Right to Information Query Modelling via Graded Response Model

Nayantara Kotoky[1] and Vijaya V Saradhi[2]

[1] Indian Institute of Technology Guwahati
`nayantara@iitg.ernet.in`
[2] Indian Institute of Technology Guwahati
`saradhi@iitg.ernet.in`

Abstract. Right to Information (RTI) Act, 2005 empowers citizens of India to access information from any governmental organization. Using this Act citizens can ask questions (through RTI applications/queries) to government offices and obtain answers. In this work we attempt to model RTI queries. Objective of modeling is to understand the latent patterns such as transparency and effectiveness of RTI Act implementation in the RTI query-reply process which are suggestive of possible amendments in the Indian Constitution. We employ Graded Response Model (GRM, a variant of Item Response Theory) for obtaining the latent patterns. A synthetic dataset corresponding to central and state educational institutions is constructed which has close characteristics to the collected RTI query dataset. From the GRM we infer that certain institutes are highly transparent in replying to citizen's questions across various categories. We also infer that RTI act's implementation is not uniform across diverse categories within a transparent institution.

Keywords: Item Response Theory, Graded Response Model, Right to Information

1 Introduction

Right to Information (RTI) Act 2005 empowers citizens of India to access information from any public institution (those institutions that are funded by government). The RTI Act came into force in October 12, 2005. Through this act citizens can inspect official documents, contracts, press releases, records, notes, certified copies by filing an RTI application/query. Each institution appoints a Public Information Officer (PIO) to implement the RTI act and reply to the questions posed by citizens. Citizens submit a hard copy of their questions to the PIO. Every RTI application costs a sum of Indian rupees ten. PIO is responsible to reply to the query within a fixed time period (typically 30 days).

RTI queries form a source of information where one can witness citizens' interaction with government establishments. Such a rich source of information when analyzed can throw light on the sensitivities of citizens and weaknesses in the implementations of laws. Certain acts have been amended using the RTI

statistics. In particular, RTI Act itself is amended through the following two examples:

1. Inclusion of Indian Postal Orders:
   For fee payment with an RTI application, the acceptable modes of payment were Banker's cheque, Demand Draft or by cash. All three modes had their own additional burdens. Both demand draft and banker's cheque had their service charges attached, and payment by cash required visiting the public institution in person. Indian Postal Order (IPO) is another convenient mode of paying fees, with a nominal charge of 10%, which is Re.1 for the fee of Rs. 10. However, IPOs were not acceptable as a mode of payment, because of which there were multiple rejections of RTI applications which were perfectly good with their content. This event was widespread enough to catch the government's eye, and there was discussion of its inclusion as a mode of payment. Ultimately changes were made to the RTI Act's scope by adding IPOs as a mode of payment. [3].

2. Exemption of political parties from being a public authority:
   Asking for source of funding for political parties is not uncommon. With the intention of understanding the inner workings of the organization, political parties are often asked to cite their source of funding. With the advent of RTI, multiple applications were filed asking their financial details. The parties argued that they are not under the direct funding of the central or state government and hence are not liable to divulge such information. Such queries were repeatedly rejected, and a notice was issued stating political parties as not being public authorities. It was finally included in the RTI amendment Bill 2013 [4].

From the above two examples it is observed that "repeated rejections" of RTI queries served as a feedback for introduction of amendments into existing RTI act. This leads us to believe that the latent patterns in the RTI query log provide potential pointers for predicting future amendments. The objective of this work is to collect the RTI queries and associated response (whether the institute has replied to the query, rejected the query or referred to third party) by institutions across India, model thus collected text data and identify latent patterns in the RTI query database.

We propose to model the RTI queries text database as a two dimensional matrix whose rows correspond to institutions and columns correspond to topics on which questions were posed to individual institutions. An entry ij in this matrix correspond to percentage of replies an institute i has given against a query topic j. This matrix is given as input to the Graded Response Model (GRM) to identify latent patterns in the RTI query-reply process.

After running GRM on our RTI data, each institution has been designated a 'transparency' value that determines how effective an institution is with respect to replying RTI queries, and indicates a difference between the central and state educational institutions. The model also identifies differences in the query-reply process along different query topics.

Contributions:

1. This is the first attempt in collecting RTI query-reply data across India.
2. A two dimensional query-reply matrix is constructed out of the given RTI query-reply text database instead of using conventional text modeling methods such as vector space model, latent semantic indexing, LDA etc.
3. We employ for the first time psychometric models in RTI query text document analysis.

2   Related work

2.1   Modelling the Political Domain

Attempts to model the legislative structure and outlook have been seen in the literature. Now and again, researchers have sought to apply mathematical models to represent affairs in the political domain. Such work opens up scope for understanding the political issues in depth. Gerrish and Blei [5] developed a probabilistic model for legislative data to identify voting patterns in specific political issues. They used the text of bills to identify the specific topics to which the bills relate to, and attempted to identify what the lawmakers' stance is with respect to bills with different topics (issues). They argued that a lawmaker's attitude cannot be captured accurately on the broad political structure since they do not exhibit enough regularity in the voting patterns. It is assumed that they have an overall (general) political stand but have different political stand on specific issues that the bills are based on. The paper introduces an issue adjusted model that identifies each lawmaker's position on individual topics, called the 'issue adjusted ideal-point model'. The adjusted model has been able to identify the lawmakers' political stand in a more realistic way, and for each issue individually.

Poole and Rosenthal [6] analysed a variant of voting patterns, namely, roll call data for legislators' votes. They took US voting data where choosers are representatives of the law or senators, and the choices are binary, that is, yes or no. They developed a unidimensional model of probabilistic roll call voting, and the methods can be applied to the analysis of voting in popular elections and other forms of political choice behavior.

2.2   Forms of Queries

From classrooms to commercial platforms and entertainment, queries are found everywhere and in all forms. Examples include e-commerce queries, customer service queries, product review queries, tourism queries, personal and rhetorical queries (natural language), queries in an Issue Tracking System, queries in medical diagnosis and of course RTI queries. Each of these query types have different models of analysis. Some of the ways of modelling are:

1. Web search engines:
   Web queries (queries put to search engines for web search) are analysed to improve user experience and search engine performance. Research has

been done to find user goals from queries [7] and temporal dynamics of query patterns have been studied [8]. [9] proposes methods for clustering similar queries together, which helps us to understand how frequent and how diverse web queries are. Traditional information retrieval mostly depended on simple term matching between queries and documents. However, it has been observed over time that understanding the meaning of the query is important in improving the precision of a search result, like certain keywords have more relevance in a given query and synonyms need to be identified. Attempts to find such hidden semantics in the queries have been made by [10].

2. Question-answer (Q/A) system:
Q/A systems do not retrieve documents, but give brief, relevant answers in short text. This speciality requires time, processing power as well as computation and understanding of the semantics of the query. In order to overcome the bottlenecks of natural language understanding, an amalgamation of statistical and representation based methods is required. Semantic information in questions and answers classification is studied in [11]. [12] has attempted to design a paraphrase component in a natural language question-answer system, whereas [13] has presented a new topology to support construction of question-answer systems.

3. Examination sets (questionnaires/test questions):
Questions are used to determine the qualification of individuals or behaviour of events. Typical examples are the survey questions under social or business context, tests for students, diagnosis of illness etc. Applications include attempts to model response behaviours and finding optimum set of questions for judgement. Examples are equating tests [14], understanding family relationships [15] etc.

3   Item Response Theory

3.1   Description

Item Response Theory (IRT) is a method for psychometric analysis. It uses statistics to analyse how people (test takers) respond to different questions and elements. The modelling of the data is done as a function that is an adjustment between two criteria:

- The persons abilities, perspective or personality traits and
- The item (question) difficulty.

The perception behind IRT is that probability of a correct response to an item is a mathematical function of person and the item parameters. IRT treats difficulty of each item as information to be incorporated in scaling items. The person parameter is interpreted as a single latent trait. Example of person parameters include intelligence, attitude etc. Likewise, we have item parameters that are taken into consideration like difficulty of the item, discrimination (slope

or correlation) representing how sharply the rate of success of persons varies with their ability, or guessing parameter which characterises certain items that even low intelligent persons can attempt to get correct response by guessing. IRT has a few presumptions. The first is that all items are independent of each other. Hence each item is modelled separately with its own set of parameters (which shall be discussed next). Second is that the response of a person to an item can be modelled by a mathematical Item Response Function (IRF). Also, the latent trait theta (θ) is assigned to each person that gives us the ability of the persons in a unidimensional scale. The main advantage of IRT is that the ability parameter (θ) and the item difficulty parameter are modelled on the same scale. We can imagine ability (intelligence of the student) and difficulty (of the questions) as two opposing parameters, both contributing to the probability of keying the correct response.

Measurement items with multiple response options also exist. In case of polytomous models, each category function must be modelled explicitly. We can imagine different response categories to be separated by boundaries. Responding in a particular category means responding between two boundaries of that response category. This gives rise to two types of conditional probabilities:

- Probability of responding in a given category
- Probability of responding positively rather than negatively at a given boundary between two categories

In case of polytomous items with multiple responses, in order to identify probability of responding in a particular category, we need to identify probability at both the boundaries. Positivity in one category boundary does not mean response to the adjacent response category. It simply means that probability is positive for all the subsequent categories, and might not refer to only the adjacent response category. Hence probability of responding to a particular category shall entail positivity at the lower category boundary and negative probability at the upper category boundary. This idea shall be exploited in the model that we shall use for our experiments.

3.2 Graded response model

The Graded Response Model (GRM) is a polytomous IRT model for ordinal response categories. It belongs to the class of Thurston/Samejima models. The GRM is an extension of the 2-Parameter Logic Model. Let θ be the latent ability underlying the response to the test items. The probability of a candidate with ability θ responding to item i in a particular category c is:

$$P_{ic}(\theta) = P_{ic}^*(\theta) - P_{ic+1}^*(\theta)$$

where

$$P_{ic}^*(\theta) = \frac{1}{1 + exp(-\alpha_i(\theta - \beta_{ic}))}$$

α_i is the Item slope parameter (one per item), β_{ic} is the Category threshold parameters and P_{ic}^* is the Category Boundary Response Function (CBRF) for item i and category c. There is one set of $\beta_{i1},...,\beta_{im}$ for each item and are ordered, where $m+1$ is the number of categories [16].

The psychological idea behind this is that in a dataset with polytomous response categories, each response category of an item exerts a level of attraction on persons taking the test. In the context of an entire item, being attracted to a category must take all prior category attractions into account. In other words, the probability of responding in any given category is a combination of being attracted through all previous categories up to the given category, but no further. In the case of ordered categories, this process means that to respond in a particular category, a person must have passed through all preceding categories. Let P_{ig} be the probability of responding in a particular category (g) to item i. If P_{ig}^* represents a CBRF in the Thurstone/ Samejima models (where both are conditional on θ), then

$$P_{ig} = P_{ig}^* - P_{ig+1}^*$$

The probability of responding in a particular category is equal to the probability of responding above (on the positive side of) the lower boundary for the category (ig) minus the probability of responding above the category's upper boundary $(ig+1)$.

3.3 Parameter Estimation

There are two types of parameters in IRT, that is, Item parameters and Person (ability) parameter. Since IRT is a trade-off between the two types, both are estimated iteratively to arrive at the best fit. For polytomous data, data is modelled by multiple dichotomizations at the category boundaries and finally using all the information to reach the final estimation of the parameters. For dichotomous data, parameter estimation is done differently for different parameters.

Estimating 'ability' parameter with known Item Parameters: To estimate an examinee's unknown ability parameter, it will be assumed that the numerical values of the parameters of the test items are known. It is an iterative process, and begins with some known values of the item parameters. The probability of the correct response to each item is then computed, and then the ability parameter is slightly adjusted so that the values closely match the observed values. The process is repeated until the adjustment becomes small enough that the change in the estimated ability is negligible.

$$\Theta_{s+1} = \Theta_s + \frac{\sum a_i[u_i - P_i(\Theta_s)]}{\sum a_i^2 P(\Theta_s)Q(\Theta_s)}$$

where Θ_s is the estimated ability of the examinee within iterations, a_i is the discrimination parameter of item i, u_i is the response given by examinee to

item i, $P_i(\Theta_s)$ is the probability of correct response to item i at ability θ and $Q_i(\Theta_s) = 1 - P_i(\Theta_s)$ is the probability of incorrect response to item i at ability θ.

Bayesian Estimation is used to estimate ability parameters given the item parameters. We have from Bayes' theorem

$$f(\Theta|u) = \frac{f(u|\Theta)f(\Theta)}{f(u)}$$

which can also be written as

$$f(\Theta|u) \propto L(u|\Theta)f(\Theta)$$

Taking logarithm of both sides,

$$lnf(\Theta|u) \propto lnL(u|\Theta) + lnf(\Theta)$$

The posterior is directly proportional to the likelihood multiplied by prior, where $f(\Theta \mid u)$ is the posterior estimate, $L(u \mid \theta)$ is the likelihood estimation and f(Θ) is the prior distribution. For each and every Θ, we can calculate the likelihood function and we also have the prior. So, we can calculate posterior distribution $P(\theta \mid u)$. The prior distribution has a bell shaped curve; hence the right hand side of the equation shall have a point with slope 0.

Estimating item parameter from response data: Let us divide examinees into J groups along the Θ scale so that all the examinees within a given group have the same ability level Θ_j, where $j = 1, 2, 3. \ldots J$. If r_j is the examinees that give correct response, then at an ability level of Θ_j, the observed proportion of correct response is $p(\Theta_j) = r_j/m_j$, where m_j is the total number of examinees in the group. Now the value of r_j can be obtained and $p(\Theta_j)$ computed for each of the j ability levels established along the ability scale. The main task now is to find an Item Characteristic Curve that best fits the observed proportions of correct responses.

For the estimation, initial values of item parameters are established. These values are then used to compute $p(\Theta_j)$ with the logistic equation. Iteratively, the item parameters are adjusted as well to find better values that reflect proximity with our observed data. This process of adjusting the estimates is continued until the adjustments get so small that little improvement in the agreement is possible. At this point, the estimation procedure is terminated and the current values of b and a are the item parameter estimates.

The method used to calculate item parameters from response data is called the marginal maximum Likelihood. Given the joint distribution of a function $f(x_1, x_2)$, we can calculate the marginal distribution of $f(x_1)$ as:

$$f(x_1) = \int_{-\infty}^{\infty} f(x_1, x_2)dx_2$$

Let y_i be the response vector for person i. $Y_i j$ shall be the response given by person i to item j. Let J be the total number of items, Θ_i be the ability of person i and Φ be the matrix of true item parameters. So we have,

$$f(y_i|\Theta, \Phi) = \prod_{j=1}^{J} Py_{ij}(\Theta_i)$$

Hence, the marginal distribution of item parameters can be given as:

$$f(y_i|\Phi) = \int f(y_i|\Theta, \Phi)g(\Theta)d(\Theta)$$

Let Y be the response matrix of each and every person and let there be n persons in total. So:

$$f(Y|\Phi) = \prod_{i=1}^{n} f(y_i|\Phi)$$

Taking logarithm of both sides for likelihood:

$$logL(Y|\Phi) = \sum_{i=1}^{n} logf(y_i|\Phi)$$

The value of Φ where likelihood function is maximised is found via Bayesian Estimation as described above.

4 Dataset

4.1 Data Collection

For the purpose of our study, we have decided to create an "RTI database" as part of our research. Our dataset consists of the RTI applications that have been posted to all public educational institutions by the citizens of India. The data collected consists of RTI applications (which include the RTI queries), date of reply of each query and the rejected queries with their grounds of rejection. This collection is going on and the database is not yet complete.

The data collection was formally started on 01.01.2015. RTI data is not found online, but have to be collected from each individual institution. Hence we resorted to filing an RTI application of our own asking for the data required, namely, all the RTI applications received by the institution, date of reply of each query and the rejected queries with their grounds of rejection. There is no facility for online RTI filing, so we had to post our RTI application to each institution. We started with the educational boards of high school and higher secondary level, and moved ahead towards universities. We shall collect RTI data from a total of 1053 educational institutions across India. Till date, we have filed RTI applications to a total of 360 institutions and have received a variety of replies to the same application from different institutions, both positive and negative.

Of the institutions that received our RTI application requesting the RTI data, 56.38% have rejected our application citing various reasons. Up to this date, we have collected data from a total of 44 institutions and 113 additional institutions have agreed to give us the data (on payment of extra money or collect data by visiting their office). The average time of receiving a reply to our application is 53.2 days. For the institutions that we have collected data from, it has taken us an average of 73.9 days to finally receive the data. This has resulted in around 35,000 RTI applications and reply stats. Each RTI application contains multiple queries (or sometimes just a single query). India is a multi-lingual country, and the queries are mostly found in the local language of the area to which the institutions belong. The data has not been processed yet, so the precise count for total queries is unavailable.

4.2 Data modelling

A citizen of India can ask an RTI query on any topic that is relevant to the institution to which the application is filed. There is also the provision of transfer of the RTI application to the appropriate department if the reply or sought document is not in the office that received the application. As a result, we find a variety of query types belonging to different topics. Upon closer inspection of the data received by us, it was observed that the queries can more or less be divided into some fixed number of topics. Some topics gets more queried, hence are popular among the masses, whereas some receive less queries. Areas of educational institutions such as Academic (marks), research, infrastructure are more targeted since people are more interested in knowing the workings of these departments. Hence analysing the RTI query-reply patterns of these specific topics is of paramount importance.

For our experiments using Graded Response Model, analysis can be done on the reply, rejection, and appeal stats etc. This shall indicate transparency among institutions and categories, probability of getting a query in a particular category accepted or rejected, etc.

- Create matrices based on queries asked, queries replied, queries rejected, queries appealed.
- Analyse behaviour patterns of institutions in answering or rejecting queries and identify the most frequently-asked topics.

The GRM models items with polytomous response categories. The model takes as input a matrix with items (questions) on one dimension and person parameters on the other. Values are filled with the response categories for every person to each item. Modelling consists of finding the optimum values of the parameters of the model that best describes the data given. For our RTI data, we can create matrices with reply stats, rejection stats and for query asked. The matrices are filled with percentage of replied queries, rejected queries and queries asked respectively. To draw an analogy between the two, the persons in GRM data is represented by the institutions in our RTI data, the items are

represented by topics to which the queries belong. The response categories are represented by percentages (0-100). Since the entries in the matrix represent percentages and they are ordinal in nature, the use of GRM is appropriate. The utility of the GRM for using on our RTI data is because of the latent patterns that it helps to identify. With respect to our data, the ability of persons shall represent 'transparency' of the institutions (with respect to answering or rejecting RTI queries), and the item difficulty shall denote the implementation of the act across institutions for each topic of query. The parallelism of the latent patterns between the typical dataset of GRM (multiple-choice questions) and the RTI data (query-reply stats) is what makes this an interesting approach.

Modelling each topic-wise statistics will give us a more in-depth picture of the dynamics of the RTI query-reply process, and capture intrinsic details hidden under an envelope of the overall performance of public institution. It is often observed that certain sections of a public body are more efficient in its performance while some others are lethargic. With targeted analysis of RTI queries divided into topics, we aim to discover specific issues or excellence regarding the different divisions of the same institution.

5   Experiment and Results

5.1   Constructing the dataset

Our dataset consists of matrices constructed from the RTI database created by us. An RTI application can have multiple queries. A survey of the data collected has shown that queries can be more or less classified into some fixed number of categories or topics, each independent of each other. A few examples of such categories are Administration, Library, Exams, Courses, Results, Academics, Admissions, Research, and Tenders etc. Each category has its own individual characteristic with respect to reply and rejection statistics. In order to dissect the RTI properties and understanding the hidden traits, analysing category-wise and institution-wise trends shall equip us with more information regarding the implementation of the RTI act. We have created matrices with topics of queries in one dimension and various institutions on the other. The matrices are created such that persons (in GRM) are represented by institutions, items are represented by topics and response categories are represented by percentages.

The RTI data collection is still going on, and only a fraction of the data is present with us. Additional tasks like translating various local languages into English, digitizing the data that was received by us in the form of photocopies etc. are yet to be undertaken. For the experiment, we have constructed a synthetic matrix of reply statistics that resembles our RTI dataset (the few RTI applications that we have collected). We have created matrices with topics of queries in one dimension (items) and various institutions on the other (person parameters). Matrix consists of ten institutions and five topics. Institute 1 to institute 6 are assumed to be central educational institutes and institute 7 to institute 10 are state institutes. Institutions are arranged in rows, and columns represent the five query topics. The matrix is filled with the percentages of the

queries replied by each institution for each of the topic. The matrix with initial values is shown in Table 1.

Table 1. Synthetic Data containing response percentages of ten institutes and five items

Inst.No.	Finance	Academic	Employment	Alumni	Medical
1	75	10	26	28	45
2	35	49	70	15	11
3	62	89	6	38	50
4	48	78	52	95	71
5	51	64	53	30	74
6	84	70	94	69	97
7	52	49	47	45	55
8	24	29	27	22	34
9	2	32	28	8	49
10	30	57	65	7	86

5.2   Transforming the dataset

Table 1 gives us the raw values of our RTI dataset. In order to fit this data into GRM, the matrix needs to be modified. We have divided the percentages into five buckets as shown in Table 2. The buckets are created so that each response category (each bucket) has a minimum amount of institutes' responses. This is done to reduce sparsity of data by clubbing together a percentage range into a single group. The response categories follow the Likert scale with 1 being the lowest and 5 representing the highest rating. This is done because GRM expects data in the form of ordinal response options. Here the five buckets represent five response options and each institution responds to one of those options corresponding to the respective reply percentages. Substituting the percentages with the values of Table 2 results in the matrix shown in Table 3.

Table 2. Percentage range of categories

Category	1	2	3	4	5
% Range	0-20	21-40	41-60	61-80	81-100

5.3   Results

In order to run GRM, we have chosen the open source platform 'R'. It has a few packages for IRT and we have used the 'ltm' package. The parameters obtained by running GRM to our synthetic data are shown in Table 4.

Table 3. Matrix created after substituting the percentages by assigned values

Inst.No.	Finance	Academic	Employment	Alumni	Medical
1	4	1	2	2	3
2	2	3	4	1	1
3	4	5	1	2	3
4	3	4	3	5	4
5	3	4	3	2	4
6	5	4	5	4	5
7	3	3	3	3	3
8	2	2	2	2	2
9	1	2	2	1	3
10	2	3	4	1	5

Table 4. Item Parameters after running the Graded Response Model on our data

Items	β_{i1}	β_{i2}	β_{i3}	β_{i4}	α_i
Finance	-1.451	-0.218	0.602	1.325	4.004
Academic	-2.559	-1.191	0.410	2.449	1.106
Employment	-5.197	-1.059	1.830	4.950	0.446
Alumni	-0.693	0.785	1.261	1.849	1.935
Medical	-2.424	-1.556	0.555	1.646	1.047

For each and every item, a graph is drawn between ability (latent trait) and the probability of responding on a particular category. Such curves, called Item Response Category Characteristic Curves for each of the five items are shown in Figures 1, 2, 3, 4 and 5.

We have used Bayesian Estimate procedure for calculating the ability parameter (θ) for each and every institute. Theta (θ) gives us the transparency of an institution. Transparency for each institution is shown in Table 5.

Table 5. Transparency Parameters after running the Graded Response Model on our data

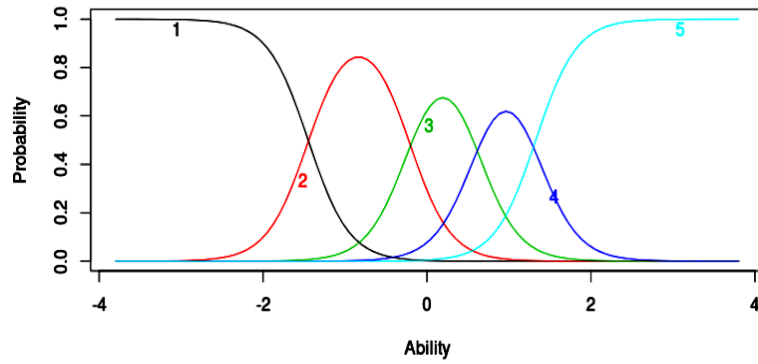| Inst. No.	Transparency (θ)
1	0.426
2	-0.906
3	0.690
4	0.589
5	0.276
6	1.623
7	0.260
8	-0.720
9	-1.590
10	-0.542

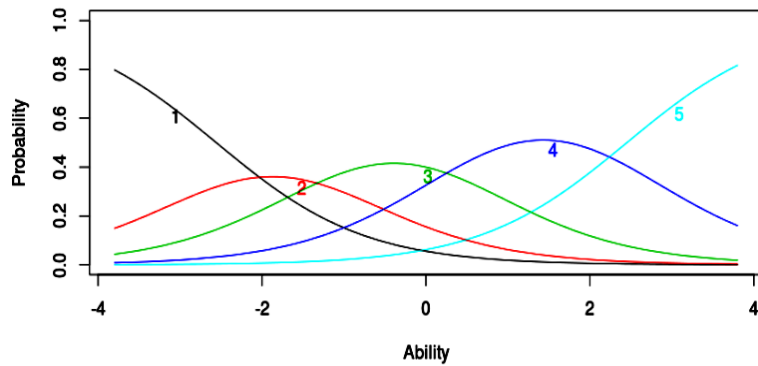**Fig. 1.** Item Response Category Characteristic Curve for item Finance

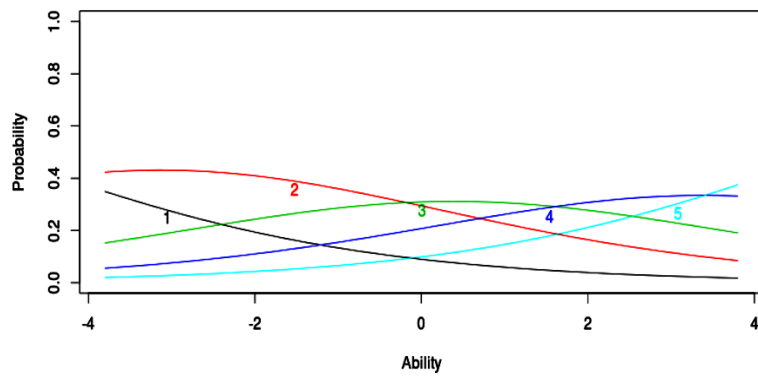**Fig. 2.** Item Response Category Characteristic Curve for item Academic

Fig. 3. Item Response Category Characteristic Curve for item Employment
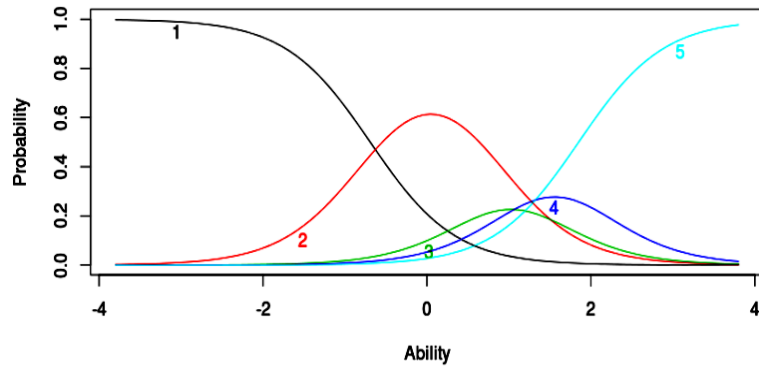
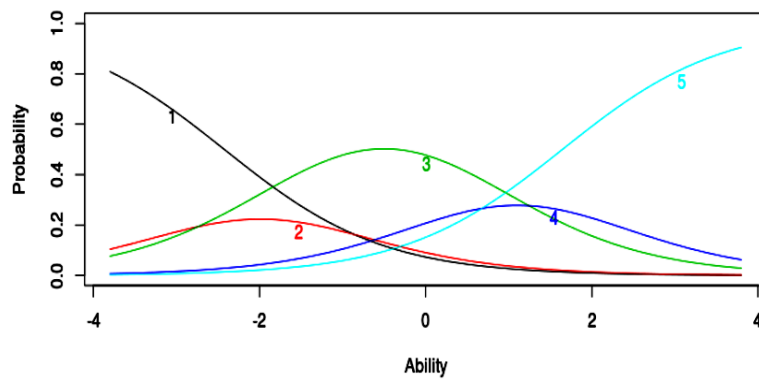**Fig. 4.** Item Response Category Characteristic Curve for item Alumni

Fig. 5. Item Response Category Characteristic Curve for item Medical

5.4 Discussion

The GRM has assigned an ability parameter to the ten institutions based on the reply stats. In the context of our dataset, the ability parameter represents the transparency of an institute. Higher the ability, more percentage of reply are given to RTI queries, hence more transparent is an institution. From Table 5 it is seen that Institute number 6 with the scores (5,4,5,4,5) has the highest ability (1.623), and Institute number 9 with the scores (1,2,2,1,3) has the lowest ability (-1.590). Arranging the institutions with respect to transparency value shows that all central institutions except institution 2 has high transparency compared to the state institutions.

Each β_{ij} is the θ-value of transition between adjacent response categories. It is the boundary at which the probability of the response falling in the previous response category (left side) becomes less than 50% and the probability of response falling on the subsequent response categories (categories on the right side) is greater than 50%. These threshold values are different for different items, indicating that each item is modelled differently and the response thresholds are not uniform across items but are dependent on the data distribution of each item.

Each and every item has a discrimination parameter. An item with high discrimination parameter can discriminate well between institutes with high ability and low ability. From our results, Finance has the highest discrimination parameter and Employment has the lowest discrimination parameter. For an item with low discrimination parameter, there is less distinction between the reply patterns of high ability and low ability. Hence, observing reply stats of 'employment' item is not enough to decide the transparency of an institute. This gives a sort of quality assessment for each item with respect to judging the RTI characteristics between institutions.

GRM also models probabilities of how each institution responds to different items, that is, query topics. It can be observed from the results that certain institutions (for example, institution 6 with θ=1.623) are very good in responding to the finance category questions (Figure 1), but not so well in responding to employment category questions (Figure 3). It means that a highly transparent institution which replies efficiently to the 'finance' related queries do not reply as efficiently to the 'employment' related queries. This reveals that there is inconsistency in RTI reply across departments of the same institution, and leads us to question as to why such inconsistencies are present.

6 Conclusion

In this paper we have modelled the RTI query-reply process via Item Response Theory (IRT). We have created a synthetic dataset that resembles our collected RTI data in its characteristics, and tried to model it in terms of inputs to an IRT model. We have selected GRM as the preferred model, and successfully run it with promising results. The novelty of our approach lies in two main points.

Firstly, such an analysis of RTI data has never been undertaken. We are collecting RTI data related to each individual, from each public educational institution and shall span multiple locations across India. Most RTI studies are limited to specific regions or specific issues in that their surveys are based to explore a fixed set of problems. Our present work of applying learning algorithms to uncover hidden traits in the RTI query-reply process is the first of its kind. Moreover, the application of GRM has been limited to the examination domain. This work is a successful attempt to extend its application scope. Secondly, the implications from the outcomes of this experiment are enormous. With this attempt, we have assigned a transparency value to the institutions with respect to the reply patterns of each and every institution. Our experiment with the synthetic data reveals that the central institutions are more transparent in replying to citizen's queries than the state institutions. A closer look into Tables 4 and 5 can help us extract further information. For example, certain institutions (for example, institution 6) are very good in responding to the 'finance' category questions (Figure 1), but not so well in responding to 'employment' category questions (Figure 3). This reveals that there is inconsistency in RTI replies across departments of the same institution, and leads us to question as to why such inconsistencies are present. This is an indication of same laws being implemented in different ways for different institutions as well as different departments within the same institution. A solution for this may be to bring some changes to the ordinances of the institution. Hence, this work of analysing RTI queries and reply statistics shall also give us strong basis for proposing amendments to the law of an institution. Once the data collection part is over, we shall be able to apply this model to our actual RTI dataset, and the conclusions from the results shall give us a clear picture of the laws and policies that govern our public institutions.

References

1. The Constitution of India, http://lawmin.nic.in/coi/coiason29july08.pdf
2. What is the Procedure of Amendment of the Constitution of India?, http://www.preservearticles.com/201012251615/procedure-of-amendment-of-the-constitution-of-india.html
3. http://ccis.nic.in/WriteReadData/CircularPortal/D2/D02rti/10_9_2008-IR26042011.pdf
4. The Right to Information (Amendment) Bill, 2013, http://www.prsindia.org/uploads/media/RTI%20%28A%29/RTI%20%28A%29%20Bill,%202013.pdf
5. Gerrish, S., Blei, D. M.: How they vote: Issue-adjusted models of legislative behavior. Advances in Neural Information Processing Systems, 2753–2761, 2012
6. Poole, K. T.,Rosenthal, H.: A spatial model for legislative roll call analysis. American Journal of Political Science, 357–384, 1985
7. Lucchese, C., Orlando, S., Perego, R., Silvestri, F., Tolomei, G.: Discovering tasks from search engine query logs. ACM Transactions on Information Systems (TOIS), vol. 31, no. 3, 2013
8. Beitzel, S.M.: On understanding and classifying web queries. Citeseer, 2006

9. Salton, G., Wong, A., Yang, C.: A vector space model for automatic indexing. Communications of the ACM, vol. 18, no. 11, 613–620, 1975

10. Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., Harshman, R.: Indexing by latent semantic analysis. Journal of the American society for information science, vol. 41, no. 6, 1990

11. Moschitti, A., Quarteroni, S., Basili, R., Manandhar, S.: Exploiting syntactic and shallow semantic kernels for question answer classification. Annual meeting-association for computational linguistics, vol. 45, no. 1, 2007

12. McKeown, K. R.: Paraphrasing using given and new information in a question-answer system. Proceedings of the 17th annual meeting on Association for Computational Linguistics, 67–72, 1979

13. Hovy, E., Hermjakob, U., Ravichandran, D.: A question/answer typology with surface text patterns. Proceedings of the second international conference on Human Language Technology Research, 247–251, 2002

14. Hovy, E., Hermjakob, U., Ravichandran, D.: Equating tests under the graded response model. Applied Psychological Measurement, vol. 16, no. 1, 87–96, 1992

15. Preston, K. S. J., Parral, S. N., Gottfried, A. W., Oliver, P. H., Gottfried, A. E., Ibrahim, S. M., Delany, D.: Applying the Nominal Response Model Within a Longitudinal Framework to Construct the Positive Family Relationships Scale. Educational and Psychological Measurement, 2015

16. Samejima, F.: Estimation of latent ability using a response pattern of graded scores. Psychometrika monograph supplement, 1969