Psychometrics Data Assignment 1

Joshua Damm

2024-09-26

1 Introduction

This report is investigating how two test questionnaires coming from a large-scale survey on social integration of young adults perform on measures of Classical Test Theory (CTT) and Item Response Theory (IRT). The two questionnaires aim at measuring impulsivity and general psychological health (GPH) through a series of Likert-scale items. In particular, the dataset utilized in this study contains data from 1,775 young adults aged 17 to 27 years in 1987. Impulsivity and GPH are crucial constructs in psychometric research as they are thought to give valuable insights into human behavior and understanding of psychopathology. In this report, we aim to evaluate the internal consistency of both tests, compute a variety of reliability estimates, and examine the measurement precision across different levels of the underlying latent traits. Moreover, aspects of construct validity, test improvement, and potential biases in test items with respect to background variables, such as gender, age, and study delay, are discussed. Finally, this reports discusses practical implications of psychometric modelling for research and clinical practice.

2 Methods

2.1 Dataset

The data for this study were drawn from a large-scale survey on social integration, conducted in 1987, with a sample of 1,775 young adults aged between 17 and 27 years. The sample included test measures of two psychological contructs, including impulsivity and general psychological health (GPH). Furthermore, demographic information were available for gender, age, and the number of years participants extended their studies (study delay). Both latent constructs were measured using two separate psychometric tests. Each test consists of multiple items rated on Likert-type scales. Impulsivity was measures on a 7-point Likert scale ranging from “not at all applicable” to “completely applicable”, including for example items about reactions in arbitrary situations that require action, speed and certainty in decision making, and dealing with the balance between decision speed and faultness. GPH was measured on a 4-point Likert scale ranging from “not at all” to “much more than usual”, including for example items about concentration, sleep quality, rumination, and perceived stress. After removing missing data, 1650 participants were included in the final analysis. For the impulsivity test, items 269 (“I feel comfortable in situations that need quick action”) and 279 (“I rather work fast with a higher chance of errors than slow but faultless”) had to be inverted because low scores indicated a high value on the latent factor. Later reliability analyses showed that these items were correlated with the first principle component.

2.2 Statistical analysis

Descriptive statistics were computed to ensure the absence of a sample bias. The internal consistency of both tests was assessed using Cronbach’s alpha and McDonald’s omega. Cronbach’s alpha was computed for both test scores to evaluate the internal consistency of the items within each test. McDonald's omega, which provides a more general measure of reliability by taking into account the factor structure of the test, was also calculated. The 90% confidence intervals for both alpha and omega were computed using a bootstrapping approach. For the investigation of the latent traits of impulsivity and GPH, a one-factor model (confirmatory) was computed. The factor variance was set to 1, following a congeneric model. To evaluate the fit of the one-factor model to the respective tests, the factor loadings and error variances were extracted. For the item-level analysis, Graded Response Models (GRM), and Partial Credit Models (PCM) were fit to allow for the assessment of item difficulty and discrimination parameters, therefore giving insight on how well the items distinguish between individuals at different levels of the latent traits. Lastly, we performed Differential Item Functioning (DIF) analysis, to examine whether any items from the impulsivity and GPH tests exhibited bias with respect to gender, age, or study delay. Age, and study delay were dichotomized using a median split.

3 Results

3.1 Classical Test Theory Analysis

When looking at the total test score, the maximum score for impulsivity is 42 (6 items x 7). The highest achieved test score was 39 and the lowest score was a 6. Participants (n = 1650) showed on average moderate impulsivity scores (mean = 25.61, SD = 5.99). For the individual items, regarding the possible score range from 1 to 7, items "S292" showed the highest average item score (mean = 4.62, SD = 1.71), whereas item "S255" showed the lowest item score (mean = 3.73, SD = 1.79). For an overview of all impulsivity item scores, also see the following table.

item	n	mean	sd	median	min	max	range	skew	kurtosis	se
S255	1650	3.73	1.79	4	1	7	6	0.15	-0.97	0.04
S264	1650	4.27	1.64	4	1	7	6	-0.19	-0.80	0.04
S269	1650	3.77	1.57	4	1	7	6	0.12	-0.59	0.04
S274	1650	4.43	1.65	4	1	7	6	-0.17	-0.81	0.04
S279	1650	4.78	1.83	5	1	7	6	-0.37	-0.94	0.05
S292	1650	4.62	1.71	5	1	7	6	-0.49	-0.68	0.04

The maximum score for general psychological health (GPH) is 44. The highest reported score on GPH was 42, and the lowest was 11. Subjects showed on average a moderate GPH (mean = 19.83, SD = 4.58). For the individual items, item "S331" showed the highest score (mean = 2.11, SD = 0.93), and item "S337" the lowest score (mean = 1.26, SD = 0.57). Interestingly, item "S337" showed the lowest item score, but also the highest skewness, indicating that many subjects scored low on this item and a few subjects showed very high scores, indicating that this item is well-suited for differentiating subjects across a subdomain of GPH. For an overview of all general psychological health item scores, also see the following table.

item	n	mean	sd	median	min	max	range	skew	kurtosis	se
S327	1650	2.11	0.58	2	1	4	3	0.55	1.39	0.01
S328	1650	1.75	0.86	2	1	4	3	0.83	-0.36	0.02
S329	1650	1.95	0.60	2	1	4	3	0.61	1.97	0.01
S330	1650	1.83	0.51	2	1	4	3	0.08	2.01	0.01
S331	1650	2.12	0.93	2	1	4	3	0.23	-1.03	0.02
S332	1650	1.70	0.81	2	1	4	3	0.97	0.23	0.02
S333	1650	1.96	0.67	2	1	4	3	0.44	0.50	0.02
S335	1650	1.84	0.87	2	1	4	3	0.68	-0.50	0.02

item	n	mean	sd	median	min	max	range	skew	kurtosis	se
S336	1650	1.40	0.66	1	1	4	3	1.55	1.71	0.02
S337	1650	1.26	0.57	1	1	4	3	2.43	6.03	0.01
S338	1650	1.93	0.55	2	1	4	3	0.33	1.71	0.01

When investigating the test reliabilities, the Cronbach-Guttman (C-G) reliabilities (alpha) were computed. For the impulsivity test in the present sample the C-G reliability is 0.62. This alpha indicates low internal consistency. The bootstrapped 90% CI for this alpha is [0.5898, 0.6487]. The C-G reliability for GPH in our sample was estimated as 0.81. The bootstrapped 90% CI for this alpha is [0.7915, 0.8239]. This indicates good internal consistency.

To investigate the item loadings and error variances, we computed a congeneric one-factor model. All items load on a single factor, and the factor has a variance of 1. In the following table you see the factor loadings and error variances for the impulsivity items. MacDonalds Omega was 0.64. This is a slightly higher estimated reliability than the C-G reliability (alpha).

The following table showes the factor loadings and error variances for the impulsivity items.

item	loadings	error
S255	0.5409152	0.7074107
S264	0.4782165	0.7713090
S269	0.3126451	0.9022530
S274	0.6320518	0.6005106
S279	0.3671040	0.8652347
S292	0.5379562	0.7106031

The following table showes the factor loadings and error variances for the GPH items.

item	loadings	error
S327	0.8612556	0.2582387
S328	0.6755192	0.5436738
S329	0.4825213	0.7671732
S330	0.4658737	0.7829617
S331	0.6762898	0.5426321
S332	0.8241274	0.3208141
S333	0.6844191	0.5315705
S335	0.8469466	0.2826815
S336	0.8041036	0.3534174
S337	0.7127335	0.4920109
S338	0.7199672	0.4816472

For general psychological health (GPH), all items load positively on the common factor. MacDonalds Omega was estimated as 0.92, which is significantly higher than Cronbach's Alpha. For the impulsivity items, removing any of the items, lowers the test reliabilty, which means that every single item is contributing to the internal consistency of the test. For the GPH items, removing any of the items does not lower the total reliabilty score either. The standard error of measurement (SEM) for the impulsivity test is 3.69. The length of the 90% CI for each test score is 12.15. Investigating the precision of the test, the standard error of measurement for the GPH test is 1.99. The length of the 90% CI for each test score is 6.55. Investigating the item constellation in both tests, it is assumed that all items are parallel (load equally on the one common factor / latent trait). For impulsivity, the Spearman-Brown prophecy formula with a desired C-G reliability of 0.80 suggests that adding 9 items to the current 6 items would reach the desired reliability, leading to a total test length of 15 items. Hypothetically, an item to add to the impulsivity test could be: “When I feel a desire for action, I am acting on it immediately.” For GPH, removing 1 item would lead to the desired reliabilty of 0.80, which is already pretty close to the current C-G reliabity of 0.81. It is difficult to decide to remove an item from the test, because no items are lowering the overall test reliabilty. However, since removing items S329 and S330 do not change alpha, it might be the most reasonable to remove them. The correlation between the observed test scores of the impulsivity and GPH test is 0.06497035. The disattenuated true score correlation between the to tests is 0.09178186.

3.2 Item-Response Analysis

Because of the ordered-categorical response categories, we need can fit a Graded response model (GRM) and Partial Credit Model (PCM) and see which fits best. For the impulsivity test, the graded respones model has a better model fit based on AIC, BIC and log-likelihood. For the GPH, GRM has a better fit as well. This makes sense as the GRM is the suited or ordered response categories where each item asks respondents to indicate their level of agreement. The discrimination parameter (a) reflects how well an item can differentiate between individuals with different levels of the latent trait. Higher values indicate better discrimination. The difficulty (threshold) parameter (b) indicates the point on the latent trait continuum where a respondent has a 50% chance of endorsing a particular response category. For polytomous models (like GRM and PCM), there will be multiple threshold parameters, one for each transition between the response categories.

The Graded Response Model (GRM) had the best fit to the items and therefore we will stick to this model for the paramter analysis. For the impulsivity test, items S255, S264, S279, and S292 rather show moderate values according to common conventions, which means that this item moderately differentiates between individuals with low and high values of the latent trait impulsivity (see also the following table. Item S269 shows low differentiabilty, whereas Item S274 shows high differentiability (slopes of the item characteristic curves). For the item difficulties, i.e. thresholds, the items show considerably differences among each other. Ranging across all category thresholds, item S269 (“I feel comfortable in situations that need quick action”) shows a high range over values of the latent trait impulsivity and big steps between the thresholds, indicating that it captures a lot of the variation in impulsivity. When comparing the other item difficulty ranges, they show much similarity. However the difficulty for the first threshold of items S255, S274, and S292 indicate that they do not adequately capture low values on impulsivity. Furthermore, items S274, S279, and S292 seem to not capture very high impulsivity values well (also see the following table).

Table 5: Parameters GRM impulsivity

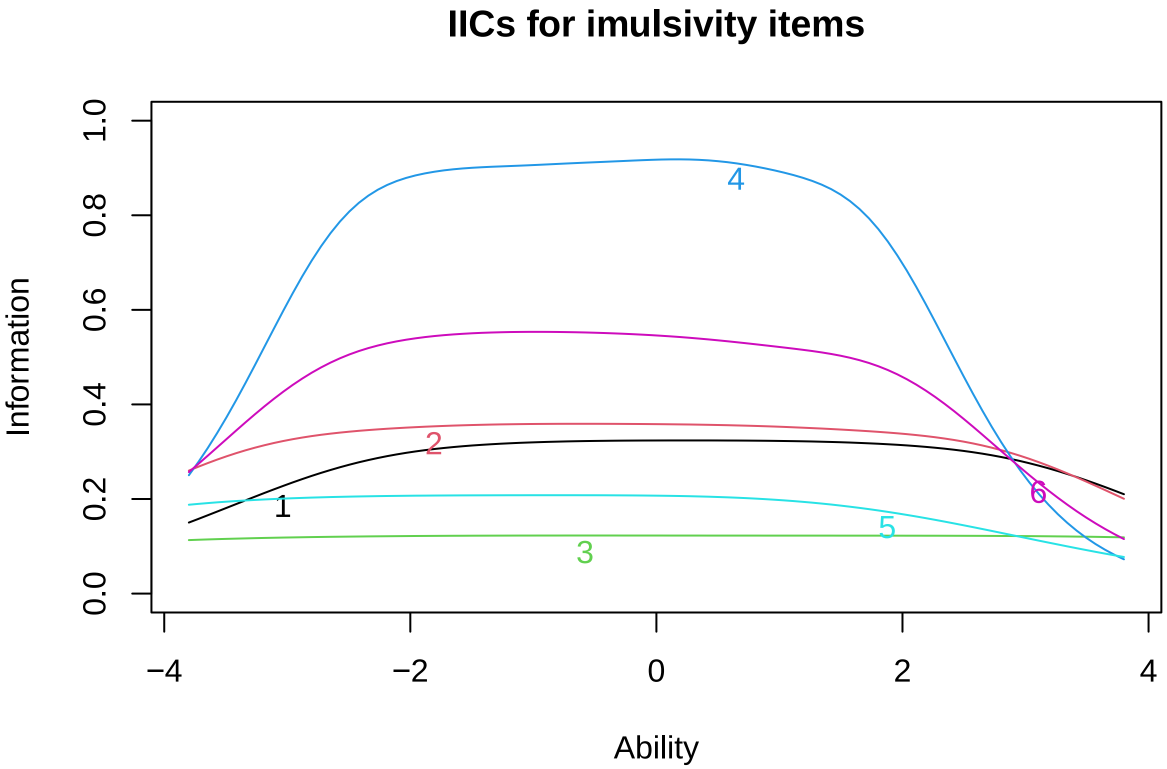
	Thres1	Thres2	Thres3	Thres4	Thres5	Thres6	Discr
S255	-2.295	-1.083	-0.226	0.768	1.628	2.884	1.001
S264	-3.162	-1.813	-0.887	0.137	1.174	2.655	1.058
S269	-4.206	-2.096	-0.554	1.339	3.145	4.843	0.619
S274	-2.478	-1.580	-0.727	0.060	0.722	1.664	1.704
S279	-3.978	-2.487	-1.591	-0.273	0.486	1.524	0.802
S292	-2.648	-1.738	-1.056	-0.333	0.537	1.853	1.310

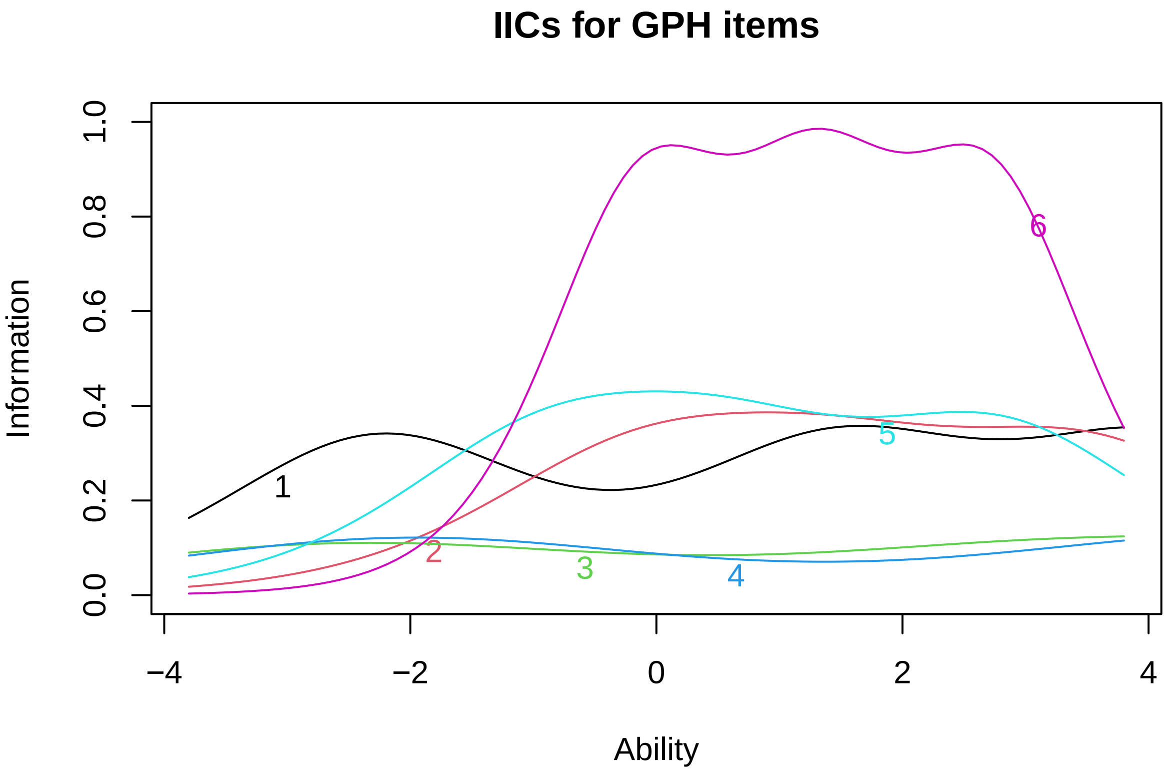
For the General psychological health (GPH) items, Items S329, S330 showed low differentiability, whereas item S335 (“Did you feel unhappy and dejected lately?”) showed the highest differentiability, followed by items S332, and S336 with high discriminative properties. Items S327, S328, S331, and S333 showed moderate differentiability. Items S337 and S338 are on the edge of conventionally high differentiablity (see also the following table). For the difficulty parameters, items S327, S329, and S330 capture a wide range of the latent trait of general psychological health. Especially the aforementioned items seem to catpure extremely high and low values of GPH.

Table 6: Parameters PCM General Psychological Health

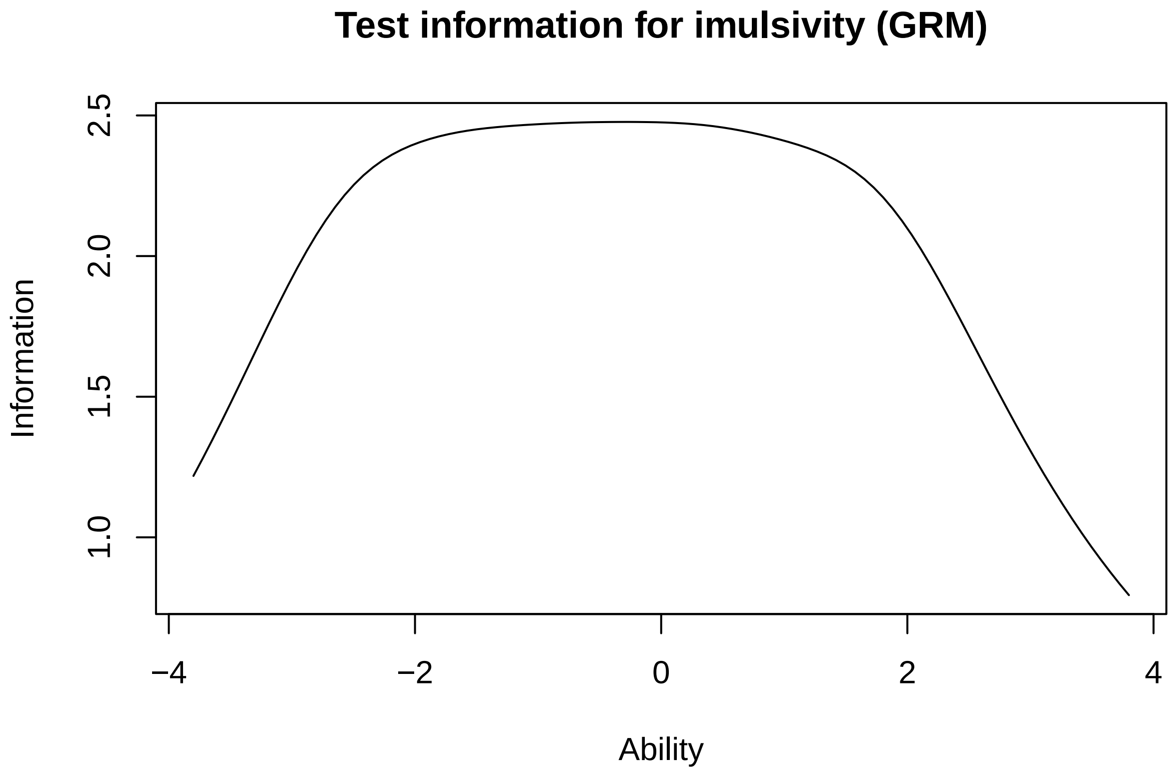
	Thres1	Thres2	Thres3	Discr
S327	-2.242	1.502	4.082	1.162
S328	-0.053	1.395	3.509	1.137
S329	-2.455	3.371	6.190	0.658
S330	-1.999	4.736	7.494	0.694
S331	-0.819	0.521	2.773	1.202
S332	-0.085	1.314	2.701	1.878
S333	-1.345	1.672	4.030	1.173
S335	-0.258	0.880	2.408	2.214
S336	0.602	1.940	3.781	1.834
S337	1.249	2.687	4.099	1.430
S338	-1.400	1.952	4.002	1.489

In the following table, the Item Information Curves and their discrimination and difficulty parameters for the 6 impulsivity items (“S255”,“S264”, “S269”, “S274”, “S279”, “S292”) can also be inspected visually.

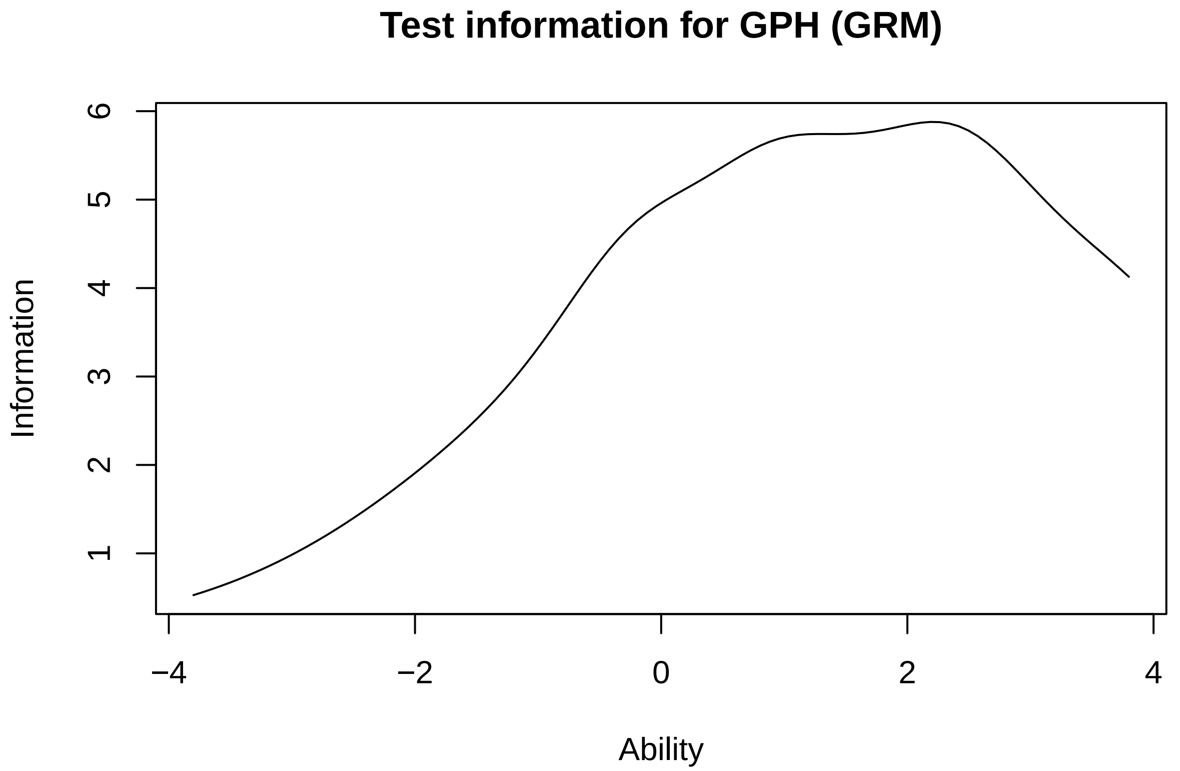




The test information curve for the impulsivity test suggests that the items are most informative for distinguishing persons with impulsivity levels roughly below and above -2 and +2 SDs above the mean. This suggests that the impulsivity items will perform well in distinguishing persons with low and high impulsivity in this range. From -2 to -4 and from +2 to +4 SDs from the mean, the test information significantly drops. Crucially, however, the test information curve peaks at 2, which indicates rather poor discriminative power for the impulsivity test generally.



For the general psychological health (GPH) items, the items show poor informity for low values of the latent trait. With increases in the latent trait, the test information function (TIF) increases proportionally. At the mean value of the estimated latent trait, the item information curve reaches a high value of informaty (TIF ~ 5). The TIF stays high until it drops again at +2 SDs from the mean.



When comparing the fitted GRM model with dependent responses due to the estimated latent trait (item responses), the independence model has higher AIC and BIC values (AIC = 36844.44) than the Graded Response Model (35842.98). The GRM has 42 degrees of freedom, the independent model has 36 estimated degrees of freedom. The LR / deviance statistic is 1013.461. The chi-squared LR statistic is highly significant, indicating that independence must be rejected in favor of dependence of the item reponses, which may be the result of an underlying latent variable.

To check if any items of the two tests are biased, we performed Differential Item Functioning (DIF) testing. For the DIF testing, the item level Eta-squared test compares three models: A model where there is only a main effect of ability (no DIF), a model where there is a main effect of ability and a main effect of subgroup (uniform DIF), and a model where there is a main effect of ability, a main effect of subgroup, and an interaction effect of subgroup (non-uniform DIF). Items that pass the significance threshold.

When investigating the impulsivity test, for gender, items 1, 3, and 5 were flagged for DIF at an alpha threshold of 0.01. Item 1, 3 and 5 show both uniform and non-uniform DIF, which is a strong indicator for biased items. Gender seems to interact with the probabilty of giving positive answers to the items, for participants with the same value on the latent trait impulsivity. For the two age groups (younger and older participants), items 1, 3, and 4 show uniform DF, but no non-uniform DF, indicating a main effect of the latent trait and a main effect of age on the probability of responding positive to this item. This means that the difference in item performance between age groups is consistent across all levels of impulsivity. Looking at study delay, no items showed uniform or non-uniform DIF, indicating that the items do not show bias between participants' duration of study.

When investigating the general psychological health (GPH) test, for gender, items 5, 8, 11 were flagged for DIF at an alpha threshold of 0.01. Items 5 and 8 show both uniform and non-uniform DIF, indicating an interactive effect of gender and the probability to give a positive response to GPH for different levels of the latent trait of GPH. Item 11 shows only uniform DIF, indicating that gender has an effect of the probability to answer positively independent of the value of the latent trait of GPH. Moreover, when looking at the subgroups of older and younger people (age), items 4, 7, and 10 were flagged for DIF. Item 5 showed non-uniform DIF, indicating an interactive effect of age on the probabilty to respond positively on items of GPH for different values of the latent trait of GPH. Items 7 and 10 only showed uniform DIF, indicating a main effect of age on the probability of answer positively, independent of the values of the latent trait. Lastly, when comparing participants with shorter and longer study delays, no DIF was detected, indicating no bias in the probabilty of answering positively for the items in this subgroup.

4 Discussion

The results of this study provide insights into the psychometric properties of both tests used to measure impulsivity and general psychological health (GPH) The GPH test demonstrated acceptable levels of internal consistency, as indicated by Cronbach's alpha and McDonald's omega. The impulsivity test, however, showed low internal consistency, indicating that items within the test do not reliably measure the same underlying construct. The test items are not highly correlated with each other, indicating that they may be capturing different constructs, rather than consistently assessing the intended trait of impulsivity. Narrow confidence intervals generated through bootstrapping suggest that the tests provide stable measurements across different samples. The one-factor analysis supported the assumption that a single latent factor underlies both tests, reinforcing the notion that impulsivity and psychological health can each be conceptualized as unidimensional constructs.

However, the Item Response Theory (IRT) analysis, both including the Graded Response Model (GRM) and the Partial Credit Model (PCM), revealed variability in item discrimination and difficulty, providing more nuanced information about how well individual items measure different levels of these traits. Items with high discrimination parameters were particularly effective in distinguishing between individuals with differing levels of the underlying trait, while items with low discrimination may benefit from revision or removal in future test iterations.

The Differential Item Functioning (DIF) analysis indicated that certain items displayed bias with respect to gender and study delay. The presence of uniform DIF suggests that some items may not function equivalently across groups, potentially skewing test results. These findings highlight the importance of ensuring that items are equally applicable across diverse populations, particularly when using these tests in both research and clinical contexts. Future test revisions should consider revising or replacing biased items to enhance fairness and validity.

Finally, considering practical implications, an IRT model is preferred a the one-factor model in research settings, although the one-factor model has some advantages like simplicity and good interpretability while still modelling some complex effects in the relationship of the items to the latent trait. The main reason to prefer IRT over the one-factor model ot sum scores is that it is able to model the most information about the items and their relationship to the latent trait. For example IRT models like the 2-PL or GRM model are able to model item-level information such as item difficulty and discrimination. It is also able to show how informative each item is across different values of the latent trait, and add these item informations up to give the total information contained in the test. The second reason for a superiority of the IRT-models is that they can handle biased items and measurement error better than one-factor models or sumscores. The one-factor model can also handle measurement error, but not for different values of the latent trait. It assumes that the measurement error is constant across all levels of the latent traits. Each item has a single error variance that is assumed to be equal for all individuals. IRT models, however, can model measurement errors for different values of the latent trait, and even give detect if certain items function differently for subgroups through Differential Item functioning (DIF) analysis.

In a clinical setting, when reporting the results to patients and decision makers, one might refer to the most simple and easy-to-work with model that allows for quick and easy interpretability. Therefore, one might refer to sumscores in this settings, because they are straightforward and fast to compute, and they are able to provide decision makers and patients with accessible and understandable information to base decisions upon. IRT models would be too complex and hard to comprehend in a short period of time to work appropriately with. Taken together, both computability and interpretability favor sumscores over more complex IRT or one-factor models.