ON COMPARISONS OF FREQUENTIST TO BAYESIAN ESTIMATION FOR ITEM RESPONSE THEORY MODELS IN THE PRESENCE OF DICHOTOMOUS RESPONSES

ADETUTU O. M.,[1] & LAWAL H. B.[2]
[1]Department of Statistics, Federal University of Technology, Minna, Nigeria.
[2]Department of Statistics and Mathematical Sciences,
Kwara State University, Malete, Nigeria
E-mail: adetutuolayiwola@gmail.com,blawal66@gmail.com
Phone: +234-803-073-7153

Abstract
Item response theory (IRT) is a family of mathematical models that attempt to explain the relationship between latent traits and their manifestations. They are widely used in education to calibrate and evaluate items in tests, questionnaires, and similar instructions, measuring abilities, attitudes, or other variables. The most frequent method for estimating latent traits called Maximum Likelihood (ML) can either fail to converge or produce biased estimates in complex latent traits models due to co-linearity of explanatory variables. Bayesian estimation approach provides a better alternative to ML (frequentist) for IRT in this case. This study compared the Bayesian against ML estimators for IRT models. The same data set were analysed using both Bayesian and ML estimations. The findings suggest that ML method is a reasonable choice for one and two-parameter logistic IRT models while Bayesian estimation is more appropriate for three-parameter IRT to circumvent non-convergence of ML estimation procedure.

Keywords: traits, Bayesian estimation, maximum likelihood, items

Introduction

The primary IRT models parameters estimation in frequentist approach was Joint Maximum Likelihood Estimation (JMLE) whereby unknown candidates' ability is considered as fixed, as in the case of analysis of variance, and has to be estimated together with item parameters, and ability parameters (Paek & Cole, 2020). Both item and ability parameters were to be estimated simultaneously thereby the maximum likelihood estimates were not consistent as the sample sizes increases. Except for one parameter model in equation (17) where a sufficient statistic was available for ability parameters which make it possible for the use of specialized procedure called conditional maximum estimation, therefore, a more efficient estimation procedure is needed for two and three parameter logistic models to avoid problem of inconsistent estimation of item parameters. As a result of shortcoming in JMLE, Bock and Lieberman (1970) came up with Marginal Maximum Likelihood Estimation (MMLE) method to remove effect of ability parameter which was considered as nuisance in JMLE by assuming that these values constitute a random sample from a population distribution and then integrate over the ability distribution. Unfortunately, the MMLE method developed caused a computational task which was only feasible for small sample size test. A reformation of MMLE was done at instance of Expectation and Maximization (EM) Algorithm by Bock and Aitkin (1981) and was implemented in BILOG computer program (Mislevy & Bock, 1984).

In this reformation, Bayesian concepts of prior and posterior distribution were used, and the underlying mathematics was higher than JMLE which was both theoretically acceptable and computationally feasible. The likelihood function, derivatives, and likelihood equations are developed for case(s) where abilities are observed along with the item responses. For individual item response data, the model is defined as:

$$P_j(\theta_i) = C_j + (1 - C_j)\ \frac{e^{a_j(\theta_i - b_j)}}{1 + e^{a_j(\theta_i - b_j)}} : \quad Q_j(\theta_i) = 1 - P_j(\theta_i) \tag{1}$$

$$P*_j(\theta_i) = \frac{e^{a_j(\theta_i - b_j)}}{1 + e^{a_j(\theta_i - b_j)}} : Q*_j(\theta_i) = 1 - P*_j(\theta_i) \tag{2}$$

Where candidate $i = 1, 2, \ldots, n$, Items $j = 1, 2, \ldots, N$, a_j is the discrimination parameter for item j, b_j is the item difficulty parameter for item j, c_j is the guessing parameter for item j, θ_i is the latent ability for candidate i, $P_j(\theta_i)$ =Prob$(y_{ij} = 1|\theta_i)$ is the probability of correct response to item j from the candidate with ability θ_i defined for a given item response theory model.

Let $P(\boldsymbol{Y_i}|\theta_i, \omega) = \prod_{j=1}^{N}[P_j(\theta_i)^{y_{ij}}\, Q_j(\theta_i)^{1-y_{ij}}]$ (3)
Where $\boldsymbol{Y_i} = (y_{1i}, y_{2i}, \ldots, y_{Ni})^1$, $y_{ij} = 0, 1$ is a response from candidate i to item j, and ω is a matrix of true item parameter.

The total likelihood of the observed item responses in equation (3) yields

$$L = \prod_{i=1}^{n}\prod_{j=1}^{N}[P_j(\theta_i)^{y_{ij}} Q_j(\theta_i)^{1-y_{ij}}] \tag{4}$$

$$LogL = \sum_{i=1}^{n}\sum_{j=1}^{N}[y_{ij} LogP_j(\theta_i) + (1 - y_{ij})Q_j(\theta_i)\,] \tag{5}$$

Under ML, the parameters estimates are the item parameter values that maximize the likelihood equation (3) given the observed responses. This is obtained by setting the first derivatives of the log- likelihood equal zero (Lord, 1980). Solving the resulting equations simultaneously, for the item estimates yields

$$\left.\begin{aligned} a_j &= (1 - c_j)\sum_{i=1}^{n}\big[y_{ij} LogP_j(\theta_i)\big](\theta_i - b_j)\gamma_{ij} = 0 \\ b_j &= (-a_j)(1 - c_j)\sum_{i=1}^{n}\big[y_{ij} LogP_j(\theta_i)\big]\gamma_{ij} = 0 \\ c_j &= (1 - c_j)^1(1 - c_j)\sum_{i=1}^{n}\big[y_{ij} LogP_j(\theta_i)\big]/P_j(\theta_i) = 0 \end{aligned}\right\} \tag{6}$$

Where $\gamma_{ij} = \frac{P*_j(\theta_i)Q*_j(\theta_i)}{P_j(\theta_i)Q_j(\theta_i)}$, assuming θ_i are known, the three-parameter for the j^{th} items are estimated simultaneously using equations (5), and (6) as a system of equations. When responses are grouped into a finite number of known ability levels, the proportion of correct response at ability level θ_i is given by $P_{ij} = \frac{r_{ij}}{n_{ij}}$, and $1 - P_{ij} = \frac{n_{ij} - r_{ij}}{n_{ij}}$ where n_{ij} is the number of candidates in the group that have ability θ_i, r_{ij} gives correct responses and $n_{ij} - r_{ij}$ gives incorrect responses. Assuming P_{ij} is binomially distributed with $E(P_{ij}) = P_{ij}$, and variance $\frac{P_{ij}Q_{ij}}{n_{ij}}$, the likelihood function of the responses observed on N items is administered to the group of candidates with abilities $\theta_1, \theta_2, \ldots, \theta_i, \ldots . \theta_q$ where q the number of groups, by conditioning on the item parameters, we have equation (7)

$$L = \prod_{i=1}^{q}\prod_{j=1}^{N}\left[\frac{n_{ij}!}{r_{ij}!(n_{ij} - r_{ij})!}\right] P_j(\theta_i)^{r_{ij}} Q_j(\theta_i)^{1-r_{ij}} \tag{7}$$

$$\text{Log}(L) = \sum_{i=1}^{q}\sum_{j=1}^{N}\{r_{ij}\text{log}P_j(\theta_i) + (n_{ij} - r_{ij})\text{log}Q_j(\theta_i)\} \tag{8}$$

Taking the first derivatives of equation (8) with respect to each of the item parameters and solved the resulting equations produce equations (9).

$$\left.\begin{aligned} a_j &= (1 - c_j)\sum_{i=1}^{q}\big[r_{ij} - n_{ij}P_j(\theta_i)\big](\theta_i - b_j)\gamma_{ij} = 0 \\ b_j &= (-a_j)(1 - c_j)\sum_{i=1}^{n}\big[r_{ij} - n_{ij}P_j(\theta_i)\big]\gamma_{ij} = 0 \\ c_j &= (1 - c_j)^1(1 - c_j)\sum_{i=1}^{n}\big[r_{ij} - n_{ij}P_j(\theta_i)\big]/cP_j(\theta_i) = 0 \end{aligned}\right\} \tag{9}$$

γ_{ij} is as defined in equation (6) and system of equations in (9) is the likelihood function for group responses for JMLE.

In Bayesian modelling framework, model parameters are treated as random variables unlike frequentist approach, and have its prior distributions which described uncertainty about the true values of the parameters before observing the responses at separate levels for the purpose of accounting for various sources of information. The model comprises of likelihood model for observed responses which described data generating process as a function of unknown parameters, and the likelihood model present the density of the responses which is conditional on the model parameters (Fox, 2010).

Let $P(y/\theta)$ denotes information about θ (student's ability) that is observed from response data y which is called likelihood function because our interest is usually on the distribution of parameters θ based on the observed data. Candidate's ability θ on our response data is presented in equation (10).

$$P(\theta|y) = \frac{P(y/\theta)P(\theta)}{P(y)} \quad (10)$$

$$\propto P(y|\theta)P(\theta) \quad (11)$$

$P(\theta|y)$ is the posterior distribution of ability θ based on the prior beliefs, and sample information. Equation (10) represents Bayes theorem and equation (11) is a product of likelihood $L(y,\theta)$ and prior. In Bayesian statistics,

$$\left.\begin{aligned} L(\theta;y) &= f(y;\theta) \\ &= \prod_{i=1}^{n} f(y_i|\theta) \end{aligned}\right\} \quad (12)$$

Where $f(y_i|\theta)$ is the probability density for y_i given θ. we assume θ has a probability distribution $P(\theta) = \pi(\theta)$ which is called prior distribution. We apply Bayes theorem in equation (10) to determine posterior distribution of θ given y since y and θ are random.

$$\left.\begin{aligned} P(\theta|y) &= \frac{P(y|\theta)P(\theta)}{P(y)} \\ &= \frac{f(y;\theta)\pi(\theta)}{m(y)} \end{aligned}\right\} \quad (13)$$

Where $m(\theta) \equiv P(y)$ is known as the marginal distribution of y, and

$$m(\theta) = \int f(y;\theta)\pi(\theta)d(\theta) \quad (14)$$

Equation (14) does not depend on the parameter of interest θ therefore can be reduced to

$$P(\theta|y) \propto L(\theta;y)\pi(\theta) \quad (15)$$

Equation (15) can be written a more computational friendly as

$$\ln\{P(\theta|y)\} = L(\theta;y) + \ln\{\pi(\theta)\} - c \quad (16)$$

Where constant $c = \ln\{m(y)\}$ and is assume to be finite for statistically valid analysis.

Every MCMC algorithm generates values from a transition kernel in such a way that the draws from that kernel converge to a pre-specified intended distribution; this simulates a Markov chain with the target distribution as stationary or equilibrium distribution of the chain. The acceptance rate of the chain and the degree of autocorrelation in the generated sample are two yardsticks to measure the efficiency of MCMC. Blocking of model parameters into two or more blocks such that MH updates are applied to each block separately in the order to achieve an effective solution are necessary however, a check for convergence of MCMC is a very important step because a valid Bayesian inference is based on the convergence of Markov chain by MCMC sample which is drawn from the desired posterior distribution. The diagnostic usually entails checking for necessary conditions for convergence (Gelman, 2014).

Aim and Objectives

This paper sought to demonstrate advantage of Bayesian over frequentist (maximum likelihood method) IRT to the test developers, analysts, and assessors in helping practitioners to determine instances in which Bayesian method might be better preferred to classical IRT. Therefore, the stated objectives are:

i. To determine the most suitable method for one, two, and three parameter logistic model.
ii. Circumvent non-convergences in three parameter logistic model.

Materials and Methods

Materials

The item comprises of a compulsory 35 statistics multiple choice items which were scored as 1 for correct response and 0 for each of the three distractors that were administered to 403 undergraduate students. Stata 16 SE on window 7 was used for the analysis. Both maximum likelihood method and MCMC algorithms for 15,000 iterations with first 5,000 as a burn-in yields results in Tables 1, 2, and 3 which contained frequentist IRT model estimates, and the posterior summaries of item properties for items 8, 11, 29, 7, 5, and 34 were presented on the basis of their performances with one, two, and three-parameter logistic models respectively.

Methods

The goal of one-parameter logistic model is to estimate a common discrimination of the items and their individual difficulties. Equation (17) as IRT model describes test items in terms of only how hard an item is perceived to achieve 0.5 probability of correct response at a given ability level (Lord, 1968; Birnbaum, 1968).

$$P_{ij}(\theta = 1|a, b_j, \theta_i) = \frac{e^{a(\theta_i - b_j)}}{1 + e^{a(\theta_i - b_j)}} \quad (17)$$

a is a common item discrimination parameter, b_j is the item difficulty parameter for item, and θ_i is the student $i's$ ability, $(\mathrm{i} = 1, 2, \ldots, \mathrm{n})$.

The likelihood of equation (17) has a generalized linear regression form of equation (18).

$$\left.\begin{array}{l} Logit\{\Pr(y_{ij} = 1|b_j, \theta_i)\} = a(\theta_i - b_j) \\ : \theta_i \sim i.i.d.N(0,1) \end{array}\right\} \quad (18)$$

Where a, b_j, and θ_i were as defined in equation (17). By re-parameterization, the discrimination parameter $\boldsymbol{a}$ can be absorbed into $\boldsymbol{b_j}$ and θ_i so that equation (18) yields

$$\left.\begin{array}{l} Logit\{\Pr(y_{ij} = 1|\overline{b_j}, \overline{\theta_i})\} = a(\overline{\theta_i} - \overline{b_j}) \\ \sigma = a, \\ \overline{b_j} = -ab_j \end{array}\right\} \quad (19)$$

Equations (3) through (6) were used to get our frequentist estimates.

A Bayesian formulation of equation (17) requires prior specifications for σ, and $\overline{b_j}$ as

$$\left.\begin{array}{l} \sigma^2 \sim \mathrm{InverseGamma}\ (0.01, 0.01) \\ \overline{b_j} \sim N(0,10) \end{array}\right\} \quad (20)$$

The random effects are assigned zero mean normal prior with σ^2 and the parameter σ^2 is assigned a non-informative inverse gamma prior with shape 0.01 and, size 0.01. Due to the large number of random effects, we exclude them from simulation results.

Two-parameter logistic model is viewed as a latent trait model in which item characteristic curve (ICC) is of two-parameters. It estimates varying items discrimination $a_j : (j = 1, 2, \ldots, N)$ that discriminate different items in relation to ability level near the infection point of an ICC, and varying item difficulty b_j ; thereby describes test items in terms of two parameters. The probability that candidate i with ability θ_i endorsed an item j correctly is given as

$$P_{ij}(\theta = 1|a_j, b_j, \theta_i) = \frac{e^{a_j(\theta_i - b_j)}}{1 + e^{a_j(\theta_i - b_j)}} \quad (21)$$

a_j discriminating parameter for item j, and i, j, b_j, and θ_i were as defined in equation (17).

Bayesian modelling of equation (21) required the following prior specifications:

$$\left.\begin{array}{r} \theta_i \sim i.i.d.N(0,1) \\ \ln(a_j) \sim N(\mu_a, {\sigma^2}_a) \\ b_j \sim N(\mu_b, {\sigma^2}_b) \\ \mu_a, \mu_b \sim N(0,1) \\ {\sigma^2}_a, {\sigma^2}_b \sim \text{Gamma}\ (1,1) \end{array}\right\} \quad (22)$$

Where a_j, b_j, and θ_i were as defined in equation (21).

In the two-parameter logistic model, a lower asymptote (pseudo-guessing) parameter $c_j\ (j = 1,2,\dots,N)$ is introduced into the model in equation (21) to produce a three-parameter logistic model equation (23) which accounts for variability in item discriminating parameter, meaning that a candidate at the lower trait level have a non-zero probability of endorsing item correctly (Wright & Stone, 1979). The probability that student i with ability θ_i endorsed item j correctly is given as

$$P_{ij}(\theta_i) = C_j + (1 - C_j)\ \frac{e^{a_j(\theta_i - b_j)}}{1 + e^{a_j(\theta_i - b_j)}} \quad (23)$$

Such that C_j is guessing parameter, a_j, b_j, and θ_i were as defined in equation (21).

The guessing parameter may be difficult to estimate using maximum likelihood; it does not converge easily. Instead of using likelihood (logit), we use likelihood (dbernoulli ()) to model the probability of success of a Bernoulli outcome directly as we have in equation (24).

$$P(Y_{ij} = 1 | a_j, b_j, c_j, \theta_i) = c_j + (1 - c_j) Invlogit\{a_j(\theta_i - b_j)\} : c_j > 0 \quad (24)$$

Bayesian modelling of equation (24) required the following prior specifications

$$\left.\begin{array}{r} a_j \sim \text{Lognormal}\ (\mu_a, var_a) \\ b_j \sim \text{Normal}\ (\mu_b, var_b) \\ c_j \sim \text{Inverse} - \text{gamma}\ (10,1) \\ \theta_i \sim \text{Normal}\ (0,1) \end{array}\right\} \quad (25)$$

These hyper prior distributions are specified for hyper-parameters

$$\left.\begin{array}{r} \mu_a, \mu_b \sim \text{Normal}\ (0,1) \\ var_a, var_b \sim \text{Inverse} - \text{gamma}\ (10,1) \end{array}\right\} \quad (26)$$

Mathematical Theory behind Prior and Hyper Prior Selection

1. If $y_i \sim$ N (μ, σ^2), $i = 1,2,\dots,n$ with σ^2 known, and $\boldsymbol{y} = (y_1, y_2, \dots, y_n)^1$ then $\mu \sim N(\mu_0, {\sigma^2}_0)$ is the prior density for μ, and posterior density

$$\mu|\boldsymbol{y} \sim \text{N}\left(\frac{\mu_0\, {\sigma_0}^{-2} + \bar{y}\frac{n^2}{\sigma}}{{\sigma_0}^{-2} + \frac{n^2}{\sigma}}, \left({\sigma_0}^{-2} + \frac{n^2}{\sigma}\right)^{-1}\right) \quad (27)$$

Equation (27) is a conjugate prior of normal data when its variance is known.

2. If $\theta = (\mu, \sigma^2)^1$ has a normal/inverse-gamma prior density with parameters μ_0, n_0, v_0, and ${\sigma_0}^2$, the posterior density of θ is also a normal/inverse-gamma density with parameters μ_1, n_1, v_1, and ${\sigma_1}^2$, where $\mu_1 = \frac{n_0\mu_0 + n\bar{y}}{n_0 + n}$, $n_1 = n_0 + n$, $v_1 = v_0 + n$,

$v_1{\sigma_1}^2 = v_0{\sigma_0}^2 + S + \frac{n_0 n}{n_0 + n}(\mu_0 - \bar{y})^2, S = \sum_{i=1}^{n}(y_i - \bar{y})^2$. Therefore,

$$\left.\begin{array}{r} \mu|\sigma, y \sim N\left(\mu_1, \frac{\sigma^2}{n_1}\right) \\ \sigma^2|y \sim \text{inverse} - \text{gamma}\left(\frac{v_1}{2}, \frac{v_1{\sigma_1}^2}{2}\right) \end{array}\right\} \quad (28)$$

Equation (28) is a conjugate prior of normal data when variance is unknown.

3. If $y_i \sim$ lognormal$(\theta, \phi), i = 1.2, \dots, n$. Then the probability density function of log-normal random variable y is

$$f(y; \theta, \phi) = \frac{1}{y\sqrt{2\pi\phi}} e^{\frac{1}{2\phi}(\ln(y) - \theta)^2}\ , y > 0, -\infty < \theta < \infty, \phi > 0 \quad (29)$$

ϕ is scale and θ is location parameter. The likelihood function is

$$L(\theta) = \frac{(2\pi\phi)^{-\frac{n}{2}}}{\prod_{i=1}^{n} y_i} e^{-\frac{1}{2\phi}\sum_{i=1}^{n}(\ln(y_i)-\theta)^3} \quad (30)$$

We assume that location parameter θ follows the normal distribution with hyper-parameters a_n, and b_n, and the prior distribution is

$$f_N(\theta) = \frac{1}{y\sqrt{2\pi b_n}} e^{\frac{1}{2b_n}(\theta-a_n)^2} \quad , b_n > 0, -\infty < \theta < \infty, -\infty < a_n < \infty \quad (31)$$

Therefore, posterior distribution of θ given data y is

$$f(\theta|y) = \frac{1}{\sqrt{2\pi{\sigma_n}^2}} e^{-\frac{1}{2{\sigma_n}^2}(\theta-\mu_n)^2, \ -\infty<\theta<\infty} \quad (32)$$

Where $\mu_n = \{b_n \sum_{i=1}^{n} \ln(y_i) + a_n\phi \}/ (nb_n + \phi)$, ${\sigma_n}^2 = \frac{\phi b_n}{\phi + nb_n}$

Results and Discussions

Items 8 and 11 were identified by one-parameter logistic model equation (17) as the most and easiest difficult items respectively, items 29 and 7 were identified by two-parameter logistic model of equation (21) as the most and least discriminating items respectively while items 5 and 34 were identified by three-parameter logistic model equation (23) as the most and least pseudo-guessing items respectively. Therefore, these items were selected for discussions due to their identification by respective models.

A conditional maximum likelihood estimation and Bayesian estimation framework of equation (17) give rise to results in Table 1 which clearly show varying items difficulty given that discrimination is held constant, standard error of estimates, and confidence interval of the estimates based on frequentist approach. On the other hand, the mean of posterior distribution for individual item, standard error of the estimates, Monte Carlo standard error which measured variations in simulation process, median, and 95% credible interval for Bayesian framework were provided for purpose of comparing the two approaches.

Table 1: Comparisons of Frequentist and Bayesian Parameters Estimates in One-parameter Logistic Model

Item	Method	Par	Std Dev	MCSE	Median	95% Conf/Cred Int	
Disc	Frequentist	0.6314	0.0335			0.5653	0.6965
	Bayesian	0.6323	0.0339	0.0017	0.6314	0.5660	0.6982
8	Frequentist	2.4988	0.2466			2.0154	2.9821
	Bayesian	2.5108	0.2486	0.0083	2.4988	2.0430	3.0196
7	Frequentist	0.8455	0.1817			0.4893	1.2017
	Bayesian	0.8482	0.1829	0.0060	0.8447	0.5030	1.2063
29	Frequentist	-1.6034	0.2034			-2.0032	-1.2036
	Bayesian	-1.6055	0.2055	0.0068	-1.6034	-2.0173	-1.2105
5	Frequentist	-3.2486	0.2935			-3.8238	-2.6734
	Bayesian	-3.2461	0.2957	0.0099	-3.2329	-3.8649	-2.7028
34	Frequentist	-4.7140	0.4200			-5.5370	-3.8908
	Bayesian	-4.7181	0.4210	0.0132	-4.6991	-5.5972	-3.9443
11	Frequentist	-5.0360	0.4556			-5.9329	-4.1471
	Bayesian	-5.0390	0.4614	0.0145	-5.0194	-5.9901	-4.1953
Var	Hyper	0.4010	0.0429	0.0021	0.3986	0.3204	0.4875

Despite the fact that impact of prior is controlled by introduction of a non-information inverse gamma hyper-prior for $\boldsymbol{var}$, estimates from both approaches are not significantly different from each other (lead to same conclusion). The results of these approaches affirmed that both Frequentist and Bayesian methods yield close results and lead to same conclusions, the most difficult item as identified by both methods is 8 (with difficulty indices 2.4988 in Frequentist, 2.5108 in Bayesian, standard error of 0.2466 in Frequentist as against

0.2486 in Bayesian, and similar confidence/credible intervals) follows by item 7 in that order as displayed in Table 1.

On the two-parameter logistic model, estimated item properties of the model presented in equation (21) using both maximum likelihood and Bayesian methods where items are arranged in descending order of discriminatory power of individual item are presented in Table 2. The methods identified item 29 as most discriminatory in terms of classifying students into different ability levels (discriminating indices of 1.6451 in Frequentist as against 1.4729 in Bayesian, difficulty indices of -0.8414 in Frequentist as against -0.8845 in Bayesian meaning moderately easy item),

Table 2: Comparisons of Frequentist and Bayesian Parameters Estimate of Two-parameter Logistic Model

Item	Method	Par	Mean	Std Dev	MCSE	Median	95% Conf/Cred Int	
29	Freq	Disc	1.6451	0.2655			1.1248	2.1654
	Bayesian	Disc	1.4729	0.2511	0.0205	1.4584	1.0093	2.0126
	Freq	Diff	-0.8414	0.1162			-1.0692	-0.6136
	Bayesian	Diff	-0.8845	0.1353	0.0110	-0.8785	-1.1693	-0.6668
34	Freq	Disc	1.4478	0.3701			0.7224	1.1731
	Bayesian	Disc	1.2607	0.2994	0.0335	1.2300	0.7401	1.9265
	Freq	Diff	-2.4717	0.4329			-4.1313	-2.0375
	Bayesian	Diff	-1.9039	0.3677	0.03356	-1.9136	-2.6433	-1.1327
11	Freq	Disc	1.1397	0.3498			0.6569	1.7669
	Bayesian	Disc	1.1016	0.2655	0.0256	1.0675	0.6569	1.7669
	Freq	Diff	-3.0927	0.7323			-4.5280	-1.6573
	Bayesian	Diff	-1.9356	0.5041	0.0429	-1.9638	-2.8008	-0.7904
5	Freq	Disc	0.5727	0.1935			0.1935	0.9520
	Bayesian	Disc	0.6361	0.1285	0.0113	0.6279	0.4243	0.9167
	Freq	Diff	-3.5352	1.0990			-5.6892	-1.3813
	Bayesian	Diff	-0.1297	0.7173	0.0868	-0.0396	-1.8174	1.0727
7	Freq	Disc	0.0570	0.1194			-0.1771	0.2911
	Bayesian	Disc	0.2679	0.0733	0.0039	0.2615	0.1429	0.4249
	Freq	Diff	8.6179	18.112			-26.8812	44.117
	Bayesian	Diff	2.6543	0.4918	0.0175	2.5901	1.8941	3.8511
Hyper		μ_a	-0.4357	0.1070	0.0047	-0.4368	-0.6443	-0.2166
		var_a	0.3053	0.0967	0.0051	0.2905	0.1590	0.5408
		μ_b	-0.8730	0.3080	0.0114	-0.8799	-1.4618	-0.2511
		var_b	3.1800	0.8983	0.0468	3.0282	1.8167	5.2871

follows by item 34 as displayed in Table 2 in that order while item 7 is perceived as the least with a questionable difficulty index of 8.6556, too high standard error, and poor discriminatory power as against Bayesian estimates. This model suggests that this item is unfit to be used except replaced, or a total item moderation is done. The hyper-parameters that controlled the effect of priors on our data were also displayed.

Attempt to determine the three item properties of the model position in equation (23) using Frequentist method encountered serious convergence problem which may affect statistical inferences except one of the item properties is constrained. Instead of modelling this data with logit link, we modelled directly with Bernoulli link as we have in equation (24) and the results are presented in Table 3.

Table 3: Comparisons of Frequentist and Bayesian Parameters Estimate of Three-parameter Logistic Model

Item	Method	Par	Mean	Std Dev	MCSE	Median	95% Conf/Cred Int	
5	Freq	Guess	0.7757	0.0545			0.6690	0.8824
	Bayesian	Guess	0.6968	0.1166	0.01433	0.7241	0.3789	0.8380
	Freq	Diff	0.3014	0.4981			-0.6749	1.2777
	Bayesian	Diff	-0.1190	0.7663	0.0906	0.0111	-1.9014	1.1831
	Bayesian	Disc	1.3945	0.4805	0.0441	1.3141	0.6972	2.5416
7	Freq	Guess	0.3643	0.0320			0.3016	0.4269
	Bayesian	Guess	0.3491	0.0334	0.0016	0.3512	0.2834	0.4124
	Freq	Diff	3.3008	1.0879			1.1686	5.4331
	Bayesian	Diff	2.5990	0.4869	0.0196	2.5472	1.8155	3.6873
	Bayesian	Disc	1.6123	0.5335	0.0402	1.5190	0.8676	2.9412
8	Freq	Guess	0.0966	0.0259			0.0458	0.1473
	Bayesian	Guess	0.1663	0.0207	0.0008	0.1644	0.1300	0.2084
	Freq	Diff	2.0544	0.2449			1.5744	2.5344
	Bayesian	Diff	2.0182	0.2819	0.0172	1.9820	1.5366	2.6644
	Bayesian	Disc	2.0182	0.2819	0.0172	1.9820	1.5366	2.6644
11	Freq	Guess	0.0033	0.0440			-0.0859	0.0865
	Bayesian	Guess	0.5073	0.1579	0.0131	0.4862	0.2485	0.8479
	Freq	Diff	-0.9016	0.2948			-1.4793	-0.3239
	Bayesian	Diff	-1.8839	0.5539	0.0482	-1.9497	-2.8757	-0.5660
	Bayesian	Disc	1.4916	0.3272	0.0236	1.4497	0.9295	2.2747
29	Freq	Guess	0.0064	0.2061			-0.3976	0.4103
	Bayesian	Guess	0.3151	0.0607	0.0034	0.3121	0.2059	0.4383
	Freq	Diff	-0.9016	0.2948			-1.4793	-0.3239
	Bayesian	Diff	-0.3126	0.1803	0.0146	-0.3242	-0.6461	0.0728
	Bayesian	Disc	1.4729	0.2511	0.0205	1.4584	1.0093	2.0126
34	Freq	Guess	0.0002	0.0013			-0.0025	0.0025
	Bayesian	Guess	0.4440	0.1251	0.0086	0.4288	0.2422	0.7179
	Freq	Diff	-2.5885	0.1781			-2.9376	-2.2395
	Bayesian	Diff	-1.8169	0.4207	0.0290	-1.8300	-2.6341	-0.9259
	Bayesian	Disc	1.5384	0.3016	0.0219	1.5212	1.0249	2.1941
	Freq	Disc	1.3742	Constrained				
Hyper	Bayesian	$m\mu_a$	0.3580	0.0893	0.0070	0.3595	0.1782	0.5267
Hyper	Bayesian	var_a	0.1006	0.0308	0.0019	0.0944	0.0569	0.1750
Hyper	Bayesian	$m\mu_b$	0.2576	0.1772	0.0072	0.2612	0.1045	0.6046
Hyper	Bayesian	var_b	1.2236	0.3226	0.0226	1.1767	0.7362	1.9874

The results of both methods are presented in Table 3 for clear comparisons. Only two item properties can be estimated (guessing and difficulty indices) for individual item with its associated standard error and 95% confidence interval were estimated while discriminatory index was constrained to circumvent non-convergence of log-likelihood using Frequentist framework. Bayesian method on the other hand allows the estimation of all the item properties (guessing, difficulty, and discriminatory indices) at the same time without convergence problems. Posterior means, medians, standard error of individual estimate, Monte Carlo standard error which assesses the variation of Monte Carlo Markov Chain simulation of our estimate, and 95% credible interval were also presented. Hyper parameters which reduced effect of prior on the observed data are also displayed in Table 3.

A more careful comparison of guessing indices from both approaches show very close results with frequentist have lower standard error of estimated parameters (guessing and

difficulty) but estimation of varying third parameter (discriminatory index) spark non-convergence of the log-likelihood hence, one of these parameters need to be constrained in frequentist. For item 5 with pseudo-guessing indices of 0.7757 in Frequentist as against 0.6968 in Bayesian, Frequentist had a better standard error of 0.0455 as against 0.1166 in Bayesian but Frequentist is limited for its inability to estimate all the three item properties at a time. However, Bayesian framework provided more parsimonious estimates for all the item properties without constrained any item property and present credible intervals that capture our current uncertainty in the location o 95% percent values and is interpreted as probabilistic statement about the parameter.

It has been discovered that when the interest of test developers or administrators is only to identify or describe how hard or easy each of the items that make up test are perceived to achieve 0.5 probability of correct response at a given ability level, the model illustrated in equation (17) yielded results in Table 1 which affirmed that frequentist (conditional maximum Likelihood) method converges and produced least standard errors compared to Bayesian method. Both approaches lead to same conclusions, and here, confidence interval from frequentist approach and credible interval from Bayesian are statistically the same.

However, when test analyst is much interested in how each item discriminate students into different ability levels, two parameters logistic IRT model positioned in equation (21) is used. The two approaches used in estimating these item properties produced the same results as shown in Table 2 and this suggests that either of the methods can be used to estimate parameters of interest but Bayesian two-parameter logistic IRT model produced least standard error with a better credible interval.

Test administrators interested in estimating varying three item properties employed IRT model displayed in equation (23). The estimates of this model on basis of frequentist and Bayesian methods are presented in Table 3. Non-convergence problem was encountered when trying to estimate varied all the three item properties using frequentist approach. Just to circumvent non-convergence of log-likelihood here, we have to constrain discrimination parameter for all the items. Bayesian framework provided the needed alternative method of estimating all the item properties. Standard error of the estimates, median and posterior means which were compromised between the prior and likelihood were presented, Monte Carlo standard error fall within the acceptable range and 0.95 credible interval. All these attested to the fact that to overcome serious problem of non-convergence when low ability students having non-zero probability of endorsing item correctly, Bayesian approach will be most appropriate.

Conclusions

This work presents basis for examination bodies, test developers, and analysts in taking right decisions on the method of analysis to be employed when confronted with modelling latent traits which cannot be observed directly. Methods to be used here will be a function of particular item response theory models under considerations as presented as follows.

1. Either Bayesian or frequentist approach may be used when our interest is to describe items that made up a test in terms of one item property (that is difficulty) or in terms of two varying item; properties (discriminations and difficulty). Both methods will lead to same conclusions but Bayesian produced smaller standard error.

2. Bayesian method is to be used when our intention to estimate varying three item properties to circumvent non-convergence of log-likelihood.

3. Suitable (conjugate) prior must be specified for our parameters of interest to obtained valid estimate and conjugate hyper-prior must be considered to reduce effect of prior on our estimates.

References

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability in Lord and Novick (Edition). Addison-Wesley.

Bock, R. D., & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: application of an EM algorithm. Psychometrika, 46, 443-459.

Bock, R. D., & Lieberman, M. (1970). Fitting a response model for n dichotomously scored items. Psychometrika, 35, 179-197.

Fox, J. P. (2010). Bayesian item response modelling theory and applications. Springer.

Gelman, A., Carlin, J. B., Stem, H. S., Dunson, D. B., Vehtari, A., & Rubin, D. B. (2014). Bayesian data analysis (3rd ´ed.). FL: Chapman and Hall/CRC.

Lord, F. M. (1968). Analysis of the verbal scholastic aptitudes test using brnbaum's three-parameter logistic model. Educational and Psychological Measurement, 28, 989-1020.

Lord, F. M. (1980). Applications of item response theory to practical testing problem. Hillsdale.

Mislevy, R. J., & Bock, R. D. (1984). BILOG: item analysis and test scoring with binary logistic models. Scientific Software.

Paek, I., & Cole, K. (2020). Using R for item response theory models and applications. Taylor and Francis Group, London.

Wright, B. D., & Stone, M. (1979). Best test design. MESA Press.