On the identification of critical questions in the PISA for Schools program*

Noelia Rico[1], Pedro Alonso[2], Laura Muñiz-Rodríguez[3], Raúl Pérez-Fernández[3,4], Luis J Rodríguez-Muñiz[3], and Irene Díaz[1]

[1] Department of Computer Science, University of Oviedo, Spain
{noeliarico,sirene}@uniovi.es
[2] Department of Mathematics, University of Oviedo, Spain palonso@uniovi.es
[3] Department of Statistics and O.R. and Mathematics Education, University of Oviedo, Spain {munizlaura,perezfernandez,luisj}@uniovi.es
[4] KERMIT, Department of Data Analysis and Mathematical Modelling, Ghent University, Belgium

Abstract. PISA for Schools is an OECD (Organization for Economic Cooperation and Development) program designed to provide results on students' performance in Mathematics, Reading and Science at school level. In order to achieve this purpose, participants are asked to answer a content-based test together with a background questionnaire. Answers are next evaluated by a group of reviewers using a coding guide defined in terms of an ordinal qualitative scale. Although guarantying consistency among reviewers is key, differences may arise on a particular question due to different interpretations of the coding guide for some specific answers. In order to identify the origin of the discrepancies and ensure consistency in the evaluation process of forthcoming editions of the program, this work aims at identifying critical questions that lead to the largest disagreements among reviewers. Ultimately, this critical questions should be examined in detail by performing some qualitative analysis of both the answers provided by the participants and the instructions on the coding guide.

Keywords: PISA · Ordinal scale · Degrees of proximity · Dispersion.

1 Introduction and context

In the context of decision-making, experts usually have to express their preferences on a set of alternatives. Many decision-making problems use ordinal qualitative scales formed by linguistic terms since very often data are expressed at an ordinal level. In fact, words are more natural than numbers [10] and more appropriate for dealing with imprecision and uncertainty in human decisions [5]. According to [8], ordinal qualitative scales used in decision-making problems are,

* Supported by Grants TIN2017-87600-P from Spanish Government and IDI/2018/000176 from the Principality of Asturias. Raúl Pérez-Fernández acknowledges the support of the Research Foundation of Flanders (FWO17/PDO/160).

in general, Likert-type scales. These scales are characterized by ordered response categories in which there is a balanced number of positive and negative categories and a numerical value is assigned to each category.

These scales are often applied to measure perceived quality, which is a task arising in many fields such as health and education. For example, in the context of education, Cerchiello *et al.* [4] propose to summarize students' perceived quality data using non parametric indices based on the observed frequency distribution that are able to exploit efficiently the ordinal nature of the analyzed variables. Thus, they obtain a ranking of the taught courses and produce indicators used to design plan actions on the organizational component and on the relationship between didactics and adequacy of the resources.

PISA is the well-known OECD Program for International Student Assessment. In particular, PISA for Schools is a voluntary assessment program that aims to provide valuable information on the learning climate within a school and to measure key components of 21st century skills mainly based on Mathematics, Reading and Science. The assessment has been successfully administered over 2200 times in schools in 11 countries around the world [9]. From these assessments valuable information for national education systems is provided. Thus, the evaluation process should be as accurate as possible. For that purpose, the pilot study is especially important and, thus, reviewers should be trained on the use of the coding guides and their interpretation. Reviewers are strongly requested to apply the coding guides with a high level of consistency. However, some answers could be difficult to interpret, and consequently they would lead to different marks if they were evaluated by different reviewers.

Clearly, PISA for Schools evaluation constitutes a decision-making problem where different reviewers express an evaluation by using an ordinal scale. How to deal with ordinal scales is somehow a complex problem. Initially, it was common to transform the scale into several dichotomous variables or to arbitrarily associate each linguistic term with a number. However, it is now known that standard measures of the spread of a distribution are inappropriate when dealing with qualitative data [1]. In this work, we follow a different approach based on dispersion measures in the context of ordinal qualitative scales [6]. More precisely, this work aims at identifying the questions that lead to the largest disagreement among reviewers, the ultimate goal being to improve the training process for reviewers in future editions of the program.

The paper is organized as follows: Section 2 describes the data used in this study. Section 3 details the method employed to identify critical questions. The method is applied to the PISA for Schools data in Section 4. Finally, in Section 5 some results and conclusions are drawn.

2 Problem description

The here-evaluated data have been retrieved from the pilot phase of PISA for Schools undertook in Spain. In the PISA for Schools program, students are asked to complete a booklet that holds different items, which are taken by a semi-randomized procedure from a previously designed dataset of PISA-based questions. Although the program was undertaken in 11 countries, booklets are the same for all of them (adapted to each national language). Different stimuli are presented in the booklet, and one or more items are related to each stimulus. Furthermore, items are classified into one of the three following domains: **Mathematics**, **Reading**, and **Science**.

These items are made up of questions, which can be classified based on the nature of their possible answer. Some of them are multiple choice questions, in which the students have to select the only correct answer from a closed set of options. These questions are considered *original* because student's genuine answer is straightforwardly translated to its mark. Obviously, these questions are out of the present study, because different marks for one pair of student/question could only obey to mistakes in typing or coding. On the other hand, there are questions that do require the students to write their own answer. These are known as *coded* questions because, in order to mark the question, the student's answer has to be previously interpreted by an expert reviewer. For some of these questions the students must write a short answer but sometimes they are required to construct a more elaborated answer or even to justify their reasoning. Hence, depending on the length of the answer, *coded* questions can be split into two different subcategories: coded questions with short answer and coded questions with long answer.

The reviewers are provided with coding guides with strict criteria and guidelines for evaluating the questions. Although *coded* questions with short answer may appear difficult to evaluate, the coding guides make clear enough how to mark them in relation to different possible answers of the students, and thus it is immediate to obtain their mark too. The most challenging questions for the reviewers are ***coded* questions with long answer**. Despite the effort on making the guidelines in the coding guide as precise as possible, sometimes there still could be subjectivity on the evaluation. Therefore, the latter are the object of interest in this work, since it is in their evaluation process where reviewers are more likely to disagree.

The PISA for Schools pilot study consists of a total of 141 questions, specifically there are 40 questions defined for Mathematics, 47 for Reading and 54 for Science. *Coded* questions with long answer requiring to be codified by experts are a total of 44 (more specifically: 7 for Mathematics, 17 for Reading and 20 for Science). Coding guides provided to the reviewers clarify how the answers of the students should be encoded and marked with one of following mutually exclusive categories: *unanswered*, *incorrect*, *semi-correct* or *correct*. Note that not all of the questions have so many categories available. More concretely, there are answers that only can be *unanswered*, *incorrect* or *correct*. The former are here referred to as Type B questions, whereas the latter are referred to as Type A questions.

These coded questions for which the students are required to write a more elaborated answer are more suitable to be misunderstood and thus their codification is thoroughly done by a total of four reviewers, which are randomly taken from a pool of reviewers following an experimental design made by the organization of the program. From now on, the term *question* will refer only to one of the coded questions with long answer, as they will be the material for our study.

Therefore, the sample size of our study consists of 44 questions which are answered by 1568 different students. Note that not all the 1568 students answered the 44 questions. More precisely, each student completed only one of the seven available booklets (whose number of questions varies between 15 and 24). To sum up, we have data of 672 different answers for each of the 44 questions and each answer is reviewed by 4 different reviewers.

3 Methodology

The purpose of this analysis is to identify critical questions in the PISA for Schools program. By providing the reviewers with a coding guide with strict instructions for marking the answers of the students, the subjectivity in the evaluation should be minimized. Unfortunately, despite the efforts to make the coding guide as explicit as possible, it seems that reviewers do not always agree on their evaluations. For this very reason, it is necessary to find a method to describe how much the reviewers disagree with regard to their evaluation of a precise question for a precise student. We aim to determine which are the questions that, despite the guidelines on the coding guide, lead to the largest disagreements among reviewers. The key points of the method proposed in this work are the following:

- Fix the considered ordinal scale and establish the degrees of proximity between the elements of this scale (see Subsection 3.1).
- Obtain all the possible **combinations** of marks given by r independent reviewers. Each combination has a vector of degrees of proximity associated, hereinafter referred to as **vector of dispersion** (see Subsection 3.2).
- Identify combinations leading to the same vector of degrees of proximity. Each set of combinations with the same vector of degrees of proximity is referred to as a **tier** (see Subsections 3.2 and 3.3).
- Establish a hierarchy of tiers (see Subsection 3.3).
- Rank the questions (see Subsection 3.3).

3.1 The ordinal scale

To model the present problem we will work with ordinal scales. An **ordinal scale** $\mathscr{L} = \{L_1, \ldots, L_\ell\}$ is a tool of measurement where the elements of $\mathscr{L}$ are linearly ordered such that $L_i \leq L_j$ if $i \leq j$. Each ordinal scale is associated with a set of degrees of proximity $\varDelta = \{\delta_1, \ldots, \delta_h\}$ representing how distant the elements of the ordinal scale $\mathscr{L}$ are with respect to each other (see [7]). Here, we do consider uniform ordinal scales, meaning that the degree of proximity between any two elements L_i and L_j is given by $|i - j| + 1$.

3.2 The vector of dispersion

Assuming that each reviewer uses the same ordinal scale $\mathscr{L}$, a **dispersion measure** on $\mathscr{L}^r$ (being r the number of reviewers evaluating each question) can be used for measuring the **disagreement among reviewers**.

One of the most prominent dispersion measures for ordinal scales is based on the Gini index [7]. The dispersion associated with the evaluations given by the r independent reviewers is obtained from comparing head-to-head all elements L_i given by each reviewer for an answer. A total of $p = \frac{r(r-1)}{2}$ degrees of proximity representing these pairwise comparisons are obtained. These degrees are gathered in a vector sorted in ascending order. Example 1 illustrates the procedure.

Example 1. Consider an answer given by a student to a question. The question is evaluated by 4 independent reviewers (R_1, R_2, R_3, R_4) using an ordinal scale $\mathscr{L} = \{L_1, L_2\}$. Consider that the degree of proximity of L_1 and L_1 is δ_1; the degree of proximity of L_1 and L_2 is δ_2; and the degree of proximity of L_2 and L_2 is δ_1.

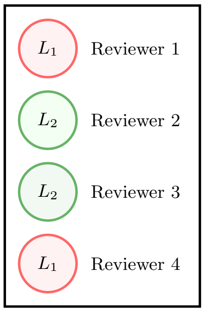

Table 1. Head-to-head degrees of proximity for the combination of marks shown in Fig. 1

Reviewer	R_1 (L_1)	R_2 (L_2)	R_3 (L_2)	R_4 (L_1)
R_1 (L_1)	-	δ_2	δ_2	δ_1
R_2 (L_2)	-	-	δ_1	δ_2
R_3 (L_2)	-	-	-	δ_2
R_4 (L_1)	-	-	-	-

Fig. 1. Example of combination of marks given by four independent reviewers using an ordinal scale $\mathscr{L} = \{L_1, L_2\}$ to a student's answer of a question.

The degrees of proximity obtained in the comparison head-to-head of the marks shown in Fig. 1 are specified in Table 1. These degrees are sorted in ascending order, thus obtaining the six-element vector of dispersion $(\delta_1, \delta_1, \delta_2, \delta_2, \delta_2, \delta_2)$.

3.3 Application of the method of majority judgement

Once the vector of dispersion is obtained, it is necessary to rank the combinations of marks according to its associated dispersion. The method selected in this work to perform this task is similar to Balinski and Laraki's majority judgment [2, 3], but with the correction described by García-Lapresta and Borge [6]. More precisely, Balinski and Laraki's majority judgment ranks (frequency) distributions of elements on an ordinal scale in terms of their couple of medians. In case two distributions have the same couple of medians, it is necessary to break the tie between them by removing this couple of each distribution and, subsequently, choosing the couple of medians of the new vector as the couple for the comparison. This procedure is repeated until all distributions are ranked. The correction described by García-Lapresta and Borge [6] considers both elements of the couple of medians (instead of the smallest among both elements, as originally proposed by Balinski and Laraki). This method is briefly described below.

- Consider an ordinal scale $\mathscr{S} = \{S_1, \ldots, S_s\}$. Consider two different vectors of elements $\mathbf{s}_1 = (s_{11}, \ldots, s_{1n}), \mathbf{s}_2 = (s_{21}, \ldots, s_{2n}) \in \mathscr{S}^n$ such that $s_{1i} \leq s_{1j}$ and $s_{2i} \leq s_{2j}$ if $i \leq j$. These vectors are ranked according to their associated couples of medians (L_i, L_j) (for $\mathbf{s}_1$) and $(L_{i'}, L_{j'})$ (for $\mathbf{s}_2$).
- The vector with couple of medians (L_i, L_j) is ranked below the vector with couple of medians $(L_{i'}, L_{j'})$ if either
  - $i + j < i' + j'$ or
  - $i + j = i' + j'$ and $j - i \leq j' - i'$.
- If $\mathbf{s}_1$ and $\mathbf{s}_2$ share the same couple of medians, it is necessary to break the tie between the vectors. To that end, the couple is removed from both vectors and the new couples of medians of the new vectors are selected to perform the comparison. The vectors are ranked according to the two rules defined in the previous step. This procedure is repeated until the two vectors are ranked.

Note that this method is applied in this work twice: firstly, to establish a hierarchy of tiers (considering as input vectors the vector of degrees of proximity associated with each combination of marks); secondly, to rank the questions (considering as input vectors the vectors of tiers associated with the evaluations for all the students). Example 2 illustrates the procedure for ranking the questions.

Example 2. Consider the ordinal scale given by the tiers $\{T_0, T_1, T_2\}$ and three questions Q1, Q2 and Q3. For this example, ten students answer to these questions Q1, Q2 and Q3 and their answers are marked by four different reviewers. The number of times that the combination of the marks falls in each tier is shown in Table 2 together with their cumulative frequencies.

These three questions share the same couple of medians (T_0, T_0). Thus, this couple is removed from the vector representing each question and a new couple of medians is computed for each question. The next three lines of Table 2 show the new couple of medians. Note that (T_0, T_0) is smaller than (T_0, T_1). Thus, $Q2$ is the least critical question. As $Q1$ and $Q3$ share the same couple of medians (T_0, T_1), this couple is removed from the combination, obtaining the pair (T_0, T_2) for $Q1$ and (T_0, T_1) for $Q3$. As (T_0, T_1) is smaller than (T_0, T_2), the resulting ranking from most critical to least critical is Q1 $\succ$ Q3 $\succ$ Q2.

4 Ranking PISA for Schools questions

The method described in Section 3 is then applied to identify critical questions in the PISA for Schools program. Recall that, as already stated in Section 2, the

Table 2. Example of answers of 10 students to 3 different questions and the distribution of the combination of their evaluation among tiers.

QuestionID	Combination	Distribution			Cumulative frequency		
		T_0	T_1	T_2	T_0	T_1	T_2
Q1	$(T_0, T_0, T_0, T_0, \mathbf{T_0}, \mathbf{T_0}, T_1, T_2, T_2, T_2)$	6	1	3	0.6	0.7	1
Q2	$(T_0, T_0, T_0, T_0, \mathbf{T_0}, \mathbf{T_0}, T_0, T_1, T_1, T_2)$	7	2	1	0.7	0.9	1
Q3	$(T_0, T_0, T_0, T_0, \mathbf{T_0}, \mathbf{T_0}, T_1, T_1, T_2, T_2)$	6	2	2	0.6	0.8	1
Removing (T_0, T_0)							
Q1	$(T_0, T_0, T_0, \mathbf{T_0}, \mathbf{T_1}, T_2, T_2, T_2)$	4	1	3	0.5	0.125	0.375
Q2	$(T_0, T_0, T_0, \mathbf{T_0}, \mathbf{T_0}, T_1, T_1, T_2)$	5	2	1	0.625	0.25	0.125
Q3	$(T_0, T_0, T_0, \mathbf{T_0}, \mathbf{T_1}, T_1, T_2, T_2)$	4	2	2	0.5	0.125	0.375
Removing (T_0, T_1)							
Q1	$(T_0, T_0, \mathbf{T_0}, \mathbf{T_2}, T_2, T_2)$	3	0	3	0.5	0	0.5
Q3	$(T_0, T_0, \mathbf{T_0}, \mathbf{T_1}, T_2, T_2)$	3	1	2	0.5	0.167	0.333

questions of interest are the ones in which the students write a long answer that the reviewers should mark according to the criteria given by a coding guide. In particular, there are two different types of questions within this group.

- Type A questions are the ones whose answers can be marked as *unanswered*, *incorrect* or *correct*.
- Type B questions are those which refine their possible marks adding an extra category so the question can be marked as *unanswered*, *incorrect*, *semi-correct* or *correct*.

Since the two types of questions have a different number of possible answers, the considered ordinal scale must differ.

On the one hand, for Type A questions we define $\mathscr{L}^A = \{L_1^A, L_2^A\}$ ($\ell = 2$), i.e., the answer of the student is marked as L_1^A if it is *incorrect* or *unanswered*; or as L_2^A if it is *correct*. Two possible degrees of proximity ($\Delta^A = \{\delta_1^A, \delta_2^A\}$) are obtained for Type A questions. The ordinal scale is illustrated in Fig. 2.

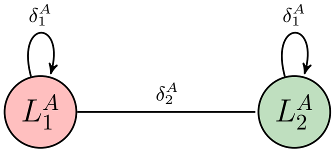

Fig. 2. Ordinal scale for Type A questions.

On the other hand, for Type B questions we define the ordinal scale $\mathscr{L}^B = \{L_1^B, L_2^B, L_3^B\}$ ($\ell = 3$), i.e., the answer of the student is marked as L_1^B if it is *incorrect* or *unanswered*; as L_2^B if it is *semi-correct*; or as L_3^B if it is *correct*. Here, we assume the scale used for Type B questions to be uniform[5]. Intuitively, this means that (for instance) for Type B questions we assume that *semi-correct* (L_2^B) is as close to being *correct* (L_3^B) as it is to being *incorrect* (L_1^B). Formally, this is the reason why the degree of proximity for L_i and L_j is given by $\delta_{|i-j|+1}$. Three possible degrees of proximity ($\varDelta^B = \{\delta_1^B, \delta_2^B, \delta_3^B\}$) are obtained for Type B questions. The resulting ordinal scale for these questions is illustrated in Fig. 3.

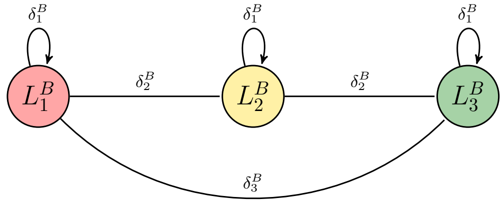

Fig. 3. Ordinal scale for Type B questions.

Once the scales are defined for each type of question, for each combination of marks provided by 4 reviewers the vector of dispersion is computed as described in Subsection 3.2 (see also Example 1). The total number of different combinations that can be obtained given ℓ different marks in a combination of length r with possible repetitions is $CR_{\ell,r} = \binom{\ell+r-1}{r}$. Thus, in our problem:

- For Type A where the scale is $\{L_1^A, L_2^A\}$ ($\ell = 2$) and $r = 4$, there are 5 different combinations of marks.
- For Type B where the scale is $\{L_1^B, L_2^B, L_3^B\}$ ($\ell = 3$) and $r = 4$, there are 15 different combinations of marks.

Following the process illustrated in Example 1 with all the 5 possible combinations for Type A questions and 15 combinations for Type B questions, it becomes clear that some of these combinations are represented by the same dispersion vector and, thus, can be considered equivalent. As the questions will be later ranked based on their dispersion, equivalent combinations with the same associated vector of dispersion are grouped into the same **tier**. Table 3 and Table 5 show the tiers obtained respectively for Type A ($\mathscr{T}^A = \{T_0^A, T_1^A, T_2^A\}$) and Type B questions ($\mathscr{T}^B = \{T_0^B, T_1^B, T_2^B, T_3^B, T_4^B, T_5^B, T_6^B\}$).

The tiers are then ranked according to the method described in Subsection 3.3. Thus, the couple of medians for each dispersion vector is computed for each tier (see Table 4 and Table 6 for Type A and Type B questions respectively).

Obtaining a ranking of tiers from least to most disperse for Type A questions is immediate since considering the subindices of the median degrees $(1+1) < (1+2) < (2+2)$ a ranking without ties is generated. Nevertheless, for Type B

[5] Note that the scale used for Type A questions is necessarily uniform since it contains only two elements.

Table 3. Combinations of the marks expressed by four different reviewers for Type A questions and their associated vector of dispersion.

Combination	Vector of dispersion
L_1^A, L_1^A, L_1^A, L_1^A L_2^A, L_2^A, L_2^A, L_2^A	$(\delta_1^A, \delta_1^A, \delta_1^A, \delta_1^A, \delta_1^A, \delta_1^A)$
L_1^A, L_2^A, L_2^A, L_2^A L_2^A, L_1^A, L_1^A, L_1^A	$(\delta_1^A, \delta_1^A, \delta_1^A, \delta_2^A, \delta_2^A, \delta_2^A)$
L_1^A, L_1^A, L_2^A, L_2^A	$(\delta_1^A, \delta_1^A, \delta_2^A, \delta_2^A, \delta_2^A, \delta_2^A)$

Table 4. Tiers of combinations for Type A questions.

Tier	Vector of dispersion	Couple of median degrees of proximity
T_0^A	$(\delta_1^A, \delta_1^A, \delta_1^A, \delta_1^A, f\delta_1^A, \delta_1^A)$	(δ_1^A, δ_1^A)
T_1^A	$(\delta_1^A, \delta_1^A, \delta_1^A, \delta_2^A, \delta_2^A, \delta_2^A)$	(δ_1^A, δ_2^A)
T_2^A	$(\delta_1^A, \delta_1^A, \delta_2^A, \delta_2^A, \delta_2^A, \delta_2^A)$	(δ_2^A, δ_2^A)

Table 5. Combinations of the marks expressed by four different reviewers for Type B questions and their associated vector of dispersion.

Combination	Vector of dispersion
L_1^A, L_1^A, L_1^A, L_1^A L_2^A, L_2^A, L_2^A, L_2^A L_3^A, L_3^A, L_3^A, L_3^A	$(\delta_1^B, \delta_1^B, \delta_1^B, \delta_1^B, \delta_1^B, \delta_1^B)$
L_1^A, L_2^A, L_2^A, L_2^A L_3^A, L_2^A, L_2^A, L_2^A L_2^A, L_1^A, L_1^A, L_1^A L_2^A, L_1^A, L_1^A, L_1^A	$(\delta_1^B, \delta_1^B, \delta_1^B, \delta_2^B, \delta_2^B, \delta_2^B)$
L_1^A, L_1^A, L_2^A, L_2^A L_3^A, L_3^A, L_2^A, L_2^A	$(\delta_1^B, \delta_1^B, \delta_2^B, \delta_2^B, \delta_2^B, \delta_2^B)$
L_1^A, L_2^A, L_2^A, L_3^A	$(\delta_1^B, \delta_2^B, \delta_2^B, \delta_2^B, \delta_2^B, \delta_3^B)$
L_1^A, L_2^A, L_2^A, L_3^A L_1^A, L_1^A, L_2^A, L_3^A	$(\delta_1^B, \delta_2^B, \delta_2^B, \delta_2^B, \delta_3^B, \delta_3^B)$
L_1^A, L_1^A, L_1^A, L_3^A L_3^A, L_3^A, L_3^A, L_1^A	$(\delta_1^B, \delta_1^B, \delta_1^B, \delta_3^B, \delta_3^B, \delta_3^B)$
L_1^A, L_1^A, L_2^A, L_2^A	$(\delta_1^B, \delta_1^B, \delta_3^B, \delta_3^B, \delta_3^B, \delta_3^B)$

Table 6. Tiers of combinations for Type B questions.

Tier	Vector of dispersion	Couple of median degrees of proximity
T_0^B	$(\delta_1^B, \delta_1^B, \delta_1^B, \delta_1^B, \delta_1^B, \delta_1^B)$	(δ_1^B, δ_1^B)
T_1^B	$(\delta_1^B, \delta_1^B, \delta_1^B, \delta_2^B, \delta_2^B, \delta_2^B)$	(δ_1^B, δ_2^B)
T_2^B	$(\delta_1^B, \delta_1^B, \delta_2^B, \delta_2^B, \delta_2^B, \delta_2^B)$	(δ_2^B, δ_2^B) and subsequently (δ_1^B, δ_2^B)
T_3^B	$(\delta_1^B, \delta_2^B, \delta_2^B, \delta_2^B, \delta_2^B, \delta_3^B)$	(δ_2^B, δ_2^B) and subsequently (δ_2^B, δ_2^B)
T_4^B	$(\delta_1^B, \delta_2^B, \delta_2^B, \delta_2^B, \delta_3^B, \delta_3^B)$	(δ_2^B, δ_2^B) and subsequently (δ_2^B, δ_3^B)
T_5^B	$(\delta_1^B, \delta_1^B, \delta_1^B, \delta_3^B, \delta_3^B, \delta_3^B)$	(δ_1^B, δ_3^B)
T_6^B	$(\delta_1^B, \delta_1^B, \delta_3^B, \delta_3^B, \delta_3^B, \delta_3^B)$	(δ_3^B, δ_3^B)

questions, tiers $\{T_2^B, T_3^B, T_4^B\}$ have the same couple of median degrees of proximity (δ_2^B, δ_2^B). To further rank these tiers it is necessary to remove this couple of the vector of dispersion, thus obtaining a new couple of median degrees of proximity. For example, for T_2 the vector of dispersion after removing the initial couple of median degrees of proximity is $(\delta_1^B, \delta_1^B, \delta_2^B, \delta_2^B)$ with couple of median degrees of proximity (δ_1^B, δ_2^B). This is repeated for each tier, resulting in the couples shown in Table 6. Fig. 4 and Fig. 5 show the combinations that belong to each tier.

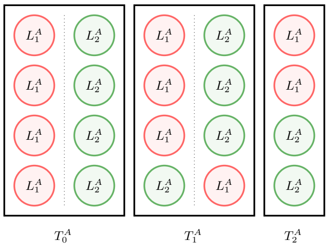

Fig. 4. Hierarchy of dispersion among reviewers for all possible combinations of elements L_i^A expressed by the $r = 4$ reviewers for Type A questions. The instances of L_1^A are pictured in red and the instances of L_2^A are pictured in green.

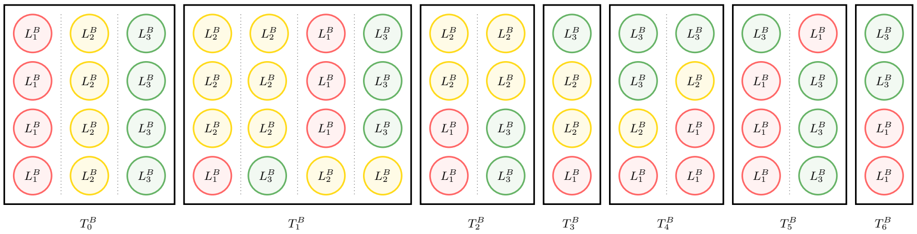

Fig. 5. Hierarchy of dispersion among reviewers for all possible combinations of elements L_i^B expressed by the $r = 4$ reviewers for Type B questions. The instances of L_1^B are pictured in red, the instances of L_2^B are pictured in yellow and the instances of L_3^B are pictured in green.

Once the tiers are ranked, it is possible to rank the questions, which is the ultimate goal of this work. Hence, for each question, the number of combinations in each tier is counted and Balinski and Laraki's majority judgement described in Subsection 3.3 is again applied. Thus, questions are ranked according to their respective distribution of frequencies of tiers. In order to rank the questions according to their disagreement among reviewers, all tiers obtained for a question throughout all students are gathered together.

Notice that, the answers of at least half of the students fall in tier T_0. This tier accumulates all the answers for which the four reviewers give the same mark. Hence, the first couple of median degrees will always be (T_0, T_0) and, still when this couple is removed, the left value of the couple will always be T_0. Thus, the question with least dispersion will be the one that changes earlier from tier T_0 to a tier associated with more dispersion. This very much simplifies the method because questions are actually ranked lexicographically according to their cumulative frequency in tier T_0, and subsequently in T_1, and so on.

Fig. 6 represents the distribution of frequencies of tiers for the questions in the three categories (Mathematics, Reading and Science) covered by the PISA for Schools program. Mathematics questions are indicated with the prefix PM, Reading questions are indicated with the prefix PR and Science questions are indicated with the prefix PS.

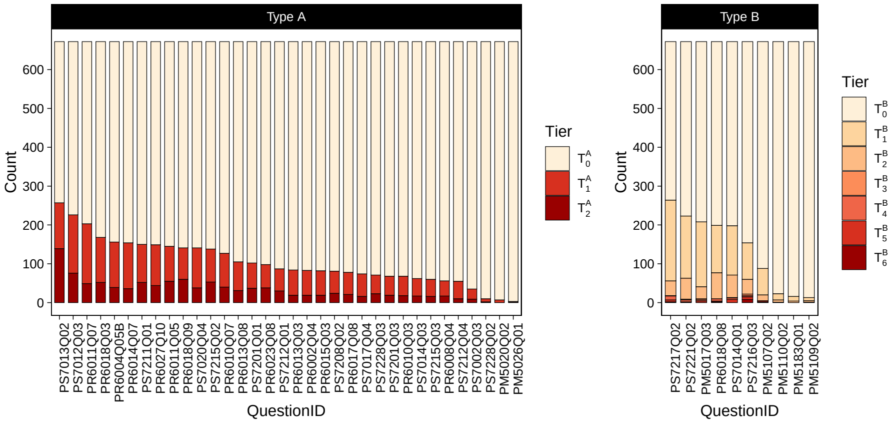

Fig. 6. Graphical representation of the distributions of frequencies for the questions of the PISA for Schools program.

At the light of the results, we infer that the obtained ranking indicates the questions for which the coding guide appears to be less precise. As it can be seen in Fig. 6 questions with more discrepancy are related to Science, whereas Mathematics questions tend to be be associated with less discrepancy, especially when the granularity of the ordinal scale is greater. In particular, the coding guide of Question PS7013Q02 should be the first one to be further improved among Type A questions and the coding guide of Question PS7217Q02 should be the first one to be further improved among Type B questions.

5 Conclusions and future work

The method proposed in this work has enabled to identify questions that lead to the largest disagreements among reviewers during an evaluation process. After using data from the pilot study of PISA for Schools program, the method is proved to be valid, since its outcome can be interpreted as a ranking of the questions from most critical to least critical in terms of the disagreement among reviewers.

We should underline that, at the present stage of our study, we are not focusing on discussions concerning the content and the coding guides that reviewers may hold after a discrepancy is detected. This is something we plan to do in a forthcoming study. Primarily, we have developed a method for detecting critical questions, that is, questions being liable to produce controversies in reviewers' interpretations. Starting by these questions identified as critical, the next step will lie in qualitatively analyzing both the answers provided by the participants and the instructions on the coding guide. The latter will provide valuable feedback in view of improving the training process for reviewers and, therefore, guarantying a higher consistency in the evaluation process of the PISA for Schools program.

References

1. Allison, R.A., Foster, J.E.: Measuring health inequality using qualitative data. Journal of Health Economics **23**(3), 505–524 (2004).
2. Balinski, M., Laraki, R.: A theory of measuring, electing and ranking. Proceedings of the National Academy of Sciences of the United States of America **104**, 8720–8725 (2007).
3. Balinski, M., Laraki, R.: Majority Judgment: Measuring, Ranking, and Electing. MIT press, Cambridge (2010).
4. Cerchiello P., De Quarti E., Giudici P., Magni C.: Scorecard models to evaluate perceived quality of academic teaching. Statistica & Applicazioni **8**(2), 145–155 (2010).
5. Fasolo, B., Bana, C.A., Costa, E.: Tailoring value elicitation to decision makers' numeracy and fluency: Expressing value judgments in numbers or words. Omega **44**, 83–90 (2014).
6. García-Lapresta, J.L., Borge, L.: Measuring Dispersion in the Context of Ordered Qualitative Scales. In: Gil E., Gil E., Gil J., Gil M. (eds) The Mathematics of the Uncertain. Studies in Systems, Decision and Control, vol 142. Springer, Cham (2018).
7. García-Lapresta, J.L., Pérez-Román, D.: Ordinal proximity measures in the context of unbalanced qualitative scales and some applications to consensus and clustering. Applied Soft Computing **35**, 864–872 (2015).
8. García-Lapresta, J.L., González del Pozo, R.: An ordinal multi-criteria decision-making procedure under imprecise linguistic assessments. European Journal of Operational Research **279**(1), 159–167 (2019).
9. PISA for Schools Brochure, (https://www.oecd.org/pisa/pisa-for-schools/PISA-for-Schools-Brochure-(Digital).pdf) . Last accessed 3 Feb 2020.
10. Windschitl, P.D., Wells, G.L.: Measuring psychological uncertainty: Verbal versus numeric methods Journal of Experimental Psychology: Applied, **2**, 343–364 (1996).