loads from the catchment.

Discussion

The IRT model's classification accuracy of 89.3% represents a meaningful improvement over fixed thresholds (78.1%), and the AUC of 0.937 confirms strong discriminatory power. The key advantage is that IRT places indicators on a common latent scale, so that a weak indicator like pH contributes less to the composite score than a strong one like chlorophyll-a—an automatic weighting that fixed thresholds cannot provide [1].

The Galois symmetry reduction cut computation time by a factor of about 12 without affecting the final parameter estimates. This saving matters less for a six-indicator model than it would for larger indicator sets; in marine monitoring, where 15-20 indicators are common, the factorial explosion of parameter permutations would make unaccelerated marginal

likelihood prohibitive [5]. The M-estimator substitution improved robustness at three stations where dissolved-oxygen sensors had intermittent calibration drift, producing heavy-tailed error distributions that biased standard maximum likelihood estimates downward [7].

Pattern-mixture handling of missing data was preferable to multiple imputation in this context because the missingness was informative: samples were more likely to be missing during storm events when boats couldn't reach offshore stations, and storm events tend to coincide with elevated turbidity and nutrient pulses. Treating the missingness as ignorable would therefore bias the trophic-state estimates downward for stormy months [8]. The mutual information analysis flagged the nitrogen-chlorophyll-a pair as partially redundant ($MI = 0.47$), suggesting that one of the two could be dropped from a cost-reduced monitoring design without much loss of classification accuracy.

Limitations include the reliance on dichotomised indicator scores, which discards information about how far above or below threshold each observation falls. A graded-response IRT model could retain this ordinal information, at the cost of

more parameters. The 14 monitoring stations provide uneven spatial coverage of a lake spanning three countries, and interpolation to unsampled locations introduces additional uncertainty.

Conclusion

This research adapted three-parameter logistic item response theory to classify monthly eutrophication states in Lake Victoria using six water-quality indicators measured at 14 stations over 2010-2023. The IRT model correctly classified 89.3% of station-months on a two-year hold-out set, outperforming fixed-threshold classification by 11 percentage points and logistic regression by 6 percentage points.

Chlorophyll-a emerged as the most discriminating indicator, while pH was the weakest. Galois symmetry reduction accelerated the estimation algorithm by a factor of 12, and M-estimators guarded against outlier-driven bias at stations with sensor drift. Pattern-mixture models handled informative missingness. The radar chart comparison between 2017 and 2022 showed a basin-wide worsening of trophic state across all six indicators. These findings suggest that IRT—a tool designed for educational testing—can sharpen environmental classification when multiple imperfect indicators must be combined into a single index. Future extensions include graded-response models that use continuous indicator values, spatial IRT models that borrow strength across stations, and real-time deployment as a lake-wide eutrophication early warning system.

Acknowledgements

Funding Sources

This research was supported by internal funds at the Ankara National Institute of Mathematical Research. No external grant number applies.

Institutional Support

The Department of Combinatorics and Optimization provided access to computing infrastructure and software.

Contributions Not Qualifying for Authorship

The authors thank Dr. Elif Yilmaz for assistance with LVEMP data extraction.

References

1. de Ayala RJ. The Theory and Practice of Item Response Theory. New York: Guilford Press; 2009.
2. Embretson SE, Reise SP. Item Response Theory for Psychologists. Mahwah, NJ: Lawrence Erlbaum Associates; 2000.
3. Hecky RE, Mugidde R, Ramlal PS, Talbot MR, Kling GW. Multiple stressors cause rapid ecosystem change in Lake Victoria. *Freshwater Biology*. 2010;55(Suppl 1):19-42.
4. Carlson RE. A trophic state index for lakes. *Limnology and Oceanography*. 1977;22(2):361-369.
5. Artin M. Algebra. Upper Saddle River, NJ: Prentice Hall; 2011.
6. Mislevy RJ. Estimating latent distributions. *Psychometrika*. 1984;49(3):359-381.
7. Huber PJ, Ronchetti EM. Robust Statistics. New York: John Wiley and Sons; 2009.
8. Little RJA. Pattern-mixture models for multivariate incomplete data. *Journal of the American Statistical Association*. 1993;88(421):125-134.
9. Samejima F. Estimation of latent ability using a response pattern of graded scores. *Psychometrika Monograph Supplement*. 1969;34(4):1-97.
10. Cover TM, Thomas JA. Elements of Information Theory. New York: John Wiley and Sons; 2006.
11. Bock RD, Aitkin M. Marginal maximum likelihood estimation of item parameters. *Psychometrika*. 1981;46(4):443-459.
12. Verschuren D, Johnson TC, Kling HJ, Edgington DN, Leavitt PR, Brown ET, Talbot MR, Hecky RE. History and timing of human impact on Lake Victoria, East Africa. *Proceedings of the Royal Society B*. 2002;269(1488):289-294.
13. Talling JF. The phytoplankton of Lake Victoria (East Africa). *Archiv für Hydrobiologie*. 1966;Suppl 25:229-256.
14. Rasch G. Probabilistic Models for Some Intelligence and Attainment Tests. Copenhagen: Danish Institute for Educational Research; 1960.
15. Lord FM. Applications of Item Response Theory to Practical Testing Problems. Hillsdale, NJ: Lawrence Erlbaum Associates; 1980.
16. Wainer H, Dorans NJ, Flaugher R, Green BF, Mislevy RJ. Computerized Adaptive Testing: A Primer. New York: Routledge; 2000.
17. Baker FB, Kim SH. Item Response Theory: Parameter Estimation Techniques. Boca Raton: CRC Press; 2004.
18. Yen WM. Using simulation results to choose a latent trait model. *Applied Psychological Measurement*. 1981;5(2):245-262.