RESEARCH ARTICLE

autopsych: An R Shiny tool for the reproducible Rasch analysis, differential item functioning, equating, and examination of group effects

Matthew G. R. Courtney[1]*, **Kevin C. T. Chang**[2], **Bing Mei**[3], **Kane Meissel**[4], **Luke I. Rowe**[5], **Laila B. Issayeva**[6]

1 Graduate School of Education, Nazarbayev University, Nur-Sultan, Kazakhstan, **2** Department of Statistics, The University of Auckland, Auckland, New Zealand, **3** School of Foreign Languages, Henan University, Henan, China, **4** Faculty of Education and Social Work, The University of Auckland, Auckland, New Zealand, **5** National School of Education, Australian Catholic University, Melbourne, Australia, **6** Center for Pedagogical Measurements, Nazarbayev Intellectual Schools, Nur-Sultan, Kazakhstan

* matthew.courtney@nu.edu.kz

Citation: Courtney MGR, Chang KCT, Mei B, Meissel K, Rowe LI, Issayeva LB (2021) autopsych: An R Shiny tool for the reproducible Rasch analysis, differential item functioning, equating, and examination of group effects. PLoS ONE 16(10): e0257682. https://doi.org/10.1371/journal.pone.0257682

Editor: Karl Bang Christensen, University of Copenhagen, DENMARK

Received: February 19, 2021

Accepted: September 7, 2021

Published: October 11, 2021

Data Availability Statement: All source code is available from Github: github.com/kcha193/AutoPsychShiny.

Funding: The authors received no specific funding for this work.

Competing interests: The authors have declared that no competing interests exist.

Abstract

In this paper, we present **autopsych**, a novel online tool that allows school assessment experts, test developers, and researchers to perform routine psychometric analyses and equating of student test data and to examine the effect of student demographic and group conditions on student test performance. The app extends current open-source software by providing (1) extensive embedded result narration and summaries for written reports, (2) improved handling of partial credit data via customizable item-person Wright maps, (3) customizable item- and person-flagging systems, (4) item-response theory model constraints and controls, (5) many-facets Rasch analysis to examine item bias, (6) Rasch fixed item equating for mapping student ability across test forms, (7) tabbed spreadsheet outputs and immediate options for secondary data analysis, (8) customizable graphical color schemes, (9) extended ANOVA analysis for examining group differences, and (10) inter-rater reliability analyses for the verifying the consistency of rater scoring systems. We present the app's architecture and functionalities and test its performance with simulated and real-world small-, medium-, and large-scale assessment data. Implications and planned future developments are also discussed.

1 Introduction

Assessment plays a central role in society. While assessment is common to the educational sector, it is also important to clinical contexts and personnel selection. In educational contexts, particular physical, intellectual, and moral states are demanded of students by the unique political, social, and educational milieu for which they are placed [1]. Correspondingly, the responsibility for measuring student ability or knowledge in a valid, reliable, and unbiased way [2] rests on the shoulders of those administering the assessments.

Recent advancements in the field of psychometrics have enabled school assessment experts, test developers, and researchers access to user-friendly psychometric software interfaces (e.g., [3]). The app presented in this paper extends current open-source software with a general focus on the traditional Rasch-based approach [4, 5] to test validation. In this paper we provide a general background to classical test theory (CTT) and Rasch modelling before providing more specific details of the ways in which the autopsych app extends current open-source software capabilities. We finish by testing the app's local and cloud-based performance under different data conditions, prior to providing some ideas for the app's future development.

1.1 Classical test theory

Classical test theory (CTT) is a traditional approach to examining the quality of an exam or scale based on the students' resultant data scoring patterns [6]. From within this framework, a question's difficulty is estimated by the number of students that respond incorrectly to that item. Likewise, student ability is estimated by the items that items that are marked correct for that student. In CTT, the quality and function of a question are assessed by the correlation between the students' response pattern for that item and the students' total scores in the test. The item-rest correlation reveals the strength of the correlation between the item-score and the test score (without the focal item), while the bi- and poly-serial correlation coefficients correct for attenuation of the strength of the correlation due to the limited number of response options in dichotomous and polytomous data, respectively [7]. Here, item-total correlations reveal the degree to which items "discriminate" between students with positive numbers viewed favorably, though negative estimates considered problematic and generally necessary to remove. Though not necessarily problematic, item-rest correlations below $r = .10$ may be considered small and flagged as lower performing [8]. Upper and lower confidence intervals (CIs) can be estimated for each correlation as an additional metric that accounts for the size of the student sample. Instances where lower 95% CIs (or more strict CIs) are below zero, though not necessarily problematic, may also be flagged as lower performing [9]. In CTT, the Cronbach's alpha reliability estimate is commonly employed to evaluate the overall internal consistency of a scale or test [10] with 0.50 = poor, 0.60 = questionable, 0.70 = acceptable, 0.80 = good, and 0.90 = excellent [11]. As an additional metric, the alpha reliability of the scale or test can also be examined upon removal of each respective item in the test. Where the reliability of the test improves (alpha increases) because of the removal of an item, though not necessarily problematic, provides another way to flag lower performing items in a test. The application of CTT provides a useful way to validate scales and tests. However, enabling users to automatically customize the flagging of items to their own specifications (i.e., item-rest correlations lower than a certain limit and the size of the CIs) would be useful.

1.2 Rasch modelling

The Rasch modelling framework was developed in the 1950s by the Danish mathematician, George Rasch. The framework was first applied to examine the quality of student achievement testing among Danish school children, and intelligence testing among Danish military personnel. Today, the framework is widely applied to the social sciences for analysis of traits and has also been adopted in clinical and public health research for the measurement of diverse outcomes [12].

The Rasch model provides a probabilistic interpretation of student competence. Such interpretations have been the basis for theoretical links between "assessment, teaching and learning, curriculum resources, and policy development" [13, p87]. In the case of dichotomous scoring outcomes (i.e., 0, incorrect, and, 1, correct), the Rasch model expresses student performance as,

$$P(X_{ni} = 1|\theta_n, \delta_i) = \frac{e^{D(\theta_n - \delta_i)}}{1 + e^{D(\theta_n - \delta_i)}} \tag{1}$$

where, θ_n is the person ability, δ_i is item difficulty (reflecting, in a binary sense, "the nature of the trait", (13, p87-88), e is the mathematical constant, and D is the scaling constant 1.702 for matching logistic and probit metrics very closely [14]; removed in subsequent formula for simplicity). To note, where $\theta_n = \delta_i$, the probability of a student's success on the item is $p = .50$, the point at which, conceivably, a student may be functioning within his or her zone of proximal development [13].

To account for non-binary conceptualizations of student proficiency (i.e., 0, 1, 2 scoring and rubric-based scoring), Rasch modelling was extended to account for partial credit scored test questions and polytomously-scored developmental criteria [15, 16]. Here, in a general sense, the dichotomous Rasch model was extended to account for polytomous scoring by representing each pair of adjacent scores as a string of ordered categories [16],

$$P(X_{nij} = 1|\theta_n, \delta_{ij}) = \frac{e^{(\theta_n - \delta_{ij})}}{1 + e^{(\theta_n - \delta_{ij})}}, \; j = 1,\ 2,\ \ldots,\ m_i \tag{2}$$

where j specifies the step to be taken by person n in item i from the lower of the adjacent score category to the higher category. Specifically, instead of defining a single item response function for an item, the partial credit model (PCM) defines mi category-response functions for each item–each function representing the conditional probability of student n completing j steps in item i given that they complete either $j – 1$ of j [16].

Theorists [13] have argued that the sets of item criteria (or items) for which students have close to a .50 probability (50:50 odds; where $\theta_n = \delta_{ij}$) of success "can be linked to research about the development of *human beings* (emphasis added) and the role that formal education plays in the process" [p90]. Specifically, this zone of proximal development (ZPD) can be defined as,

> a state of readiness in which a student will be able to make certain kinds of conceptual connections, but not others; anything too simple for the student will quickly become boring; anything too difficult will quickly become demoralising.
>
> [17 p122]

Here, it is argued that test scores (i.e., theta, θ_n) should be interpreted as a starting point for instructional intervention, a ZPD point where students may be able to optimally improve with additional support. This reconceptualization of the test score as a point of intervention and instructional scaffolding can also be used to identify and build appropriate teaching resources and curricula policy [13, p90]. It has been common for these methodologies to be beyond the interest of the classroom teacher [4, p91]. This has been largely due to an inability to overcome the burden of purchasing proprietary software and the learning the programming language associated with learning how to use the software [18, p107].

However, today, psychometric software has become more user-friendly and ubiquitously accessible (see, for example [3]). Moreover, it is likely that schoolteachers with a background in Maths, Physics, Engineering, or Computer Science committed to improving the quality of assessments would be able to grasp and make use of such psychometric tools in an applied way

—more certainly after being provided with a careful exposition of the fundamental principles and methodology.

1.3 Current psychometric software systems

Multiple psychometric software packages exist today and can be classified as either commercial with ongoing licenses [e.g., 19, 20] or non-commercial [e.g., 21, 22]. Currently, only several commercially available systems, such as XCalibre [23], offer automatic embedded narration, that is conditionally worded descriptive paragraphs of the results of the analysis in output technical reports. Currently, the freely available cloud-based ShinyItemAnalysis [3] provides automated conventional psychometric analyses (classical test theory, CTT) alongside the implementation of 1, 2, 3, and 4PL item-response theory (IRT) models and PDF and HTML report options (For review of the alternative 2, 3, and 4PL modeling approaches, see [24]). The development of such open-source cloud-based psychometric software is an emerging interdisciplinary field, defined here as “Shiny Psychometrics”, encompassing (1) psychometrics (e.g., CTT, IRT), (2) data science (i.e., the implementation of algorithms to extract knowledge and insights from structured data), (3) computer science (e.g., cloud-based software development), and (4) learning sciences (e.g., the design of learning innovations for the improvement of instructional methodologies).

The **ShinyItemAnalysis** architecture (https://shiny.cs.cas.cz/ShinyItemAnalysis/) was created from what can be described as an *open-source R Shiny development framework* which enables (a) the immediate integration of efficient cutting-edge statistical and graphical functionality, (b) automatic cloud-based software version updates, and (c) dynamic front-end and comprehensive report rendering capabilities via **rmarkdown**. Considering the general trend and ongoing need in the industry, we chose to develop **autopsych** as a freely available, cloud-based psychometric software.

2 Materials and methods

The **autopsych** app point of difference is that it focuses on Rasch modelling, is more accessible to stakeholders, and provides multiple extended functionalities and user-customizations.

2.1 Rasch model focus

Like the **ShinyItemAnalysis**, the **autopsych** Shiny app presented here also provides the methodological exposition, analysis, and reproducible reporting of CTT and IRT-based analysis. However, **autopsych** has a particular focus on the PCM Rasch model given its flexibility, practicality to handle different data types, simplicity to provide sufficient item and person statistics (as opposed to 2PL and other IRT models [24]; and the restrictive rating scale model [25]), and broad utility for the measurement of student performance and growth. Masters and Wright [16] provide an eloquent description of the advantages of such models,

> The consequence of modelling operating curves to have the same slope is that a unit is defined which enables *all* parameter estimates to be expressed with respect to a common interval scale (which). . . supports the quantitative study of growth. . . (and that) there is no place in these models for schemes which try to assign best weights to items or to response categories [as in the 2PL]. Instead, these models provide coherent and verifiable support for the traditional measurement practice of forming raw scores by counting events. What these models add is control over this traditional (measurement) practice.
>
> [16 p542]

The **autopsych** app adds *ubiquitously accessible practitioner control* over some fundamental aspects of this traditional measurement practice. The app also provides a proof-of-concept for a cloud-based psychometric research platform dedicated to supporting high quality educational assessment and research into the role of individual differences and instructional practice on student learning.

2.2 Broad accessibility for developing countries

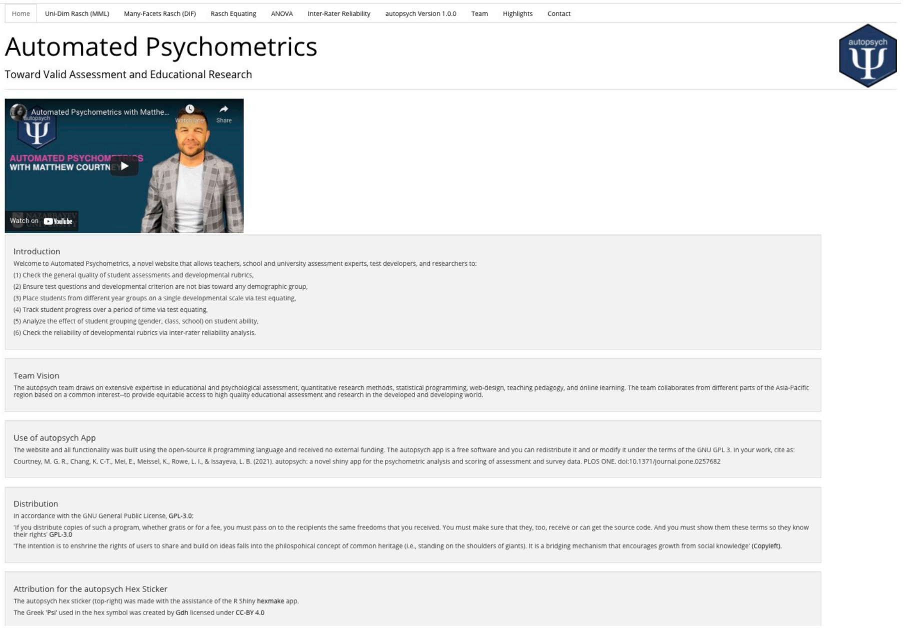
Even though psychometrics was born more than 300 years ago [26], its modern implementations are not ubiquitous worldwide. There are some countries where the application of psychometric methods is yet to be employed or has only recently been employed. For instance, in 2012, advanced psychometric methods were first applied for the Student Performance Monitoring system for Mathematics [27]. These innovations were made possible with the support of commercially provided psychometric consultancy. However, today there is a shift toward providing open-source software so that individuals and institutions can retain rights to the software and eventually continue to develop the software themselves.

2.3 Extended functionality

Specifically, **autopsych** contributes to the growing open-source R Shiny development framework by providing the following extended capabilities:

1. Comprehensive embedded result narration *and* summaries of key outcomes at the start of reports.
2. User-customised item and person flagging systems for identifying anomalous question and person response patterns.
3. Handling and exposition of dichotomous and partial credit data in user-customized Wright maps.
4. User-customized Rasch model constraints and controls.
5. Automated point-biserial orderedness analysis for a detailed examination of utility of polytomous (partial credit) scoring response categories.
6. Single multi-tabbed spreadsheet output providing immediate options for secondary data analysis.
7. Many-facets Rasch analysis for examining differential item functioning (item bias, or ‘invariance’ by groups of interest).
8. Fixed item equating option for mapping students onto different test forms via link items (for developing single ability scales across grades, and for the analysis of growth).
9. A one-way ANOVA tab for examining between-group effects on student ability (e.g., school and gender).
10. Inter-rater reliability analysis option for examining rater consistency in different rater scoring contexts.

The **autopsych** app is built on 31 packages listed in Table A1 in S1 Appendix. In Table A1 in S1 Appendix, for each package, the title, application in autopsych, and the license is also provided. Given that all of the dependent packages have a form of an open-source license, the authors of this paper also decided to ascribe the more recent GNU GPL v3 [28] license to **autopsych**.



Automated Psychometrics

Toward Valid Assessment and Educational Research

Introduction

Welcome to Automated Psychometrics, a novel website that allows teachers, school and university assessment experts, test developers, and researchers to:

(1) Check the general quality of student assessments and developmental rubrics,

(2) Ensure test questions and developmental criterion are not bias toward any demographic group,

(3) Place students from different year groups on a single developmental scale via test equating,

(4) Track student progress over a period of time via test equating,

(5) Analyze the effect of student grouping (gender, class, school) on student ability,

(6) Check the reliability of developmental rubrics via inter-rater reliability analysis.

Team Vision

The autopsych team draws on extensive expertise in educational and psychological assessment, quantitative research methods, statistical programming, web-design, teaching pedagogy, and online learning. The team collaborates from different parts of the Asia-Pacific region based on a common interest--to provide equitable access to high quality educational assessment and research in the developed and developing world.

Use of autopsych App

The website and all functionality was built using the open-source R programming language and received no external funding. The autopsych app is a free software and you can redistribute it and or modify it under the terms of the GNU GPL 3. In your work, cite as:

Courtney, M. G. R., Chang, K. C-T., Mei, E., Meissel, K., Rowe, L. I., & Issayeva, L. B. (2021). autopsych: a novel shiny app for the psychometric analysis and scoring of assessment and survey data. PLOS ONE. doi:10.1371/journal.pone.0257682

Distribution

In accordance with the GNU General Public License, **GPL-3.0**:

'If you distribute copies of such a program, whether gratis or for a fee, you must pass on to the recipients the same freedoms that you received. You must make sure that they, too, receive or can get the source code. And you must show them these terms so they know their rights' **GPL-3.0**

'The intention is to enshrine the rights of users to share and build on ideas falls into the philosopohical concept of common heritage (i.e., standing on the shoulders of giants). It is a bridging mechanism that encourages growth from social knowledge' **(Copyleft)**.

Attribution for the autopsych Hex Sticker

The autopsych hex sticker (top-right) was made with the assistance of the R Shiny **hexmake** app.

The Greek **'Psi'** used in the hex symbol was created by **Gdh** licensed under **CC-BY 4.0**

Fig 1. Homepage of autopsych.

https://doi.org/10.1371/journal.pone.0257682.g001

2.4 Exposition of the autopsych app

In this section, we give an exposition of the **autopsych** UI, customizable, reproducible analysis and reporting options for the uni-dimensional Rasch analysis, many-facets Rasch analysis, Rasch equating, one-way ANOVA, and inter-rater reliability analysis options. To complement this exposition, the **autopsych** app itself can also be accessed here: https://autopsych.shinyapps.io/version_1_0_0/.

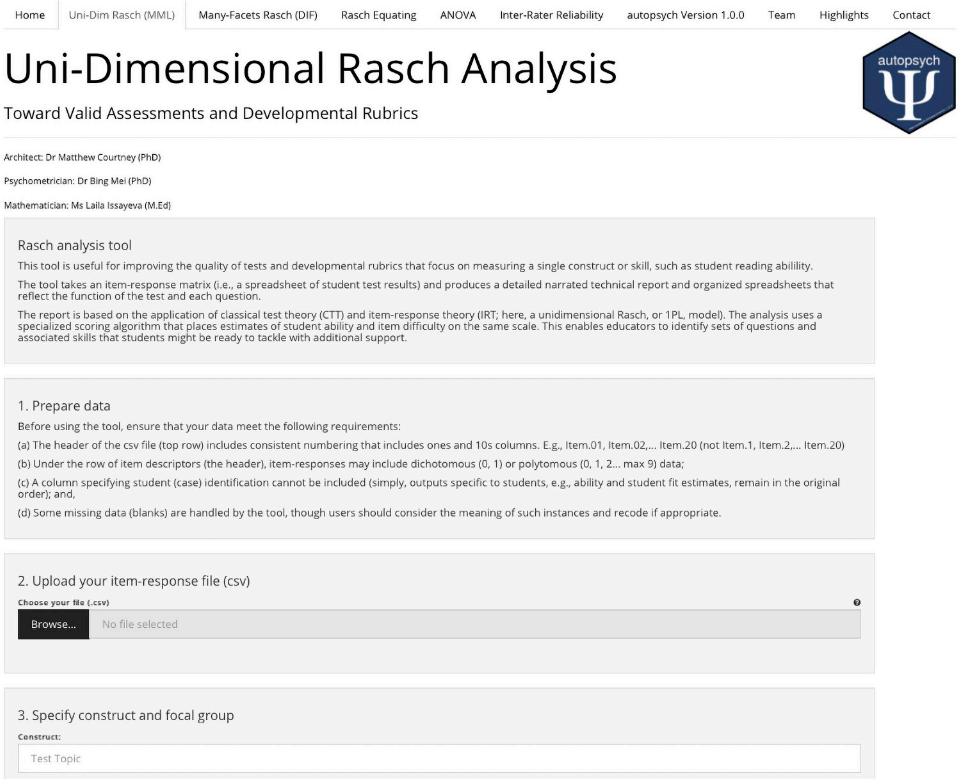
2.5 Home page and introduction to app

The home page of the **autopsych** R Shiny app (Fig 1) introduces six general functions of the app and the co-authors' vision for making quality assessment and educational research accessible to the developed and developing world.

2.6 Uni-dimensional Rasch analysis

After a general introduction and instructions on how users should prepare their data, the uni-dimensional Rasch tab (Fig 2) allows users to upload their item-response data. In the current Rasch framework, marginal maximum likelihood estimation (MML) is used. For this analysis, it is assumed that the items constitute a single uni-dimensional construct. For foundational empirical work on various tests of uni-dimensionality in IRT, see Hattie [29]; to implement several modern tests of uni-dimensionality in common statistical programs, see Courtney [30].

After uploading data, users specify the construct (e.g., "Numeracy") and focal group (e.g., "Auckland students"). Thereafter, users are able to specify the settings for the CTT and Rasch analysis of their data. Users can pre-specify settings for flagging items (panel 4 sliders) and persons (panel 5, "Flag cases. . ."). Rasch modelling options also include controls over whether or not the model imposes a constraint on cases (persons) or items (questions). Convergence criteria for the model and maximum number of iterations can also be controlled.



Uni-Dimensional Rasch Analysis

Toward Valid Assessments and Developmental Rubrics

Architect: Dr Matthew Courtney (PhD)

Psychometrician: Dr Bing Mei (PhD)

Mathematician: Ms Laila Issayeva (M.Ed)

Rasch analysis tool

This tool is useful for improving the quality of tests and developmental rubrics that focus on measuring a single construct or skill, such as student reading abilility.

The tool takes an item-response matrix (i.e., a spreadsheet of student test results) and produces a detailed narrated technical report and organized spreadsheets that reflect the function of the test and each question.

The report is based on the application of classical test theory (CTT) and item-response theory (IRT; here, a unidimensional Rasch, or 1PL, model). The analysis uses a specialized scoring algorithm that places estimates of student ability and item difficulty on the same scale. This enables educators to identify sets of questions and associated skills that students might be ready to tackle with additional support.

1. Prepare data

Before using the tool, ensure that your data meet the following requirements:

(a) The header of the csv file (top row) includes consistent numbering that includes ones and 10s columns. E.g., Item.01, Item.02,... Item.20 (not Item.1, Item.2,... Item.20)

(b) Under the row of item descriptors (the header), item-responses may include dichotomous (0, 1) or polytomous (0, 1, 2... max 9) data;

(c) A column specifying student (case) identification cannot be included (simply, outputs specific to students, e.g., ability and student fit estimates, remain in the original order); and,

(d) Some missing data (blanks) are handled by the tool, though users should consider the meaning of such instances and recode if appropriate.

2. Upload your item-response file (csv)

Choose your file (.csv)

Browse... No file selected

3. Specify construct and focal group

Construct:

Test Topic

Fig 2. Uni-dimensional Rasch tab.

https://doi.org/10.1371/journal.pone.0257682.g002

Users are also able to specify the graphical settings of the report and make their own notes (perhaps after reviewing the results of an initial report).

The methodology and results in the output technical report (PDF) adapt to account for the user settings. For example, the color and bin width of the Wright map (Fig 3) has been customized in Fig 4. To note, the Wright map provides an exposition of both dichotomous and partial credit scoring thresholds (using the S10 File).

The Wright map suggests that persons and items are matched quite well, though some more difficult item categories could conceivably be created to separate the very top ability students.

Beyond the illustration of relative student ability and item difficulty via the Wright map, a more formal exposition of test targeting and precision is provided by the test information and standard error ($se[\theta_n]$) curve graph (Fig 5, Appendix 5 of *output* PDF report). In this case of the dichotomous Rasch model, when θ_n denotes the maximum likelihood estimate,

$$se[\theta_n] = [\sum\nolimits_{i=1}^{I} \Pr(X_{ni} = 1|\theta_n, \delta_i)(1 - \Pr(X_{ni} = 1|\theta_n, \delta_i))]^{-1} \tag{3}$$

Conversely, when δ_i denotes the maximum likelihood estimate, $se[\delta_i]$ is given by

$$se[\delta_i] = [\sum\nolimits_{n=1}^{N} \Pr(X_{ni} = 1|\theta_n, \delta_i)(1 - \Pr(X_{ni} = 1|\theta_n, \delta_i))]^{-1} \tag{4}$$

With this important information, we can see that the accuracy of student ability estimates may be improved by (a) increasing the number of items in a test, and (b) improving test

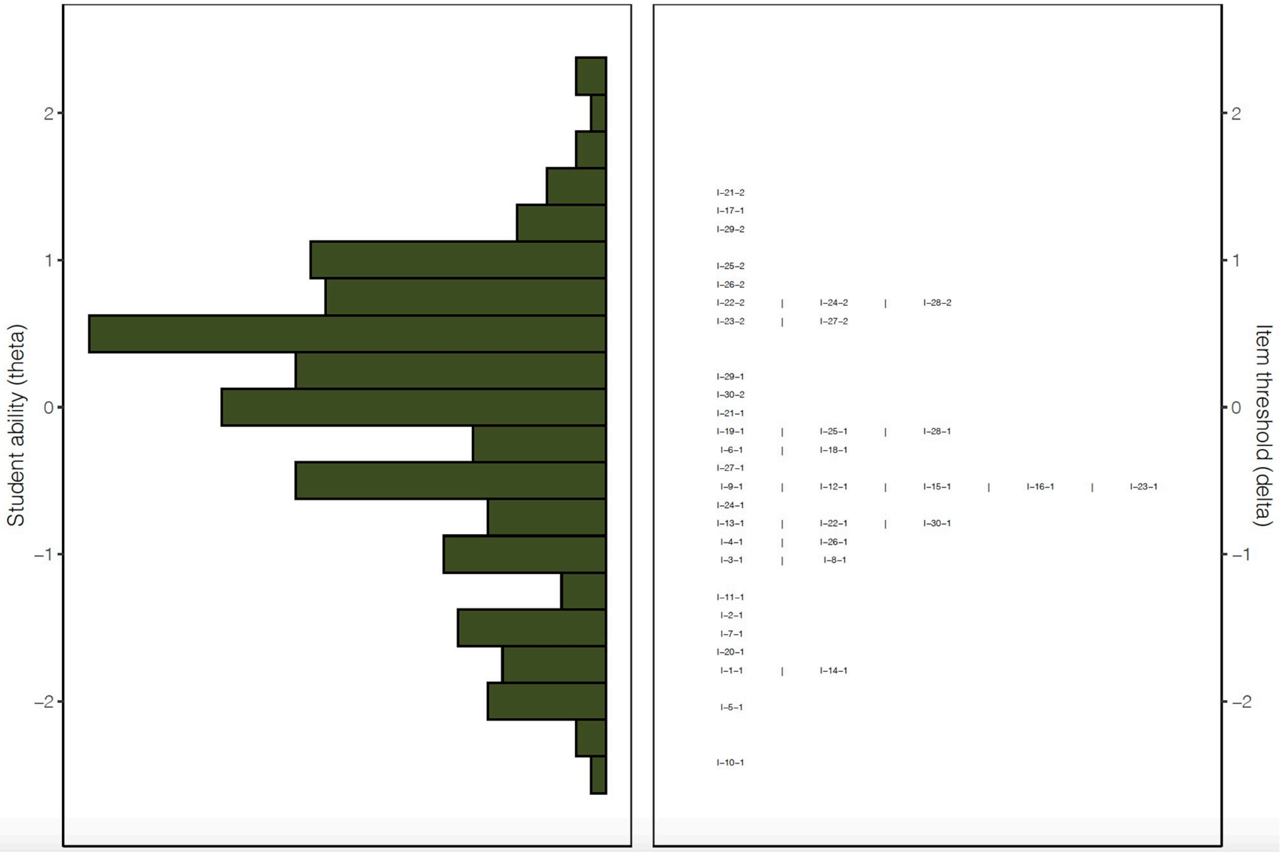


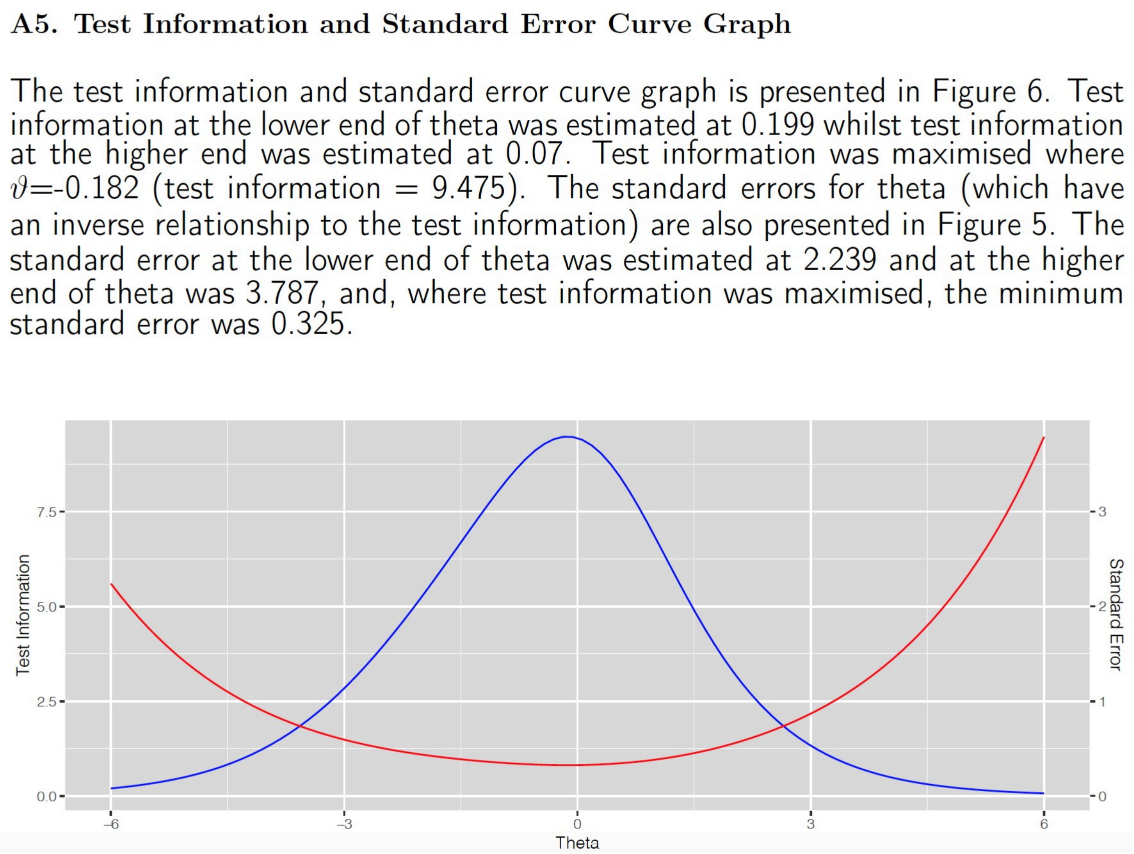
Fig 3. Exposition of dichotomous and partial credit thresholds in Wright map. *Note*. I-21-2 represents the item difficulty threshold for the partial credit scores of 0, 1, and 2.

https://doi.org/10.1371/journal.pone.0257682.g003

targeting for all ability levels. Conversely, the precision of item difficulty estimates, and associated resolution into how learning likely unfolds in a particular domain of interest, is improved with a larger and more developmentally-aligned calibrating student sample (see Fig 8, Item Results tab for results of **autopsych** implementation). These simple insights about test information and precision provide continuity between small- and large-scale assessment, and arguably, a more accessible and unifying theory of knowledge and measurement [1 p17].

The slight visual asymmetry in test targeting revealed by the Wright map is supplemented by the test information and standard error curve in Fig 5. The information (provided in Appendix 5 of the *output* PDF report produced for the software user) also includes automated embedded narration vis-à-vis test targeting and accuracy. In this instance, larger standard errors are exhibited at the upper end of the ability spectrum suggesting that the inclusion of slightly more difficult items in subsequent test forms may enable improved precision for higher performing students. For the current test, due to the relatively low number of items (20 dichotomous, 10 polytomous items; S10 File), person standard errors tend to be quite large (min = 0.325 to max 3.787), which may also prompt test designers to include more questions in subsequent tests.

The report also includes the orderedness of item-category point-biserial correlations and mean ability (theta) for each scoring category (e.g., 0, 1, 2). In this instance, the disorderedness in point-biserial correlations is automatically flagged red. However, in this instance, there is no disorderedness, therefore no categories are flagged. As an example, the final three rows of



A5. Test Information and Standard Error Curve Graph

The test information and standard error curve graph is presented in Figure 6. Test information at the lower end of theta was estimated at 0.199 whilst test information at the higher end was estimated at 0.07. Test information was maximised where ϑ=-0.182 (test information = 9.475). The standard errors for theta (which have an inverse relationship to the test information) are also presented in Figure 5. The standard error at the lower end of theta was estimated at 2.239 and at the higher end of theta was 3.787, and, where test information was maximised, the minimum standard error was 0.325.

Fig 4. Embedded narration, test information, and standard errors.

https://doi.org/10.1371/journal.pone.0257682.g004

the table give -0.467, 0.07, and 0.367 (Item OTV30 score categories 0, 1, and 2). The correlation between instances of the score 0 (coded "1") and student theta is, understandably negative (-0.467), while the correlation between instances of the score 2 (coded "1") is understandably positive (0.367). Consequently, for this item, we would expect the correlation between instances of the middle score category, score 1 (coded "1"), to lie somewhere between -0.467 and 0.367. Therefore, in this case, the item is not flagged. Mean ability for each scoring group is also given in the final column and flagging of disorderedness (reflecting more serious model violations) is also automated.

In addition to adaptive methodology and results, a summary of the settings of the report and the general results of the analysis are also provided in an appendix (Fig 6). The output here also includes a measure of overall model deviance with smaller values reflective of improved global model fit [31]. This will be useful when comparing the results of other models in future versions of the app.

The automatically output.xlsx file includes a collation of all key statistical results from the CTT and Rasch analysis and presents these results via 13 tabs (Fig 7).

Note that total score is a sufficient statistic for theta in the Rasch model (see highlight, Fig 7). Also, the S1 and S2 Files includes tabs for estimates of item difficulty and the full initial item-response matrix. Both of these data become important in the equating procedure illustrated in the Rasch equating section below.

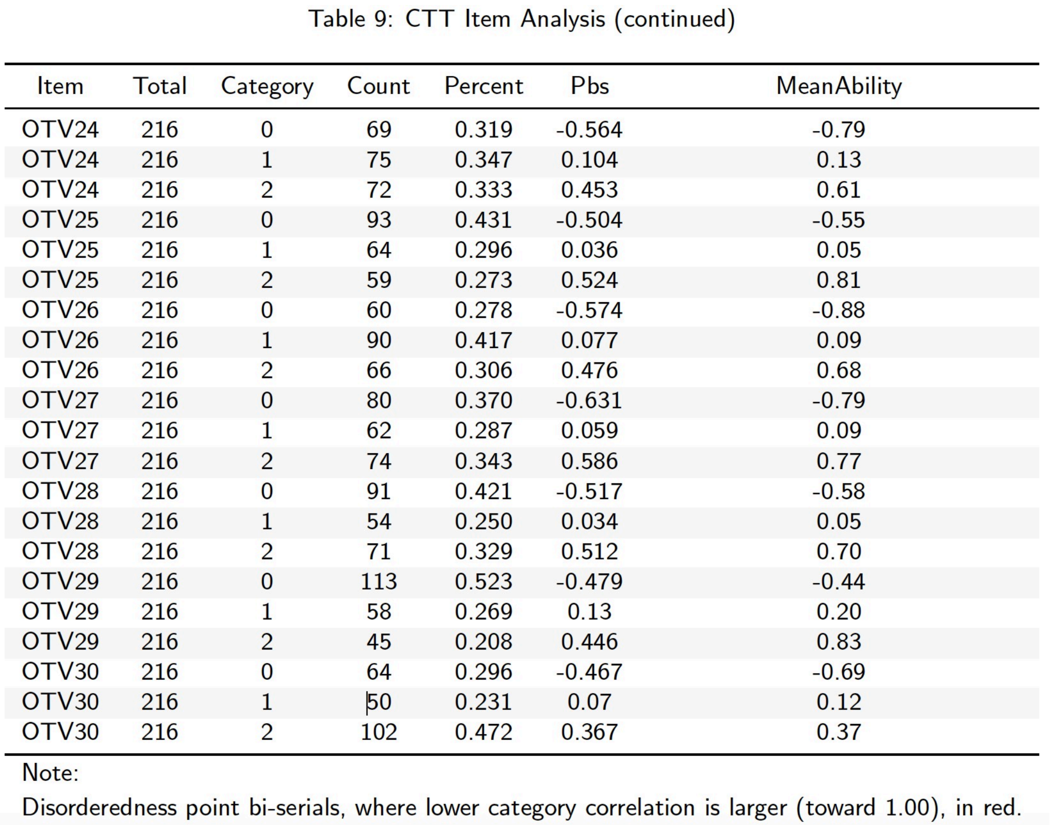


Table 9: CTT Item Analysis (continued)

Item	Total	Category	Count	Percent	Pbs	MeanAbility
OTV24	216	0	69	0.319	-0.564	-0.79
OTV24	216	1	75	0.347	0.104	0.13
OTV24	216	2	72	0.333	0.453	0.61
OTV25	216	0	93	0.431	-0.504	-0.55
OTV25	216	1	64	0.296	0.036	0.05
OTV25	216	2	59	0.273	0.524	0.81
OTV26	216	0	60	0.278	-0.574	-0.88
OTV26	216	1	90	0.417	0.077	0.09
OTV26	216	2	66	0.306	0.476	0.68
OTV27	216	0	80	0.370	-0.631	-0.79
OTV27	216	1	62	0.287	0.059	0.09
OTV27	216	2	74	0.343	0.586	0.77
OTV28	216	0	91	0.421	-0.517	-0.58
OTV28	216	1	54	0.250	0.034	0.05
OTV28	216	2	71	0.329	0.512	0.70
OTV29	216	0	113	0.523	-0.479	-0.44
OTV29	216	1	58	0.269	0.13	0.20
OTV29	216	2	45	0.208	0.446	0.83
OTV30	216	0	64	0.296	-0.467	-0.69
OTV30	216	1	50	0.231	0.07	0.12
OTV30	216	2	102	0.472	0.367	0.37

Note:
Disorderedness point bi-serials, where lower category correlation is larger (toward 1.00), in red.

Fig 5. CTT item analysis.

https://doi.org/10.1371/journal.pone.0257682.g005

Important to note is that the ability (theta) estimates for each student can then be matched with items or item categories that are comparable. In this instance, students would have a 50:50 odds of completing in the tab to the left, the ordered Thurstonian threshold tab (Fig 8). Essentially, the ordered Thurstonian threshold [32 p170] provides insight into the ordering and reasonable classification of skills in certain domains of learning, and can assist in the development of teaching resources. However, the development of structured teaching resources and curriculum policy becomes more viable when a large number of cases provide more certainty about the actual level of difficulty of skills, and, therefore, how learning likely unfolds in a particular domain (see Wu [33]) on forms of measurement, sampling, and equating errors.

As an example, student 8 who scored 26 out of 30 (theta = 0.29; Fig 8, Participant ID 8), with support from a teacher or peer, may be ready to work toward mastery of continuing decimalized counting patterns (I-23-2, Threshold = 0.59) and, in general, developing in the area of distinguishing fractional problems. Teaching resources and pedagogical approaches for different developmental levels might also be included in another column to the right.

2.7 Many-facets Rasch analysis

The many-facets Rasch analysis provides app users with the opportunity to test for item bias against particular student groups (e.g., females), or test for rater effects when the identity of the rater is known. This is done via the addition of an item by group interaction term in the Rasch

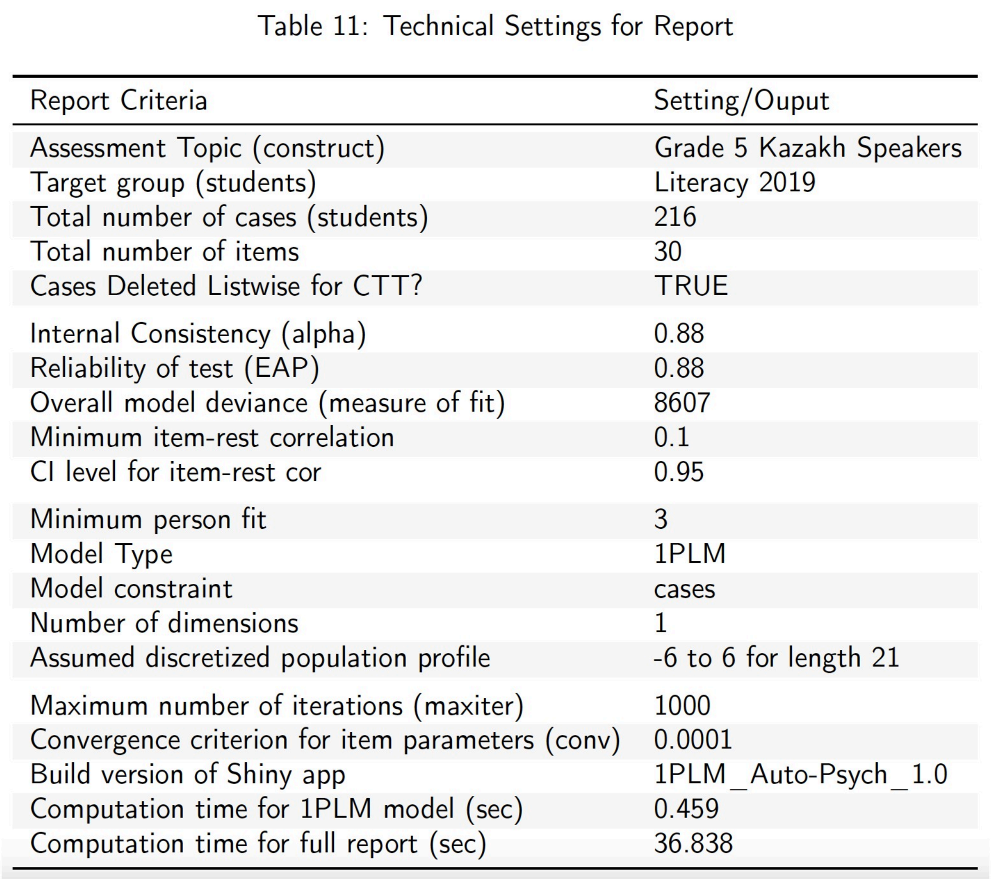


Table 11: Technical Settings for Report

Report Criteria	Setting/Ouput
Assessment Topic (construct)	Grade 5 Kazakh Speakers
Target group (students)	Literacy 2019
Total number of cases (students)	216
Total number of items	30
Cases Deleted Listwise for CTT?	TRUE
Internal Consistency (alpha)	0.88
Reliability of test (EAP)	0.88
Overall model deviance (measure of fit)	8607
Minimum item-rest correlation	0.1
CI level for item-rest cor	0.95
Minimum person fit	3
Model Type	1PLM
Model constraint	cases
Number of dimensions	1
Assumed discretized population profile	-6 to 6 for length 21
Maximum number of iterations (maxiter)	1000
Convergence criterion for item parameters (conv)	0.0001
Build version of Shiny app	1PLM_Auto-Psych_1.0
Computation time for 1PLM model (sec)	0.459
Computation time for full report (sec)	36.838

Fig 6. Summary of technical settings and results.

https://doi.org/10.1371/journal.pone.0257682.g006

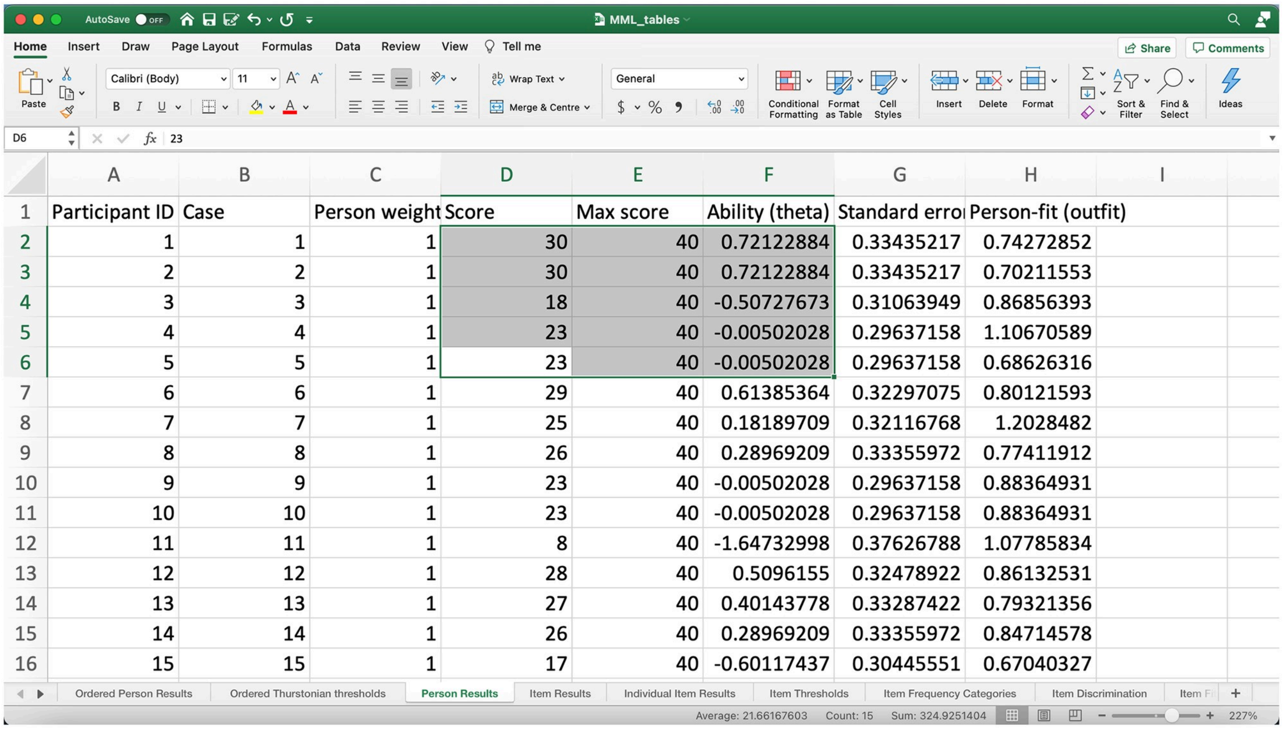
model. In the case of the dichotomous Rasch model, the facet model applied is given by,

$$P(X_{ni} = 1|\theta_n, \delta_i) = \frac{e^{(\theta_n - (\delta_i + G_g + D_{gi}))}}{1 + e^{(\theta_n - (\delta_i + G_g + D_{gi}))}} \quad (5)$$

where G_g represents the overall group effect on item difficulty, D_{gi} represents the student group (g) by item (i) interaction term with Eq 5 basically specifying that the probability of student success on each item depends on an adjustment to the difficulty of the item as a consequence of group membership. For example, if the item favors female students, the D_{gi} element (representing membership to the female group) will be negative, reflecting the fact that the item is easier for females.

Users upload the same type of item-response data to the many-facets Rasch tab though the first column of the csv file should include group membership (e.g., 1 = male, 2 = female). Additional user controls (Fig 9) in the facet tab UI provide options for setting statistical and practical significance of interest for the facets analysis.

In addition to all of the standard PDF and.xlsx outputs provided in the uni-dimensional Rasch MML tab, the Rasch facets tab provides dynamic tabular results (Fig 10) with highlighted elements of interest. In the case of the S5 File data (N = 1000), using standard statistical and practical settings, item I0006 is one of the items flagged (blue for practical and red for statistical significance) for being biased against male students (female1 = male, female2 = female; Fig 11). This result may warrant further investigation by the app user. Item



Participant ID	Case	Person weight	Score	Max score	Ability (theta)	Standard erro	Person-fit (outfit)
1	1	1	30	40	0.72122884	0.33435217	0.74272852
2	2	1	30	40	0.72122884	0.33435217	0.70211553
3	3	1	18	40	-0.50727673	0.31063949	0.86856393
4	4	1	23	40	-0.00502028	0.29637158	1.10670589
5	5	1	23	40	-0.00502028	0.29637158	0.68626316
6	6	1	29	40	0.61385364	0.32297075	0.80121593
7	7	1	25	40	0.18189709	0.32116768	1.2028482
8	8	1	26	40	0.28969209	0.33355972	0.77411912
9	9	1	23	40	-0.00502028	0.29637158	0.88364931
10	10	1	23	40	-0.00502028	0.29637158	0.88364931
11	11	1	8	40	-1.64732998	0.37626788	1.07785834
12	12	1	28	40	0.5096155	0.32478922	0.86132531
13	13	1	27	40	0.40143778	0.33287422	0.79321356
14	14	1	26	40	0.28969209	0.33355972	0.84714578
15	15	1	17	40	-0.60117437	0.30445551	0.67040327

Fig 7. Selection of tabs in output single multi-tabbed spreadsheet. Note. Student Score, Max score, and Ability (theta) is highlighted.

https://doi.org/10.1371/journal.pone.0257682.g007

I0010 is also flagged for being biased against females (though this effect does not reach statistical significance). A decision on retaining items exhibiting large DIF should be made carefully [32 p217-223].

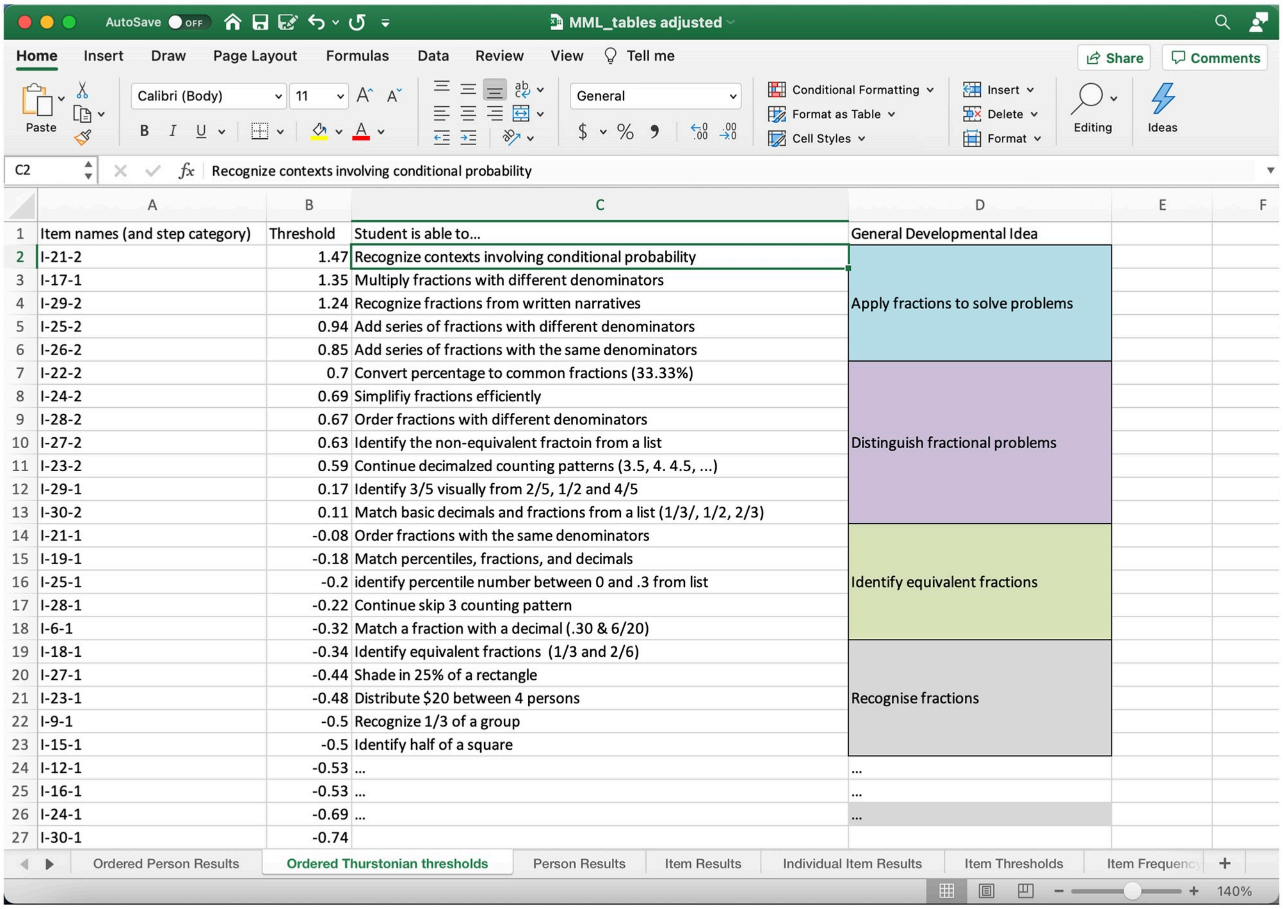
While the many facets tab helps users identify potentially biased items, the Rasch equating tab enables users to ensure that students who take different tests receive a fair score.

2.8 Rasch equating

Test equating is commonly carried out when two (or more) test forms are administered to different groups of students. For example, imagine a 40 item Numeracy test (Form A) is administered to a group of Grade 3 students. At the same time, another 40 item Numeracy test (Form B) is administered to a group of Grade 4 students. In order for both groups of students to receive a fair score on a single scale, the test designers built in some overlap where 10 link items (questions) are delivered in both Test Form A and B assessments (with link items generally a little difficult for Form A students, and easy for Form B students). In order to provide all of the students with a fair score on a single unified scale, one needs to carry out test equating.

Test equating is also carried out when you are tracking student progress across two time periods. Imagine delivering Test Form A at the start of a school year and Test Form B at the conclusion of a school year. Your aim is to provide stakeholders with an understanding of the extent to which each student improved for the given period. To provide students with a fair score for each time period on a unified scale, one needs to carry out test equating.

Here, we make fixed-anchor equating, a common and flexible form of equating, automatically accessible. To illustrate this, we make use of two datasets: S1 and S2 Files. These are simply re-labelled output files from two independent uni-dimensional Rasch analyses. We might



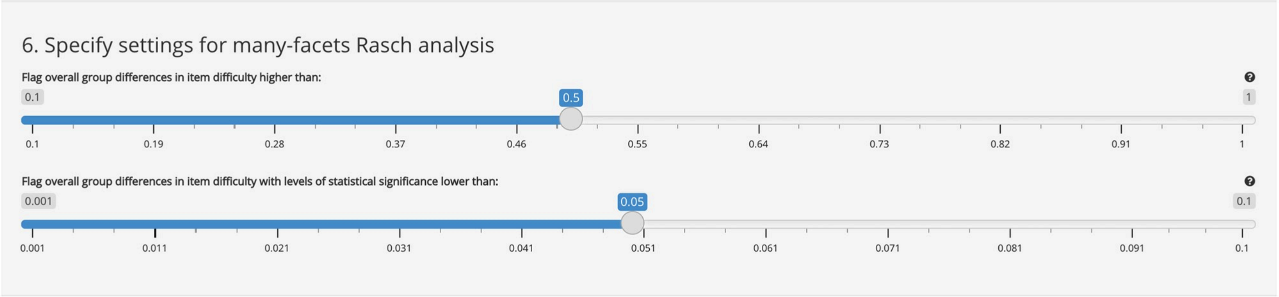
	A	B	C	D
1	Item names (and step category)	Threshold	Student is able to...	General Developmental Idea
2	I-21-2	1.47	Recognize contexts involving conditional probability	Apply fractions to solve problems
3	I-17-1	1.35	Multiply fractions with different denominators	
4	I-29-2	1.24	Recognize fractions from written narratives	
5	I-25-2	0.94	Add series of fractions with different denominators	
6	I-26-2	0.85	Add series of fractions with the same denominators	
7	I-22-2	0.7	Convert percentage to common fractions (33.33%)	Distinguish fractional problems
8	I-24-2	0.69	Simplifiy fractions efficiently	
9	I-28-2	0.67	Order fractions with different denominators	
10	I-27-2	0.63	Identify the non-equivalent fractoin from a list	
11	I-23-2	0.59	Continue decimalzed counting patterns (3.5, 4. 4.5, ...)	
12	I-29-1	0.17	Identify 3/5 visually from 2/5, 1/2 and 4/5	
13	I-30-2	0.11	Match basic decimals and fractions from a list (1/3/, 1/2, 2/3)	
14	I-21-1	-0.08	Order fractions with the same denominators	Identify equivalent fractions
15	I-19-1	-0.18	Match percentiles, fractions, and decimals	
16	I-25-1	-0.2	identify percentile number between 0 and .3 from list	
17	I-28-1	-0.22	Continue skip 3 counting pattern	
18	I-6-1	-0.32	Match a fraction with a decimal (.30 & 6/20)	
19	I-18-1	-0.34	Identify equivalent fractions (1/3 and 2/6)	Recognise fractions
20	I-27-1	-0.44	Shade in 25% of a rectangle	
21	I-23-1	-0.48	Distribute $20 between 4 persons	
22	I-9-1	-0.5	Recognize 1/3 of a group	
23	I-15-1	-0.5	Identify half of a square	
24	I-12-1	-0.53	...	...
25	I-16-1	-0.53	...	...
26	I-24-1	-0.69	...	...
27	I-30-1	-0.74		

Fig 8. Entering developmental descriptors for item categories. Note. Based upon an analysis of the skills necessary to complete the questions, app users can complete columns C pertaining to the skill associated with each question item of step, and may also like to categorise general developmental ideas, as in column D.

https://doi.org/10.1371/journal.pone.0257682.g008

imagine that Form A was associated with Grade 3 students while form B was associated with a Grade 4 group (with the link items labelled the same). For this analysis, users simply upload both.xlsx files, complete the settings, then run the analysis. The resultant PDF output produces a visualization to identify item difficulty invariance across test forms (Fig 11).

By inspecting Fig 11, we note that Item 37 may be non-invariant across test forms. For a more formal assessment of invariance, we can assess standardized differences in


6. Specify settings for many-facets Rasch analysis

Flag overall group differences in item difficulty higher than: 0.5

Flag overall group differences in item difficulty with levels of statistical significance lower than: 0.05

Fig 9. Additional user controls for facets analysis.

https://doi.org/10.1371/journal.pone.0257682.g009

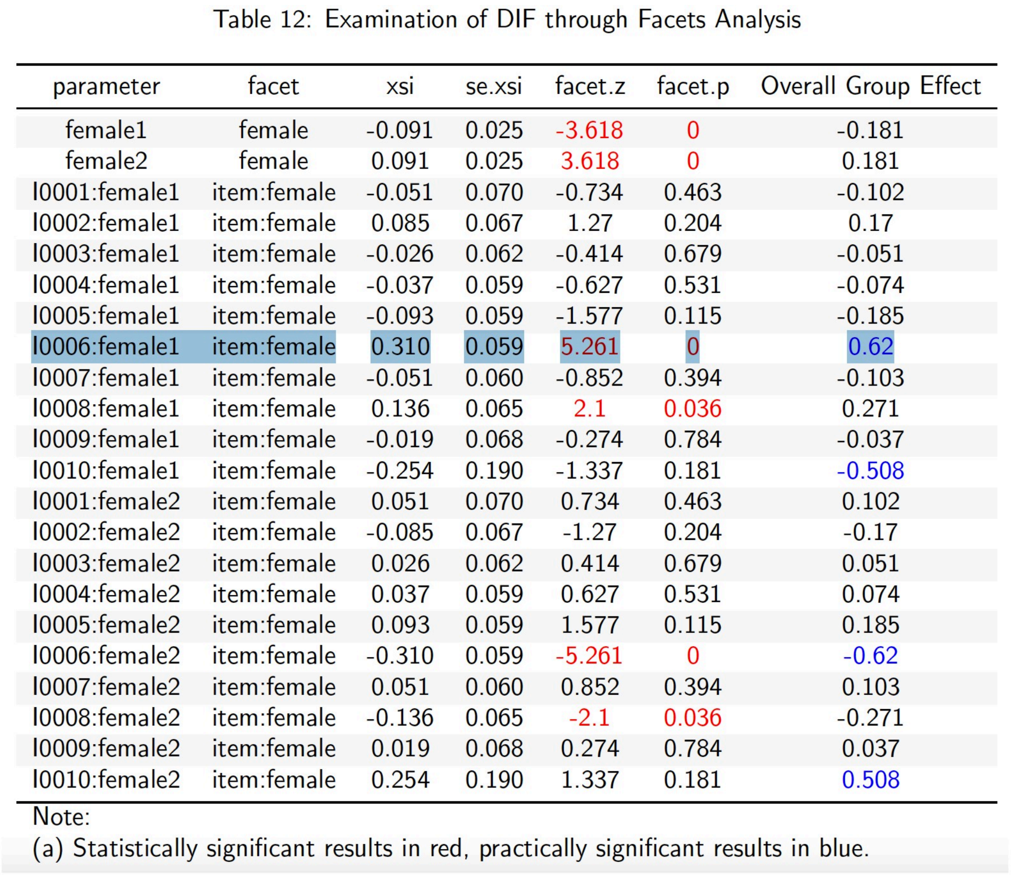


Table 12: Examination of DIF through Facets Analysis

parameter	facet	xsi	se.xsi	facet.z	facet.p	Overall Group Effect
female1	female	-0.091	0.025	-3.618	0	-0.181
female2	female	0.091	0.025	3.618	0	0.181
I0001:female1	item:female	-0.051	0.070	-0.734	0.463	-0.102
I0002:female1	item:female	0.085	0.067	1.27	0.204	0.17
I0003:female1	item:female	-0.026	0.062	-0.414	0.679	-0.051
I0004:female1	item:female	-0.037	0.059	-0.627	0.531	-0.074
I0005:female1	item:female	-0.093	0.059	-1.577	0.115	-0.185
I0006:female1	item:female	0.310	0.059	5.261	0	0.62
I0007:female1	item:female	-0.051	0.060	-0.852	0.394	-0.103
I0008:female1	item:female	0.136	0.065	2.1	0.036	0.271
I0009:female1	item:female	-0.019	0.068	-0.274	0.784	-0.037
I0010:female1	item:female	-0.254	0.190	-1.337	0.181	-0.508
I0001:female2	item:female	0.051	0.070	0.734	0.463	0.102
I0002:female2	item:female	-0.085	0.067	-1.27	0.204	-0.17
I0003:female2	item:female	0.026	0.062	0.414	0.679	0.051
I0004:female2	item:female	0.037	0.059	0.627	0.531	0.074
I0005:female2	item:female	0.093	0.059	1.577	0.115	0.185
I0006:female2	item:female	-0.310	0.059	-5.261	0	-0.62
I0007:female2	item:female	0.051	0.060	0.852	0.394	0.103
I0008:female2	item:female	-0.136	0.065	-2.1	0.036	-0.271
I0009:female2	item:female	0.019	0.068	0.274	0.784	0.037
I0010:female2	item:female	0.254	0.190	1.337	0.181	0.508

Note:
(a) Statistically significant results in red, practically significant results in blue.

Fig 10. Dynamic tabular results for facet analysis.

https://doi.org/10.1371/journal.pone.0257682.g010

corresponding items in accordance with,

$$z_i = \frac{\delta_i - \acute{\delta}_i}{\sqrt{se[\delta_i]^2 + se[\acute{\delta}_i]^2}} \tag{6}$$

where, z_i is the standardised delta difference for each corresponding item, δ_i is the item parameter for item i in Test Form A, and $\acute{\delta}_i$ is the corresponding item parameter for item i in Test Form B. Note that $\sum_{i=1}^{L}\left(\delta_i - \acute{\delta}_i\right) = 0$, where L is the total number of link items. Fig 12 illustrates the output associated with the formal check.

Further, the standardized error of equating can be calculated in accordance with,

$$\varepsilon = \frac{\sqrt{\frac{\sum_{i=1}^{L}(\delta_i - \acute{\delta}_i)^2}{L-1}}}{\sqrt{L}} \tag{7}$$

where ε is the equating error, L is the number of link items, δ_i is the item parameter for item i in Test Form A, and $\acute{\delta}_i$ is the corresponding item parameter for item i in Test Form B. Using the provided, the PDF report states that the standard error of equating is 0.042 logits.

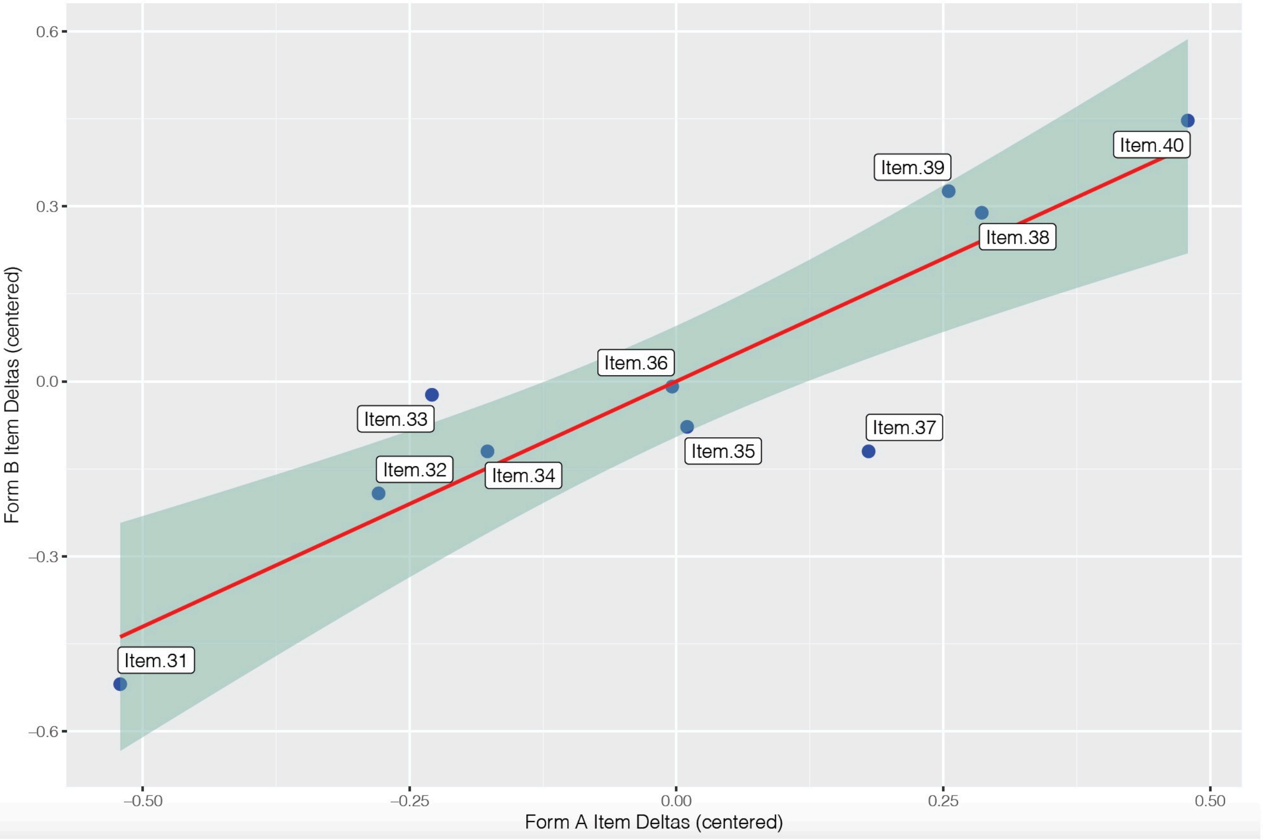


Fig 11. Item difficulty invariance visualization.

https://doi.org/10.1371/journal.pone.0257682.g011

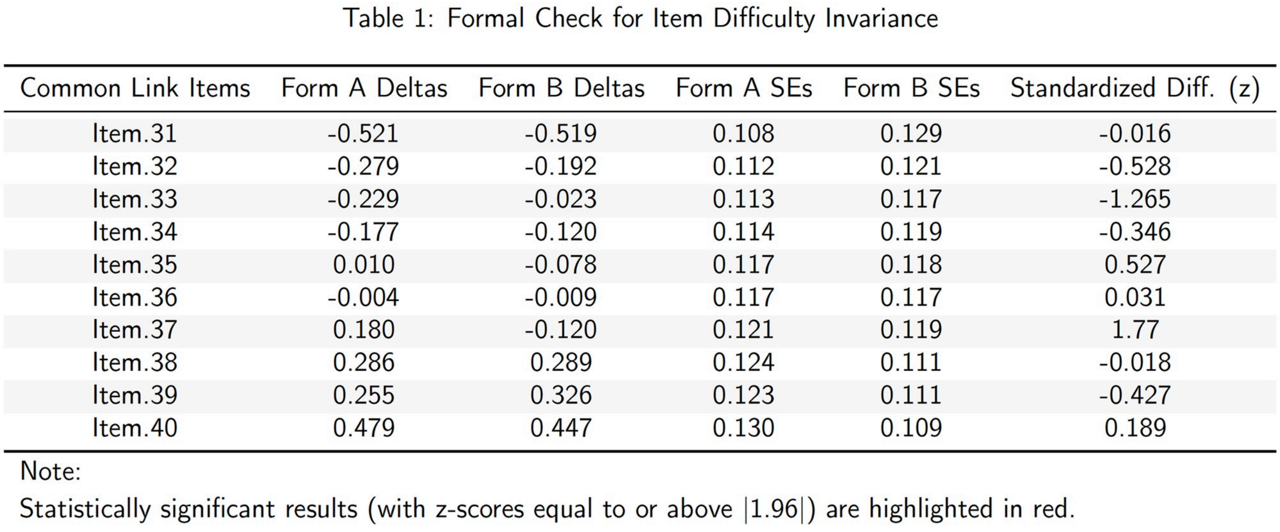


Table 1: Formal Check for Item Difficulty Invariance

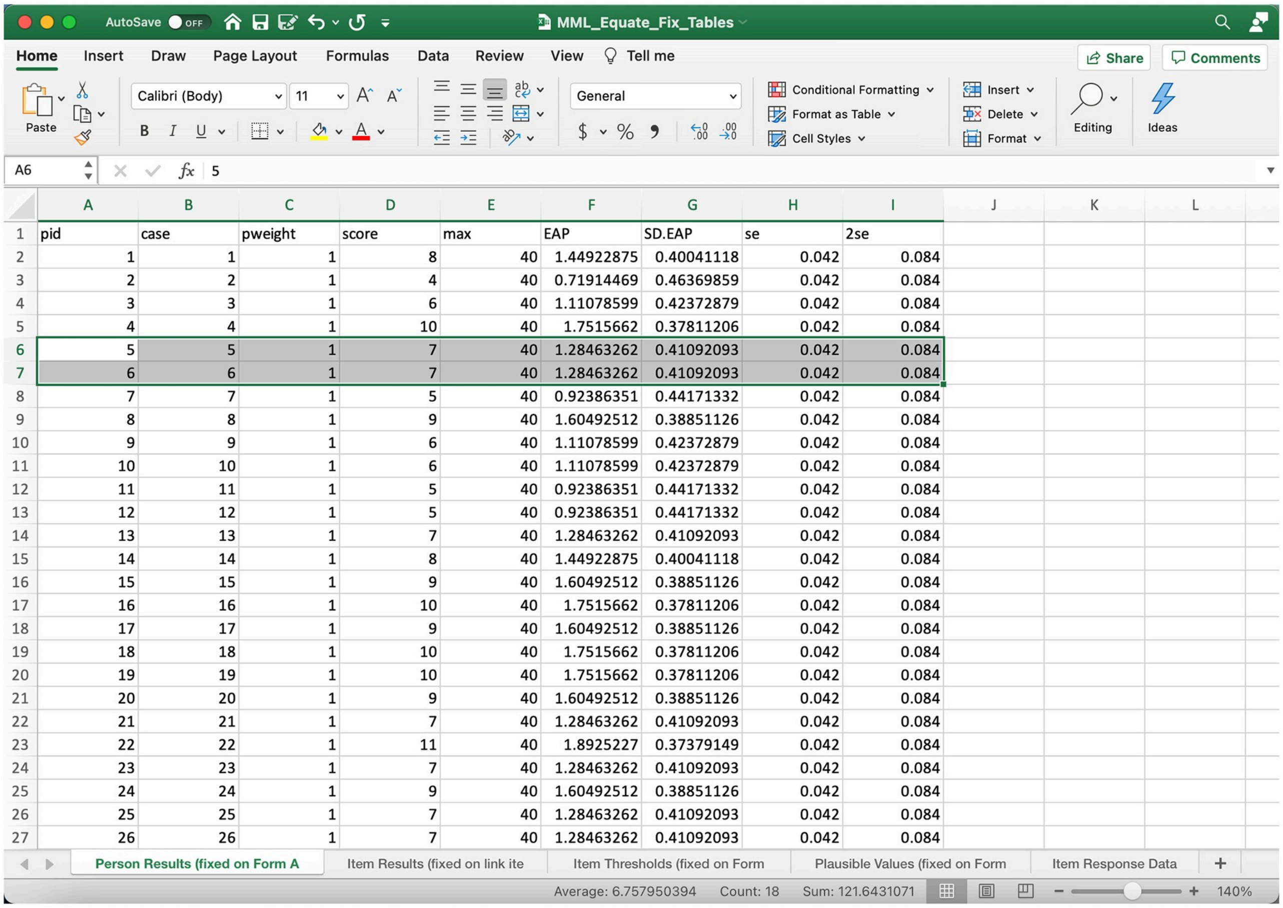
Common Link Items	Form A Deltas	Form B Deltas	Form A SEs	Form B SEs	Standardized Diff. (z)
Item.31	-0.521	-0.519	0.108	0.129	-0.016
Item.32	-0.279	-0.192	0.112	0.121	-0.528
Item.33	-0.229	-0.023	0.113	0.117	-1.265
Item.34	-0.177	-0.120	0.114	0.119	-0.346
Item.35	0.010	-0.078	0.117	0.118	0.527
Item.36	-0.004	-0.009	0.117	0.117	0.031
Item.37	0.180	-0.120	0.121	0.119	1.77
Item.38	0.286	0.289	0.124	0.111	-0.018
Item.39	0.255	0.326	0.123	0.111	-0.427
Item.40	0.479	0.447	0.130	0.109	0.189

Note:

Statistically significant results (with z-scores equal to or above |1.96|) are highlighted in red.

Fig 12. Formal test of link item invariance table.

https://doi.org/10.1371/journal.pone.0257682.g012



pid	case	pweight	score	max	EAP	SD.EAP	se	2se
1	1	1	8	40	1.44922875	0.40041118	0.042	0.084
2	2	1	4	40	0.71914469	0.46369859	0.042	0.084
3	3	1	6	40	1.11078599	0.42372879	0.042	0.084
4	4	1	10	40	1.7515662	0.37811206	0.042	0.084
5	5	1	7	40	1.28463262	0.41092093	0.042	0.084
6	6	1	7	40	1.28463262	0.41092093	0.042	0.084
7	7	1	5	40	0.92386351	0.44171332	0.042	0.084
8	8	1	9	40	1.60492512	0.38851126	0.042	0.084
9	9	1	6	40	1.11078599	0.42372879	0.042	0.084
10	10	1	6	40	1.11078599	0.42372879	0.042	0.084
11	11	1	5	40	0.92386351	0.44171332	0.042	0.084
12	12	1	5	40	0.92386351	0.44171332	0.042	0.084
13	13	1	7	40	1.28463262	0.41092093	0.042	0.084
14	14	1	8	40	1.44922875	0.40041118	0.042	0.084
15	15	1	9	40	1.60492512	0.38851126	0.042	0.084
16	16	1	10	40	1.7515662	0.37811206	0.042	0.084
17	17	1	9	40	1.60492512	0.38851126	0.042	0.084
18	18	1	10	40	1.7515662	0.37811206	0.042	0.084
19	19	1	10	40	1.7515662	0.37811206	0.042	0.084
20	20	1	9	40	1.60492512	0.38851126	0.042	0.084
21	21	1	7	40	1.28463262	0.41092093	0.042	0.084
22	22	1	11	40	1.8925227	0.37379149	0.042	0.084
23	23	1	7	40	1.28463262	0.41092093	0.042	0.084
24	24	1	9	40	1.60492512	0.38851126	0.042	0.084
25	25	1	7	40	1.28463262	0.41092093	0.042	0.084
26	26	1	7	40	1.28463262	0.41092093	0.042	0.084

Fig 13. Output.xlsx file from fixed equating procedure.

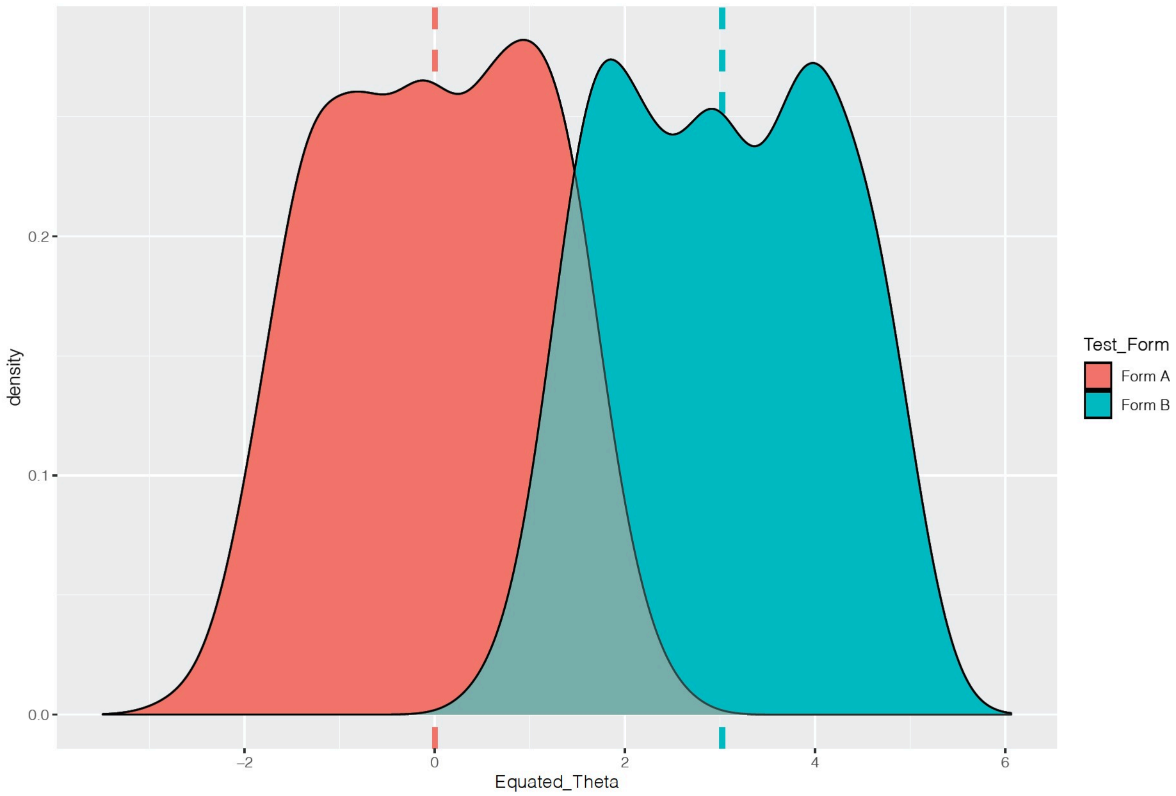
https://doi.org/10.1371/journal.pone.0257682.g013

The output.xlsx file (Fig 13) includes the person ability and item difficulty estimates for Test Form B students (fixed onto test Form A). As a consequence of test equating, the student ability estimates from both test forms can now be compared across test forms. Fig 14 provides a density plot for the Form B student ability estimates (aqua) equated with the ability estimates from Test Form A (red). This gives insight into the distributional properties of student ability form both test forms, and the breadth of student ability measured by both test forms.

2.9 One-way ANOVA

The app also provides an automated one-way ANOVA to examine the effect of student grouping on the student ability outcome of interest. The tab takes two data inputs: the first is the.xlsx output from the uni-dimensional Rasch tab; the second contains an.xlsx file containing the grouping variable(s) of interest. The one-way ANOVA function takes the ability theta variable from the person tab (example data: S4 File) and another file containing the grouping variables (S3 File). The S3 File (Fig 15) contains any number of grouping variables with each element corresponding to the same participant in the S4 File dataset.

The ANOVA function provides users with immediate results and options to examine the effects of the student grouping variable of interest on the student ability (theta). Fig 16 provides



Note. Dashed lines represent respective group means.

Fig 14. Density plot for equated student ability (theta).

https://doi.org/10.1371/journal.pone.0257682.g014

an example of the output boxplots in addition to normality checks, ANOVA tables, Estimated marginal means, and Tukey pairwise comparisons.

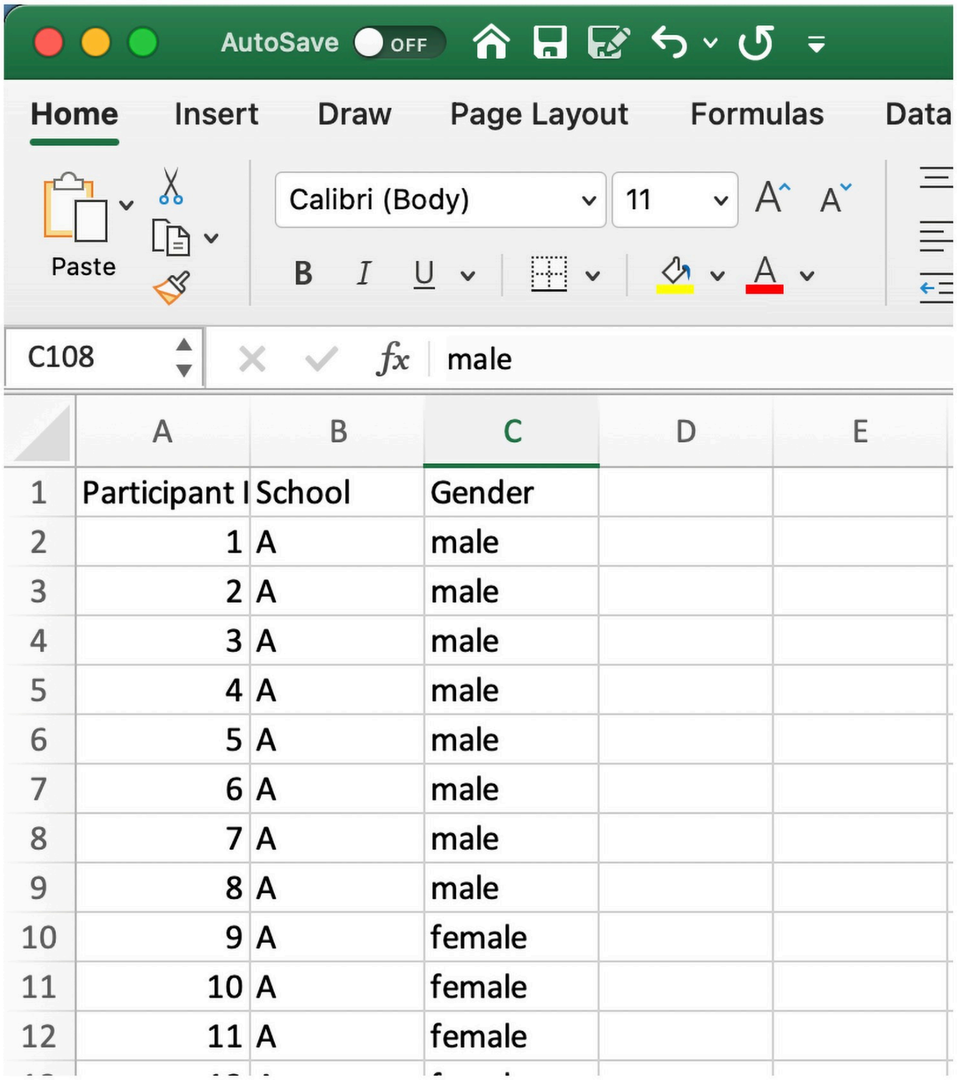
2.10 Inter-reliability analysis

Student products, such as essays and other written or oral performances, are often judged by multiple raters. One way to determine the consistency of marking (and, perhaps the utility of the instrument) is to estimate the level of inter-rater reliability from a set of data. One versatile way to determine rater consistency is to use the intra-class correlation coefficient. The **autopsych** app also provides this option (Fig 17).

Depending on the data conditions and focus of the analysis, users may specify the model, type, unit of analysis, and confidence interval for the intra-class correlation (ICC) statistic. Using the provided S8 File, and standard settings, a rendered report is produced. The report provides a technical summary (Fig 18) with all results in the first section while a more detailed description of the methodology adopted by the user is given in the latter sections.

3. App performance testing

App testing involved the use of simulated item-response data with different conditions using an *I* by *N* design (with *I* = items and *N* = student sample size). Four total item number conditions included 20, 40, 60, and 80 dichotomous items reflective of quite short school-based



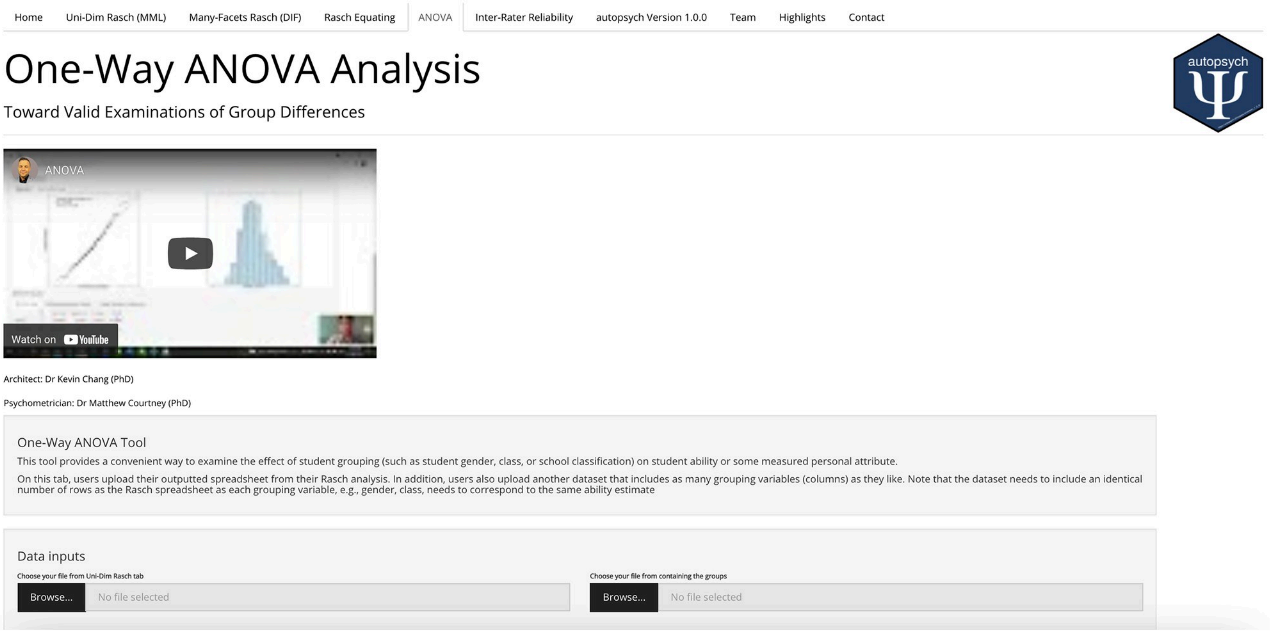
	A	B	C	D	E
1	Participant I	School	Gender		
2	1	A	male		
3	2	A	male		
4	3	A	male		
5	4	A	male		
6	5	A	male		
7	6	A	male		
8	7	A	male		
9	8	A	male		
10	9	A	female		
11	10	A	female		
12	11	A	female		

Fig 15. The Test_ANOVA_Group.xlsx file.

https://doi.org/10.1371/journal.pone.0257682.g015

assessments to more prolonged external assessments. Sample sizes included 50 (an approximate minimum for exploratory work), 500, 10,000, 50,000, 100,000, to 500,000 (similar to PISA). This meant that a total of 24 conditions were tested: 4(item) by 6(person) conditions (see S2 Appendix of this document for code used to simulate these data). Performance testing was not carried out for the many-facets and equating tab as these involved minor procedural extensions to the uni-dimensional tab and likely similar. In addition, the ANOVA and IRR functionality were not tested as these procedures are far less computationally intensive.

The first tests were run locally on a 2.2 GHz Quad-Core Intel Core i7 processor with 16GB 1600 MHz DDR3 Ram using an Intel Iris pro 1536 MB graphics card. The second tests (Table 1, in brackets) were run on the online version: https://autopsych.shinyapps.io/version_1_0_0/ (with 8GB instance size, datasets with $I = 60$ $N = 50{,}000$ ran successfully, while larger



One-Way ANOVA Analysis

Toward Valid Examinations of Group Differences

Architect: Dr Kevin Chang (PhD)

Psychometrician: Dr Matthew Courtney (PhD)

One-Way ANOVA Tool

This tool provides a convenient way to examine the effect of student grouping (such as student gender, class, or school classification) on student ability or some measured personal attribute.

On this tab, users upload their outputted spreadsheet from their Rasch analysis. In addition, users also upload another dataset that includes as many grouping variables (columns) as they like. Note that the dataset needs to include an identical number of rows as the Rasch spreadsheet as each grouping variable, e.g., gender, class, needs to correspond to the same ability estimate

Data inputs

Choose your file from Uni-Dim Rasch tab

Browse... No file selected

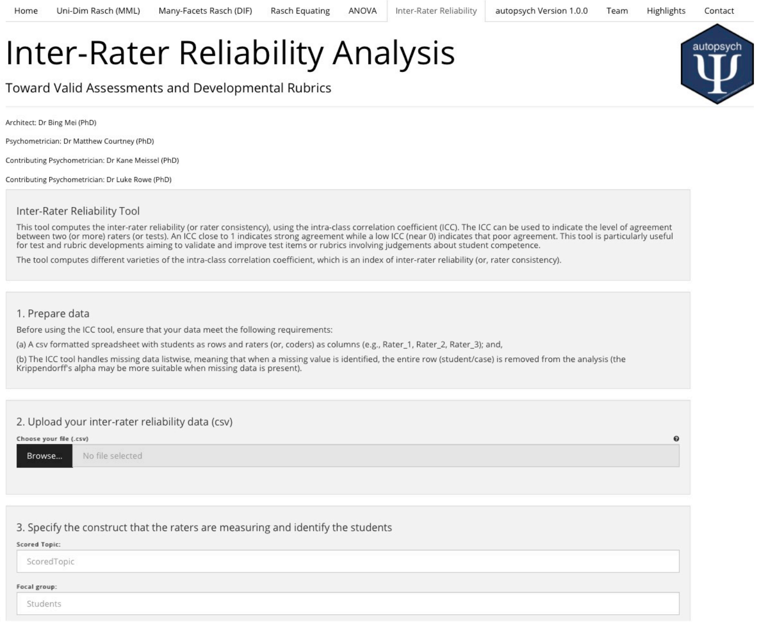
Choose your file from containing the groups

Browse... No file selected

Fig 16. One-way ANOVA function.

https://doi.org/10.1371/journal.pone.0257682.g016

datasets could only be processed locally with 16GB ram). All performance tests were run with standard settings. Results in minutes are presented in Table 1. Results suggest that the online version when *N* was less than 10,000, reports take less than two minutes to produce, though computation time balloons to over an hour and a half with 80 items and 500,000 students.



Inter-Rater Reliability Analysis

Toward Valid Assessments and Developmental Rubrics

Architect: Dr Bing Mei (PhD)

Psychometrician: Dr Matthew Courtney (PhD)

Contributing Psychometrician: Dr Kane Meissel (PhD)

Contributing Psychometrician: Dr Luke Rowe (PhD)

Inter-Rater Reliability Tool

This tool computes the inter-rater reliability (or rater consistency), using the intra-class correlation coefficient (ICC). The ICC can be used to indicate the level of agreement between two (or more) raters (or tests). An ICC close to 1 indicates strong agreement while a low ICC (near 0) indicates that poor agreement. This tool is particularly useful for test and rubric developments aiming to validate and improve test items or rubrics involving judgements about student competence.

The tool computes different varieties of the intra-class correlation coefficient, which is an index of inter-rater reliability (or, rater consistency).

1. Prepare data

Before using the ICC tool, ensure that your data meet the following requirements:

(a) A csv formatted spreadsheet with students as rows and raters (or, coders) as columns (e.g., Rater_1, Rater_2, Rater_3); and,

(b) The ICC tool handles missing data listwise, meaning that when a missing value is identified, the entire row (student/case) is removed from the analysis (the Krippendorff's alpha may be more suitable when missing data is present).

2. Upload your inter-rater reliability data (csv)

Choose your file (.csv)

Browse... No file selected

3. Specify the construct that the raters are measuring and identify the students

Scored Topic:

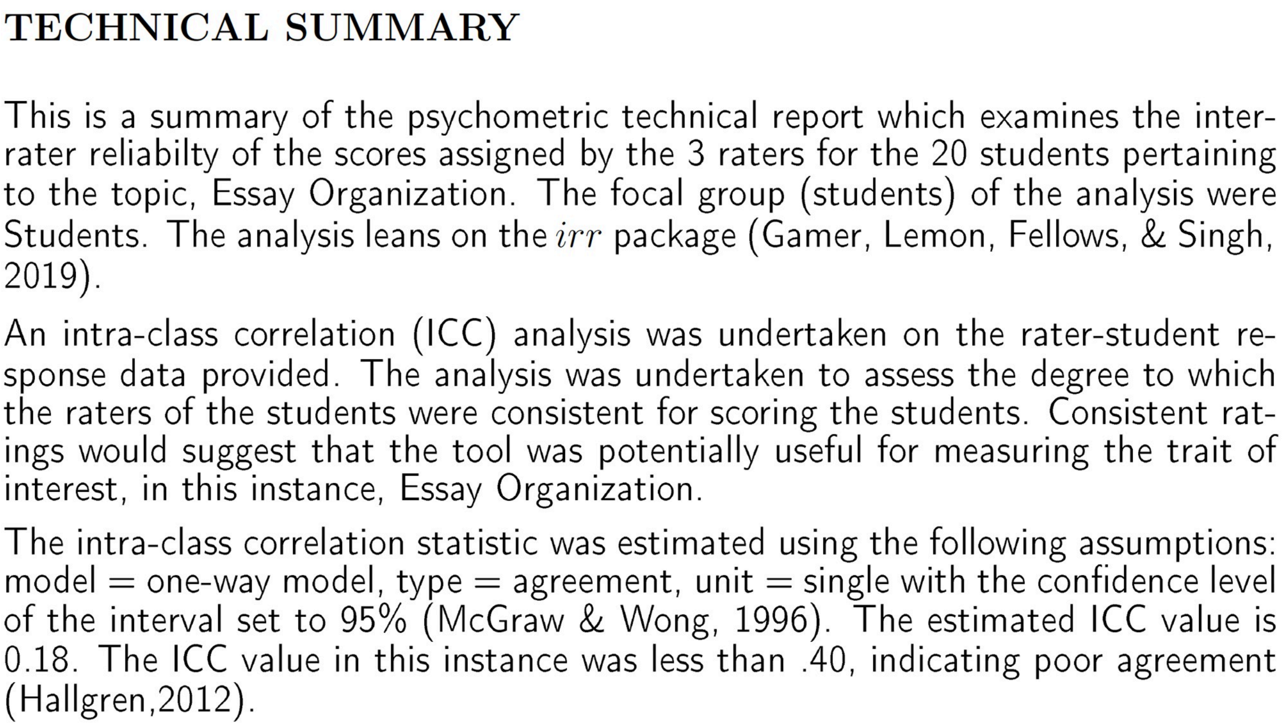
ScoredTopic

Focal group:

Students

Fig 17. Inter-rater reliability function.

https://doi.org/10.1371/journal.pone.0257682.g017



TECHNICAL SUMMARY

This is a summary of the psychometric technical report which examines the inter-rater reliabilty of the scores assigned by the 3 raters for the 20 students pertaining to the topic, Essay Organization. The focal group (students) of the analysis were Students. The analysis leans on the irr package (Gamer, Lemon, Fellows, & Singh, 2019).

An intra-class correlation (ICC) analysis was undertaken on the rater-student response data provided. The analysis was undertaken to assess the degree to which the raters of the students were consistent for scoring the students. Consistent ratings would suggest that the tool was potentially useful for measuring the trait of interest, in this instance, Essay Organization.

The intra-class correlation statistic was estimated using the following assumptions: model = one-way model, type = agreement, unit = single with the confidence level of the interval set to 95% (McGraw & Wong, 1996). The estimated ICC value is 0.18. The ICC value in this instance was less than .40, indicating poor agreement (Hallgren,2012).

Fig 18. Inter-rater technical summary example as provided by autopsych.

https://doi.org/10.1371/journal.pone.0257682.g018

Table 1. Uni-dimensional Rasch report computational time (in minutes).

Item/N Conditions	*N* = 50	*N* = 500	*N* = 10,000	*N* = 50,000	*N* = 100,000	*N* = 500,000
I = 20	0.36 (0.25)	0.39 (0.21)	0.62(0.52)	2.20(1.62)	4.18	47.24
I = 40	0.63 (0.39)	0.67 (0.37)	1.16(.84)	3.48(2.65)	7.23	61.23
I = 60	0.50 (0.48)	0.56 (0.52)	1.64(1.23)	5.35(3.85)	11.97	86.54
I = 80	1.17 (0.63)	1.23 (0.68)	1.97(1.52)	6.77	13.47	94.80

Note. Test times not in brackets pertain to locally run shiny app; test times in brackets pertain to those done online with the R Shiny app; all tests completed successfully; see Appendix B in S2 Appendix to replicate item-response data.

https://doi.org/10.1371/journal.pone.0257682.t001

4. Conclusions

The **autopsych** app promotes and makes accessible high quality educational assessment and related research into student learning. The platform makes use of CTT and Rasch-based modelling to (a) provide continuity between classroom-based and large-scale assessments, (b) make information about test quality immediately accessible, (c) provide teachers and learners with immediate feedback about what might be useful to teach/learn next, (d) support the establishment of unified test forms, and (e) enable the examination of the effects of student grouping on student ability. Though the app is certainly not comprehensive—point-and-click software will always be a step ahead due to significant licensing costs that help support development but also reduce access. However, the app provides a proof of concept for a more sophisticated and intelligent assessment system that places learners and the teachers as "the primary consumers and benefactors of the information derived from assessment" [34 p17].

4.1 Future directions

The app has the potential to evolve in a number of directions. The app's utility lies in its capacity to make powerful psychometric procedures, those typically only available proprietarily, ubiquitously accessible. Therefore, it is envisaged that future developments and derivative works will focus on expanding the app's Rasch-based and related functionality to account for nested data conditions common to educational research.

> A fundamental rule in technology says that whatever *can* be done *will be* done.
>
> [35 p46]

The convergence between the disciplines of psychometrics, data science, computer science, and the learning sciences is inevitable. However, these improved efficiencies also carry inherent risk. Potential blind user reliance on outputs and strict adherence to rules-of-thumb need to be countered providing users with not only source code, but also clear methodological expositions and referenced reading and learning material. The app **autopsych** app makes an attempt to do this.

4.2 Final thoughts

It should be noted that quality assessment starts at the early design stage with a foundational understanding of the various forms of validity, and the construction of items and developmental criteria that adequately, and representatively, sample the content area to be measured [2]. Part 1 of this important book [2] presents key concepts associated with validation, precision and errors of measurement, and fairness in testing—insights about tests worth making universally accessible. To this end, an *open-source R Shiny development framework* provides a state-of-the-art ecosystem for the on-going co-creation of a suite of user-friendly tools that contribute positively toward an expanded psychometric tool-box. Finally, it is the authors' view that the more ubiquitous application of improved measurement practice—beyond the fields of educational and health assessment [36] to the psychological sciences—may offer an important way out of the replicability crisis [37]. Early examples provide a useful food for thought [38].

Supporting information

S1 Appendix.
(DOCX)

S2 Appendix.
(DOCX)

S1 File.
(XLSX)

S2 File.
(XLSX)

S3 File.
(XLSX)

S4 File.
(XLSX)

S5 File.
(CSV)

S6 File.
(CSV)

S7 File.
(CSV)

S8 File.
(CSV)

S9 File.
(CSV)

S10 File.
(CSV)

Acknowledgments

The first author dedicates this paper to his teachers and thanks them for their encouragement and nudges in the right direction.

The **autopsych** app is intended to make high quality assessment and educational research accessible to the developed and developing world. All source code is available at the following repository and can be considered a public good. It can be shared, copied, and adapted in accordance with GNU GPL 3 [28].

Author Contributions

Conceptualization: Matthew G. R. Courtney.

Investigation: Matthew G. R. Courtney, Kevin C. T. Chang.

Methodology: Matthew G. R. Courtney, Kevin C. T. Chang, Bing Mei, Kane Meissel, Luke I. Rowe, Laila B. Issayeva.

Project administration: Matthew G. R. Courtney, Kevin C. T. Chang.

Software: Matthew G. R. Courtney, Kevin C. T. Chang, Bing Mei, Kane Meissel, Luke I. Rowe, Laila B. Issayeva.

Visualization: Matthew G. R. Courtney, Kevin C. T. Chang, Bing Mei.

Writing – original draft: Matthew G. R. Courtney, Kevin C. T. Chang, Bing Mei, Kane Meissel, Luke I. Rowe, Laila B. Issayeva.

Writing – review & editing: Matthew G. R. Courtney, Kevin C. T. Chang, Bing Mei, Kane Meissel, Luke I. Rowe, Laila B. Issayeva.

References

1. Ballantine J. The Sociology of Educational Assessment: A Systematic Analysis. 2nd ed. New Jersey, U.S., PrenticeHall, Inc.; 1983.
2. American Educational Research Association, American Psychological Association, National Council on Measurement in Education, editors. Standards for educational and psychological testing. Lanham, MD: American Educational Research Association; 2014.
3. Martinková P, Drabinová A. ShinyItemAnalysis for teaching psychometrics and to enforce routine analysis of educational tests. R J. 2018 Dec; 10(2): 503–515.

4. Rasch G. Probabilistic models for some intelligence and attainment tests. Copenhagen: Neilson & Lydiche; 1960.
5. Rasch G. Some probabilistic models for the measurement of attainment and intelligence. Chicago: MESA Press; 1980.
6. Cappelleri JC, Lundy JJ, Hays RD. Overview of classical test theory and item response theory for the quantitative assessment of items in developing patient-reported outcome measures. Clin. Ther. 2014 April; 36(5): 648–662. https://doi.org/10.1016/j.clinthera.2014.04.006 PMID: 24811753
7. Olssen U, Drasgow F, Dorans NJ. The polyserial correlation coefficient. Biometrika. 1982; 47: 337–347.
8. Cohen J. A Power Primer. Psychological Bulletin. 1992; 112: 155–159. https://doi.org/10.1037//0033-2909.112.1.155 PMID: 19565683
9. Cohen J, Cohen P, West SG, Aiken LS. Applied Multiple Regression/Correlational Analysis for the Behavioral Sciences. 3rd ed. Mahwah, NJ: Lawrence Erlbaum; 2003.
10. Cronbach LJ. Coefficient alpha and the internal structure of tests. Psychometrika. 1951; 16(3):297–334.
11. DeVellis RF. Scale Development: Theory and Applications. Thousand Oaks, LA: Sage; 2012.
12. McHorney CA, Monahan PO. Postscript: Applications of Rasch Analysis in Health Care. 2004; 42(Supplement): 1–73.
13. Griffin P. The comfort of competence and the uncertainty of assessment. Studies in Educational Evaluation. 2007 March; 33(1): 87–99.
14. Lord FM, Novick MR. Statistical theories of mental test scores. Reading, MA: Addison-Wesley; 1968.
15. Masters GN. A Rasch model for partial credit scoring. Psychometrika. 1982 June; 47: 149–174.
16. Masters GN, Wright BD. The essential process in a family of measurement models. Psychometrika. 1984 December; 49: 529–544.
17. Burbules NC. Savage inequalities: An interview with Jonathan Kozol. Education Theory. 1993 March; 43(1):43–55.
18. Kline T. J. B. (2005). Psychological testing: A practical approach to design and evaluation, Thousand Oaks, CA: Sage.
19. Linacre JM. Winsteps Rasch measurement computer program. Beaverton, Oregon: Winsteps.com; 2020.
20. Adams RJ, Wu ML, Cloney D, Wilson MR. ACER ConQuest: Generalised Item Response Modelling Software (Computer software). Version 5. Camberwell, Victoria: Australian Council for Educational Research; 2020.
21. Meyer JP. Applied Measurement with jMetrik. New York: Routledge; 2014.
22. Hanson BA. IRT Command Language (IRT). 2020. http://www.openirt.com/b-a-h/software/irt/icl/.
23. Assessment Systems Corporation. User's manual for the XCALIBRE marginal maximum-likelihood estimation program. St. Paul (MN): Assessment Systems Corp; 2021.
24. Magis D, Yan D, von Davier AA. Computerized Adaptive and Multistage Testing with R: Using Packages catR and mstR. Cham, Switzerland: Springer; 2017.
25. Andrich DA (1978). A rating formulation for ordered response categories. Psychometrika 43: 561–573.
26. Fullmer S, Daniel D. Psychometrics. In R. Kimmons & S. Caskurlu (Eds.), The Students' Guide to Learning Design and Research. EdTech Books; 2020. https://edtechbooks.org/studentguide/psychometrics.
27. Issayeva LB, Dieteren NMA, Crans SE. Assessment tool validation research at Nazarbayev Intellectual Schools: Student Performance Monitoring system for Mathematics. Paper presented at: The 18th Association for Educational Assessment in Europe Conference "Assessment cultures in a globalised world"; 2017 Nov 8–11; Prague, the Czech Republic. https://www.researchgate.net/publication/335033501.
28. GNU General Public License version 3. 2007 June. https://www.gnu.org/licenses/gpl-3.0.en.html.
29. Hattie J. An empirical study of various indices for determining unidimensionality. Multivariate Behav Res. 1984; 19(1): 49–78. https://doi.org/10.1207/s15327906mbr1901_3 PMID: 26776067
30. Courtney MGR. Determining the number of factors to retain in EFA: Using the SPSS R-Menu v2.0 to make more judicious estimations. 2013 April; 18(8): 1–15.
31. Wilson M, De Boeck P, Carstensen CH, "Explanatory Item Response Models: A Brief Introduction" in *Assessment of Competencies in Educational Contexts* edited by Hartig J., Klieme E., & Leutner D., 83–110. Göttingen: Hogrefe & Huber Publishers, 2008.

32. Wu M, Tam HP, Jen T-H. Educational Measurement for Applied Researchers: Theory into Practice. Singapore: Springer Nature; 2016.

33. Wu M. Measurement, sampling, and equating errors in large-scale assessments. *Educational Measurement*: *Issues and Practice*. 2010 Winter; 29(4): 15–27.

34. Pellegrino JW. The Evolution of Educational Assessment: Considering the Past and Imagining the Future. Princeton, NJ: Educational Testing Service; 2004.

35. Grove AS. Only the paranoid survive: How to exploit the crisis points that challenge every company and career. New York: Currency Doubleday; 1996.

36. Christensen KB, Kreiner S, Mesbah M. Rasch models in health. Hoboken, NJ: John Wiley & Sons; 2012.

37. Anvari F, Lakens D. The replicability crisis and public trust in psychological science. Comprehensive Results in Psychology. 2018 July; 3(3): 266–286.

38. Zanon C, Hutz CS, Yoo HH, Hambleton RK. An application of item response theory to psychological test development. Psicologia: Reflexão e Crítica. 2016 April; 29(18): 1–10.