# Identifying Annotator Bias: A new IRT-based method for bias identification

**Jacopo Amidei** and **Paul Piwek** and **Alistair Willis**
School of Computing and Communications
The Open University
Milton Keynes, UK

## Abstract

A basic step in any annotation effort is the measurement of the Inter Annotator Agreement (IAA). An important factor that can affect the IAA is the presence of annotator bias. In this paper we introduce a new interpretation and application of the Item Response Theory (IRT) to detect annotators' bias. Our interpretation of IRT offers an original bias identification method that can be used to compare annotators' bias and characterise annotation disagreement. Our method can be used to spot outlier annotators, improve annotation guidelines and provide a better picture of the annotation reliability. Additionally, because scales for IAA interpretation are not generally agreed upon, our bias identification method is valuable as a complement to the IAA value which can help with understanding the annotation disagreement.

## 1 Introduction

In computational linguistics (CL), since Carletta's influential paper (Carletta, 1996), the standard approach to check the reliability in annotation efforts is the Inter Annotator Agreement (IAA) – that is the extent to which different annotators independently make the same annotation decisions. For the purpose of calculating the IAA, Carletta proposed the use of a family of statistics related to Cohen's $\kappa$ coefficient of agreement (Cohen, 1960), which she collectively refers to as the *Kappa statistic*.[1]

An important factor that can affect the IAA is the presence of annotator bias – that is differences between annotator preferences for subjective reasons. Recent work – for example Sampson and Babarczy (2008), Lommel et al. (2014), Joshi et al. (2016) and Amidei et al. (2018) – show that annotators diverge in language annotation tasks due to a range of ineliminable factors such as background knowledge, preconceptions about language and general educational level. Although clear annotation schemes with effective guidelines and annotator training aim to reduce annotation bias, some individual differences persist. A method that allows for a better identification and understanding of individual bias could be valuable in annotation efforts. For example, it could be used to:

- display differences in annotators' behaviour. Annotations that show a markedly different pattern from the other annotations can either be removed or further analysed. This can help to spot outlying annotators and help with improving annotation guidelines and reduce annotation disagreement.[2]

- provide a better picture of the annotation reliability. Indeed, once identified, the annotator bias could be used to explain and understand annotation disagreement and accordingly the IAA values. For instance, as we will see in Section 2.3, such a method could be used to show in which respect an annotator shows more strict annotation behaviour than another annotator.



[1] Some examples of the Kappa statistic are Cohen's $\kappa$ (Cohen, 1960), Fleiss' $\kappa$ (Fleiss, 1971) and Krippendorff's $\alpha$ (Krippendorff, 1980).

[2] It is important to note that for annotation tasks which involve a high level of subjectivity, there is a boundary to the reduction of annotation disagreement. For example, Amidei et al. (2018) present evidence for such boundaries in the area of Natural Language Generation evaluation.



*Proceedings of the 28th International Conference on Computational Linguistics*, pages 4787–4797
Barcelona, Spain (Online), December 8-13, 2020

In some human annotation tasks, where a high level of subjectivity is involved — for example annotation that concerns the quality of a generated sentence (on a range of dimensions including syntax, semantic and pragmatics) — what determines the human annotators' decisions cannot be measured directly. It is not a physical dimension, as for example weight or distance. Nevertheless, we can study the annotation behaviour, as directly observable, in order to have a better understanding of the unobservable decision process behind the human annotation. A first step in this direction is to analyse the frequency of the categories used by the annotators. The frequency gives us a first approximation of the annotator's behaviour. Nevertheless, the frequency of the categories used is a raw image. Ideally, we would like to be able to compare the annotators' annotations based on their perception of the phenomena being annotated. For this reason, in this paper we introduce a way to analyse the unobservable decision process behind the human annotation.[3]

Our method is based on a new interpretation and use of the Item Response Theory (IRT) (Gulliksen, 1950). IRT is a psychometric theory used for analysing and designing tools — for example, surveys and tests — for measuring abilities or attitudes. We will use two examples to explain the novelty of our interpretation: mathematical ability (a traditional use of IRT), and natural language generation (our novel use).

IRT has traditionally been used, for example, to determine the validity of a test, say a test of mathematical ability. Such a test is administered to a number of students. On each test item, each student will achieve a certain score. Ideally, students with a strong mathematical ability should receive high scores on the test, and students with weak mathematical ability should receive low scores. There should also be a range in between these extremes. Based on data of a large group of students completing such a test, IRT can extract from the data a model that gives us, for a hypothetical student with a specific level of mathematical ability, the probability of specific test scores. Ideally, a high level of mathematical ability should be associated with a high probability for a high test score. Importantly, mathematical ability is not observed directly. It is only available via the student's performance on the test. And this coupling of the trait and the test performance is not going to be perfect (some good students may on occasion not do well, hence the use of a probability function). In IRT, mathematical ability is modelled as a latent (not directly observable) trait of individuals. It is essential to note that a latent trait assumes a context, e.g. that of a conventional mathematical education. It exists against a background of assumption about culture, regional variance, personal preferences, schooling/training, etc.

In our IRT-based method to detect annotator bias, the annotation exercise involves annotating a corpus data (for example a set of sentences of images) with linguistic or other information. For the sake of simplicity, let's suppose that the annotation is an evaluation of NLG systems. In this case the annotation exercise involves a set of sentences. Annotators are asked to judge each of these sentences according to one or more criteria. Let's assume the criterion in question is the fluency of the sentence. *Our proposal is to treat fluency as a latent trait of sentences: it cannot be directly observed or measured, but we can get hold of it via the judgements of the annotators*. In an annotation, the annotators will not always agree: for instance, some may be more severe in their judgements than others. We can now apply IRT for individual annotators. This time, IRT can extract from an individual's annotation the probability of a category's answer given a hypothetical sentence of a certain level of fluency to that annotator. For each of the fluency scores which the annotator can choose from when annotating, IRT gives us the probability that that annotator will choose that particular of fluency score. This allows us to see how the annotator's behaviour (the score they assign) relates to the latent trait (the level of fluency of a hypothetical sentence). Also in this case the latent trait assumes a context. It is a standard of fluency which is implicit in the judgements of a (more or less homogeneous) language community. The language community does distinguish between more and less fluent sentences. However, on specific occasions, individual members (or subgroups) will display smaller or larger differences in judgement. IRT allows

[3]We note that a decision process is affected by several factors. It is not feasible (and unrealistic) to attempt to control all the factors behind a decision. Nevertheless, we can perform a post-decision analysis to arrive at a better understanding of the decision itself. The decision (in our case the annotation), because observable, represents the observation unit. The method we propose can analyse the annotation to identify annotator bias, but it does not explain the factors behind those biases. Nevertheless, the identification of annotator bias allows us to better understand the disagreement between the annotators.

| | Ambiguity | | | Variety | | | Fluency | | | | Relevance | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 1 | 2 | 3 | 1 | 2 | 3 | 1 | 2 | 3 | 4 | 1 | 2 | 3 | 4 |
| **J1** | 0.68 | 0.19 | 0.11 | 0.20 | 0.04 | 0.74 | 0.43 | 0.26 | 0.22 | 0.07 | 0.91 | 0.01 | 0.07 | 0 |
| **J3** | 0.55 | 0.32 | 0.11 | 0.23 | 0.04 | 0.71 | 0.40 | 0.29 | 0.14 | 0.14 | 0.83 | 0.05 | 0.07 | 0.02 |

Table 1: Frequency of the scores used by Judge 1 (J1) and Judge 3 (J3).

us to quantify and identify these differences.

We present our proposal with an example taken from the case of intrinsic human evaluation of NLG systems (Gatt and Krahmer, 2018). For this reason, in what follows we use the term *judge* instead of the term *annotator*. We choose such terminology to emphasize the evaluation aim.

## 2 Analyzing raters bias: From frequency to Item Response Theory

By way of example, in this paper, we use the QG-STEC evaluation dataset (Rus et al., 2012)[4]. The QG-STEC evaluation dataset is composed of questions generated from four systems that participated in the QG-STEC Task B, that is, the task of generating a question from an input sentence. Each question is evaluated based on five criteria: *"Question Type"* (on a scale from 1 to 2), *"Ambiguity"* and *"Variety"* (on a scale from 1 to 3), *"Relevance"* and *"Syntactic Correctness and Fluency"* (on a scale from 1 to 4).[5] Six judges took part in the evaluation. They judged batches of sentences independently. Each batch was evaluated by two judges. We decided to analyse the data collected from Judge 1 and Judge 3.[6] This pair was randomly selected from all pairs of judges.[7] We limit our investigation to the criteria ambiguity and variety (on a scale from 1 to 3) and syntactic correctness and fluency (for the sake of simplicity, in what follows, we refer to this criterion as fluency) and relevance (on a scale from 1 to 4).

### 2.1 Frequency

We are looking for a way to characterise the annotation behaviour of different judges. This can help us to understand better the source of disagreement in an annotation. A way to do this is by analysing the frequency of the categories used by the judges. Indeed, the frequency gives us a first approximation of the judges' annotation behaviour. The comparison between different judges' decisions can show sources of difference that explain the disagreement in the annotation. Table 1 shows the frequency of the scores used by Judge 1 and Judge 3.[8] It shows that Judge 3 tends to give slightly higher scores than Judge 1 for the ambiguity, fluency and relevance criteria. Indeed, Judge 3 tends to be more cautious in giving the extreme score 1 than Judge 1, preferring the more neutral score 2. This trend is inverted for the extreme score 4. In this case Judge 1 tends to be more cautious and prefers lower scores than score 4. For example, we note that Judge 1 does not use the score 4 in the relevance criterion. In the case of variety, Judge 1 tends to score higher than Judge 3, preferring high scores whereas Judge 3 prefers low scores.

The frequency gives us a first approximation of the judges' annotation behaviour. Nevertheless, it is a raw image. Ideally, we would like to be able to compare the judges based on their perception of the quality of a sentence which is behind their annotation decision. That is, how much a judge feels that a

[4]The QG-STEC evaluation dataset is available at: `http://computing.open.ac.uk/coda/resources/qg_form.html`.

[5]The evaluation guidelines are available at: `http://computing.open.ac.uk/coda/resources/qg_form.html`.

[6]Although our example uses only two judges, the method we propose can be applied to any number of judges. The only requirement is that the judges annotate the same set of sentences.

[7]Although randomly chosen, the data annotated by Judges 1 and 3 represent an excellent dataset to introduce our proposal. Indeed, the criteria annotated by Judges 1 and 3 reach a different level of agreement. This allows us to show how our proposal deals with different levels of agreement between the judges.

[8]The small differences are due to the fact that the frequency analysis considers how many times a category was chosen by a judge. It does not consider when the categories were used. That is, the information provided in Table 1 does not consider how the judges' annotations differ in the cases of disagreement.

sentence satisfies a criterion at hand. We propose the use of IRT for this aim. IRT provides a probabilistic analysis which allows inferences to be drawn about the judge's annotation behaviour.

### 2.2 Item Response Theory

IRT is used to measure various types of latent trait which are investigated by the use of item tests. For example it can be abilities, such as mathematical ability, or it can be a behavioral attitude, such as tendency to make particular purchases. The main aim of the theory is to evaluate and adjust test items and score examinees based on their latent traits such as abilities or attitudes.

In this section we are going to use IRT to describe judges' annotation behaviour in order to better understand the disagreement that arises from their annotation. We will see that IRT, as well as the frequencies, can be used to study the judges' preference of the scores. Furthermore, in contrast to the frequency analysis, the IRT analysis uses a probabilistic model that allows us to have an insight where these differences take place in relation to the latent trait.

To do so, we are going to give a new interpretation of the IRT. For this reason, it is important to explain the traditional use of IRT first. We will then present the ways in which our interpretation deviates from the traditional use.

#### 2.2.1 The traditional use of IRT

Traditionally a test (or a survey) is designed in such a way that there are few items and many respondents. IRT is based on the assumption that each respondent answers each item in line with their level of the latent trait. In IRT it is assumed that the latent trait can be measured on a scale having a midpoint of zero and it can take any real number from $-\infty$ to $+\infty$. For practical considerations usually the range is limited to the interval [-4, 4]. It is important to keep in mind that this is just for simplicity and other intervals can be used.

Once a scale of measurement is given, it is possible to define a probability function of the possible answer as a function of the latent trait. Standard IRT uses a logistic model for this purpose. Logistic models have an S-shaped probability function such as the one depicted in Figure 1.

From a mathematical point of view the model, and more specifically Rasch's logistic model (Rasch, 1960), can be expressed by the following equation:

$$P(x_{im} = 1|z_m) = \frac{e^{(z_m - \beta_i)}}{1 + e^{(z_m - \beta_i)}}$$

where 1 is a label that represents the correct response, $x_{im}$ represents the response of the $m$th respondent for the $i$th item and $z_m$ represents the latent trait of the $m$th respondent. Finally, $\beta_i$ is a parameter that takes into account the difficulty of the $i$th item. In other words, $P(x_{im} = 1|z_m)$ is the probability that respondent $m$ correctly answers the item $i$ given his/her latent trait $z_m$. Intuitively, the higher the latent trait $z_m$ the higher the probability of correctly answering the $i$th item.

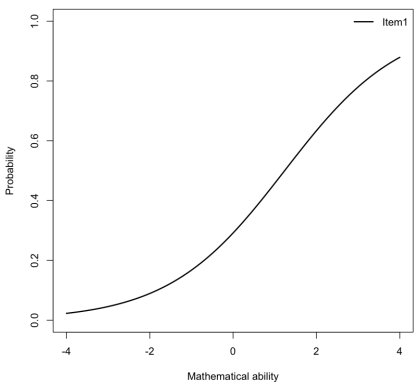



Figure 1: IRCCC example for a mathematics exam.

Each IRT model allows an Item Response Category Characteristic Curve (IRCCC) to be defined for each item (Figure 1 depicts an example of an IRCCC for Rasch's model). The IRCCC assesses the relationship between the latent trait and the scores (in the case of ability, for example mathematical ability) or each chosen item category (in the case of behaviour attitude, such as in a Likert scale survey). It shows the likelihood of a respondent receiving a score (in the case of ability) or selecting a certain category (in the case of behaviour attitude) at various levels of the latent trait. Concretely, the IRCCC gives a graphic representation of the probability function determined by an IRT model.

Let us give an example that involves a mathematics exam and the Rasch's IRT model. In this case there are few items, let's say 4 mathematical questions and many students, let's say 35 students. All the students have to answer the 4 items. We assume that each student has mathematical ability which contributes to the answers that they give. Based on the students' answers, both the items' difficulty and the students' mathematical ability can be measured. Once the test is done, all the answers are marked as correct or incorrect. Based on such binary test scoring, Rasch's model can be used to define a probability function that describes the expected score the respondents should receive based on their mathematical ability. As said before, the probability function can be graphically represented by the IRCCC.

Figure 1 depicts the IRCCC of the first item (item 1). Accordingly, Figure 1 shows the probability that respondents will answer item 1 correctly, based on their latent trait, which represents their mathematical ability. Suppose for example that the mathematical ability of a respondent $E_0$ is 0. In this case Figure 1 suggests that $E_0$ has a low probability to answer correctly. To check this we imagine a straight line from point 0 of the latent trait to identify at which point it intersects with the line depicts in the graph. Figure 1 shows that this happens for a probability value which is slightly higher than 0.2. In other word, the probability of $E_0$ to answer correctly is slightly higher than 0.2. Now, suppose that we have a respondent $E_4$ whose mathematical ability is 4. In this case the IRCCC suggest that $E_4$ has a high probability of answering item 1 correctly. Indeed, following the same procedure as we did for $E_0$, we can see that the probability of $E_4$ to answer item 1 correctly is slightly higher than 0.8.

In the mathematical test example, we were working at a binary or dichotomous setting. We assumed that the answers to the items could be either correct or incorrect. However, there are also IRT models of non-dichotomous ordinal categories tests. One of these models is the Graded Response Model (GRM) (Samejima, 1969). GRM is a variant of IRT developed to analyse tests that use polytomous categories, that is tests that use more than two categories. This is particularly appropriate for the analysis of rating and Likert scales. GRM can be applied to gather information about how change in the latent trait affect observed questionnaires' items response.

#### 2.2.2 IRT for judges' bias detection

The traditional application of IRT focuses on modelling the relationship between the observable respondents' answers and their unobservable latent trait, for example mathematical ability or behavioural attitude. In the case of a mathematical test, the respondents' performance on the test items should vary according to their mathematical ability: ideally, higher mathematical ability results in higher item test scores. The IRCCC is used to visualise the relationship between item test scores and the latent trait for a specific test item. It can tell us whether the test item indeed has the property that higher mathematical ability (of the person taking the test item) results in higher scores.

In order to apply IRT for identifying bias among annotators, we introduce the following twist in the application of IRT: rather than model a latent trait of the respondents, the latent trait in question now is conceived of as a property of linguistic items, that is sentences. One such property is sentence fluency: one sentence can be more fluent than another, but this is not a property that can be measured directly. Rather, it is a latent trait of sentences that we can uncover only via the judgements of language users. The IRT helps with modelling how different judges respond to different levels of a latent trait, such as fluency. Individual biases, which may be due to a wide variety of factors (e.g. regional variance, personal preferences, schooling/training) that skew application of the shared norms or standards (that the language community as a whole has adopted) can thus be laid bare. The IRCCC is now used to visualise for individual judges the relationship between fluency scores and the latent trait (fluency) of sentences. Judges can be compared with each other by putting their IRCCCs next to each other.

### 2.3 The use of GRM to analyse judges' bias: An example from NLG evaluation

In this section, we present our IRT-based method for the ambiguity and variety criteria used in Section 2.1.[9] We use GRM to analyse the judges' decision bias. For each criterion studied, we present two analyses.

[9]The analysis for the fluency and relevance criteria can be found via the link `https://bit.ly/3lT2m2f`.

The first analysis is based on the IRCCC graphs. The IRCCC graphs provide an informal analysis of the judges decision behavior through the evaluation. They have the advantage of being easy to interpret and they allow for a quick insight about the judges' disagreement. Nevertheless, in order to have a more refined analysis we will introduce the concept of an extremity parameter.

The extremity parameters show the latent trait score at which judges have a $50\%$ chance of selecting certain categories and $50\%$ chance of selecting the remaining categories. For this reason they can be used to suggest the spectrum of the latent trait where we can find the main source of disagreement.

**The extremity parameters: an explanatory example**

Suppose that two Judges (let's say $J1$ and $J2$) are performing an evaluation about the fluency of a set of sentences, and assume that the evaluation is based on the four scores: 1, 2, 3, and 4. Supposing we are interested in the extremity parameter that suggests the $50\%$ chance of selecting scores 1 and 2 (for sake of simplicity let us denote this as $Ex2$). $Ex2$ expresses the latent trait level at which $J1$ and $J2$ have a probability of $0.5$ of selecting either score 1 or score 2 and a probability of $0.5$ of selecting either score 3 or score 4. Let us suppose the $Ex2$ level for $J1$ is $0.1$, whereas the $Ex2$ level for $J2$ is $0.6$. Under our interpretation of IRT, this means that, given a sentence of latent trait (fluency in our example) of $0.1$ the probability that $J1$ selects categories 1 or 2 is $0.5$. Likewise, the probability that $J1$ select categories 3 or 4 is $0.5$. On the other hand, $J2$ has a probability of 0.5 of selecting categories 1 or 2 for sentences of latent trait of $0.6$. Likewise, for sentences of latent trait of $0.6$ the probability that $J3$ select categories 3 or 4 is $0.5$.

The interval $[0.1, 0.6]$ defines a probabilistic analysis of the disagreement between $J1$ and $J2$. In such an interval, the IRT analysis suggests the disagreement mainly comes about because $J2$ tends to give lower scores than $J1$. Indeed, approaching the latent trait score $0.6$, $J1$ has a higher probability of selecting scores 3 and 4 than scores 1 and 2, whereas $J2$ has a slightly higher probability of selecting scores 1 and 2 than scores 3 and 4.

**GRM-based method for our analysis**

As we have seen in the traditional use of IRT, the test is designed in such a way that there are few items and many respondents. In this case, for each item the IRCCC and the extremity parameters are defined based on the respondents' answers to that item. In the case of intrinsic human evaluation of NLG systems we have few judges and several items. Usually each item is a different instance of few criteria that are analysed. For example, given the criterion fluency, there are several items making up with the same question about fluency but with different generated sentences to be evaluated.

In the present paper we used the GRM model in the following way. Given a judge $J$ and a criterion $C$, we collect all the items annotated by $J$ aimed at evaluating $C$. Let us call $I^n_C$ the $n$th item that aims to evaluate $C$. In this case, given a judge $J$, for each criterion $C$ the IRCCC and the extremity parameters are defined based on $J$'s answers to $I^n_C$ for $n = 1 \dots m$, where $m$ is the number of sentences to be evaluated. In what follows, given a criterion $C$ we collect all the items $I^n_C$ under the name of $C$.

Our GRM-based method provide a probabilistic analysis of the judge annotation behavior for each criterion. Because our method analyse one judge at the time, it allows us to compare different judges' annotation behavior. Such comparison allows to determine the judge bias and to better understand the disagreement between judges.

**Interpreting the criteria**

GRM considers the scores in increasing order. Conversely, in the QG-STEC evaluation dataset the scores are considered in decreasing order. For this reason, in what follows, we have to pay attention to the interpretation of the IRCCC. More specifically, the relevance, the fluency and the variety criteria consider 1 as the best score — best from a criterion quality point of view. For this reason we have to interpret the positive latent trait as the sentence irrelevance, non fluency and non variety. The situation for the ambiguity criterion is different. It is stated in term of unambiguously (that is, lack of ambiguity). In this case we have to interpret the positive latent trait for the ambiguity criterion as the sentence ambiguity.

**Software used for the analysis**

In what follows, we carry out our IRT analysis with the statistical software *R*. We used the library *ltm* (Rizopoulos, 2018). This library was developed to provide researchers with a flexible framework to perform IRT analyses. More specifically, we used the function *grm()* with the parameter *constrained* set to *TRUE*.

Regarding the use of the coefficient of agreement, our aim is to validate the final human annotation. Following Artstein and Poesio (2008) we use a coefficient of agreement based on the interpretation of chance agreement as "Individual coder distributions". To this end we use the Conger's $\kappa$ (Conger, 1980) with ordinal weight, as defined by Gwet (2014). We used the weighed Conger's $\kappa$ because it is a generalization of Cohen's $\kappa$. To measure Conger's $\kappa$ we used the library *irrCAC* provided by the *R* software.[10] More specifically, we used the function *conger.kappa.raw* with the variable weights set to "ordinal".

### 2.3.1 The ambiguity criterion

The ambiguity criterion reaches a Conger's $\kappa$ of 0.21, indicating a low level of agreement. Such a value suggests that the judges rarely have the same score decision.

**The IRCCC graphs analysis:** The scores frequency analyses in Section 2.1 suggest that the disagreement emerges mainly due to the fact that Judge 1 prefers score 1 whereas Judge 3 prefers score 2. We can use GRM to investigate such a divergence in more depth.

The IRCCC for the ambiguity criterion shows that Judge 1 (see Figure 2(a)) was less cautious in choosing score 1 than Judge 3 (see Figure 2(b)). Indeed, we can see that:

- For positive levels of the latent trait (approximately the interval [0, 2]) the peak of the curve of score 2 (red line) is higher for Judge 3 than for Judge 1.
- At the same time, we can see how in the case of Judge 1 the curve of score 2 (red line) intersects that of score 1 (black line) for higher latent trait levels than is the case for Judge 3.

These facts together suggest the following. On one hand, for the latent trait levels that is approximately in the interval [0, 2], Judge 3 has a higher probability of selecting score 2 than Judge 1. On the other hand, for the latent trait levels that is approximately in the interval [0, 1], Judge 1 has a higher probability of selecting score 1 than Judge 3. From the IRCCC graphs, we can conclude that the main divergence between Judge 1 and Judge 3 takes place approximately in the interval [0, 2] of the latent trait. We can now use the extremity parameter to refine such analysis.

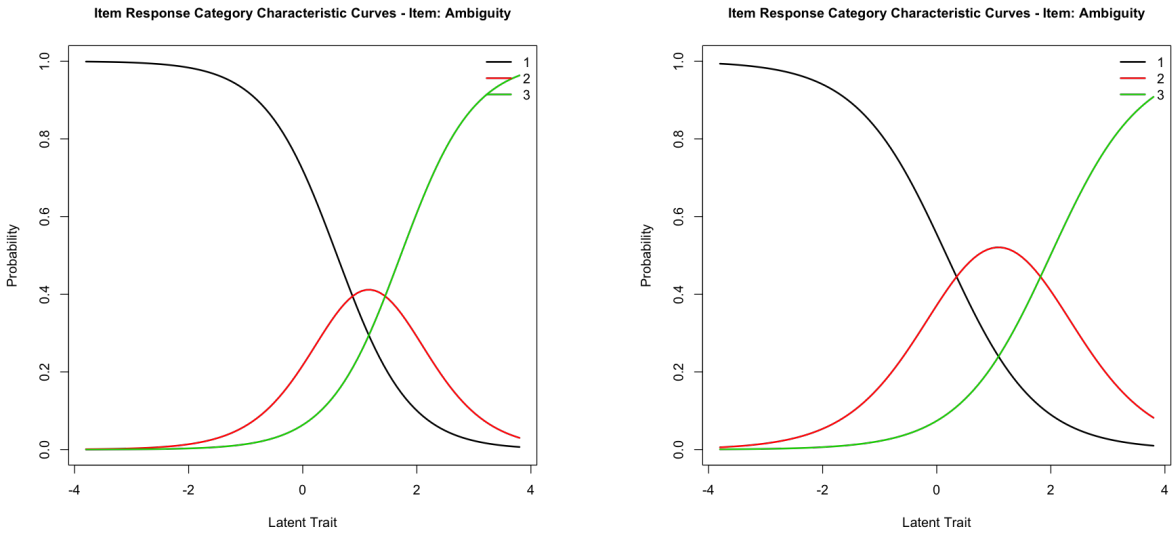
**The extremity parameter analysis:** The extremity parameter for the ambiguity criterion suggests that:

- Judge 1 has a $50\%$ chance of selecting the score 1 with a latent trait level of 0.601 and a $50\%$ chance of selecting the scores 2 or 1 with a latent trait level of 1.717.
- On the other hand, Judge 3 has a $50\%$ chance of selecting the score 1 with a latent trait level of 0.173 and a $50\%$ chance of selecting the scores 2 or 1 with a latent trait level of 1.993.

From these levels, it follows that:

- For the latent trait levels between the interval [0.173, 0.601], Judge 1 has a higher probability of selecting score 1 than Judge 3. Whereas for the same interval, Judge 3 has a higher probability of selecting scores 2 and 3 than score 1.
- At the same time we can see that between the latent trait of [1.717, 1.993] Judge 1 has a higher probability of selecting score 3 than score 1 and 2 whereas Judge 3 has a higher probability of selecting score 1 and 2 than score 3 (this is due to the fact that for 3 sentences Judge 3 gives score 2 whereas Judge 1 selects score 3).

[10] All the details can be found via the link `https://rdrr.io/cran/irrCAC/`.




(a) (b)

Figure 2: IRCCC for the ambiguity criterion for Judge 1 (a) and Judge 3 (b). The numbers 1, 2 and 3 represent the scores the judges can choose from. The graphs show as the main source of disagreement can be found in the latent trait interval [0, 2] mainly relative to score 2 (red line).

The IRT analysis suggests that the main source of disagreement can be found in the latent trait intervals [0.173, 0.601] (let's denote it as $I^1_{am}$) and [1.717, 1.993] (let's denote it as $I^2_{am}$).

Based on the evaluation guideline provided for the QG-STEC task B for the ambiguity criterion, and under the assumption that a question can be more ambiguous than another one, we can draw the following conclusions.

The questions that fall in $I^1_{am}$ can be interpreted as questions that, if stated out of the blue, are slightly ambiguous in that they are missing some information. A couple of examples from the dataset are:

> "How many gunners who died should there be a fitting public memorial to?" or "Where did the trust employ over 7,000 staff and manage another six sites?".

In $I^1_{am}$, we can expect Judge 3 to present a more strict annotation behaviour than Judge 1. The questions that fall in $I^2_{am}$ can be interpreted as questions that, if stated out of the blue, have high probability of being perceived as ambiguous. A couple of example from the dataset are:

> "What is the axiom in Euclidian Geometry?" or "Why tend accidents to be relatively minor ?".

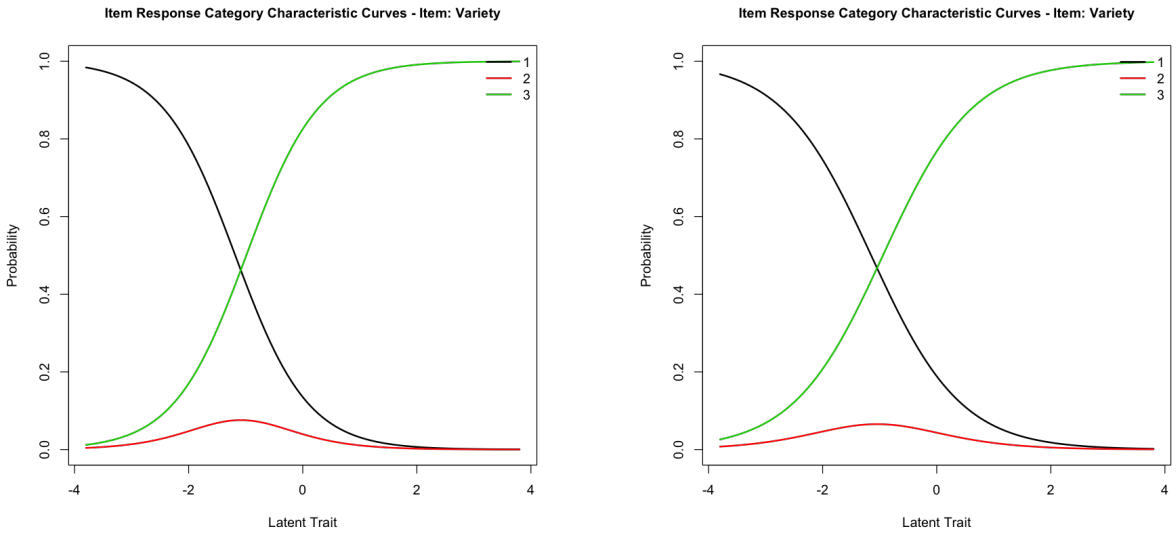
In $I^2_{am}$, we can expect Judge 1 to present a more strict annotation behaviour than Judge 3.

### 2.3.2 The variety criterion

The variety criterion aims to measure, given two questions, the extent of their difference. This measures the ability of a system to generate a variety of different questions given the same input. The variety criterion reaches a high Conger's $\kappa$ agreement coefficient of 0.93. This value shows a tendency of Judges 1 and 3 to reach the same score decision.

**The IRCCC graphs analysis:** The scores frequency analysis in Section 2.1 shows that there is little difference between the judges' annotation. A deeper analysis shows that the disagreement emerges in three pairs of questions where Judge 1 chose the scores 2, 3, 3 whereas Judge 3 chose the scores 1, 1, 2. This fact is captured by the IRCCC depicted in Figure 3. More precisely, Figure 3 shows that:

- The curves in both the figures are highly similar.




(a) (b)

Figure 3: IRCCC for the variety criterion for Judge 1 (a) and Judge 3 (b). The numbers 1, 2 and 3 represent the scores the judges can choose from. The graphs are very similar which shows a low level of disagreement between the judges.

- The main difference is about the score 1 (black line). We notice that the line moves slightly towards the positive trait (that is, it meets the green line slightly forward the positive trait) for Judge 3, who indeed chose such scores more than Judge 1.

We note also that the low peak for the score 2 (red line) shows that both the judges tend to avoid the score 2 in favour of the extreme scores 1 and 3. Indeed, the frequency for the variety criterion depicted in Table 1 shows that both the judges choose the score 2 $4.5\%$ of the time.

**The extremity parameter analysis:** The extremity parameters for the variety criterion show that:

- Judge 1 has a $50\%$ chance of selecting the score 1 with a latent trait level of -1.183 whereas Judge 3 with a latent trait level of -1.150.
- Similarly, Judge 1 has a $50\%$ chance of selecting the scores 1 and 2 with a latent trait level of -0.990 whereas Judge 3 with a level of -0.944.

The small differences between the extremity parameters explain the small disagreement between Judge 1 and Judge 3. In this case, the GRM shows that Judge 3 was slightly more cautious (indeed this happens in just 3 cases) in giving high scores than Judge 1 for a latent trait level at approximately the levels -1 and 0.

## 3 Related work

In this paper, we suggest considering IRT as a method to visualize and understand annotators' subjective bias. To the best of our knowledge this is an original contribution. In Natural Language Processing (NLP), IRT has been used by Lalor et al. (2016). Lalor et al. place their work inside the traditional use of IRT, generating a gold-standard test-set for the Recognizing Textual Entailment task and providing more insight into models' performance. In contrast, in this paper, we present a new interpretation of IRT with the aim of improving and analysing annotation tasks. We propose to use it as a method for spotting outlier annotators, for improving annotation guidelines or for interpreting data reliability (for example, as a complement to the scale of interpretation for IAA values).

Regarding the use of IAA in corpus annotation tasks, and more specifically the task of linguistic annotation, we refer to Palmer and Xue (2010), Pustejovsky and Stubbs (2013) and Artstein and Poesio (2008). These provide extensive descriptions of how to perform an annotation task.

Regarding a general description and study of IRT we refer to Embretson and Reise (2013) and Rizopoulos (2006). Embretson and Reise (2013) give an extensive presentation of IRT. Rizopoulos (2006), gives a presentation of IRT along side examples and information about the use of the *R* package *ltm*, a package developed for the application of IRT.

## 4 Conclusion

Going beyond the analysis of category frequency, in this paper we have introduced a way to visualize and identify annotators' bias by the use of IRT, more specifically the GRM. Our interpretation allows us to use IRCCC and the extremity parameters to gain an insight into the judges' annotation bias, based on a common latent trait scale of measurement. The use of IRT sheds light on the annotation disagreement. Accordingly, it can be used to have a better picture of the annotation reliability. As an example, in this paper we have showed how it can be used to accompany IAA values in order to explain the annotation disagreement. In particular, IRT can show in which respect an annotator shows a stricter or more severe annotation behaviour than another annotator. Our proposal can also be used as a standard procedure in pilot studies. Indeed, it can help in taking some actions for the main annotation. For example, our proposal can be used to remove outlying annotators or it can be used for isolating where disagreement occurs. In the latter case, one option is then to discuss the disagreement with the annotators to improve the annotation guidelines. Alternatively, identifying the areas of disagreement could be used to understand where to intervene for training the annotators.[11]

## References


J. Amidei, P. Piwek, and A. Willis. 2018. Rethinking the agreement in human evaluation tasks. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 3318–3329.

R. Artstein and M. Poesio. 2008. Inter-coder agreement for computational linguistics. *Computational Linguistics*, 34(4):555–596.

J. Carletta. 1996. Assessing agreement on classification tasks: The kappa statistic. *Computational Linguistics*, 22(2):249–254.

J. Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46.

Anthony J Conger. 1980. Integration and generalization of kappas for multiple raters. *Psychological Bulletin*, 88(2):322.

S. E. Embretson and S. P. Reise. 2013. *Item response theory*. Psychology Press.

J. L. Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological bulletin*, 76(5):378.

A. Gatt and E. Krahmer. 2018. Survey of the state of the art in natural language generation: Core tasks, applications and evaluation. *Journal of Artificial Intelligence Research*, 61:65–170.

H. Gulliksen. 1950. *Theory of mental tests*. Wiley: New York.

Kilem L Gwet. 2014. *Handbook of inter-rater reliability: The definitive guide to measuring the extent of agreement among raters*. Advanced Analytics, LLC.

A. Joshi, P. Bhattacharyya, M. Carman, J. Saraswati, and R. Shukla. 2016. How do cultural differences impact the quality of sarcasm annotation?: A case study of indian annotators and american text. In *Proceedings of the 10th SIGHUM Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities*, pages 95–99.


[11] Where an IRT analysis needs to included with a paper and a page limit prevents this, we suggest presenting the analysis in an appendix or referring to supplementary materials posted to an online repository.

K. Krippendorff. 1980. *Content Analysis: An Introduction to Its Methodology*. Sage Publications, Beverly Hills, CA.

John P Lalor, Hao Wu, and Hong Yu. 2016. Building an evaluation scale using item response theory. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, volume 2016, page 648. NIH Public Access.

A. Lommel, M. Popović, and A. Burchardt. 2014. Assessing inter-annotator agreement for translation error annotation. *In Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC), 26-31 May. Reykjavik, Iceland*.

M. Palmer and N. Xue. 2010. Linguistic annotation. *The Handbook of Computational Linguistics and Natural Language Processing*, pages 238–270.

J. Pustejovsky and A. Stubbs. 2013. *Natural Language Annotation for Machine Learning*, volume 1. Published by O'Reilly Media, Gravenstein Highway North, Sebastopol, CA.

Georg Rasch. 1960. Studies in mathematical psychology: I. probabilistic models for some intelligence and attainment tests.

Dimitris Rizopoulos. 2006. ltm: An R Package for Latent Variable Modeling and Item Response Theory Analyses. *Journal of statistical software*, 17(5):1–25.

Dimitris Rizopoulos. 2018. Latent Trait Models under IRT, Retrieved from: `https://cran.r-project.org/web/packages/ltm/ltm.pdf`.

V. Rus, B. Wyse, P. Piwek, M. Lintean, S. Stoyanchev, and C. Moldovan. 2012. A detailed account of the first question generation shared task evaluation challenge. *Dialogue & Discourse*, 3(2):177–204.

Fumiko Samejima. 1969. Estimation of latent ability using a response pattern of graded scores. *Psychometrika monograph supplement*.

G. Sampson and A. Babarczy. 2008. Definitional and human constraints on structural annotation of english. *Natural Language Engineering*, 14(4):471–494.