Personality Measurement of Students Using Item Response Theory Models: Stability Responses from Nigerian Institutions

Olawale Ayoola Ogunsanmi, Temitope Babatimehin, and Yejide Adepeju Ibikunle

Abstract

Item Response Theory (IRT) is utilised to detect bias in assessment tools and address issues such as faked or manipulated responses, enhancing the reliability and stability of conclusions in personality assessment. This article examines the item parameter estimates of a scale and the effectiveness of one-, two-, and three-parameter logistic models in analysing response stability in personality measurement from repeated administration. Three hundred undergraduate students at three tertiary institutions in Nigeria were sampled using a multi-stage sampling procedure. Data was collected using an adapted version of the Big Five Inventory (BFI) with a reliability coefficient of 0.85. The results showed that the item parameter estimates (mean threshold) are within the recommended benchmarks. A comparison of the three IRT models based on the Likelihood ratio (InL), Akaike Information Criterion (AIC), and Bayesian Information Criterion (BIC) values revealed that the two-parameter logistic model best fit the personality data among undergraduates from repeated administration. It is recommended that, rather than relying solely on a statistical decision-making process, IRT fit and model comparison should be applied to gain insight into the functioning of items and tests.

Key words: response stability, personality traits, personality measurement, Item Response Theory

ABOUT THE AUTHORS: OLAWALE AYOOLA OGUNSANMI, Obafemi Awolowo University, Ile-Ife, Nigeria, email: ogunsanmiwale2014@gmail.com, TEMITOPE BABATIMEHIN, Obafemi Awolowo University, Ile-Ife, Nigeria, email: temitopebabatimehin@gmail.com and YEJIDE ADEPEJU IBIKUNLE, Lead City University, Ibadan, Nigeria, email: adepejuyejide@gmail.com

Résumé:

La théorie de la réponse à l'item (TRI) est utilisée pour détecter les biais dans les outils d'évaluation et traiter des questions telles que les réponses truquées ou manipulées, améliorant ainsi la fiabilité et la stabilité des conclusions dans l'évaluation de la personnalité. Cet article examine les estimations des paramètres d'une **échelle** et l'efficacité des modèles logistiques à un, deux et trois paramètres dans l'analyse de la stabilité des réponses dans la mesure de la personnalité à partir d'une administration répétée. Trois cents **étudiants** de premier cycle de trois **établissements** d'enseignement supérieur au Nigeria ont **été échantillonnés** à l'aide d'une procédure d'échantillonnage à plusieurs degrés. Les données ont **été** collectées à l'aide d'une version adaptée de l'inventaire Big Five (BFI) avec un coefficient de fiabilité de 0,85. Les résultats ont montré que les estimations des paramètres des items (seuil moyen) se situent dans les limites des repères recommandés. Une comparaison des trois modèles IRT basée sur le rapport de vraisemblance (InL), le critère d'information d'Akaike (AIC) et le critère d'information bayésien (BIC) a révélé que le modèle logistique à deux paramètres correspondait le mieux aux données de personnalité chez les **étudiants** de premier cycle à partir d'une administration répétée. Il est recommandé, plutôt que de s'appuyer uniquement sur un processus de prise de décision statistique, d'appliquer l'ajustement IRT et la comparaison de modèles pour mieux comprendre le fonctionnement des items et des tests.

Mots clés: stabilité des réponses, traits de personnalité, mesure de la personnalité, théorie de la réponse aux items.

Introduction

Personality measurement is significant in psychological research as it promotes comprehension of human behaviour and individual differences. This is crucial in tertiary education settings like Nigeria, where the transition from adolescence to early adulthood may significantly impact academic achievement, social interactions, and overall well-being. Advanced statistical models such as Item Response Theory (IRT) have gained traction among researchers seeking to unravel the complexities surrounding personality stability among students at tertiary institutions (Alexander et al., 2020; Yang et al., 2023; Zhu et al., 2021). While helpful, traditional assessment methods often fall short of capturing the subtle dynamics of personality traits over time (Nadkarni and Herrmann, 2010; Riaz et al., 2012). In contrast, IRT models offer a comprehensive framework to simultaneously assess item attributes and latent trait stability, thus providing invaluable insights into the performance of individual items in personality questionnaires and the stability of latent personality traits among students over time. Nigeria's heterogeneous tertiary education landscape, which includes universities, polytechnics, and colleges, plays a vital role in educational experiences that may influence personality development and stability. Moreover, the tertiary student population's transitional nature provides a unique opportunity to explore potential changes in personality traits during this critical stage of personal growth and development.

Measurement of personality and attitudes has historically shaped the progression of psychology and remains pivotal in empirical studies. However, recent decades have seen limited progress in refining the statistical methodologies underpinning the development of measurement scales in this field. Psychological constructs, including personality traits, are often intangible and inferred, raising questions about the value of quantifying them using physical features (Cuthbert and Kozak, 2013; Yang et al., 2023). Despite these challenges, it is important to effectively measure these theoretical elements in order to gain a comprehensive understanding of human behaviour (Smith, 2005; Seidman, 2013; Stoughton et al., 2013).

Various assessment strategies, including peer reports, life outcomes data, and self-reported data, contribute to the diverse landscape of personality studies (Kelley et al., 2016). However, concerns persist within the academic community regarding the potential for skewed, faked, or manipulated responses in personality assessments (Morizot et al., 2007; Revelle and Wilt, 2013; Paulhus, 2014). In response to these challenges, IRT has emerged as a valuable tool, supplementing classical test theory (CTT) methods and enhancing the reliability and stability of personality evaluations (Waller et al., 1996; Ibikunle, 2021). It has also been instrumental in identifying and addressing bias in assessment instruments, contributing to more equitable and accurate evaluations (Adedoyin, 2010; Ogunsanmi and Faleye, 2021; Adediwura and Asowo, 2020, 2021).

Item Response Theory serves as a comprehensive statistical framework to evaluate item and test performance, facilitating deeper understanding of the relationship between performance and the abilities tested (Hambleton and Jones, 2013). Its application extends beyond cognitive data to potentially benefit the study of personality data, offering a promising avenue to advance research in this domain. The development, evaluation, and scoring of tests, questionnaires, and other instruments to gauge mental prowess or psychometric features all benefit from the use of IRT (DeMars, 2010; Chalmers, 2012; Hambleton and Jones, 2013; Zanon et al., 2016). Through its nuanced approach to assessing character qualities, IRT enhances the reliability and validity of personality assessments, thereby contributing to more robust conclusions in psychological research (Kubinger, 2002; Benson and Campbell, 2007; Hambleton and Jones, 2013).

Item Response Theory models

Mathematical models can be used to establish a connection between the latent variables of interest and the probability of responding to an assessment question. These connections can be employed to predict the evaluation's outcome. One-, two-, and three-parameter logistic models are widely used in IRT (Hambleton et al., 1991; Carvalho, Primi and Baptista, 2015; Annan-Brew, 2020; Gyamfi and Acquaye, 2023). In contrast to more holistic approaches to modelling, IRT-based models focus on analysing test takers' responses to specific questions. Item-level modelling offers more versatility for a wide range of uses, including but not limited to development testing; evaluating differential item functionality; deploying computer-adaptive testing; and aggregating score summaries.

The Rasch Model

The Rasch Model developed by psychologist Georg Rasch is a paradigm in the field of psychometrics that has attained near-universal acceptance. The term "Item Response Theory" is often used interchangeably with "one-parameter model." The Rasch Model is widely used to analyse students' answers in reading comprehension tests used for statistical purposes in a wide range of contexts, including reading evaluations.

Our study investigated the correlation between participants' aptitudes, attitudes, or personality traits and the level of difficulty associated with the items evaluated. The logistic function is used to build a correlation between the probability of a correct response and the scale of ability. The study's primary emphasis was the difficulty parameter, while maintaining a constant value of 1.0 for the discrimination parameter, indicated as "a". Nevertheless, it is worth noting that, as stated by De Ayala (2009) and Gyamfi and Acquaye (2023), the difficulty parameter, represented as b, has the potential to fluctuate across different values. The one-parameter model postulates that the score is only influenced by questions' level of difficulty and the latent trait. The equation for the one-parameter model is as follows:

$$P\ (\theta) = \frac{1}{1+e^{-D(è-b)}} \qquad (1)$$

Where θ = latent trait, b = difficulty parameter, è = ability level

Two-Parameter Logistic Model (2PLM)

The emergence of the two-parameter model can be attributed to the limitations of the one-parameter model. One of the disadvantages of this approach is its failure to include the variability in the discriminating power of items, which could lead to erroneous conclusions in terms of model fit. The 2PL model is used to estimate the likelihood of a correct answer to a given test item based on the individual's level of ability and two specific item attributes. The primary distinction with regard to the first-person plural (1PL) paradigm is the substitution of the term exp (è - bi) with exp[ai(è - bi)]. Similar to the first-person logistic (1PL) model, the parameter bi represents the level of difficulty. The newly-introduced parameter, denoted as ai, is often referred to as the discrimination parameter. The equation for the two-parameter logistic model is as follows:

$$P\ (\theta) = \frac{1}{1+e^{-Dai(è-bi)}} \qquad (2)$$

Where, θ = latent trait, a = discriminating parameter b = difficulty parameter è = ability level

Three-Parameter Logistic Model (3PLM)

While the 2PL model is an expansion of the Rasch Model, which is also known as the 1PL model, other models can be seen as expansions of the 2PL model. The inclusion of an additional item parameter is a distinguishing characteristic of the three-parameter logistic model (3PL). A notable phenomenon in the field of testing is that examinees have the potential to answer correctly by guessing. Hence, the likelihood of providing the right answer includes a small factor attributable to random guessing. Guessing was not taken into account in the two preceding models. The 3PL model has an additional parameter, denoted as c, that is often referred to as the "pseudo-chance" or "pseudo-guessing" parameter. Skrondal and Habe-Hesketh (2004) and Gyamfi and Acquaye (2023) defined the concept of three-parameter logistic IRT (3PL IRT), which accounts for the possibility of examinees responding correctly to items due to chance or guessing. This parameter introduces a lower asymptote to the item characteristics curve (ICC). The 3PL can be expressed as follows:

$$P\ (\theta) \ = \frac{c+(1\text{-}c)1}{1+e^{-Da(è-b)}} \qquad (3)$$

Where θ = latent trait, a = discriminating parameter, b = difficulty parameter, è = ability level c = guessing parameter.

In theoretical terms, it can be posited that as the degree of talent or attribute diminishes towards zero, the likelihood of producing a right answer should also move towards zero. Nevertheless, individuals with very low scores on a particular attribute may have the ability to accurately infer the correct response. Consequently, examinees with the lowest and highest abilities have an equal likelihood of answering the question correctly by random guessing. The parameter c is theoretically bounded within the range of 0 ≤ c ≤ 1.0. Initially, IRT models were formulated to accommodate dichotomous replies, namely, binary responses characterised by values of 0 (indicating wrong) and 1 (indicating correct). However, contemporary advancements have led to the development of models capable of accommodating a wide range of educational and psychological data (De Ayala, 2009; Gyamfi and Acquaye, 2023).

Personality questionnaires continue to serve as a crucial tool to assess personality traits. However, several issues are associated with their utilisation. The literature highlights that this includes the potential for faking (where individuals may withhold objectively honest information due to fear of being misjudged), manipulation (where they may present themselves as having a different personality than their own), distortion, and psychometric challenges. Psychologists and researchers thus seek to comprehend why individuals alter their responses to identical item stimuli on two separate occasions. They also aim to determine the item properties that contribute to response stability or more frequent inconsistent responses. Utilisation of a sophisticated mathematical model such as IRT is necessary to address these concerns (Chalmers, 2012). However, there is limited empirical research on the stability of responses in personality measurement, specifically with regard to inconsistent or changing responses to the same item during repeated administration. These studies have primarily focused on a descriptive analysis of the relationships between item and examinee characteristics and the stability of item responses. Furthermore, most have utilised CTT as their mathematical framework to demonstrate a curvilinear association between the fraction of endorsement and the stability of personality assessment.

Moreover, the existing body of research on model fit has mainly focused on two specific IRT models (MacDonald and Paunonen et al., 2003; Wyatt, 2016). Consequently, several essential inquiries on this topic have yet to be addressed. The appropriateness of the 3PL model in comparison to the 1PL and 2PL models for personality data remains uncertain. Therefore, it is necessary to evaluate the stability of responses and analyse the model fit of the 1PL, 2PL, and 3PL models when applied to personality data. Against this background, our research aimed to identify the most appropriate application of the IRT model in analysing personality data obtained from many administrations. The following research questions were formulated to achieve this objective:

i. What are the item parameter estimates of the personality scale from repeated administration?
ii. Which of the one-, two and three-parameter logistic IRT models is more effective in analysing the stability of responses from the personality scale?

Methodology

This study adopted a survey research design. Three hundred undergraduate students from three tertiary institutions in Osun State, Nigeria, were selected using a multi-stage sampling technique. The three senatorial districts in Osun State include Osun Central, Osun East, and Osun West. Three Local Government Areas (LGAs) in these senatorial districts were selected using a purposive sampling technique. Three tertiary institutions (a university, a polytechnic, and a college of education) were selected from the three LGAs using purposive sampling. A hundred undergraduate students residing in the hostels of each chosen tertiary institution were purposefully selected to participate in the study. This selection method was employed to ensure consistency in the sample group across the initial and subsequent administrations of the assessment instrument. Students accommodated in hostels were specifically chosen to facilitate access to the second administration of the assessment tool. The study utilised an adapted research instrument known as the Big Five Inventory (BFI) initially developed by Goldberg (1993). The original BFI based on the 1999 version by John and Srivastava comprises 44 items designed to assess an individual's personality across the dimensions of Extraversion, Openness, Neuroticism, Agreeableness, and Conscientiousness. However, for this study, a modified version of the BFI was employed consisting of 40 items. This was created by removing four items to ensure an equal distribution of eight items across each dimension. The reliability coefficient of the instrument using Cronbach's Alpha yielded a value of 0.85. The test-retest interval was two weeks. Data were analysed using Bilog-MG and SPSS statistical software.

Results

Research Question One: What are the item parameter estimates of the personality scale from repeated administration?

Table 1: Item Parameter Estimates for 1PL Model for Time 1 (T1) and Time 2 (T2)

Item	Intercept	Slope	Threshold	Loading	Asymptote	CHISQ	
	S.E.	S.E.	S.E.	S.E.	S.E.	(Prob)	DF
PM01 T1	1.191	0.470	-2.535	0.425	0.000	224.9	8.0
	0.026*	0.004*	0.054*	0.003*	0.000*	(0.0000)	
T2	1.112	0.470	-2.365	0.425	0.000	194.7	9.0
	0.025*	0.004*	0.054*	0.003*	0.000*	(0.0000)	
PM02 T1	3.767	0.470	-8.015	0.425	0.000	376.0	8.0
	0.070*	0.004*	0.149*	0.003*	0.000*	(0.0000)	
T2	3.454	0.470	-7.349	0.425	0.000	571.3	9.0
	0.061*	0.004*	0.130*	0.003*	0.000*	(0.0000)	
PM03 T1	1.788	0.470	-3.804	0.425	0.000	434.6	8.0
	0.031*	0.004*	0.067*	0.003*	0.000*	(0.0000)	
T2	1.765	0.470	-3.754	0.425	0.000	463.2	9.0
	0.031*	0.004*	0.066*	0.003*	0.000*	(0.0000)	
PM04 TI	0.948	0.470	-2.017	0.425	0.000	320.5	8.0
	0.024*	0.004*	0.052*	0.003*	0.000*	(0.0000)	
T2	0.979	0.470	-2.082	0.425	0.000	237.2	9.0
	0.024*	0.004*	0.052*	0.003*	0.000*	(0.0000)	
PM05 T1	2.880	0.470	-6.128	0.425	0.000	77.4	7.0
	0.046*	0.004*	0.099*	0.003*	0.000*	(0.0000)	
T2	2.883	0.470	-6.133	0.425	0.000	65.9	7.0
	0.047*	0.004*	0.099*	0.003*	0.000*	(0.0000)	
PM06 T1	1.527	0.470	-3.249	0.425	0.000	527.9	8.0
	0.028*	0.004*	0.059*	0.003*	0.000*	(0.0000)	
T2	1.558	0.470	-3.314	0.425	0.000	467.3	9.0
	0.028*	0.004*	0.059*	0.003*	0.000*	(0.0000)	
PM07 T1	1.769	0.470	-3.763	0.425	0.000	168.8	8.0
	0.031*	0.004*	0.066*	0.003*	0.000*	(0.0000)	
T2	1.703	0.470	-3.622	0.425	0.000	154.9	9.0
	0.030*	0.004*	0.064*	0.003*	0.000*	(0.0000)	
PM08 T1	2.690	0.470	-5.723	0.425	0.000	128.1	8.0
	0.043*	0.004*	0.091*	0.003*	0.000*	(0.0000)	
T2	2.656	0.470	-5.650	0.425	0.000	124.7	9.0
	0.043*	0.004*	0.090*	0.003*	0.000*	(0.0000)	
PM09 T1	2.668	0.470	-5.677	0.425	0.000	48.3	7.0
	0.042*	0.004*	0.090*	0.003*	0.000*	(0.0000)	
T2	2.705	0.470	-5.755	0.425	0.000	40.0	7.0
	0.043*	0.004*	0.092*	0.003*	0.000*	(0.0000)	
PM10 T1	1.205	0.470	-2.565	0.425	0.000	337.2 8.0	
	0.026*	0.004*	0.056*	0.003*	0.000*	(0.0000)	
T2	1.209	0.470	-2.571	0.425	0.000	309.0	9.0
	0.026*	0.004*	0.056*	0.003*	0.000*	(0.0000)	
+ +	+	+	+	+	+ +	+	+
+ +	+	+	+	+	+ +	+	+
+ +	+	+	+	+	+ +	+	+
PM30 T1	0.920	0.470	-1.957	0.425	0.000	377.6	8.0
	0.025*	0.004*	0.053*	0.003*	0.000*	(0.0000)	
T2	0.804	0.470	-1.711	0.425	0.000	344.7	8.0
	0.024*	0.004*	0.052*	0.003*	0.000*	(0.0000)	
PM31 T1	2.795	0.470	-5.947	0.425	0.000	179.7	7.0
	0.045*	0.004*	0.096*	0.003*	0.000*	(0.0000)	
T2	2.533	0.470	-5.390	0.425	0.000	124.4	9.0
	0.040*	0.004*	0.086*	0.003*	0.000*	(0.0000)	
PM32 T1	1.679	0.470	-3.571	0.425	0.000	78.6	8.0
	0.030*	0.004*	0.063*	0.003*	0.000*	(0.0000)	
T2	1.419	0.470	-3.019	0.425	0.000	93.7	9.0
	0.030*	0.004*	0.058*	0.003*	0.000*	(0.0000)	
PM33 T1	-0.245	0.470	0.522	0.425	0.000	1117.8	8.0
	0.022*	0.004*	0.046*	0.003*	0.000*	(0.0000)	
T2	-0.183	0.470	0.390	0.425	0.000	872.9	9.0
	0.022*	0.004*	0.046*	0.003*	0.000*	(0.0000)	
PM34 T1	0.948	0.470	-2.016	0.425	0.000	580.9	8.0
	0.025*	0.004*	0.054*	0.003*	0.000*	(0.0000)	
T2	0.969	0.470	-2.062	0.425	0.000	626.1	9.0
	0.025*	0.004*	0.054*	0.003*	0.000*	(0.0000)	
PM35 T1	0.340	0.470	-0.724	0.425	0.000	318.2	8.0
	0.023*	0.004*	0.049*	0.003*	0.000*	(0.0000)	
T2	0.307	0.470	-0.653	0.425	0.000	392.6	8.0
	0.023*	0.004*	0.049*	0.003*	0.000*	(0.0000)	
PM36 T1	0.392	0.470	-0.834	0.425	0.000	251.8	8.0
	0.023*	0.004*	0.063*	0.003*	0.000*	(0.0000)	
T2	0.358	0.470	-0.762	0.425	0.000	344.9	8.0
	0.023*	0.004*	0.058*	0.003*	0.000*	(0.0000)	
PM37 T1	1.086	0.470	-2.311	0.425	0.000	504.3	8.0
	0.025*	0.004*	0.052*	0.003*	0.000*	(0.0000)	
T2	1.098	0.470	-2.337	0.425	0.000	415.2	9.0
	0.025*	0.004*	0.052*	0.003*	0.000*	(0.0000)	
PM38 T1	-0.128	0.470	0.272	0.425	0.000	109.7	7.0
	0.022*	0.004*	0.047*	0.003*	0.000*	(0.0000)	
T2	-0.155	0.470	0.329	0.425	0.000	86.8	7.0
	0.022*	0.004*	0.047*	0.003*	0.000*	(0.0000)	
PM39 T1	-0.187	0.470	0.397	0.425	0.000	45.5	8.0
	0.022*	0.004*	0.047*	0.003*	0.000*	(0.0000)	
T2	-0.108	0.470	0.229	0.425	0.000	53.1	9.0
	0.022*	0.004*	0.047*	0.003*	0.000*	(0.0000)	
PM40 T1	0.194	0.470	-0.413	0.425	0.000	178.9	7.0
	0.023*	0.004*	0.048*	0.003*	0.000*	(0.0000)	
T2	0.274	0.470	-0.583	0.425	0.000	179.0	9.0
	0.023*	0.004*	0.048*	0.003*	0.000*	(0.0000)	

Time1		* STANDARD ERROR	
LARGEST CHANGE =	0.000764 (0.0000)	49746.3 290.0	
Time2		* STANDARD ERROR	
LARGEST CHANGE =	0.000764 (0.0000)	49746.3 290.0	

PARAMETER	N	MEAN	STD. DEV.	ADJUSTED MEAN
TIME: 1	40			
THRESHOLD		-2.823	2.673	0.000
TIME: 2	40			
THRESHOLD		-2.758	2.586	0.064

Note: S.E. = Standard Error, CHISQ = Chi-square, Prob = Probability, Df = Degree of Freedom

Table 1 shows the parameter estimates of the 1PLM for Time 1 (T1) and Time 2 (T2), respectively. The INTERCEPT column contains the estimated item intercept, the SLOPE column contains the "a" parameter (discrimination), the THRESHOLD column contains the "b" parameter (difficulty), and the ASYMPTOTE column includes the "c" parameter (pseudo-guessing). The parameter table shows the relationship between responses on each item and the latent trait. The mean threshold value is at (T1= -2.823, T2= 2.673) and the adjusted threshold mean value at (T1= 0.000, T2= 0.064). These results indicated that the items possessed adequate item difficulty index.

Table 2: Item Parameter Estimates for 2PL Model for Time 1 (T1) and Time 2 (T2)

ITEM	INTERCEPT	SLOPE	THRESHOLD	LOADING	ASYMPTOTE	CHISQ	
	S.E.	**S.E.**	**S.E.**	**S.E.**	**S.E.**	**(PROB)**	**DF**
PM01 T1	1.135	0.047	-24.113	0.047	0.000	173.3	9.0
	0.025*	0.010*	5.308*	0.010*	0.000*	(0.0000)	
T2	1.061	0.470	-22.543	0.047	0.000	139.1	9.0
	0.025*	0.010*	4.946*	0.010*	0.000*	(0.0000)	
PM02 T1	4.852	2.320	-2.092	0.918	0.000	58.4	9.0
	0.177*	0.203*	0.117*	0.081*	0.000*	(0.0000)	
T2	4.549	2.320	-1.961	0.918	0.000	29.4	9.0
	0.163*	0.203*	0.110*	0.081	0.000*	(0.0005)	
PM03 T1	3.012	2.725	-1.105	0.939	0.000	9.3	9.0
	0.066*	0.087*	0.019*	0.030*	0.000*	(0.4124)	
T2	3.011	2.725	-1.105	0.939	0.000	15.0	9.0
	0.065*	0.087*	0.019*	0.030*	0.000*	(0.0919)	
PM04 T1	0.903	0.075	-12.105	0.074	0.000	294.2	9.0
	0.024*	0.014*	2.353*	0.014*	0.000*	(0.0000)	
T2	0.936	0.075	-12.550	0.074	0.000	213.0	9.0
	0.024*	0.014*	2.430*	0.014*	0.000*	(0.0000)	
PM05 T1	8.857	8.202	-1.080	0.993	0.000	346.3	9.0
	1.421*	1.607*	0.039*	0.194*	0.000*	(0.0000)	
T2	8.865	8.202	-1.081	0.993	0.000	185.0	9.0
	1.420*	1.607*	0.039*	0.194*	0.000*	(0.0000)	
PM06 T1	1.457	0.018	-81.865	0.018	0.000	607.4	9.0
	0.028*	0.004*	18.868*	0.004	0.000*	(0.0000)	
T2	1.487	0.018	-83.560	0.018	0.000	483.0	9.0
	0.028*	0.004*	19.233*	0.004*	0.000*	(0.0000)	
PM07 T1	1.986	1.047	-1.898	0.723	0.000	101.0	9.0
	0.046*	0.062*	0.086*	0.043*	0.000*	(0.0000)	
T2	1.950	1.047	-1.863	0.723	0.000	48.0	9.0
	0.044*	0.062*	0.086*	0.043*	0.000*	(0.0000)	
PM08 T1	4.286	3.080	-1.391	0.951	0.000	17.1	8.0
	0.123*	0.140*	0.030*	0.043*	0.000*	(0.0294)	
T2	4.265	3.080	-1.385	0.951	0.000	163.0	9.0
	0.125*	0.140*	0.029*	0.043*	0.000*	(0.0000)	
PM09 T1	2.730	0.687	-3.971	0.567	0.000	106.1	9.0
	0.052*	0.069*	0.358*	0.056*	0.000*	(0.0000)	
T2	2.791	0.687	-4.059	0.566	0.000	62.5	9.0
	0.053*	0.069*	0.365*	0.056*	0.000*	(0.0000)	
PM10 T1	2.029	2.246	-0.903	0.914	0.000	31.7	9.0
	0.057*	0.083*	0.017*	0.034*	0.000*	(0.0002)	
T2	2.070	2.246	-0.922	0.914	0.000	23.0	9.0
	0.057*	0.083*	0.018*	0.034*	0.000*	(0.0061)	
+ +	+	+	+ +	+	+	+	+
+ +	+	+	+ +	+	+	+	+
+ +	+	+	+ +	+	+	+	+
PM30 T1	2.176	3.12	-0.690	0.952	0.000	95.6	9.0
	0.051*	0.065*	0.010*	0.020*	0.000*	(0.0000)	
T2	2.064	3.126	0.660	0.952	0.000	134.8	9.0
	0.048*	0.065*	0.009*	0.020*	0.000*	(0.0000)	
PM31 T1	3.329	1.560	-2.135	0.842	0.000	7.7	8.0
	0.095*	0.123*	0.119*	0.066*	0.000*	(0.4626)	
T2	3.086	1.560	-1.979	0.842	0.000	60.5	9.0
	0.088*	0.123*	0.109*	0.066*	0.000*	(0.0000)	
PM32 T1	1.752	0.717	2.444	0.583	0.000	230.2	9.0
	0.031*	0.024*	0.082*	0.020*	0.000*	(0.0000)	
T2	1.521	0.717	-2.122	0.583	0.000	201.1	9.0
	0.028*	0.024*	0.074*	0.020*	0.000*	(0.0000)	
PM33 TI	-0.231	0.023	9.916	0.023	0.000	247.9	9.0
	0.022*	0.005*	2.451*	0.005*	0.000*	(0.0000)	
T2	-0.172	0.023	7.406	0.023	0.000	244.9	9.0
	0.022*	0.005*	1.945*	0.005*	0.000*	(0.0000)	
PM34 T1	1.140	1.072	-1.064	0.731	0.000	141.2	9.0
	0.037*	0.057*	0.039*	0.039*	0.000*	(0.0000)	
T2	1.202	1.072	-1.122	0.731	0.000	130.0	9.0
	0.038*	0.057*	0.041*	0.039*	0.000*	(0.0000)	
PM35 T1	0.324	0.096	-3.382	0.096	0.000	101.4	9.0
	0.022*	0.015*	0.570*	0.015*	0.000*	(0.0000)	
T2	0.298	0.096	-3.103	0.096	0.000	126.1	9.0
	0.022*	0.015*	0.521*	0.015*	0.000*	(0.0000)	
PM36 T1	0.374	0.085	-4.371	0.085	0.000	90.2	9.0
	0.022*	0.014*	0.770*	0.014*	0.000*	(0.0000)	
T2	0.346	0.085	-4.044	0.085	0.000	97.9	9.0
	0.022*	0.014*	0.708*	0.014*	0.000*	(0.0000	
PM37 T1	1.084	0.460	-2.355	0.418	0.000	173.9	9.0
	0.025*	0.024*	0.127*	0.022*	0.000*	(0.0000)	
T2	1.119	0.460	-2.430	0.418	0.000	117.7	9.0
	0.026*	0.024*	0.125*	0.022*	0.000*	(0.0000)	
PM38 T1	-0.120	0.083	1.438	0.083	0.000	320.7	9.0
	0.022*	0.014*	0.355*	0.014*	0.000*	(0.0000)	
T2	0.142	0.083	1.698	0.083	0.000	265.2	9.0
	0.022*	0.014*	0.394*	0.014*	0.000*	(0.0000)	
PM39 T1	-0.177	0.279	0.634	0.269	0.000	98.5	9.0
	0.022*	0.018*	0.089*	0.017*	0.000*	(0.0000)	
T2	-0.085	0.279	0.304	0.269	0.000	124.0	9.0
	0.022*	0.018*	0.083*	0.017*	0.000*	(0.0000)	

PM40	T1	0.199 0.022*	0.413 0.019*	-0.482 0.057*	0.381 0.017*	0.000 0.000*	64.0 (0.0000)	8.0
	T2	0.301 0.022*	0.413 0.019*	-0.729 0.060*	0.381 0.017*	0.000 0.000*	81.4 (0.0000)	8.0

Time1 * STANDARD ERROR
LARGEST CHANGE = 0.078502 36878.5 214.0
(0.0000)

Time2 * STANDARD ERROR
LARGEST CHANGE = 0.078502 36878.5 214.0
(0.0000)

PARAMETER	N	MEAN	STD. DEV.	ADJUSTED THRESHOLD
SLOPE		1.993	3.101	
LOG(SLOPE)		-0.402	1.737	
TIME: 1	40			
THRESHOLD		-3.746	13.544	0.000
TIME: 2	40			
THRESHOLD		-3.794	13.687	-0.048

Note: S.E. = Standard Error, CHISQ = Chi-square, Prob = Probability, Df = Degree of Freedom

Table 2 shows the parameter estimates of the 2PLM for T1 and T2, respectively. The INTERCEPT column contains the estimated item intercept, the SLOPE column contains the "a" parameter (discrimination), the THRESHOLD column contains the "b" parameter (difficulty), and the ASYMPTOTE column contains the "c" parameter (pseudo-guessing). The parameter loading column refers to the relationship between responses on each item and the latent trait. The mean threshold value at (T1= -2.823, T2= -2.758) and the adjusted threshold mean value at (T1= 0.000, T2= -0.048) indicate no response change for T1 and T2. The items also possess adequate item difficulty and discrimination indices.

Table 3: Item Parameter Estimates for 3PL Model for Time 1 and Time 2

		THRESHOLDS		SLOPES		ASYMPTOTES	
ITEM		MU	SIGMA	MU	SIGMA	ALPHA	BETA
PM01	T1	-	-	1.000	1.649	5.00	17.00
	T2	-	-	1.000	1.649	5.00	17.00
PM02	T1	-	-	1.000	1.649	5.00	17.00
	T2	-	-	1.000	1.649	5.00	17.00
PM03	T1	-	-	1.000	1.649	5.00	17.00
	T2	-	-	1.000	1.649	5.00	17.00
PM04	T1	-	-	1.000	1.649	5.00	17.00
	T2	-	-	1.000	1.649	5.00	17.00
PM05	T1	-	-	1.000	1.649	5.00	17.00
	T2	-	-	1.000	1.649	5.00	17.00
PM06	T1	-	-	1.000	1.649	5.00	17.00
	T2	-	-	1.000	1.649	5.00	17.00
PM07	T1	-	-	1.000	1.649	5.00	17.00
	T2	-	-	1.000	1.649	5.00	17.00
PM08	TI	-	-	1.000	1.649	5.00	17.00
	T2	-	-	1.000	1.649	5.00	17.00
PM09	T1	-	-	1.000	1.649	5.00	17.00
	T2	-	-	1.000	1.649	5.00	17.00
PM10	T1	-	-	1.000	1.649	5.00	17.00
	T2	-	-	1.000	1.649	5.00	17.00
+	+	+	+	+	+	+	+
+	+	+	+	+	+	+	+
+	+	+	+	+	+	+	+
PM30	T1	-	-	1.000	1.649	5.00	17.00
	T2	-	-	1.000	1.649	5.00	17.00
PM31	TI	-	-	1.000	1.649	5.00	17.00
	T2	-	-	1.000	1.649	5.00	17.00
PM32	TI	-	-	1.000	1.649	5.00	17.00
	T2	-	-	1.000	1.649	5.00	17.00
PM33	T1	-	-	1.000	1.649	5.00	17.00
	T2	-	-	1.000	1.649	5.00	17.00
PM34	TI	-	-	1.000	1.649	5.00	17.00
	T2	-	-	1.000	1.649	5.00	17.00
PM35	TI	-	-	1.000	1.649	5.00	17.00
	T2	-	-	1.000	1.649	5.00	17.00
PM36	TI	-	-	1.000	1.649	5.00	17.00
	T2	-	-	1.000	1.649	5.00	17.00
PM37	TI	-	-	1.000	1.649	5.00	17.00
	T2	-	-	1.000	1.649	5.00	17.00
PM38	TI	-	-	1.000	1.649	5.00	17.00
	T2	-	-	1.000	1.649	5.00	17.00
PM39	TI	-	-	1.000	1.649	5.00	17.00
	T2	-	-	1.000	1.649	5.00	17.00
PM40	T1	-	-	1.000	1.649	5.00	17.00
	T2	-	-	1.000	1.649	5.00	17.00

Table 3 shows the parameter estimates of the 3PLM for T1 and T2, respectively. The INTERCEPT column contains the estimated item intercept, the SLOPE column contains the "a" parameter (discrimination), the THRESHOLD column contains the "b" parameter (difficulty), and the ASYMPTOTE column contains the "c" parameter (pseudo-guessing). The LOADING column refers to the relationship between responses on each item and the latent trait. For the 3PLM, there is no calculated threshold mean value, the slopes are (MU = 1.0, SIGMA = 1.649), and the asymptotes (ALPHA = 5.00, BETA = 17.00). The findings suggest that the 3PLM is not suitable for the personality data since the threshold value used as the benchmark of model fit is not computed.

Research Question Two: Which of the one-, two- and three-parameter logistic IRT models is more effective in analysing the stability of responses from the personality scale?

To determine the best-fit model among the three IRT models (1PLM, 2PLM, and 3PLM), their loglikelihood and goodness of fit values were estimated and compared (see Table 4).

Table 4: Likelihood-based Values and Goodness of Fit Statistics for IPLM, 2PLM, 3PLM from Repeated Administration

Statistics based on Goodness of Fit	1PLM	2PLM	3PLM
-2loglikelihood:	655999.59	637115.27	647386.26
Akaike Information Criterion (AIC):	2993.9	2853.6	2878.9
Bayesian Information Criterion (BIC):	3072.4	3010.7	3114.5

Table 4 presents Likelihood-based Values and Goodness of Fit Statistics for the 1PLM, 2PLM, and 3PLM models, respectively. The 1PLM yielded a 2loglikelihood value of 655999.5925, an AIC value of 2993.9, and a BIC value of 3072.4. Similarly, the 2PLM produced a 2loglikelihood value of 637115.2710, an AIC value of 2853.6, and a BIC value of 3010.7. Additionally, the 3PLM displayed a -2loglikelihood value of 647386.269, an AIC value of 2878.9, and a BIC value of 3114.5. To ascertain the efficacy of the one-, two- and three-parameter logistic IRT models in analysing response stability, their -2loglikelihood, AIC, and BIC were evaluated and compared (see Table 5).

Table 5: Comparison of overall fit for models (1PLM, 2PLM and 3PLM)

Model	No of Sample	- 2InL	AIC	BIC
IPLM	300	655999.59	2993.9	3072.4
2PLM	300	637115.27	2853.6	3010.7
3PLM	300	647386.26	2878.9	3114.5

Note: 1PLM = one-parameter model; 2PLM = two-parameter model; 3PLM = three-parameter model

Table 5 above shows the overall model fit of the IRT models (1PLM, 2PLM, and 3PLM). The values obtained were -2loglikelihood (655999.59), AIC = 2993.9, and BIC = 3072.4, for the 1PLM, -2loglikelihood = 637115.27, AIC = 2853.6, and BIC = 3010.7 for the 2PLM and -2loglikelihood = 647386.26, AIC = 2878.9, and BIC = 3114.5 for the 3PLM, respectively. From these results, the -2InL, AIC, and BIC values for the 1PLM, 2PLM, and 3PLM were compared, and the results showed that the 2PLM had the lowest values of -2InL, AIC, and BIC, indicating that it is the model of best fit.

Discussion

The purpose of this research was to evaluate the item parameter estimates and the goodness of fit of IRT models, namely the 1PLM, 2PLM, and 3PLM when applied to personality data obtained through repeated administration of the modified BFI. The results obtained from the estimation of item parameters for the 1PLM and 2PLM revealed that both models exhibited satisfactory difficulty and discrimination indices from repeated administration. These findings implied that the level of challenge or ease presented by individual items within the test and the respondent's ability to endorse or respond correctly to the item over repeated administration are satisfactory, indicating the stability of the instrument over time. The findings also implied that the items could differentiate between individuals who scored high and low on the trait measured from repeated administration, indicating the stability of the personality test over time.

These findings suggest that item difficulty and discrimination contribute to the overall effectiveness of personality tests. They offer standard metrics to compare items across different personality domains and ensure that the test accurately captures the distinctions of the trait assessed. Ludewig et al. (2023) emphasise that item difficulty reflects the proportion of individuals capable of answering the item correctly; thus, items with a high level of difficulty are more challenging and may require deeper self-reflection or introspection to answer accurately, supporting the importance of this metric at an acceptable benchmark. Furthermore, Date et al. (2019) recommend that those who construct the tests aim for acceptable levels of item difficulty and discrimination, underscoring the significance of these factors in ensuring the validity and reliability of the assessment. The results also indicated a distinct correlation between the estimations of item parameters and the fraction of observed changes in item responses when the same instrument was administered for a second time. Nevertheless, the 3PLM failed to provide sufficient estimates for item parameters due to the absence of a computed threshold mean value in the model. This indicates that the 3PLM is not suitable for the personality data, i.e., an individual's personality traits should not be guessed. It concurs with Ahmad and Mokshein's (2016) assertion that when tests involve guessing, the 3PLM produces robust parameter estimates. Moreover, the results of the optimal model demonstrated that the 2PLM exhibited the lowest values in terms of -2InL, AIC, and BIC. These findings suggest that the 2PLM provided the most suitable fit for the current dataset when compared to other models. They align with the American Association of Educational Research, American Psychological Association, and National Council on Measurement in Education's (2014) recommendation that evidence of model-data-fit be established when employing an IRT model to draw inferences from a real dataset, as per the standards for educational and psychological testing. The findings also support those of MacDonald and Paunonen et al.'s (2003) study that revealed that the 2PLM had the best match when applied to personality traits, as well as those of Kose (2014) who asserted that the 2PLM is superior to other IRTs. In contrast, Ahmad and Mokshein, (2016) and Nye et al. (2019) concluded that the 3PLM and Mixed Model (2 and 3PLM) were the best fit. However, the 3PLM produces the least model fit, which could explain its infrequent use for personality data in the scholarly literature. From a psychologist's perspective, it is argued that the 3PLM is not an appropriate framework to analyse personality data as personality tests should not involve guessing. This assertion aligns with the findings of Morizot, Ainsworth, and Reise (2007) and Zanon et al. (2016) that when it comes to achievement statistics, the c parameter is crucial. Nevertheless, IRT estimation should employ different models if the test consists of items with many responses such as in personality assessment.

Conclusion and Recommendations

In conclusion, both the 1PLM and 2PLM exhibited satisfactory item parameter estimates, reflecting adequate item difficulty and discrimination indices. The 2PLM demonstrated the best fit among the three models based on -2InL, AIC, and BIC values. Based on the findings of this study, researchers and practitioners in personality assessment should consider employing the 2PLM, as it demonstrated superior fit compared to the 1PLM and 3PLM in analysing personality data from repeated administration. It is essential for researchers to continuously evaluate the effectiveness and appropriateness of assessment tools such as personality inventories using advanced statistical techniques like IRT. Regular assessments ensure the reliability and stability of conclusions drawn from personality assessments, especially in dynamic environments like tertiary education settings. The limitations of different IRT models and their suitability for specific datasets should be borne in mind as this assists in selecting the most appropriate model to accurately analyse psychological constructs.

Implications of the Findings

Our findings demonstrate the applicability of IRT models to assess item functioning for non-achievement assessments. They also provide valuable insights into item and test performance, enhancing the reliability and validity of personality assessments. By employing the most suitable IRT model, researchers and practitioners can achieve more accurate interpretations of individual differences in personality traits among undergraduate students. Robust personality assessments that accurately capture students' traits and behaviours over time will enable policymakers, educators, and psychologists to make more informed decisions regarding student support programmes, counselling services, and academic interventions.

Limitations and Suggestions for Future Research

The study was limited to undergraduate students from three tertiary institutions in Osun State, Nigeria. Students who reside in the hostels made up the sample. Furthermore, while several personality inventories and scales are available, the study only employed the BFI.

Drawing on the findings and conclusions, the following suggestions are made for future research:

1. The use of recent or newly-developed personality instruments/scales could provide stronger evidence of response stability in the personality domain.
2. It would be of interest to compare dichotomous and polytomous IRT models' fit in the area of personality measurement.

References

Adediwura, A. A., and Asowo, P. (2021). Examining the Nature of Item Bias on Students' Performance in National Examinations Council (NECO) Mathematics Senior School Certificate Dichotomously Scored Items in Nigeria. *International Journal of Contemporary Education* 5(16).

Adedoyin, O. O. (2010). Investigating the invariance of person parameter estimates based on classical test and item response theories. *International Journal of Education Science 2*(2), 107-113.

Ahmad, H., and Mokshein, S. E. (2016). Is 3PL item response theory an appropriate model for dichotomous item analysis of the anatomy and physiology final examination? *Journal Pendidikan Sains and Matematik Malaysia 6*(1), 13. https://doi.org/10.17576/JPSM-2016-0601-13

Alexander, L., Mulfinger, E., and Oswald, F. L. (2020). Using Big Data and Machine Learning in Personality Measurement: Opportunities and Challenges. *European Journal of Personality 34*(5), 632-648. https://doi.org/10.1002/per.2305

Annan-Brew, R. (2020). Differential Item Functioning of West African Senior Secondary Certificate Examination in Core Subjects in Southern Ghana. PhD Thesis, UCC, Ghana.

Benson, M. J., and Campbell, J. P. (2007). To be or not to be linear: An expanded representation of personality and its relationship to leadership performance. *International Journal of Personality and Assessment (15)*2, 232-249.

Chalmers, R. P. (2012). MIRT: A multidimensional item response theory package for the R environment. *Journal of Statistical Software 48*(6), 1-29. Retrieved from http://www.jstatsoft.org/v48/i06.

Carvalho, L. F., Primi, R., and Baptista, M. N. (2015). IRT Application to verify psychometric properties of the Beck Depression Inventory (BDI) University Psychological. *Bogotá, Colombia 14*(1), 91-102.

Costa, P. T., Jr., and McCrae, R. R. (1992). *Revised NEO Personality Inventory (NEO PI-R™) and NEO Five-Factor Inventory (NEO-FFI) professional manual*. Odessa, FL: Psychological Assessment Resources.

Cuthbert, B. N., and Kozak, M. J. (2013). Constructing constructs for psychopathology: The NIMH research domain criteria. *Journal of Abnormal Psychology 122*(3), 928-937. https://doi:10.1037/a0034028

Date, A. P., Borkar, A. S., Badwaik, R. T., Siddiqui, R. A., Shende, T. R., and Dashputra, A. V. (2019). Item analysis as tool to validate multiple choice question bank in pharmacology. International *Journal of Basic & Clinical Pharmacology 8*(9), 1999-2003.

De Ayala, R. J. (2009). *The Theory and Practice of Item Response Theory*. New York: Guilford Press.

DeMars, C. (2010). *Item response theory. Understanding statistics measurement*. Oxford University Press.

Gyamfi, A., and Acquaye, R. (2023). Parameters and models of item response theory (IRT): A review of literature. *Acta Educationis Generalis 13*(3). https://doi.org/10.2478/atd-2023-0022

Hambleton, R., Swaminathan, H., and Rogers, J. (1991). *Fundamentals of Item Response Theory*. Newbury Park California: Sage publications.

Kelley, S. Edens, J., and Morey, L. (2016). Convergence of Self-Reports and Informant Reports on the Personality Assessment Screener. Assessment. https://doi10.1177/1073191116636450.

Kubinger, K. D. (2002). Psychology's challenge when personality questionnaires are applied for individual assessment. *Psychologische Beiträge 44*, 3-9.

Ludewig, U., Alscher, P., Chen, X., and McElvany, N. (2022). What makes domain knowledge difficult? Word usage frequency from SUBTLEX and dlexDB explains knowledge item difficulty. *Behavioural Research* 55, 2621-2637. https://doi.org/10.3758/s13428-022-01918-0

Morizot, J., Ainsworth, A. T., and Reise, S. P. (2007). Toward modern psychometrics: Application of item response theory models in personality research. In R. W. Robins, R. C. Fraley, and R. F. Krueger (eds) *Handbook of Research Methods in Personality Psychology* (pp. 407-423). New York: Guilford.

Nye, C., Joo, S., Zhang, B., and Stark, S. (2019). Advancing and evaluating IRT model data fit indices in organizational research. *Organizational Research Methods 23*. https://doi.org/10.1177/1094428119833158

Ogunsanmi, O. A., and Faleye, B. A. (2021). Effect of language manipulation on differential item functioning of West African Examinations Council's Physics Items among Osun State Secondary School Students. *Journal of Evaluation 6*(1), 131-140.

Paunonen, S. V., Haddock, G., Forsterling, F., and Keinonen, M. (2003). Broad Verses, Narrow Personality Measures and the Prediction of Behaviour Across Cultures. *European Journal of Personality 17,* 413-433.

Paulhus, D. L. (2014). Toward a taxonomy of dark personalities. *Current Directions in Psychological Science 23*(6), 421-426.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Copenhagen, Denmark: Danish Institute for Educational Research.

Revelle, W., and Wilt, J. (2013). The general factor of personality: A general critique. *Journal of Research in Personality 47*(5), 493-504.

Seidman, G. (2013). Self-presentation and belonging on Facebook: How personality influences social media use and motivations. *Personality and Individual Differences 54*(3), 402-407.

Stoughton, J. W., Thompson, L. F., and Meade, A. W. (2013). Big Five personality traits reflected in job applicants' social media postings. *Cyberpsychology, Behavior, and Social Networking 16*(11), 800-805.

Waller, N. G., Tellegen, A., McDonald, R. P., and Lykken, D. T. (1996). Exploring nonlinear models in personality assessment: Development and preliminary validation of a negative emotionality scale. *Journal of Personality 64,* 545-576.

Yang, K., Lau, R. Y., and Abbasi, A. (2023). Getting personal: a deep learning artifact for text-based measurement of personality. *Information Systems Research 34*(1), 194-222. https://doi.org/10.1287/isre.2022.1111

Zanon, C., Hutz, C., Yoo, H., and Hambleton, R. (2016). An application of item response theory to psychological test development. *Psicologia: Reflexão e Crítica 29*. https://doi.org/10.1186/s41155-

016-0040-x

Zhu, G., Zhou, Y., Zhou, F., Wu, M., Zhan, X., Si, Y., and Wang, J. (2021). Proactive personality measurement using item response theory and social media text mining. *Frontiers in Psychology 12,* 705005. https://doi.org/10.3389/fpsyg.2021.705005