UNIVERSITY OF PADOVA

DEPARTMENT OF GENERAL PSYCHOLOGY

PHD COURSE IN PSYCHOLOGICAL SCIENCES
XXXVI CYCLE

Flexible Rating Models: Bayesian parametric and nonparametric approaches

PhD Course Coordinator
Lucia Regolin

Supervisor
Andrea Spoto

Co-supervisor
Antonio Calcagnì

PhD Candidate
Giuseppe Mignemi

Flexible Rating Models: Bayesian parametric and nonparametric approaches

Abstract

In several observational contexts where different raters evaluate a set of items, it is common to assume that all raters draw their scores from the same underlying distribution. However, plenty of scientific works have evidenced the relevance of individual variability in different types of rating tasks. To address this issue the intra-class correlation coefficient (ICC) has been used as a measure of variability among raters within the Hierarchical Linear Models approach. A common distributional assumption in this setting is to specify hierarchical effects as independent and identically distributed from a normal with the mean parameter fixed to zero and unknown variance. The present work aims to overcome this strong assumption in the inter-rater agreement estimation by placing a Dirichlet Process Mixture over the hierarchical effects' prior distribution. A new nonparametric index λ is proposed to quantify raters' polarization in the presence of group heterogeneity. The model is applied to a set of simulated experiments and real-world data. Possible future directions are discussed.
The statistical framework introduced in Nucci et al. (2021) is here generalized. This generalization concerns three different features. First, the specification of *cross-classified* observations, i. e. two independent sources of redundancy are modelled. This is the case in which the same set of items are evaluated independently by different raters. Second, the heteroscedasticity among different raters. The independent and identically normally distributed assumption over the residuals across all the observations might be relaxed. This allows us to capture some systematic differences in rating behaviour among the raters. Some of them might be more consistent than others, this implies a smaller residual variance across their ratings. On the contrary, some raters might be less consistent, as a result, the variance across their ratings is larger. The third generalization feature concerns the rating scale. We generalize the previous framework to the ordinal data case. This implies a flexible modelling in which both the ordinal and the continuous rating data might be analyzed under the same framework. Under this general framework, an *approximate intra-class correlation coefficient* (*ICC$_a$*) is proposed.
In some cases, when the objects of the evaluation are people it might be possible to have a "bidirectional" rating scheme. More specifically, under this scheme people rate each other, a person evaluates other people and, in turn, he/she is evaluated by others as well. People might have a *twofold* role, one as a rater and another as an object of rating, that is they are evaluated. This is a valuable rating solution in situations of peers, for instance in the educational contexts in which each student is evaluated by the other students. As a consequence, they might be regarded both as examinees and as graders (i.e., raters). In this regard, in the last part of the thesis, a peer grading model is proposed – a system in education where each student's work is assessed by several other students. This system is widely used in massive open online courses (MOOCs) as well as classroom settings. While peer grading substantially reduces teachers' burden in grading coursework and may also facilitate students' learning, there are reliability concerns on the measurement caused by the heterogeneous grading behaviours among the students. To address these concerns, we introduce a general statistical framework for peer grading data. The naïve average score may be

inaccurate due to the biases and variances of the individual grades, and thus, the proposed framework provides an optimal scoring rule. Additionally, this framework provides a way to assess the performance of each student as a grader, which may be used to identify a pool of reliable graders or generate feedback to help students improve their grading. Our model can also provide insights into the system by answering questions such as whether a student who performs better in the coursework also tends to be a more reliable grader. Finally, for longitudinal peer grading data, our framework allows latent growth modelling that characterises student progress. The effectiveness of our model is shown via simulation studies and a real-world application.

Contents

List of Figures

List of Tables

Chapter 1

Psychological and psychometric reasons behind raters heterogeneity statistical modeling

It is the difference of opinion that makes horse races

Mark Twain

1.1 Introduction

The hetero-evaluation, i.e. the assessment of an object or a subject [1] made by a rater, is a widely common practice across scientific fields (Bartoš et al., 2019; Walter et al., 2019; Gerke et al., 2016; Gisev et al., 2013). The rater, who is generally a trained expert, is asked to provide feedback on some specific aspect of the object.

The first step of the rating process is the observation of the object, in this phase, the rater focuses his/her attention on some target aspects of the object. The second step is the formulation of the rating in which the rater evaluates some rubrics and rules. Quantitative scales are widely used for

[1] These two terms, i.e. object and subject, are considered synonym throughout the present work. For clarity purpose, the first will be mostly used. We prefer the term object because it better highlights that it is referred to as the object of the rating process.

this aim. Consequently, the essential elements of every rating process are four:

- The **object**, i.e. the target of the evaluation;
- The **rater**, i.e. the person who rates, is sometimes referred to as the *observer*;
- The crucial **attributes** of the object, i.e. the object's aspects that matter and have to be assessed for the rating formulation;
- The **rating scale**, i.e. the well-defined metric used to explicit the evaluation.

From a cognitive perspective, this might be seen as an iterative process in which the formulation of the rating arises spontaneously during the observation of the target characteristics of the object (Anderson and Crawford, 1980; Ormrod, 1999). The process ends when the rater explicitly reports the evaluation on a rating scale (Jia and Zhang, 2023).

This procedure is crucial when the attributes of the object are not directly available or when the context is characterized by a high level of complexity (Nelson and Edwards, 2015). Experts' judgments are helpful in these situations and shed light on the object's attributes and as a result, a decision is often made on the base of raters evaluations. The outcome of the rating process is often followed by a relevant decision-making process, as the rater's opinion is referred to as a notable reference. For this reason, their ratings are often considered as ground truth or a gold standard. Considering that they are required to provide a rate and to formulate an evaluation of some features of the object, their specific point of view matters; it is an important part of the rating process and it might be considered as a tool. The subjective thoughts of the rater regarding the target aspects of the object are of main interest. This is one of the main reasons behind the choice of this assessment procedure; when the raters are experts conceiving their judgments as ground truth might seem more than reasonable. Some implications of this kind of assumption are discussed below.

The rating process as described above is widely adopted in several scientific realms. For explanatory purposes few examples are reported below. In natural sciences, ethologists are asked to evaluate animal behaviours and their interactions (Giammarino et al., 2021; Pfeifer et al., 2019; Vieira et al., 2018). In medical sciences, pathologists are required to assess the severity of a disease (Alavi et al., 2022; Nelson et al., 2020; Nelson and Edwards, 2015; Beach et al., 1995). It is common in educational contexts that the teachers are asked to evaluate the students on some ability scale (Gamage et al., 2021; Gere et al., 2019; Casabianca et al., 2015; Cho and Rabe-Hesketh, 2011; DeCarlo, 2008). In all the quality control fields the experts rate the production chain on a risk scale (Cassata et al., 2022). In emergency rescue contexts the operators are asked to evaluate the request and assign a congruent severity code (Hörlin et al., 2022; Dalwai et al., 2018; Noveanu et al., 2017). One of the most prominent examples is the ambulance system (Alotaibi et al., 2021). The dispatch operator receives the call and, based on some crucial information, rates the request on an urgency scale (Bohm and Kurland, 2018; Deakin et al., 2017; Snooks et al., 2009).

Raters are commonly assumed to be exchangeable [2], that is, their rating behaviour is assumed to be the same. Considering the above-mentioned elements of the rating process, this assumption implies that given the same object, the same set of crucial attributes and the same rating scale, the output of the process doesn't depend on the person who makes it. It is supposed to be the same across different raters. This consideration implies that the result of the rating process is the same, regardless of the rater who carried it out. In this case, it makes no difference whether the assessment is carried out by a rater or by another. All the possible individual differences among raters in expressing their judgments are generally neglected. In the previous example, the help request received at the ambulance dispatch centre should be, in principle, assigned systematically to the same server level by all the possible operators who might address the call.

It is reasonable to expect the rating process to be led and inspired by some standard protocol since the raters are supposed to be well-trained. Notwithstanding, It is widely recognised that different raters might differ in their rating behaviours (Gisev et al., 2013), and a certain amount of heterogeneity among them might result in different ratings; this is especially the case when the evaluation of the attributes is not straightforward and the context is highly uncertain (Nelson et al., 2020). The evaluation in the rating context is always somehow holistic, that is, it regards several elements of the object, and the rater is asked to provide a quite comprehensive evaluation (Gwet, 2008). As the *whole is greater than the sum of the parts*, the evaluation of the same crucial attributes of the object might be slightly different across raters. More specifically, each rater might

[2] This concept will be deepened from a psychometric point of view below.

have a different systematic bias in rating the objects (Nelson and Edwards, 2015). This implies that some of them might be more severe or more lenient than others. Given the same object and even using the same rubrics for the evaluation, they might disagree on the rating. This systematic heterogeneity might be due to several reasons. On the object side, irrelevant characteristics of the object might affect the evaluation of some raters even if this is not supposed to happen. On the raters' side, their experience or some individual characteristic might result in different rating behaviours.

The role of the rater and the related effect on the evaluation process needs to be carefully considered and heterogeneity among raters might pose a threat to the whole rating process and undermine the aim of the procedure. This issue needs to be addressed at different times and from different perspectives and it has to be taken into account during the choice and the training of the raters. They have to be prepared and aware that their judgments may sometimes differ from one another, nonetheless, they need to attempt to be homogeneous in their ratings and reduce any possible divergence. In this sense a practical perspective might be helpful to make raters more similar in their evaluation, for instance, recurrent training courses and meetings might be useful for this aim. A statistical perspective is also needed to quantify each rater's effect, it is a valuable solution that allows us to analyze and control possible systematic biases and then to get rid of them.

Recently, Nucci et al. (2021) pointed out the relevance of raters heterogeneity in assessment procedures. They focus on the dichotomous classification of the object and discuss the role of the rater's threshold in this kind of task. They deepen the concept of agreement from a theoretical point of view. A prominent consideration behind their work is that raters, even if they agree on the general evaluation of the object, might differ in the use of the rating scale. More precisely, they might even share the same thoughts regarding the object, but they end up with different classifications. This is a crucial issue in rating processes. They are supposed to share the same meaning of the rating scale. The reference points for their ratings are supposed to be the very same. Otherwise, any comparison between raters or ratings might be misleading.

Theoretical remarks. Some remarks regarding the concept of heterogeneity among raters are needed. In a broader sense, it refers to the different rating behaviours among raters. For instance, some are more severe and others are more lenient; some are more consistent in their ratings than others. These very general elements of diversity among raters need to be thoughtfully addressed. A substantive issue arises at the intersection of the above-mentioned perspectives, it regards the nature of the diversity among raters. Under the *methodological* perspective, the raters are encouraged to find a common view on the rating task, they attempt to find a shared meaning of the target attributes of the object. This is a crucial step in the rating process: given the same set of attributes to be evaluated, if the raters disagree on the meaning and the relative importance of each of them the rating process is not homogeneous, there is a low consensus rate. The standardization of the protocol might be coupled with the same shared meaning and opinion on the various aspects of the rating. From a psychological point of view, a single meeting or a training course on some rating protocol is not enough to make some experts a group of homogeneous raters.

The central matter is whether the heterogeneity has to be considered as noise or not, the question is about the nature of this diversity. When raters are most eminent experts, they have achieved remarkable knowledge and experience on a specific topic. Therefore, their opinion is grounded and well reasoned and any possible divergence of opinions or any difference in the rating task needs to be carefully considered. They need to delve into the substantive matter in depth. The consensus among raters, that is, the same shared meaning regarding the rating task, is a complex concept. It might be low because of a high level of noise and uncertainty among raters. They do not share a single clear opinion. On the other hand, the consensus might be low due to the presence of a few strong divergent opinions. In this case, they have to be addressed and the reason for their divergence deepens. The two scenarios are way different, they imply separate substantive considerations that need to be addressed.

These considerations are closely related to the concept of *inter-rater agreement*, i.e. the extent to which different evaluators agree on the assessment of an object. The rating, as stated above, might focus on some features, behaviour or ability of the object. The congruence between the evaluation of different raters is a notable issue. It might undermine the quality of the rating process itself. For this reason, the analysis of the inter-rater agreement before the data collection is deemed a good practice. To this aim, the group of raters that are involved in studies are asked to rate the same set of objects on the same scale. Several statistical solutions are available to analyse this sort of data. They aim to quantify the level of agreement and provide some useful information about the rating process. In the most desirable scenario, the output of the rating is equal across raters; the maximum agreement is achieved. In these cases not many considerations are needed, the rating of each object is unequivocal. The level of uncertainty about the ratings is very low. Unfortunately, this is a very rare scenario. It is common to achieve this level of agreement when the rating of the objects is quite trivial. In cases when the object's features are very defined and the application of the rating protocol is straightforward. This happens very rarely. The request for a rater evaluation is frequently driven by a not clear and uncertain context. It might result in a non-trivial task for the raters.

The inter-rater agreement analysis as a preliminary procedure is not always a feasible solution, in more general observational contexts, i.e. when rating data are not collected under experimental and controlled conditions, any analysis on the rating process is conducted afterwards (Zenk et al., 2007). In these cases the possible rater heterogeneity has to be carefully addressed, it might help a better understanding of the rating and might bring novel knowledge in the respective fields.

Raters' heterogeneity needs to be deepened and taken into account during all the phases of the rating process and it has to be considered more broadly, not only noise to get rid of. If properly addressed and analyzed it might be a useful source of information and may highlight how similar raters are and reveal some latent clusters among raters. It might also point out some theoretical or practical discrepancies among raters. Given these premises, a flexible and careful consideration of raters heterogeneity is needed both from a *methodological* and a *statistical* perspective.

1.2 Raters heterogeneity in psychological sciences

The rating process has a well-known appeal within the psychological sciences. Given the observational nature of this procedure, it has been widely used in these fields. One of the most prominent examples is the semi-structured interviews, they are very common in several psychological fields due to their flexibility. This is a case of the rating process, the interviewer is asked to assess the interviewed (i.e., the subject) on a continuous or ordinal scale about some target attributes. There are plenty of these examples in educational (Sun and Cheng, 2014; Cho and Rabe-Hesketh, 2011), work (Martinková et al., 2023; Schneider et al., 2019) or clinical psychology (Levinson et al., 2017; Lobbestael et al., 2011).

It is common for the teacher to assess students' essays or their oral exams (Casabianca et al., 2015; DeCarlo, 2008). It might be done by a single teacher (e.g., in the classroom setting) or by a panel (e.g., admission procedures or final exams), in both cases the final grade is rather relevant for students and a fair process must be guaranteed. A very high level of consensus on the grading process is needed across raters (i.e. teachers) to provide examinees with comparable and fair grades. This is a not trivial issue when different groups of students are graded by different groups of teachers, a standard choice is that every class is evaluated by their teachers. In these cases the expected heterogeneity among raters is very high as every teacher has his/her point of view and interpretation of the grading scale, they might have various training and experience (Randall and Engelhard, 2009; Sun and Cheng, 2014; Brookhart, 1993). If neglected, this diversity might be detrimental to the grading procedures as students are supposed to be graded through a very fair system. Nevertheless, once analysed and deepened, this kind of heterogeneity might reveal the individual rater's biases and different ways to look at the students' essays or oral exams (Quinn, 2020; Drake et al., 2019). In this case, a clear quantification of the different grading behaviours might be an important point of reflection. It might be seen as a needed analysis for future training programs for teachers.

Several studies in developmental psychology rely on rating data (Hou et al., 2020; Barbot et al., 2014; Elliot et al., 1993). Using hetero-evaluation data is very common when the participants are toddlers, in these cases the caregivers are asked to evaluate the behaviour of their kids. For similar reasons, this procedure is rather common also in disability psychology field (Lee et al., 2022; Sheaffer et al., 2021; Swanson and Vaughn, 2010; Araújo et al., 2009). In all these cases each subject is rated by a different rater (or by a different pair of raters if both parents are involved) who is not even an expert (Hou et al., 2020). Also in these contexts, the level of expected heterogeneity among caregivers is very high. Notwithstanding, this might be meaningful information for the psychologists or the researchers. It might reveal some aspect of diversity in their parental style or the way they judge their kids' behaviour. For instance, more permissive caregivers are expected to over-rate their kids' behaviour. On the contrary, more severe caregivers tend to under-rate it.

One of the most common examples of rating in clinical psychology is when the diagnosis data are collected by the therapist or the hospital specialists (Levinson et al., 2017; Lobbestael et al., 2011; Möller, 2000). The patient is usually assessed by a psychologist on an ordinal scale in which each step corresponds to a level of disease severity. Generally in these studies, each rater (psychologist) assesses a different small group of patients. Here the raters are supposed to be well trained on this task and the expected consensus is high. Even though, the different settings in which the data are collected and the various experiences of the psychologist might result in a moderately heterogeneous rating style.

The hetero-evaluations are becoming increasingly used in workplace studies (Schneider et al., 2019; Strahl et al., 2019). Work and organizational psychologists use this procedure to evaluate the job performance of participants (Salgado and Moscoso, 2019; Rothstein, 1990) or, for instance, the leadership style of the supervisors (Heimann et al., 2020; Alonderiene and Majauskaite, 2016). In the former case, the supervisors rate the job performance of the collaborators, in the second case the latter evaluates the supervisors. Generally, one supervisor has many collaborators and, as for the previous examples, the expected heterogeneity among different raters is high. They are not used to these tasks and each of them might have a specific point of view of the co-worker's performance or leadership style.

In personnel-selection procedure, semi-structured interviews are very common (Tippins et al., 2018; Salgado and Moscoso, 1996). The raters (the selectors) are asked to evaluate the candidates' responses. Given the typical freedom of response of these interviews, the rating system might be quite complicated. It might result in a heterogeneous rating system. Even in front of the same *person-specification* selectors might differ in terms of severity or consistency (Van Trappen, 2022; Levin et al., 2005).

The rating process is also used in developing new psychological tests (Spoto et al., 2023; Nucci et al., 2021; Mokkink et al., 2010). More specifically, it is common to ask some experts for an opinion on the theoretical relevance of some items. It is often referred to as *content validity analysis*. In these contexts, raters evaluate the pertinence of a pool of items regarding some psychological construct. The frequent slight overlapping between different psychological variables makes this rating critical and not trivial. Different raters might have very different opinions on the relevance of an item in measuring a specific variable. In the case of psycho-diagnostic tests, every item refers to one or more particular symptoms. It is an operationalization of the latter. Occasionally, this correspondence is not straightforward and disputed. A careful consideration of raters heterogeneity might shed light on the reasons behind the disagreement. This is a case in which the consensus is supposed to be very high. Any divergence should be resolved during the item generation procedure.

When heterogeneity is analysed considering some covariate or social-demographics variable, it might be useful to unveil the varying role of the former on the rating process. Some variables might affect the rating behaviour of some raters. The latter might be more lenient or more severe concerning some characteristic of the object that is supposed to be not relevant. For instance, it is common in educational or work contexts that the rating is affected by some racial bias (Dee, 2005). Even if the race of the subjects must be an irrelevant factor for their rating, sometimes it plays a role (Childs and Wooten, 2023). In these cases, these biases need to be quantified and removed.

The considerations above highlight the wide use of the rating practice in psychological sciences. The focus on raters' heterogeneity is supported by all the issues exposed above. The natural differences in rating tasks between different raters, experts or naïve, need to be addressed both from a practical and a statistical perspective. The present work aims to provide a valuable solution under the latter. Through the general statistical framework proposed below, we hope to contribute to a better understanding of different kinds of rating processes.

1.3 Psychometrics solutions to raters' heterogeneity

All the points and the issues raised above have been differently addressed in psychometrics during the last decades [3]. The relevance of possible raters effects is well known and several statistical solutions have been proposed. Some of them aim for a better estimation of the inter-rater agreement, others focus on modelling raters' effects. As a result, plenty of inter-rater agreement indices a raters' effects models are now available (Martinková et al., 2023; Nelson et al., 2020; Gwet, 2008; Nelson and Edwards, 2008).

Ignoring the possible specific systematic bias of each rater might result in very biased and inaccurate estimates. From a practical point of view, this means a lack of information and a misleading interpretation of the results. It is widely known that each rater, although well trained, might have a small or large systematic bias in rating (Gisev et al., 2013) and they may also have some response style in this task; succinctly, they may differ in this task and the full lack of consideration of this might be very detrimental (Gelman et al., 2013).

The most common statistical framework used to address raters' heterogeneity is the hierarchical modelling (Martinková et al., 2023; Agresti, 2015; Nelson and Edwards, 2015; Gelman et al., 2013). The rating is decomposed into two parts: the effect due to the object (traditionally referred to as *the true score*) and a systematic bias due to the rater. When a continuous rating scale is used also the residual term is added and the error component is modelled. Each rater has his/her effect [4] and some covariates might be considered in the model.

In the parametric hierarchical models, raters heterogeneity is implied by the distributional assumptions on the raters' effect. In the univariate case, i.e. when the intercept is indexed by the rater (the varying intercept models), the varying effects are assumed to be independent and identically distributed according to a normal distribution (Gelman et al., 2013). For identifiability purposes, the mean is commonly fixed to zero. The interest here is in the variance parameter. It indicates to which extent the systematic biases of the raters differ from one another (Agresti, 2015). This measure of variability is crucial information related to the heterogeneity among raters. Since each

[3] The technical details of these models and of the related literature are broadly discussed in the next chapters.

[4] Also known as random effect or varying effect.

rater corresponds to a unique parameter (the rater effect, his/her varying intercept), the estimate of the variability between these quantities might be seen as a measure of heterogeneity (Nelson and Edwards, 2008).

Nonetheless, this is only a strong assumption which implies a very mild heterogeneity among raters. Their effects are assumed to be drawn from the same normal distribution. It has some limits. The distribution is assumed to be normal, consequently, possible asymmetry particularly shapes is not learned by the model. This might affect the accuracy of the estimates and implies a lack of information. Every difference between them is absorbed only by the variance parameter. Under this assumption, it is not possible to identify possible latent clusters of raters or raters' behaviours.

More flexible models might be specified to get more information about different kinds of heterogeneity. Considering the above-mentioned example, it might be interesting to discriminate cases in which raters' biases are uniformly distributed from cases in which they are strongly polarized. Under the common normal assumption on raters effects it is not possible to capture anything about the polarization of these parameters.

Another important point that might be addressed is the heteroscedasticity of the residual terms. When the rating scale is continuous it is common to model the errors or the residual terms as independent and identically distributed following a zero-mean normal distribution. The variance parameter is assumed to be the same for all the ratings across raters (it is the well-known homoschedastic assumption). It assumes that the raters are equally consistent in their ratings. It is reasonable to assume some differences among them also in this aspect. We might expect that some raters might be more consistent than others. The homoschedastic assumption might be relaxed and a different variance parameter might be estimated for each rater. It might borrow some important information about the consistency of each single rater.

The more flexible the rating models, the more information we might get (James et al., 2013). As for every statistical model a good trade-off might be achieved between parsimony and complexity. In this attempt, a well-reasoned choice has to be made in analysing rating data. Raters' evaluations are not to be considered, beyond any reasonable doubt, as a ground truth, nor as negligible opinions. Raters' opinions and their rating patterns have to be carefully considered and emerge from the data.

1.4 Conclusion

The points raised in the previous sections of this chapter highlight both the theoretical and the practical need to address the heterogeneity among raters. It might be due to unaddressed differences in their point of view, moreover, it might be a result of a non-homogeneous and poor training program. In this light, raters heterogeneity might bring some important information in terms of their evaluations, moreover taking into consideration raters' differences in their evaluations is an important step also from a practical perspective. The quantification of their heterogeneity might lead to a more accurate analysis of their judgments and a better awareness of the whole rating process.

The next chapters focus on the modelling and the quantification of raters' heterogeneity. Beyond the theoretical considerations raised about the nature of the raters' heterogeneity, this work aims to provide a more informative psychometric framework for the analysis of their evaluations and a more flexible tool for the quantification of raters' heterogeneity.

Chapter 2

Mixture polarization in inter-rater agreement analysis: a Bayesian nonparametric index

This Chapter is adapted from Mignemi, G., Calcagnì, A., Spoto, A., Manolopoulou, I. (in press). Mixture polarization in inter-rater agreement analysis: a Bayesian nonparametric index. Statistical Methods and Applications. http://arxiv.org/abs/2309.15076.

2.1 Introduction

In several contexts, decision-making relies heavily (or exclusively) on expert ratings, especially in situations where a direct quantification of the quality of an object or a subject is either impossible or unavailable. Examples include applicant selection procedures, grading of student assignments in education, or risk evaluation in emergencies, all of which rely on observational ratings made by experts. For ease of exposition, throughout this paper, we will refer to the evaluation of students' work in an educational context as the primary example. To ensure consistency across different teachers, harmonization of marking criteria is often used to improve inter-rater agreement and homogeneity (Gisev et al., 2013; Gwet, 2008); however, discrepancies between grades assigned by different teachers may persist (Bygren, 2020; Barneron et al., 2019; Makransky et al., 2019; Zupanc and Štrumbelj, 2018), reflecting each teacher's approach to evaluation. Therefore, statistical models that capture inter-rater agreement (or disagreement) can shed light on heterogeneity between teachers and aid the mark moderation process (Bygren, 2020; Barneron et al., 2019; Crimmins et al., 2016).

The specific context that we are considering in this chapter is the observational setting where a set of raters are evaluating different sets of items, commonly referred to as *subjects* in the rating context, out of a total population of items; these sets may be completely disjoint (i.e., each item is evaluated by exactly one rater). Each item might be associated with a set of covariates. Within a hierarchical statistical model, a common assumption is that raters (who may or may not be associated with covariates) may each be characterized through a latent variable capturing e.g. whether an evaluator is generous or how they assess different aspects of the work. In the simplest setting, in an evaluation context where there is no space for subjectivity, these latent variables will be identical for all raters, in the sense that their view of the item is identical and as a result, their evaluation style is assumed to be the same. However, it is well-known in many scientific fields, e.g., cognitive neuroscience (Barneron et al., 2019; Makransky et al., 2019; Briesch et al., 2014), statistics (Agresti, 2015; Gelman et al., 2013) and psychometrics (Bartoš et al., 2019; Nelson and Edwards, 2015; Hsiao et al., 2011), that individual variability in rating tasks (Wirtz, 2020) needs to be accounted for when aggregating or interpreting individual raters' recommendations.

Different teachers, even if well-trained, might have different views and divergent opinions. This possible heterogeneity among them might result in an unfair marking process.

Considering an observational setting, each rater $i = 1,\dots,I$ evaluates a different set of items $\mathscr{J}_i \in \mathscr{J}$, where $\mathscr{J}_i \cap \mathscr{J}_{i+1} = \emptyset$. It might be assumed that the elements of $\mathscr{J}$ (i.e., the items) are independent and identically distributed (i.i.d.), they are drawn from the same population. It is common to assume that also raters are drawn from the same population. It means that the parameters which account for the correlation among the ratings nested within the same rater (i.e., raters hierarchical effects) are i.i.d. This might be a strong assumption in very heterogeneous samples of raters. Existing works account for heterogeneity among raters through a latent variable within a mixed-effects model (Martinková et al., 2023; Bartoš et al., 2019; Nelson and Edwards, 2015, 2008). In other words, a regression model is used where the rating is modelled conditionally on covariates with a random effect that varies across raters. However, the distribution of the latent variable is typically assumed to be unimodal, and cannot capture, for instance, polarisation or clustering of rater types. In this chapter, we will show a proposal to extend these models to account for clustered variability between raters. Through a Bayesian approach, a Dirichlet process mixture prior is placed over the hierarchical rater effects in a linear model. This flexible prior naturally accommodates different clusters among raters (i.e., different distributions for the rater effects). A multiple-level model is specified in which observations (i.e., ratings) are nested within raters, and in turn, these are nested within clusters. These clusters reflect distinct groups of raters in terms of their decision-making and can be used to characterise the level of (dis)agreement. The level of multimodality (i.e., how separated the latent group densities are) quantifies the polarization of the latent groups. For instance, a large variance between teacher scores might be due to both the presence of two main divergent latent trends among them or to a high level of noise in their assessment (Koudenburg and Kashima, 2022). It is important to differentiate the two cases and quantify the group polarization both for theoretical and practical purposes. Differentiate systematic differences of opinion against high levels of noise might be needed (Koudenburg et al., 2021). They are two very different cases and much attention must be paid in distinguishing one another. The former is a case of high group polarization (Esteban and Ray, 1994): two different teachers' clusters emerge with a small within-cluster variance and a large variance between different clusters. In the second case, only one cluster emerges with a large variance. It might be argued that in the first case, even though the overall agreement might be quite low it is due to the presence of two different opinions (Tang et al., 2022). The raters that share the same opinion show a high agreement. Assuming the latent agreement among raters as the degree of latent similarity in rating, an index regarding the polarization of the different possible groups of raters might be informative (Koudenburg and Kashima, 2022; Tang et al., 2022; Koudenburg et al., 2021). In this chapter, we introduce a novel index to quantify the latent polarization among raters through the posterior distribution of the hierarchical effects (DiMaggio et al., 1996). This index naturally derives from the nonparametric model and overcomes some strong assumptions (e.g., the number of latent groups, the ratings ’distribution) of the previous indices (Koudenburg et al., 2021; Esteban and Ray, 1994). This nonparametric index, referred to as λ, is based on the shape of the posterior distribution of the hierarchical effects. It connects two different research lines: it relates the works on distribution polarization of opinions (Koudenburg and Kashima, 2022; Tang et al., 2022; Koudenburg et al., 2021) with those about the inter-rater agreement analysis (Martinková et al., 2023; Bartoš et al., 2019; Nelson and Edwards, 2015; Gisev et al., 2013; Gwet, 2008; Nelson and Edwards, 2008). The chapter proceeds as follows: Section 2.2 is devoted to the general psychometric framework, the key concepts of inter-rater agreement and inter-rater reliability are introduced; the statistical model is specified in Section 2.3, while the adopted Gibbs sampler in Section 2.4; the novel rater similarity index is described in Section 2.5; simulation studies are reported in Section 2.6, as an illustrative example, real data analysis is described in Section 2.8; it is followed by interim conclusion and future directions in Section A.

2.2 Existing work in inter-rater agreement and hierarchical effects models

Several methods and statistical models that aim to account for inter-rater variability have appeared in the literature (Nucci et al., 2021; Nelson and Edwards, 2015; Gwet, 2008; Cicchetti, 1976). Nucci et al. (2021) discussed the importance of raters’ heterogeneity in their evaluation and provided a novel method to improve the inter-rater agreement estimation. Models such as the Cultural Consensus Theory (Oravecz et al., 2014), which explores individuals’ shared cultural knowledge, have been proposed to capture unobserved agreement and similar trends in groups of raters (Dressler et al., 2015). Two related but different concepts have been introduced: *inter-rater agreement* and *inter-rater reliability*. The former refers to the extent to which different raters’ evaluations are concordant (i.e., they assign the same value to the same item), whereas the latter refers to the extent to which their evaluations consistently distinguish different items (Gisev et al., 2013). In other words, while the inter-rater agreement indices quantify the *observed* concordance, the inter-rater reliability indices aim to quantify the *consistency* of their evaluations (e.g., despite assigning different values, the distinction among the items is the same). The present chapter focuses on latent agreement intended as homogeneity in the evaluators' point of view (Tang et al., 2022; Esteban and Ray, 1994).

Several methods are available to quantify both inter-rater agreement and inter-rater reliability. Indices for pairs (Nelson and Edwards, 2008; McHugh, 2012) or multiple raters (Jang et al., 2018), for binary (Gwet, 2008), polytomous (Nelson and Edwards, 2015) or continuous (Liljequist et al., 2019) ratings are commonly used in different contexts. Recent developments using the framework of Hierarchical Linear Models (i.e., HLMs) provide a more accurate estimation of inter-rater reliability accounting for different sources of variability (Martinková et al., 2023).

Despite the popularity of work on this issue, less attention has been paid to possible latent similarities of the raters (Wirtz, 2020). From a psychometric point of view, it can be appealing to assess the extent to which different raters might be heterogeneous in their ratings (Martinková et al., 2023; Bartoš et al., 2019; Koudenburg et al., 2021; Casabianca et al., 2015; Nelson and Edwards, 2015; Gisev et al., 2013; DeCarlo, 2008; Gwet, 2008; Nelson and Edwards, 2008).

There are certain situations in which the subjective opinion of the raters is very informative; as a simple example, the type of teachers' training or experience can be thought of as latent states which affect a range of evaluations differently (Childs and Wooten, 2023; Barneron et al., 2019; Bonefeld and Dickhäuser, 2018; Dee, 2005). Sometimes the major interest is not in the mere consistency between raters but in their actual evaluation. For instance, in a selection process, the students' scores are very relevant for their admission (Zupanc and Štrumbelj, 2018). Even if a strict standardization of teachers' evaluation is not feasible, some statistical methods can tackle these issues. In all these contexts the assessment of uniformity among raters could be useful and would provide further information about the rating process.

To this aim, existing works (e.g.Martinková et al. 2023; Nelson and Edwards 2015; Casabianca et al. 2015; Hsiao et al. 2011; Cao et al. 2010; DeCarlo 2008), adopts a hierarchical approach where correlations between ratings are naturally captured through a hierarchical Bayesian model.

Each rater $i = 1,..,I$ is assumed to be rating a different set of items $\mathscr{J}_i \in \mathscr{J}$, $\mathscr{J}_i \cap \mathscr{J}_{i+1} = \emptyset$[1]. The rating y_{ij} of the item $j \in \mathscr{J}_i$ carried out by rater $i = 1,..,I$, is modelled as follows:

$$y_{ij} \quad = \quad \mathbf{x}'_{ij}\beta + \mathbf{z}'_{ij}\mathbf{u}_i + \varepsilon_{ij}, \;\; i = 1,..,I, \; j \in \mathscr{J}_i. \tag{2.1}$$

Here $\mathbf{x}_{ij}$ and $\mathbf{z}_{ij}$ are, respectively, $1 \times p$ and $1 \times q$ vectors of distinct explanatory variables of rating y_{ij}; β is a $p \times 1$ vector of non-varying effects and $\mathbf{u}_i$ is a $q \times 1$ vector of the hierarchical effects of rater i.

In the standard HLM formulation, the following distribution is specified for the rater effects:

$$\mathbf{u}_i \quad \sim \quad N_q(\mathbf{0},\mathbf{\Sigma}), \quad i = 1,..,I.$$

Here $N_q(\cdot)$ stands for a q-variate normal distribution; Here $\mathbf{0}$ is a $q \times 1$ zero vector and $\mathbf{\Sigma}$ is a $q \times q$ positive semi-definite covariance matrix. For the hierarchical normal linear model $\varepsilon_{ij} \sim N(0, \sigma^2_\varepsilon)$, with $\mathbf{u}_i$ and ε_{ij} typically assumed independent. The distribution of each vector-valued hierarchical effects $\mathbf{u}_i$ is then assumed to follow some distribution and capture variability across different raters. In the above-mentioned example, y_{ij} is the score given to student $j \in \mathscr{J}_i$'s essay by teacher i. Since an observational approach is adopted (i.e., each rater rates a different set of students), the effect of the student is not identifiable (each student is rated only by one rater). Assuming that students' effects are i.i.d., their variance is added to that of the residuals. In the univariate case (i.e., when $z_i = 1$, varying intercept model) and ignoring any covariate for the effect β [2], the relevance of the raters' effect $u_i \sim N(0, \sigma^2_u)$, where $\sigma^2_u > 0$ is the variance, might be quantified through the *intraclass correlation coefficient* (i.e., ICC):

$$ICC = \frac{\sigma^2_u}{\sigma^2_u + \sigma^2_\varepsilon}.$$

The ICC is the ratio between the variance of the raters' effect and the total variability of the model, i.e., the proportion of variance of the score due to the teacher, which reflects the correlation of two

[1] The multiple rating case (i.e., raters rate the same set of items, $\mathscr{J}_i = \mathscr{J}$, $i = 1,\dots,I$) is addressed in Appendix.

[2] It will be consistent also in the next chapter

ratings given by the same rater. Smaller values of ICC indicate a small effect of the rater on the student's score. When some covariate is included in the model, the *ICC* formula is different if one is interested in accounting also for the variance due to these covariates. This case will not be considered in this thesis as the focus here is on the proportion between the variance of the effect u_i and that of the residuals.

2.3 Dirichlet Process Mixture and Hierarchical Effects

The HLM assumption regarding the distribution of the hierarchical effects is crucial in characterising different possible clusters or latent patterns of heterogeneity among raters (Dorazio, 2009). The common Gaussian assumption for the distribution of these effects may obscure the skewness and multimodality present in the data. A more flexible specification of the hierarchical effects distribution can help capture more complex patterns of variability. Models that account for skew-normal (Lin and Lee, 2008), skew-normal-Cauchy (Kahrari et al., 2019), multivariate t (Wang and Lin, 2014), extreme values (McCulloch and Neuhaus, 2021) effects distributions have been proposed (Schielzeth et al., 2020). Nevertheless, they poorly account for the possible presence of multimodality in those distributions. In this regard, a mixture distribution has been proposed as a potential solution (Heinzl and Tutz, 2013; Kyung et al., 2011; Kim et al., 2006). Each mode can then correspond to a cluster with a similar pattern (e.g., the same deviation from the population mean). Several works have explored this issue in the past two decades (Villarroel et al., 2009; Tutz and Oelker, 2017). For instance, Verbeke and Lesaffre (1996) proposed standard normal mixture distributions for the hierarchical effects [3]. James and Sugar (2003) explored this approach in the context of functional data. De la Cruz-Mesia and Marshall (2006) proposed a mixture distribution for non-linear hierarchical effects in modelling continuous-time autoregressive errors [4]. A heteroscedastic normal mixture model in the hierarchical effects distribution was considered in linear (Komárek et al., 2010) and generalized hierarchical linear (Komárek and Komárková, 2013) models. Despite the breadth of specifications for the mixture model, in all the aforementioned models,

[3] Hierarchical models in which the observations are modelled to follow a mixture (i.e., a collection) of normal distributions. This implies that each observation is assumed to follow different distribution with different probabilities.

[4] Models in which a temporal dependency is specified between errors

the number of mixture components needs to be specified. Although this may not be a critical assumption in certain contexts, it may be questionable or detrimental in settings with a lack of a priori information on the level of multimodality, especially in cases where the characterisation of multimodality is of direct interest.

When the number of components of the mixture is unknown, a Dirichlet Process Mixture (hereafter DPM) for the hierarchical effects is a natural extension (Gill and Casella, 2009; Navarro et al., 2006; Verbeke and Lesaffre, 1996). This nonparametric extension allows the model to capture an unknown marginal distribution of the hierarchical effects through the Dirichlet Process (Antoniak, 1974; Ferguson, 1973). Modelling the hierarchical effect $\mathbf{u}_i$ as an infinite mixture of some distribution family (e.g., Normal) enables the model to account for possible multimodality without specifying the number of mixture components. Some existing works adopted this nonparametric approach and pose a DPM prior over the hierarchical effects (e.g., Heinzl and Tutz (2013); Heinzl et al. (2012); Kyung et al. (2011)).

The HLM of Equation (2.1) is then specified in the same way as before through:

$$y_{ij} \quad = \quad \mathbf{x}_{ij}'\beta + \mathbf{z}_i'\mathbf{u}_i + \varepsilon_{ij}, \qquad i = 1,..,I, \quad j \in \mathscr{J}_i.$$

The following hierarchical prior distribution [5] is placed over the raters' effects:

$$\begin{aligned} \mathbf{u}_i|\mu_i, \mathbf{Q}_i &\sim N_q(\mu_i, \mathbf{Q}_i) \\ \mu_i, \mathbf{Q}_i|G &\overset{iid}{\sim} G \\ G &\sim DP(\alpha, G_0) \end{aligned}$$

where μ_i and $\mathbf{Q}_i$ are, respectively, the $q \times 1$ a location parameter vector [6] and the $q \times q$ positive semi-definite covariance matrix [7] for the hierarchical effects $\mathbf{u}_i$ of rater $i = 1, \ldots, I$. Here $\varepsilon_{ij} \sim$

[5] To make more into account the uncertainty about the rating process, we specify a "prior over the prior", that is a hierarchical prior, over the raters' effects $\mathbf{u}_i$.

[6] The vector containing the respective means of the effects.

[7] This is a condition for any covariance matrix. it might be seen as a generalization of the fact that the variance must be equal or greater than zero.

$N(0,\sigma_\varepsilon^2)$, $i=1,..,I$, $j \in \mathscr{J}_i$; $\mathbf{u}_i$ and ε_{ij} are assumed independent as before.

This hierarchical prior is addressed and broadly discussed in the next section, the mathematical details are provided below. Even though, from an intuitive point of view, *DP* might be seen as a "distribution of distributions" and G is the sampled distribution. The parameter α is proportional to the variance of *DP* and G_0 is the expected value of *DP*. For these reasons, they are respectively referred to as *precision parameter* and *base measure*.

These concepts are addressed more rigorously in the next section and appendix A.

2.3.1 DPM as a generative process for the hierarchical effects

Here, $DP(\alpha, G_0)$ is a DPM with $\alpha > 0$ *precision parameter* and *base measure* G_0. These specify the mixing distribution G (Heinzl and Tutz, 2013), so that each realization of G is almost surely a discrete probability measure on the measurable space $(\Omega, \mathscr{F})$ (Ferguson, 1973). Where Ω is a separable space and $\mathscr{F}$ is a σ-field on Ω [8]. Thus, since the DPM is a discrete generative process with a non-zero probability of ties, some of the realizations might be identical to each other with probability determined by the precision parameter α. Therefore, specifying this hierarchical model on the components' location parameters μ induces a clustering in the hierarchical effects (i.e., the raters) (Kyung et al., 2011); hierarchical effects belonging to the same c-th cluster with location parameter μ_c are then independent and identically distributed. In other words, in the context of the HLM, the DPM specifies the component-specific location parameter μ_c, so that each rater has each has their own unique hierarchical effects value $\mathbf{u}_i$ (Heinzl and Tutz, 2013).

The DPM is a generative process commonly used in conjunction with a parametric family of distributions (e.g., Normal, Poisson), and the base measure parameter G_0 denotes this specified distribution. Thus, for any element A_n, $n=1,..,N$, of $\mathscr{A}$, a finite measurable partition of Ω,

$$(G(A_1), G(A_2), ..., G(A_N)) \sim Dir(\alpha G_0(A_1), \alpha G_0(A_2), ..., \alpha G_0(A_N))$$

where $Dir(\cdot)$ stands for the Dirichlet distribution, and G_0 defines the expectation of G[9], there-

[8] To put it more simply, Ω is the sample space and $\mathscr{F}$ is the event space.

[9] Considering the partition (A, A^c) of Ω and thus that $G(A) \sim Be(\alpha G_0(A), \alpha G_0(A^c))$ the expectation of $G(A)$ is defined as: $\mathbb{E}[G(A)] = \frac{\alpha G_0(A)}{\alpha G_0(A)+\alpha G_0(A^c)} = \frac{\alpha G_0(A)}{\alpha(G_0(A)+G_0(A^c))} = G_0(A)$.

fore they have the same support. The parameter α, a multiplicative constant of the vector-valued Dirichlet parameter, determines the probability of a new realization of the process to be different from the previous ones (Blackwell and MacQueen, 1973). In other words, it governs the probability that the DPM generates a new cluster. Formally, the generative property of the DPM is that, for $i = 1, ..., I$, with I being for instance the total number of raters:

$$G \sim DP(\alpha, G_0), \quad \mu_i | G \sim G$$

the probability that the new I-th realization μ_I of G assumes a different value than the previous ones is described by the so-called *Pólya Urn Model*:

$$\mu_I | \mu_1, \mu_2, ..., \mu_{I-1}, \alpha \sim \tfrac{\alpha}{\alpha+I-1} G_0 + \tfrac{1}{\alpha+I-1} \sum_{c=1}^{C} r_c$$

with $C \in \mathbb{N}$ being the number of already observed distinct clusters among the realizations of G (i.e., the number of the different values of μ already observed, in other words, the number of clusters) and r_c counts the elements in the c-th cluster. Since G is a discrete probability measure, the C clusters represent different point masses (or different sets of point masses in the multivariate case) and r_c is the frequency of each of them. Considering the conditional distribution of μ_I as a mixture distribution, the probability that μ_I is a new point mass sampled from G_0 is proportional to α, whereas the probability that it is equal to the already observed c-th point mass is proportional to r_c. This notation highlights that α is proportional to the probability that a new (not already observed) value of μ_I (i.e., a new point mass, a new cluster) is sampled from G_0. Given the same number of observations, larger values of α correspond to a larger probability of observing a new cluster (i.e., a not already observed value for μ_I).

To this regard, Sethuraman (1994) described a *stick-breaking construction* of the DP [10]. In this formulation, G is equivalent to:

$$G = \sum_{c=1}^{\infty} \pi_c \delta_{\mu_c}$$

where δ is the Dirac measure on μ_c and $\mu_c \overset{iid}{\sim} G_0$ is assumed. The weights $\{\pi_c\}_{c=1}^{\infty}$ of the infinite mixture result from the stick-breaking procedure as follows:

[10] Other stick-breaking representations might be used, e.g., Rigon and Durante (2021); Stefanucci and Canale (2021); Rodriguez and Dunson (2011).

$$\pi_c = \nu_c \prod_{l<c}(1-v_l)$$

$$v_c \overset{iid}{\sim} Be(1,\alpha)$$

with $Be(\cdot)$ indicating the Beta distribution and $\{\nu_c\}_{c=1}^{\infty}$ being reparameterized weights. It is even more explicit in this construction that the random measure G is a mixture of point masses. The distribution of the random weights π (i.e., the probability of different allocation to the clusters) is governed through the stick-breaking process by the precision parameter α. Further details are given in the Appendix.

In practice, one of the established approximations of the stick-breaking process is to truncate the infinite number of components to a large, finite value:

$$G = \sum_{c=1}^{R} \pi_c \delta_{\mu_c}$$

for large enough values of R (Tutz and Oelker, 2017; Gelman et al., 2013).

In summary, using the above-introduced notation, the hierarchical effects distribution considering a stick-breaking construction of the DPM might be then specified as follows:

$$\begin{array}{rcl}
\mathbf{u}_i|\mu,\mathbf{Q}, & \overset{iid}{\sim} & \sum_{c=1}^{R} \pi_c N_q(\mu_c, \mathbf{Q}_c), \quad i = 1,\ldots,I \\
\mu_c, \mathbf{Q}_c & \overset{iid}{\sim} & G_0 \\
\pi_c & = & v_c \prod_{l<c}(1-v_l), \text{ where} \\
v_c & \overset{iid}{\sim} & Be(1,\alpha), \; c = 1\ldots,R
\end{array}$$

With this nonparametric model specification, latent common tendencies among raters might emerge through the components of the model (Heinzl and Tutz, 2013; Heinzl et al., 2012; Kyung et al., 2011). The Bayesian approach allows us to characterize the shape of the distribution of the rater effects, as well as to explore the effect of uncertainty on these (Gelman et al., 2013). For example, in an applied context, strict vs. accommodating are very common latent states that drive students' essay grading process (Zupanc and Štrumbelj, 2018; Briesch et al., 2014; Dee, 2005).

2.4 Prior distributions and estimation procedure

The DPM mixture model has been well studied in the literature in a variety of different settings, especially within Bayesian inference (Canale and Prünster, 2017; Müller et al., 2015). Several sampling schemes have been proposed both in the Bayesian context (e.g., Canale and Dunson (2011); Dahlin et al. (2016); Kyung et al. (2011)) and in the frequentist one (e.g., Tutz and Oelker (2017)). Within the Bayesian framework, Gibbs sampling (Dahlin et al., 2016), slice sampler (Kyung et al., 2011; Walker, 2007), Sequential Monte Carlo algorithms (Ulker et al., 2010), split-merge algorithms (Bouchard-Côté et al., 2017), have been proposed among others.
In this chapter, the model specification permits the use of conjugate priors, so that a blocked Gibbs sampling can be used (Heinzl and Tutz, 2013; Heinzl et al., 2012; Kyung et al., 2011)., with details shown below.

2.4.1 Prior specification

Several of the parameters in the model have conjugate prior distributions which allow easier computation.

- For the effects β the following hierarchical prior is assigned:

$$\begin{aligned}
\beta|\mathbf{b}_\beta, \mathbf{B}_\beta &\sim N_p(\mathbf{b}_\beta, \mathbf{B}_\beta) \\
\mathbf{b}_\beta &\sim N_p(\mathbf{b}_0, \mathbf{S}_0) \\
\mathbf{B}_\beta &= diag(\sigma^2_{\beta_1}, ..., \sigma^2_{\beta_p}) \\
\sigma^2_{\beta_m} &\overset{iid}{\sim} IG(a_{\beta_0}, b_{\beta_0})
\end{aligned}$$

for $m = 1, ..., p$, where p is the number of covariates associated with the effects β. Here, $IG(\cdot)$ stands for inverse-gamma with shape parameters $a_{\beta_0} > 0$ and rate parameters $b_{\beta_0} > 0$. Where $\mathbf{b}_0$ and $\mathbf{S}_0$ are, respectively, the $p \times 1$ vector of location parameters and the $p \times p$ positive semi-definite covariance matrix of $\mathbf{b}_\beta$ (i.e, the location parameter vector of the non-varying effect β); $\mathbf{b}_\beta$ and $\mathbf{S}_\beta$ are, respectively, the $p \times 1$ location parameter and the $p \times p$ positive semi-definite covariance matrix for β (i.e., the non-varying effect). The set $\{\mathbf{b}_0, \mathbf{S}_0, a_{\beta_0}, b_{\beta_0}\}$ of the hyperparameters are specified by the user. A diagonal matrix is suggested for $\mathbf{S}_0$ as showed by Heinzl et al. (2012).

- A diagonal structure for the $q \times q$ prior covariance matrix $\mathbf{Q}_r$ for the hierarchical effects is specified as follows for each mixture component $r = 1, ..., R$ and each related covariate $d = 1, ..., q$:

$$\mathbf{Q}_r \quad = \quad diag(\sigma^2_{Q_{1r}}, ..., \sigma^2_{Q_{qr}})$$

For the base measure G_0[11] and the precision parameter α of the DP mixture model the following priors are specified:

$$\begin{aligned} G_0 &= N_q(\mu_0, \mathbf{D}_0) \times IG(a_{Q_0}, b_{Q_0})^q \\ \mu_0 &\sim N_q(\mathbf{m}_0, \mathbf{W}_0) \\ \mathbf{D}_0 &= diag(\sigma^2_{D_{01}}, ..., \sigma^2_{D_{0q}}) \\ \sigma^2_{D_{0d}} &\sim IG(a_{D_0}, b_{D_0}) \\ \alpha &\sim Ga(c_0^\alpha, C_0^\alpha) \end{aligned}$$

for $d = 1, ..., q$, where q is the number of covariates associated with the hierarchical effects $\mathbf{u}_i$, and for $a_{D_0}, a_{Q_0} > 0$ and $b_{Q_0}, b_{D_0} > 0$. Here $Ga(\cdot)$ stands for Gamma distribution with $c_0^\alpha > 0$ and $C_0^\alpha > 0$ respectively the shape and the rate parameters. Where $\mathbf{m}_0$ and $\mathbf{W}_0$ are, respectively, the $q \times 1$ location parameter vector and the $q \times q$ positive semi-definite covariance matrix of μ_0 (i.e. the location parameter of the base measure G_0); μ_0 and $\mathbf{D}_0$ are, respectively, the $q \times 1$ location parameter vector and the $q \times q$ positive semi-definite

[11] Assuming independence between the location and the scale parameters of each mixture component, and between all the scale parameters for each covariate $d = 1, \ldots, q$, G_0 is then the product of the q-variate normal and the q inverse gamma distributions.

covariance matrix of the base measure G_0. The set $\{a_{Q_0}, b_{Q_0}, \mathbf{m}_0, \mathbf{W}_0, a_{D_0}, b_{D_0}, c_0^{\alpha}, C_0^{\alpha}\}$ of the hyperparameters need to be fixed. A diagonal structure is suggested for $\mathbf{W}_0$ as above.

- The following prior is assigned to the noise variance:

$$\sigma_{\varepsilon}^2 \sim Ga(a_{\varepsilon}, b_{\varepsilon})$$

with $a_{\varepsilon} > 0$ and $b_{\varepsilon} > 0$ hyperparameters fixed by the user as well.

2.4.2 Posterior sampling

Since most of the parameters in the model have conjugate prior distributions, a blocked Gibbs sampling algorithm was used for the posterior sampling (Ishwaran and James, 2001) .

The parameter vector for the model is $\theta = \{\beta, \mathbf{b}_{\beta}, \mathbf{B}_{\beta}, \mu_0, \mathbf{D}_0, \mathbf{Q}, \sigma_{\varepsilon}, \alpha, \pi, \mathbf{c}\}$ which is updated at each state of the Markov chain of the Gibbs sampling. Here $\mathbf{c} = (c_1, \ldots, c_I)$ is the allocation parameter of the raters to the clusters. At each step of the sampling scheme (which is reported below and the steps are enumerated), a component of the parameter vector is updated. That is, each parameter is sampled from its conditional posterior distribution at a specific step given the other parameters. The notation is consistent with that used in the previous sections.

Further details on the following sampling are given in the Appendix. The closed-form posteriors are as follows.

1. Update parameters referring to effects β (which is the non-varying, ”fixed”, effect):

$$\beta | \mathbf{b}_{\beta}, \mathbf{B}_{\beta}, \mathbf{u}, \sigma_{\varepsilon}, \mathbf{y} \sim N_p(\mathbf{b}_{\beta}^*, \mathbf{B}_{\beta}^*)$$

For each covariate $m = 1, ..., p$ associated with a non-varying effect β_m,

$$\begin{aligned} b_{\beta_m}|\sigma^2_{\beta_m}, \beta_m &\sim N\left(\left(\frac{1}{\sigma^2_{\beta_m}} + \frac{1}{s^2_{0_m}}\right)^{-1}\left(\frac{\beta_m}{\sigma^2_{\beta_m}} + \frac{m_{\beta_m}}{s^2_{0_m}}\right), \left(\frac{1}{\sigma^2_{\beta_m}} + \frac{1}{s^2_{0_m}}\right)^{-1}\right) \\ \sigma^2_{\beta_m}|b_{\beta_m}, \beta_m &\sim IG\left(a_{\beta_0} + \frac{1}{2}, b_{\beta_0}\frac{1}{2}(\beta_m - b_{\beta_m})^2\right) \end{aligned}$$

2. Update parameters referring to hierarchical effects (that are the parameters related to the raters):

- For each rater $i = 1, ..., I$:

$$\mathbf{u}_i|\mu_{c_i}, \mu_0, \mathbf{Q}_0, \beta, \sigma_\varepsilon, \mathbf{y}_i \sim N_q(\mu^*_{c_i}, \mathbf{Q}^*_{c_i}),$$

  where μ_{c_i} is the location parameter vector (i.e., the vector of the means) of the cluster where the i-th rater is allocated.

- For each component $r = 1, ..., R$ of the truncated mixture (the mixture implied by the stick-breaking construction):

  - If $\nexists i : c_i = r$ (if no rater is currently allocated into cluster r), for each covariate $d = 1, ..., q$ associated to a hierarchical effect (independently):

$$\begin{aligned} \mu_r|\mu_0, \mathbf{D}_0 &\sim N_q(\mu_0, \mathbf{D}_0) \\ \sigma^2_{Q_{dr}} &\sim IG(a_{Q_0}, b_{Q_0}) \end{aligned}$$

  - If $\exists i : c_i = r$ (if at least one rater is assigned to component r), for each covariate

$d = 1, ..., q$ associated to a hierarchical effect (independently):

$$\begin{aligned}
\mu_{r_d}|\sigma^2_{Q_m}, \mu_{0_r}, \sigma^2_{D_{0_m}}, \mathbf{u}, \mathbf{c} &\sim N(\mu^*_{0_r}, \sigma^{2*}_{D_{0_m}}) \\
\sigma^2_{Q_{dr}}|\mu, \mathbf{u} &\sim IG\left(a^*_{Q_0}, b^*_{Q_0}\right)
\end{aligned}$$

Essentially, at each iteration t, if the r-th cluster is empty the component location parameters μ_r are sampled from the prior as suggested by (Gelman et al., 2013), otherwise they are drawn from the above-mentioned closed-form posterior.

- Each rater $i = 1, ..., I$ is re-allocated into a cluster:

$$c_i|\pi, \mu, \mathbf{Q}, \mathbf{u}_i \sim Cat(\omega^*_i)$$

where $Cat(\cdot)$ stands for Categorical distribution, and ω^*_i is reported in the Appendix. A truncated approximation for the DPM mixture model was used (Gelman et al., 2013; Heinzl et al., 2012) for a large value of R. The stick-breaking construction was used to generate the mixture weights $\pi_{1:R}$.

- For each component $r = 1, ..., R-1$:

$$v_r|\phi, \alpha \sim Be\left(1 + c_r, \alpha + \sum_{l=c+1}^{R} r_l\right)$$

and $v_R = 1$ for the last cluster. Here c_r is the number of raters assigned to the cluster r, and r_l is the number of raters assigned to the cluster l; v_r is the reparameterized weight of cluster r.

- The precision parameter is updated as follows:

$$\alpha|v_1, ..., v_{R-1} \sim Ga\left(R - 1 + a_\alpha, b_\alpha - \sum_{c=1}^{R-1} ln(1 - v_r)\right)$$

- For each covariate $d = 1, ..., q$ associated with a hierarchical effect (the effect associated with the raters) the base measure parameters are updated:

$$\mu_{0_d}|\sigma^2_{D_{0_d}},\mu \quad \sim \quad N\left(\left(\frac{I}{\sigma^2_{D_{0_d}}}+\frac{1}{\sigma^2_{W_{0_d}}}\right)^{-1}\left(\frac{I}{\sigma^2_{D_{0_d}}}\overline{\mu}_d+\frac{m_{0_d}}{\sigma^2_{W_{0_d}}}\right),\left(\frac{I}{\sigma^2_{D_{0_d}}}+\frac{1}{\sigma^2_{W_{0_d}}}\right)^{-1}\right)$$

  where $\overline{\mu}_d$ is the mean of the location parameters (i.e., the mean of the means) related to the d-th covariate over all the clusters.

$$\sigma^2_{D_{0_m}}|\mu_{0_m},\mu \quad \sim \quad IG\left(a_{D_0}+\frac{I}{2}, b_{D_0}+\frac{1}{2}\sum_{i=1}^{I}(\mu_{i_m}-\mu_{0_m})^2\right)$$

3. Update the error variance (that is the variance of the residuals):

$$\sigma^2_{\varepsilon}|\beta,\mathbf{u},\mathbf{y} \quad \sim \quad IG\left(a_{\varepsilon}+\frac{1}{2}I|\mathscr{J}|, b_{\varepsilon}+\frac{1}{2}\sum_{i=1}^{I}\sum_{j\in\mathscr{J}_i}\left(y_{ij}-\mathbf{x}'_{ij}\beta-\mathbf{z}'_{ij}\mathbf{u}_i\right)^2\right).$$

Here $|\mathscr{J}|$ is the cardinality of the set of all the rated items $\mathscr{J}$, it equals the number of observations.

2.5 The nonparametric λ index

The marginal posterior distribution of the hierarchical effects in the model outlined above captures information about the polarization or disagreement among raters (on the assumption that the model captures the data adequately). The ICC might adequately quantify inter-rater variability if the normal distributional assumption of the rater hierarchical effect holds. Two assumptions are made when computing the standard ICC considering a normal distributed hierarchical effect. Firstly, the raters are sampled from the same population. Secondly, possible different latent trends among raters are not interesting or eventually regarded as disagreement ratings. This might be a good first approximation of the rating process. Nevertheless, when more detailed considerations are needed, or subtle heterogeneity among raters is expected, the standard ICC might be less informative and inaccurate. Besides the latter issue, further information about the shape of the posterior might be quantified. For instance, in the presence of a bimodal hierarchical effects distribution with two very distant modes (for example, when opinions are polarised), considering the posterior distribution of σ_u as an index of variability among raters might be misleading.

Several indexes have been proposed to quantify group opinion polarization (e.g., (Tang et al., 2022; Koudenburg and Kashima, 2022; Koudenburg et al., 2021; Esteban and Ray, 1994)) and to measure distribution bimodality (e.g., Ashman's D (Forchheimer et al., 2015) or the bimodal separation index (Zhang et al., 2003)). The strong assumptions behind their use limit them to be valid options only in the parametric context or when the number of clusters is known. A model-based nonparametric index is here proposed to overcome these limitations.

Whether the density around the two modes is very high might be interesting to use this information to quantify the latent agreement among raters. The two modes might indicate the two major different opinions not considered in the model. In this regard, an exploration of the shape of the marginal posterior density might be a fruitful option. The univariate case (e.g., a varying intercept specification) is considered.

To this end, the full estimated distribution of $\mathbf{u}$ resulting from the model might be useful. At each iteration t, the density of $\mathbf{u}$ is given by the corresponding mixture model given the parameters at iteration t. Following the formulation of (Gelman et al., 2013) , the set of modes and antimodes (i.e., the lowest frequent value between two modes) is identified. When the distribution of $\mathbf{u}$ is multimodal, the latent polarization (disagreement) λ is then defined as the log ratio between the mean density of the modes and that of the anti-modes, it is zero when it is unimodal:

$$\lambda = \begin{cases} \log\left(\dfrac{\frac{1}{M}\sum\limits_{m=1}^{M} f_u(\gamma_m)}{\frac{1}{M-1}\sum\limits_{m=1}^{M-1} f_u(\zeta_m)}\right), & \text{if} \quad M > 1 \\ \\ 0, & \text{otherwise.} \end{cases}$$

Where M is the number of modes γ_m, $m = 1,\ldots,M$ and the number of antimodes ζ_m, $m = 1,\ldots,M-1$ of the density of $\mathbf{u}$; $f_u(\cdot)$ denotes the density at a specific point. Larger values of λ indicate a strongly multimodal distribution of the hierarchical effects, whereas smaller values are evidence of weak multimodality, thus the estimated hierarchical effects are less concentrated. As it is shown in Figure 2.1 larger values of λ indicate distribution polarization, whereas smaller values indicate a less concentrated and more spread density distribution. The λ index is strongly affected by both the location and scale parameters of the mixture components. For this reason, it might be very informative in the presence of multimodal distributions. Assuming such a raters' group polarization as a result of low latent agreement among raters, the λ index might be a useful diagnostic tool.

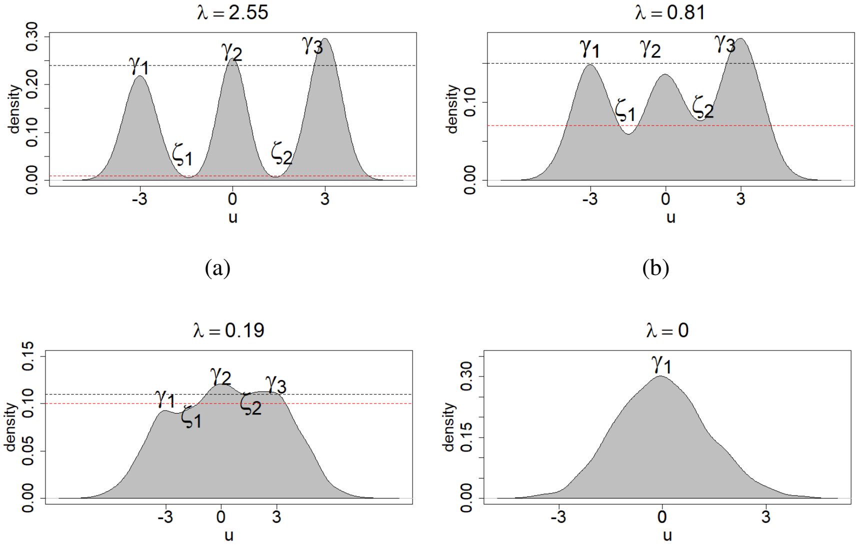
2.6 Simulation studies

The following simulations aim to describe how the values of λ vary across different polarization settings. The first simulation investigates the role of the precision parameter α and the variance of the mixture components in determining the values of λ. The second one shows the complementary role of λ in the inter-rater agreement analysis and how this index varies across different settings.

2.6.1 Simulation 1: DPM and λ

Simulation setting

The first simulation study explores the role that the precision parameter of the Dirichlet Process and the variance of the components have in determining the values of the log-density index λ. For simplicity purposes, the mixture components are assumed to have the same variance Q in this simulation, so the component subscription will be omitted. The objective is to study the effect of α and Q, on λ conditional on all the other variables. Since the former has a crucial role in the determination of the point masses of G, and thus the concentration of its realizations, an inverse relation between α and λ is expected if Q is fixed. Likewise, an inverse relation between Q and λ



(a)

(b)

(c)

(d)

Figure 2.1: Different values of λ indicate different polarization levels. Three different values of λ were computed for three different mixture distributions, respectively. The realizations of these distributions are here referred to as u. Black dotted lines indicate the mean mode density and red dotted lines indicate the mean antimode density. (a) High polarization: the mixture components are highly and separate, and the mean density values of the modes are far larger than the mean density value of the antimodes; the log-density ratio between these two quantities is $\lambda = 2.55$ (b) Medium polarization: the mixture components are separated, but the mean density values of the modes are closer to the mean density value of the antimodes; the log-density ratio between these two quantities is $\lambda = 0.81$ (c) Low polarization: the mixture components are not separated, the mean density values of the modes are very close to the mean density value of the antimodes; the log-density ratio between these two quantities is $\lambda = 0.19$. (d) No polarization: the mixture distribution has only one mode (i.e., γ_1) and $\lambda = 0$ since the number of modes is not greater than one.

is expected if α is fixed. The index λ is interpretable as an index of the sharpness of the modes. For this reason, both the precision parameter of the DPM and the variance of its components are expected to affect λ. Controlling for Q (i.e., keeping it fixed), the expected relation is: the smaller α, i.e. the precision of the DPM mixture, the larger λ, i.e. the relative density around the modes; controlling for α (i.e., keeping it fixed), the expected relation is: the smaller Q, i.e. the variance of the components of the DPM, the larger λ. The parameters of the base measures G_0 have a non-negligible role in determining λ, so in this section focus is devoted to the relation between the precision parameter α, the mixture components variance Q and the index λ. Indeed, in all the study simulations the values of the other parameters involved in the DPM have been kept fixed across the scenarios.

Data generating process The experimental design is as follows. For 4 different values of $\alpha = (0.1, 1, 5, 20)$ and 2 different values of $Q = (0.1, 1.5)$ a set of independent observations $u = 1, \ldots, n$ are drawn from the following DPM:

$$\begin{aligned}
u_i|\mu, Q, &\overset{iid}{\sim} \sum_{c=1}^{R} \pi_c N(\mu_c, Q), \quad i = 1, \ldots, n \\
\mu_c &\overset{iid}{\sim} G_0 \\
\pi_c &= v_c \prod_{l<c} (1 - v_l), \text{ where} \\
v_c &\overset{iid}{\sim} Be(1, \alpha), \; c = 1 \ldots, R.
\end{aligned}$$

Where μ_c and π_c are the location parameter and the mixing proportion of the component c, respectively; G_0 is the base measure; and v_c is the parameter of the stick-breaking. Following the above-mentioned truncated stick-breaking construction, here R is the maximum number of observable clusters. Across the eight scenarios, the following quantities are assigned: the number of observations $n = 500$, the maximum number of clusters $R = 50$, and the base measure G_0 :

$U(-6,6)$. Here, $U(\cdot)$ stands for uniform distribution. The use of these distributions in the present experimental context aims to highlight the effect of different values of α and Q on λ in a more evident and interpretable manner.

Results

As shown in Tables 2.1 and 2.2 as α increases, and so the number of point masses of G increases as well, λ decreases. The density of the observations $u = 1,\ldots,n$ is concentrated around a few point masses (few modes) for lower values of α and is spread out the larger. Note also the change of density of the antimodes. As expected, it is proportional to the precision parameter in a positive fashion. As the observations are more spread as α increases, there are fewer intervals in the support with relatively small density: λ index decreases at larger values of α (column-wise Table 2.1 and Table2.2). A similar proportional relation is observed between the variance of the mixture components Q and λ when α is kept fixed (row-wise Table 2.1 and Table2.2). Smaller values of both α and Q result in a high polarized distribution of $u = 1,\ldots,n$ and correspond to larger values of λ. Whereas larger values of both α and Q result in a low polarized distribution and correspond to smaller values of λ. It is an index of how spread the density is over the support of the hierarchical effects.

From an interpretative point of view, λ indicates the degree of overlap between the infinitely many clusters. It might be informative of the separation between them. Since this quantification is based on a nonparametric density, λ is not directly related to the number of groups or the cluster location. It indicates the degree to which the independent observations drawn from a DPM overlap; the variance of the cluster Q also plays a crucial role. The index thus quantifies the combined effect of the parameters to assess the extent to which possible different opinions (i.e., the modes) might be strongly shared among the raters (i.e., the modes are sharp pick of density). In this regard, λ is a polarization index in the presence of heterogeneity. The higher the polarization levels, the larger the values of the index. The practical interpretation and the operational decisions must be guided by the field of application.

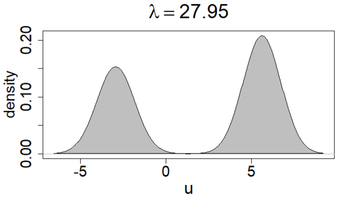
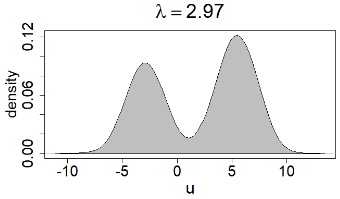
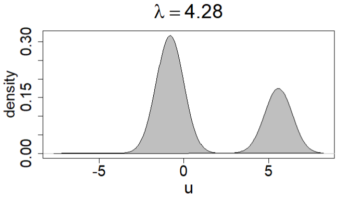
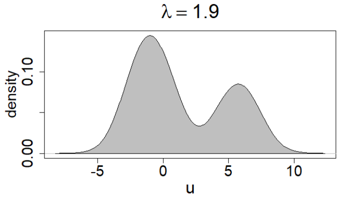
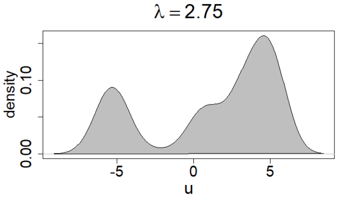
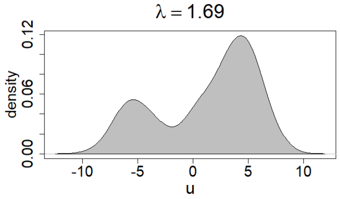
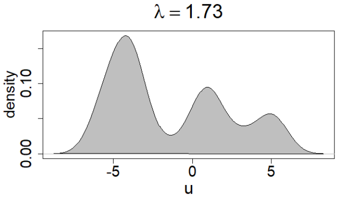
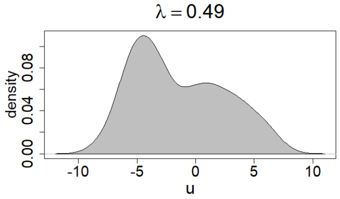
	$Q_c = 0.1$	$Q_c = 1.5$
$\alpha = 0.1$		
$\alpha = 1$		
$\alpha = 5$		
$\alpha = 20$		

Table 2.1: The eight scenarios correspond to a DPM with different values of the precision parameter α and the variance of the components Q. Each scenario corresponds to a specific combination of these two parameters. All the other quantities are fixed across the scenarios. The realizations of the DPM are here indicated as $u = 1, \ldots, n$. Different combinations of α and Q result in different values of λ. For fixed values of Q (column-wise), a proportional relation is shown between α and λ: when the first increases, the second decreases. Similarly, for fixed values of α (row-wise), a proportional relation is shown between Q and λ: when the first increases, the second decreases. Smaller values of both α and Q result in a high polarized distribution of $u = 1, \ldots, n$ and correspond to larger values of λ. Whereas larger values of both α and Q result in a low polarized distribution and correspond to smaller values of λ.

	α	Q	λ
Scenario 1	0.1	0.1	27.95
Scenario 2	0.1	1.5	2.97
Scenario 3	1	0.1	4.28
Scenario 4	1	1.5	1.9
Scenario 5	5	0.1	2.75
Scenario 6	5	1.5	1.69
Scenario 7	20	0.1	1.73
Scenario 8	20	1.5	0.49

Table 2.2: Parameters values at each scenario. Each of them corresponds to a DPM with different values of the precision parameter α and the variance of the components Q. Both α and Q affect the distribution polarization of the realizations of the DPM. As a result, different values of λ are observed.

2.6.2 Simulation 2: Inter-rater agreement and λ

Simulation setting

The following simulation study aims to highlight the complementary role of λ as an additional summary metric in inter-rater agreement analysis. The varying intercept parametrization is hereafter adopted as univariate case for the raters' effects. To this aim the standard modelling approach (i.e., the normal distributed varying intercept and the resulting ICC) is compared with the nonparametric one proposed above (i.e., the DPM prior over the varying intercept and λ). The experiment evaluates the performance of both standard *ICC* and λ in the presence of heterogeneity between raters' evaluations due to a multimodal distribution of the hierarchical effects.

Data generating process Three experimental scenarios were planned, in which a different clustering on the raters' intercept parameter was specified in the generative model. In each scenario, the rater's intercept u_i was generated from a bimodal Gaussian mixture. The location parameters of the mixture components were fixed across the scenarios, $\mu_1 = -3$ and $\mu_2 = 3$; whereas decreasing values $(1, 0.5, 0.1)$ were assigned to the components scale parameters Q_1 and Q_2 (see Table 2.3). This resulted in different polarization scenarios. The mixture components were kept equiprobable ($\pi_c = 0.5$), $c \in \{1,2\}$ throughout. The number of raters $I = 100$ and the number of items $J = 250$ were fixed across the scenarios. One continuous covariate x_{ij} with an effect $\beta = 2$ was used and it was the same across the scenarios.

Scenario 1	$u_i \overset{iid}{\sim} 0.5 \cdot N(-3,1) + 0.5 \cdot N(3,1)$
Scenario 2	$u_i \overset{iid}{\sim} 0.5 \cdot N(-3,0.5) + 0.5 \cdot N(3,0.5)$
Scenario 3	$u_i \overset{iid}{\sim} 0.5 \cdot N(-3,0.1) + 0.5 \cdot N(3,0.1)$

Table 2.3: True raters' hierarchical effects distribution across different scenarios. A Gaussian mixture is specified as a distribution of the hierarchical effects $u_i = 1, \ldots, I$. The location parameters of two components of the mixture are kept fixed across the scenarios and decreasing values were assigned to the respective scale parameters.

Standard model approach The following priors were specified for the standard hierarchical effect model (i.e., the varying intercepts are assumed to be i.i.d. normal distributed):

$$\begin{aligned} \beta &\sim N(0,5), \\ \sigma_\varepsilon &\sim Exp(0.2), \\ \sigma_u &\sim Exp(0.2), \\ u_i &\overset{iid}{\sim} N(0,\sigma_u), \end{aligned}$$

for $i = 1, ..., I$; $Exp(\cdot)$ stands for the exponential distribution and β is the non-hierarchical effect, σ_u and σ_ε are the hierarchical effects and the noise variances parameters, respectively. A logic of complexity penalization was used in the choice of the above-mentioned prior distributions (Simpson et al., 2017). The posterior of each standard hierarchical effect model was sampled using NUTS-Hamiltonian MCMC in Stan language (Stan Development Team, 2022).

Nonparametric model approach The set of priors introduced in section 2.4 were elicited for the DPM models with the following hyperparameters as suggested by (Heinzl et al., 2012): $\mathbf{m}_0 = \mathbf{0}, \mathbf{W}_0 = 100\mathbf{I}_q, a_{D_0} = 0.5, b_{D_0} = 0.5, a_{Q_0} = 0.001, b_{Q_0} = 0.001, a_\alpha = 2, b_\alpha = 2, a_\varepsilon = 0.005, b_\varepsilon = 0.005$. Note that in this study only one covariate is considered for the non-varying effect β, thus the mean b_β and the variance σ_β of this parameter are not identifiable. As a result, they are fixed $b_0 = 0, \sigma_\beta = 5$ to the same values as in the standard model.

As a result of some preliminary analysis, a dense grid of 481 equally-spaced values from -12 to 12 (i.e., with a fixed interval of 0.05) was used to monitor the mixture density of the nonparametric varying intercept u_i at each iteration. The posterior distribution of the nonparametric hierarchical effect is obtained as the set of the mean density of each point of the grid over the iterations (Gelman et al., 2013).

In all the computations for both models 55,000 iterations with 5,000 burn-ins were used, and the Markov chains were thinned by a factor of 50, resulting in samples of size 1000 (Heinzl et al., 2012).

Results

As shown in Table 2.4 the standard model (i.e., that in which hierarchical raters intercepts are assumed to be i.i.d. normally distributed) due to the rigid distributional assumption of the hierarchical parameters is not able to capture the possible multimodal distribution and it resulted in a large value of the hierarchical effect variance σ_u (see Table 2.4 and Figure 2.4). As a result, the ICC didn't capture almost any difference among the three different scenarios (see Table 2.4 and Figure 2.5). On the contrary, the DPM model, due to the flexible nonparametric specification of the intercepts prior, showed a good performance. As evident from Figure 2.4 and Table 2.4 the DPM model was far more able to reproduce the data generating process. The different mixtures used to generate the data emerged clearly from the posterior of the grid adopted to monitor $\mathbf{u}$. Since the DPM model properly learns the multi-modalities of the raters intercepts density, the index λ, being based on the ratio between the mean density of the modes and that of the antimodes present in the grid at each iteration, showed to be able to differentiate the three different polarization scenarios. It gives some interesting information regarding the shape of the nonparametric mixture distribution. The 95% credible interval of λ (see Table 2.5 and Figure2.2) as estimated in the three different scenarios highlighted different degrees of amplitude and separation (i.e., different degrees of polarization) along them. The index λ is computed as a logarithm of the ratio of the average mode density against the density of the antimodes at each iteration t of the posterior sampler. So, in the first scenario, the values of the 95% credible interval are smaller, indicating that in most of the iterations, the difference between the mean density at the modes and that of the antimodes was very small. In terms of the third scenario, λ assumed rather larger values along the iterations as evidence that the mode density is far larger than that of the antimodes. In other words, the rater clusters were separated and distinct. The parameters of the DPM are the most influential concerning λ. Specifically, the location parameters μ_c, $c = 1, \ldots, R$, and the scale parameters Q_c, $c = 1, \ldots, R$, showed to have a combined effect of the proposed index. The 95% HDP intervals of the parameters of both the DPM prior model and that with normal distributional assumption are reported in tables 2.6 and 2.7, respectively.

	σ_u	Grid density
Scenario 1	$(2.80, 3.75)$	$(-4.60, -1.60) \cup (1.40, 4.60)$
Scenario 2	$(2.65, 3.55)$	$(-3.95, -1.85) \cup (1.75, 4.35)$
Scenario 3	$(2.57, 3.46)$	$(-3.70, -2.30) \cup (2.30, 3.60)$

Table 2.4: 95% HPD intervals of the hierarchical effects variance σ_u from the standard models (i.i.d. normal distributed varying intercepts) and of the Grid density of the hierarchical effect in DPM models.

	ICC	λ
Scenario 1	$(0.952, 0.973)$	$(1.22, 6.33)$
Scenario 2	$(0.948, 0.970)$	$(1.14, 10.69)$
Scenario 3	$(0.945, 0.969)$	$(0.05, 31.94)$

Table 2.5: 95% HPD intervals of the ICC from the standard models (i.i.d. normal distributed varying intercepts) and of λ from the DPM models.

	Scenario 1	Scenario 2	Scenario 3
β	$(1.87, 2.12)$	$(1.87, 2.13)$	$(1.87, 2.12)$
μ_0	$(-0.16, 1.16)$	$(0.12, 1.33)$	$(-0.49, 0.70)$
σ_{D_0}	$(6.63, 12.63)$	$(6.09, 11.30)$	$(6.52, 11.68)$
σ_ε	$(0.43, 0.46)$	$(0.42, 046)$	$(0.42, 0.46)$
α	$(8.05, 18.06)$	$(8.15, 18.09)$	$(8.12, 18.03)$

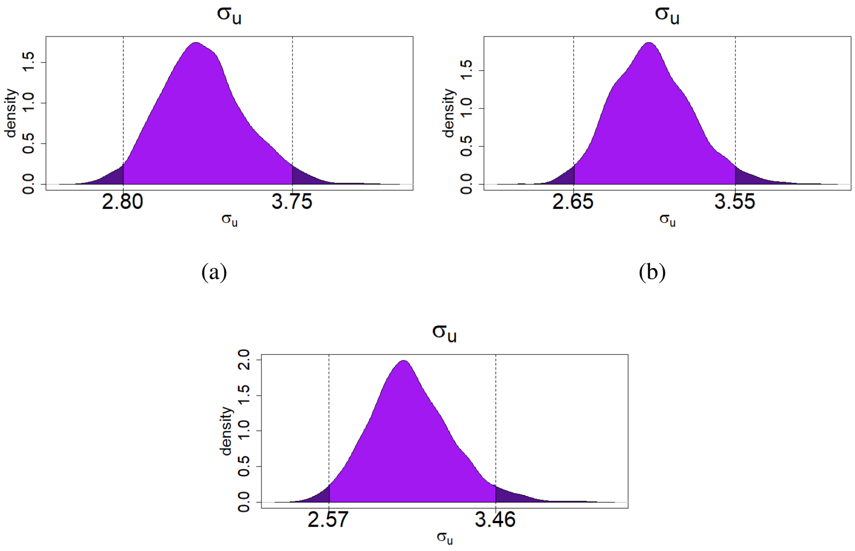
Table 2.6: 95% credible intervals of β, DPM and residuals-related parameters. Here β is the non-varying effect, b_β and σ_β are, respectively, the related location and scale hyperparameters; μ_0 and σ_{D_0} are the location and scale parameters of the base measure G_0, respectively. The precision parameter α and the residual standard deviation σ_ε are also reported.

	Scenario 1	Scenario 2	Scenario 3
β	$(1.99, 2.00)$	$(1.98, 2.01)$	$(1.99, 2.01)$
σ_ε	$(0.60, 0.64)$	$(0.62, 0.63)$	$(0.62, 0.63)$

Table 2.7: 95% HPD intervals of the other parameters of the standard models (i.i.d. normal distributed varying intercepts)

2.7 Large-scale performance assessment

The heterogeneity evaluation of teachers is a long-standing issue in psychometrics (Uto, 2022; Shirazi, 2019; Bonefeld and Dickhäuser, 2018; Casabianca et al., 2015; DeCarlo, 2008). Highly

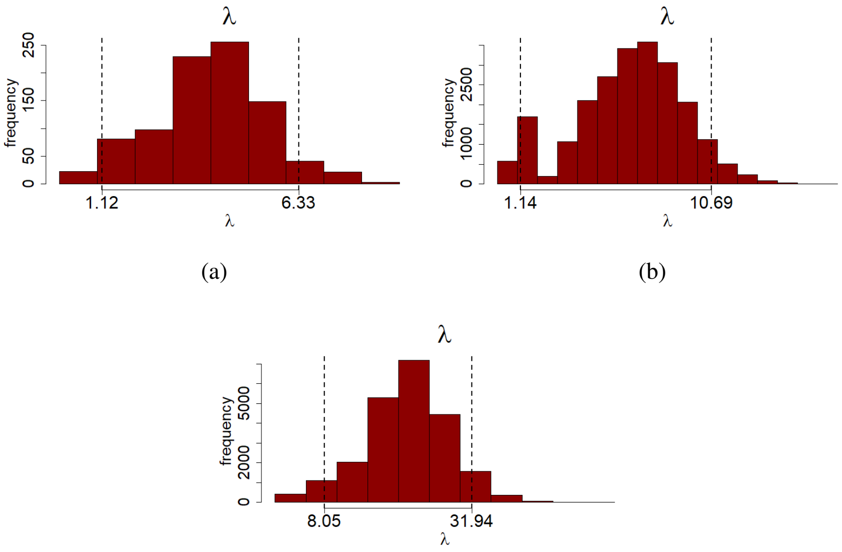


(a)

(b)

(c)

Figure 2.2: 95% HPD intervals of σ_u from the standard models (i.i.d. normal distributed varying intercepts): (a) Scenario 1: lower polarization (b) Scenario 2: medium polarization (c). Scenario 3: higher polarization. Scenario 3: higher polarization. These models, due to their rigid distributional assumption, poorly differentiate the three different polarization scenarios.

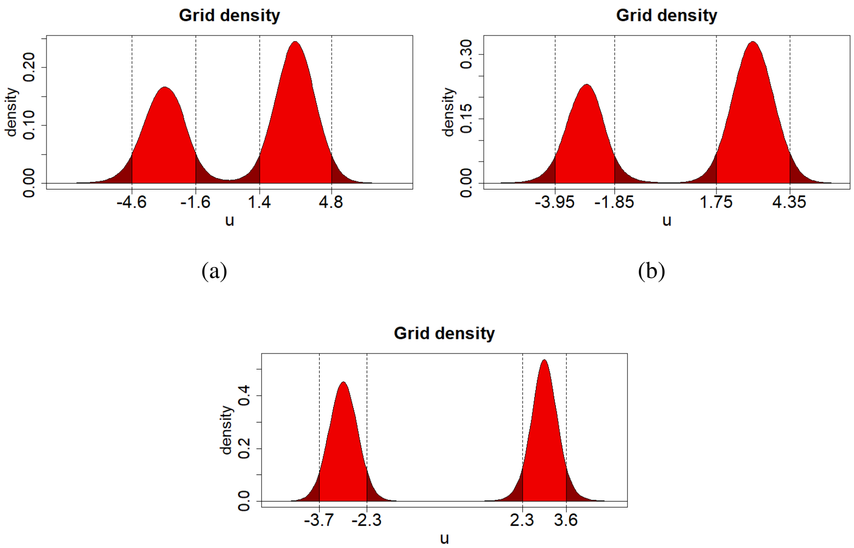


(a)

(b)

(c)

Figure 2.3: Posterior distribution of λ: (a) Scenario 1: lower polarization (b) Scenario 2: medium polarization (c). Scenario 3: higher polarization. The black dotted lines stand for 95% credible intervals.

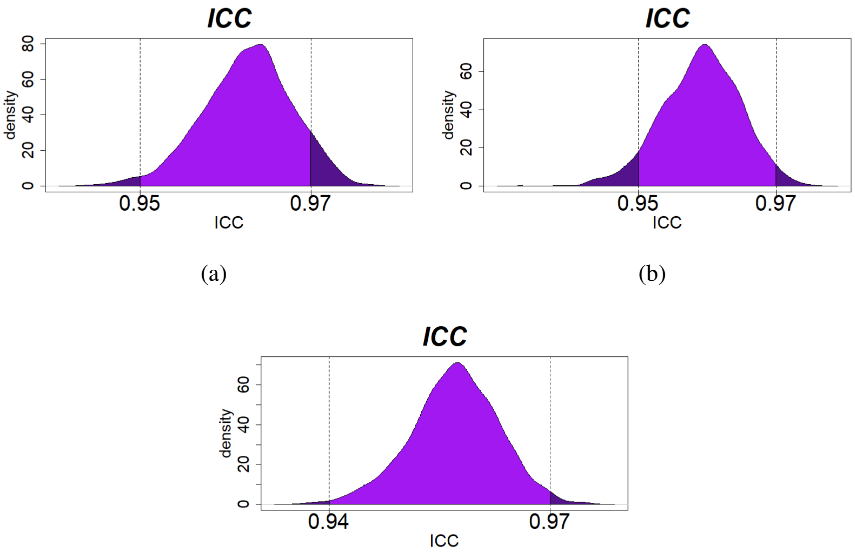


(a)

(b)

(c)

Figure 2.4: 95% HPD intervals of the hierarchical effect from the DPM. The different mixtures used to generate the data emerged clearly from the posterior of the grid adopted to monitor **u**. (a) Scenario 1 (b) Scenario 2 (c). Scenario 3



(a)

(b)

(c)

Figure 2.5: 95% HPD intervals of ICC from the standard models (i.i.d. normal distributed varying intercepts): (a) Scenario 1: lower polarization (b) Scenario 2: medium polarization (c). Scenario 3: higher polarization.

biased scores might have a detrimental effect on students' proficiency and education (Chin et al., 2020; Paredes, 2014; Cooper, 2003). The proposed nonparametric model and the index λ might be valuable tools to address this issue. They might help to shed light on very biased assessment contexts and to provide fairer scores. The estimated hierarchical effect of each teacher (which may be interpreted as the teacher's bias) might be used to adjust the observed score. A similar procedure is adopted by Nucci et al. (2021) to adjust the inter-rater agreement. The index λ might quantify teachers' polarization in their grading.

The Matura data set As an illustrative real data application, a large-scale performance assessment data set was analysed (Zupanc and Štrumbelj, 2018). The DPM model was applied to large-scale essay assessment data obtained during the nationwide external examination conducted by the National Examination Centre in upper secondary schools in Slovenia also known as *Matura* and analyzed in Zupanc and Štrumbelj (2018). These data were related to the spring term argumentative essays for years between 2010 and 2014. Particular attention is devoted to the distinction between two main aspects of essay writing: language correctness (i.e., the presence of grammatical or syntactic errors) and good argumentation of the content (i.e., a good and clear presentation of all the arguments). Regarding the data structure, students are nested within the teachers. So that each student's essay is evaluated by one trained teacher, who is asked to grade it using two different rubrics. An essay can receive a score between 0 and 20 for the language-related rubric and between 0 and 30 for the content-related one. Prior analysis of these data (Zupanc and Štrumbelj, 2018) revealed that heterogeneity among teachers was broadly down to two types: strict and lenient. The two different trends might be captured by the model and their polarization quantified by the λ index.

For this reason, N=2616 students' essays, each scored by one of I=18 different teachers, were considered for the analysis [12]. The objective of this application is to analyze teachers' differences in scoring the essay content, controlling for its language correctness. How lenient or strict they are in scoring the quality of an essay content, without the effect of the language correctness.

[12] For illustrative purposes, only the variables related to the first teachers were considered.

The content score is commonly regarded to as more susceptible to idiosyncrasies or biases of the teacher than the language-related score, which is generally more objective (Childs and Wooten, 2023; Zhu et al., 2021; Shirazi, 2019). Accordingly, the content-related score was specified as the outcome variable and the language-related score as a covariate with a non-varying effect. A DPM hierarchical prior was specified over the teachers' intercepts. All the scores were re-scaled for this analysis to get an easier parameters value interpretation [13] (Gelman et al., 2013).

2.7.1 Results

The language-related score showed a posterior mean effect of 0.27 on the content-related score, with a (0.16, 0.38) 95% credible interval. The language correctness of the essay writing had a moderate role in predicting the evaluation of its content. As shown by Figure 2.6(a) the DPM model learned the presence of two main trends from the data. The bimodal nonparametric distribution over the grid suggested that the teachers were rather heterogeneous in the essay scoring process. More precisely, they seemed to be slightly polarized around two main tendencies. Some teachers showed a slightly more lenient or more severe than the others (i.e., who had a larger or smaller hierarchical effect posterior mean, respectively), see Figure 2.6. The λ index showed a posterior mean of 1.87 which suggested a low polarization. The λ 95% HPD interval was (0.0, 6.83) which indicated a non-negligible occurrence of quite high values of λ. All the other parameters' credible intervals are reported in Table 2.8

Assuming this latent group polarization as a low latent agreement among raters, the λ index might be used in a diagnostic manner. Considering the present application, some solutions might be suggested for a fairer assessment process. Firstly, assuming a very negligible noise term, the teacher's estimated bias might be removed from the actual score. Another practical solution might be the implementation of *ad hoc* training.

Several considerations might be formulated regarding the evaluation bias of the teachers. According to the theoretical elements raised in the previous chapter, a high level of heterogeneity among teachers might be addressed and in such cases resolved. Exam evaluations need to be as accurate

[13] The following transformation was applied to standardize the score: $f(x) = \frac{x-\bar{x}}{\hat{\sigma_x}}$, where $\bar{x} = \frac{1}{N}\sum_{n=1}^{N} x_n$ was the sample mean and $\hat{\sigma_x} = \sqrt{\frac{1}{N}\sum_{n=1}^{N-1}(x_n - \bar{x})^2}$ the sample standard deviation.

and fair as possible. The way teachers evaluate or grade students must be homogeneous and, as much as possible, standardized. The moderate degree of polarization, that is the heterogeneity, that emerged among teachers' biases might be deepened and further investigated.

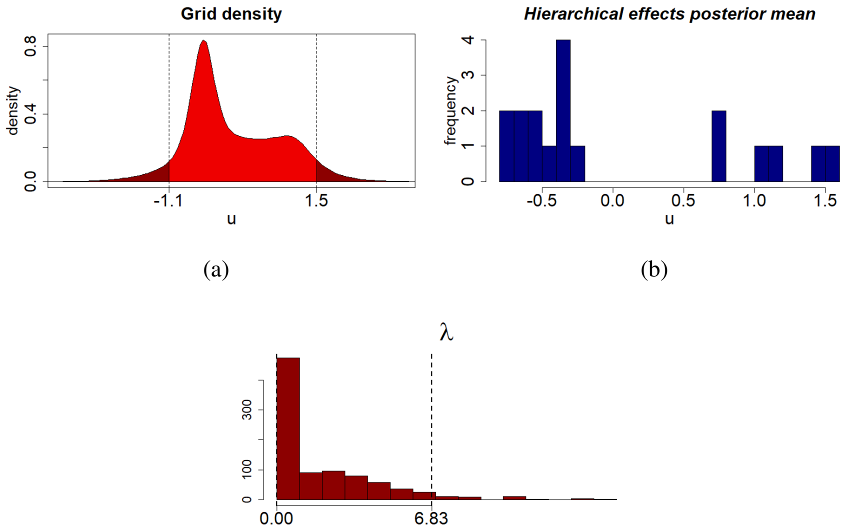
β	$(0.16, 0.38)$
μ_0	$(-0.33, 0.34)$
σ_{D_0}	$(0.19, 1.03)$
σ_ε	$(0.13, 0.15)$
α	$(3.00, 4.17)$

Table 2.8: 95% credible intervals of β, DPM and residuals-related parameters. Here β is the non-varying effect, b_β and σ_β are, respectively, the related location and scale hyperparameters; μ_0 and σ_{D_0} are the location and scale parameters of the base measure G_0, respectively. The precision parameter α and the residuals' standard deviation σ_ε are also reported.

2.8 Conclusions

Most of the statistical models commonly used to analyze data from such observational contexts haven't shown to be very flexible to certain types of heterogeneity among raters. The common HLMs with a normal (or unimodal) distributional assumption for the hierarchical effects cannot capture any possible latent clusters, i.e. any multimodality. Indeed, the residual covariance modelled through the hierarchical effects might be informative about different latent similarities among raters. In this regard, incorporating a DPM prior over the hierarchical effects distribution is a flexible choice to address this issue.

Consequently, the estimation of the agreement among the raters should take into account the possible multimodal distribution of the hierarchical effects. Interest might not be exclusively on the proportion of variance attributable to the hierarchical effects over the total variance (i.e., the main interpretation of the ICC); instead, it might be more appealing to explore the entire multimodal density. Since the DPM naturally accommodates clusters among hierarchical effects (i.e., among raters), it is natural to consider the extent to which the mixture components are separated. Since λ



(a)

(b)

(c)

Figure 2.6: (a) 95% HPD of the monitoring grid for the teachers' hierarchical effects $u = 1, \ldots, I$ (b) Posterior mean of the hierarchical effect of each teacher. Two different clusters emerged from the analysis, as expected: the more lenient (to the right-hand side) and the more severe (to the left-hand side). (c) λ' 95% credible intervals. It indicates a moderate polarized posterior distribution of the posterior hierarchical effects.

is based on the density approximated through the grid approach $f(\mathbf{u})$, it reflects both the clustering induced by the Dirichlet process and the variance of the mixture components. Due to the particular information carried by λ, it might be more informative about the latent agreement among raters than the solely ICC. The latter is very useful when the normal distributional assumption of the hierarchical effects holds. However, in the presence of multimodality the estimate of the variance of the hierarchical effect σ_u is not accurate (it might be over-estimated) and the related ICC might be non-informative.

In contexts in which strong prior beliefs about the exact number of clusters are present, a hierarchical model with a prior finite mixture distribution over the hierarchical effects is expected to have comparably good performance as well. The parametric variance of a mixture might be taken into account in the ICC formula in these cases. For the above-mentioned reasons, added flexibility and the shrinkage property the DPM was here preferred.

Many other studies are needed to fully understand the performance of λ across different combinations of the Dirichlet process parameters. Future works might highlight the role of λ when the rating is either expressed on a dichotomous or a polytomous scale. Further studies might highlight the computation of λ when multivariate hierarchical effects are specified. Comparisons between this index and the others widely used in these cases Tang et al. (2022); Forchheimer et al. (2015); Zhang et al. (2003) might be a focus of future studies. Further application of λ in more general nonparametric contexts might be studied.

The method discussed in this chapter might be compared with others already proposed in the literature for the analysis of rating data (Martinková et al., 2023; Nucci et al., 2021; Nelson and Edwards, 2015). For instance, further theoretical and practical similarities might be found between this method and that introduced by Nucci et al. (2021). The belonging threshold (BT) concept discussed in their work might be seen as the univariate rater effect we present in this chapter. Nucci et al. (2021) considered the dichotomous rating context, whereas the continuous rating is here addressed. Even though, we generalize our method to account for polytomous data in the next chapter.

Chapter 3

Generalized Bayesian nonparametric heteroschedastic hierarchical models

This Chapter is based on a joint work with Professor Ioanna Manolupoulou.

3.1 Introduction

The statistical framework introduced in the previous chapter is here generalized. This generalization concerns three different features. First, the specification of *cross-classified* observations, i. e. two independent sources of redundancy are modelled. This is the case in which the same set of items [1] are evaluated independently by different raters. Second, the heteroscedasticity among different raters. The independent and identically normally distributed assumption over the residuals across all the observations might be relaxed. This allows us to capture some systematic differences in rating behaviour among the raters. Some of them might be more consistent than others, this implies a smaller residual variance across their ratings. On the contrary, some raters might be less consistent, as a result, the variance across their ratings is larger. The third generalization feature concerns the rating scale. We generalize the previous framework to the ordinal data case. This implies a flexible modelling in which both the ordinal and the continuous rating data might

[1] According to the previous section here the term *item* is meant to be whatever object of evaluation, e.g. a subject, an object.

be analyzed under the same framework.

This three-fold generalization is crucial in this field. It is widely common that raters are asked to score the objects on an ordinal scale. In these cases, a model misspecification might be extremely detrimental to the estimates. As a consequence, any inference upon them might be wrong or not informative. Furthermore, it is also very frequent in rating contexts that different raters are asked to evaluate the same set of items (objects). In these cases, some specific characteristic of the object might introduce some dependency among ratings. This allows us to decompose the observed rating as a sum of the effects due to the object (which is generally supposed to be the *true score*), and the effects due to the rater (sometimes referred to as the noise or error terms). To capture the rating behaviours of the raters, the heteroscedastic modelling allows us to know who is the more consistent or reliable among them. This also gives us the possibility to assess the assumption that raters are equally reliable. It is a very common assumption made in several contexts.

Under this general framework, an *approximate intra-class correlation coefficient* (ICC_a) is proposed. This is commonly used as an index of reliability in rating contexts (Martinková et al., 2023; Nelson and Edwards, 2015; Gwet, 2008) as it might be interpreted as the proportion of variance of ratings due to the rater effects.

For the sake of clarity, the same notation of the previous chapter is here adopted.

For illustrative purposes, a motivating example might be considered for the framework introduced in the present chapter. Let us consider the case in which each member of a teachers' committee has to evaluate the students' assignment of a class. The assignment of each student is then evaluated independently by all the teachers. A single observed rating y_{ij} is the grade given by the teacher i to the student j. Since we generalize the previous model for continuous ratings to the ordered ratings one, here the rating is assumed to be on a linearly ordered scale. Each item, that is each student, might have his/her ability level, which might be seen as the univariate effect of the item. Accordingly, each teacher might have his/her specific effect in evaluating the students. For instance, someone might be more severe or more lenient than others.

3.2 Bayesian nonparametric HETOP model

The multiple rating scheme is here addressed as a general case. Each item $j \in J$ is assumed to be rated by different raters, $\mathscr{J}_i \subseteq \mathscr{J}$, $\mathscr{J}_i \cap \mathscr{J}_{i'} \neq \emptyset$, $i \neq i'$,$i = 1, \ldots, I$. For illustrative purposes, all the raters are hereafter assumed to be rating the same set of items, $\mathscr{J}_i = \mathscr{J}$, $i = 1, \ldots, I$. For instance, when the same students' assignments are evaluated by different teachers. The statistical and computational implications are the same for both schemes[2].

Modelling the observed rating y_{ij}. The rating $y_{ij} \in \{1, \ldots, K\} \subset \mathbf{N}$ of the item $j \in \mathscr{J}$ carried out by rater $i = 1, .., I$, is modelled as follows:

$$
\begin{aligned}
y_{ij} &\sim Cat(\pi_{ij}) \\
\pi_{ij1} &= F\left(\frac{\gamma_1 + \delta_j + \mathbf{x}'_{ij}\beta + \mathbf{z}'_{ij}\mathbf{u}_i}{\sigma_{\varepsilon,i}}\right) \\
\pi_{ijk} &= F\left(\frac{\gamma_k + \delta_j + \mathbf{x}'_{ij}\beta + \mathbf{z}'_{ij}\mathbf{u}_i}{\sigma_{\varepsilon,i}}\right) - F\left(\frac{\gamma_{k-1} + \delta_j + \mathbf{x}'_{ij}\beta + \mathbf{z}'_{ij}\mathbf{u}_i}{\sigma_{\varepsilon,i}}\right), \quad k = 2, \ldots, K-1, \\
\pi_{ijK} &= 1 - F\left(\frac{\gamma_{K-1} + \delta_j + \mathbf{x}'_{ij}\beta + \mathbf{z}'_{ij}\mathbf{u}_i}{\sigma_{\varepsilon,i}}\right)
\end{aligned}
$$

with $\gamma_0 = -\infty < \gamma_1 < \cdots < \gamma_{K-1} < \gamma_K = \infty$ and $F(\cdot)$ being the standard Normal cumulative distribution function. Here $\mathbf{x}_{ij}$, $\mathbf{z}_{ij}$ are, respectively, $1 \times p$ and $1 \times q$ vectors of distinct explanatory variables of rating y_{ij}; β is a $p \times 1$ vector of non-varying effects and $\mathbf{u}_i$ is a $q \times 1$ vector of the hierarchical effects of rater i; δ_j is the varying intercept of the item. This effect models the correlation between the ratings of the same item. The rater-specific parameter $\sigma_{\varepsilon,i} > 0$ is a scale parameter. It might be seen as the residuals' standard deviation of the underlying continuous variable $Y^*_{ij} \sim N(\delta_j + \mathbf{x}'_{ij}\beta + \mathbf{z}'_{ij}\mathbf{u}_i, \sigma^2_{\varepsilon,i})$ (Agresti, 2015; Gelman et al., 2013; Kyung et al., 2010; Gu et al., 2009). This parameter is indexed by the rater i which implies heteroscedastic residual variance.

[2] More details are discussed below

See Shear and Reardon (2021); Reardon et al. (2017); Albert and Chib (1993) for more details on this method.

Modelling the item effect δ_j. We assume that the item's varying intercepts are independent and identically distributed from a Normal distribution:

$$\delta_j \quad \sim \quad N(0, \sigma_\delta^2), \quad j \in \mathscr{J}.$$

Here $\sigma_\delta > 0$ is the scale parameter of the normal distribution. The location parameter is fixed to zero for identifiability purposes.

Modelling the rater effects $\mathbf{u}_i$. According to the previous chapter, a DPM prior is placed over the rater's effects $\mathbf{u}_i$:

$$\begin{aligned} \mathbf{u}_i | \mu_i, \mathbf{Q}_i &\sim N_q(\mu_i, \mathbf{Q}_i) \\ \mu_i, \mathbf{Q}_i | G &\overset{iid}{\sim} G \\ G &\sim DP(\alpha, G_0) \end{aligned}$$

where μ_i and $\mathbf{Q}_i$ are, respectively, the $q \times 1$ a location parameter vector and the $q \times q$ positive semi-definite covariance matrix for the hierarchical effects $\mathbf{u}_i$ of rater $i = 1, \dots, I$. Here $\mathbf{u}_i$ and ε_{ij} are assumed independent. The same stick-breaking construction discussed above might be here used as well (Sethuraman, 1994).
This model belongs to the class of HETOP models, i.e. heteroscedastic ordered probit models.

3.3 Prior specification and posterior sampling scheme

Some conjugate priors might be placed over the new parameters. As a consequence, the Block Gibbs sampling introduced for the previous model might be easily extended to account for the new parameters of the HETOP model.

Prior specification The same priors of the previous model (Chapter 2) are here placed over the parameters that are the same across the two models.

The generalization concerns three features of the model: the crossed-effects, the heteroscedasticity and the modelling of the ordinal response data.

Concerning the latter, the common underlying variable approach might be used (Agresti, 2015; Gelman et al., 2013; Bartholomew et al., 2011; Albert and Chib, 1993). For each observed rating $y_{ij} \in \mathbf{N}$, a continuous underlying variable is assumed, $Y^*_{ij} \in \mathbf{R}$. The probit specification implies a normal distribution for this variable, $Y^*_{ij} \sim N(\delta_j + \mathbf{x}'_{ij}\beta + \mathbf{z}'_{ij}\mathbf{u}_i, \sigma^2_{\varepsilon,i})$. In the present model, the location parameter is given by the linear predictor and the scale parameter has to be estimated [3]. Since $\sigma_{\varepsilon,i}$, $i = 1,\dots,I$, is indexed by the raters, it captures the amount of variability among the ratings given by i. Larger values imply a larger variability in the choice of the ordered rating categories in the assessment of the object. On the contrary, smaller values imply a less variable rating behaviour. According to the present model specification, Y^*_{ij} is assumed to be *discretized* into K ordered categories by $K-1$ cutoffs, $\gamma_1,\dots,\gamma_{K-1}$. We observe $Y_{ij} = k$, if $\gamma_{k-1} < Y^*_{ij} \leq \gamma_k$, with $\gamma_0 = -\infty < \gamma_1 < \cdots < \gamma_{K-1} < \gamma_K = \infty$ (Albert and Chib, 1993). As discussed by Lockwood et al. (2018), for any $K \geq 3$, two cutoffs need to be fixed for identifiability purposes. A diffuse prior might be assigned to the other cutoffs (Gill and Casella, 2009; Albert and Chib, 1993).

A natural conjugate prior might be placed over the variance of the item effect σ^2_δ:

$$\sigma^2_\delta \sim IG(a_\delta, b_\delta).$$

We might use the same conjugate prior for the raters' specific error variance parameters $\sigma_{\varepsilon,i}$:

$$\sigma^2_{\varepsilon,i} \sim IG(a_e, b_e), \quad i = 1\dots,I.$$

[3] Under the homoschedastic assumption the scale parameter is commonly fixed to 1 (Agresti, 2015; Kyung et al., 2010)

Posterior computation We make use of a data augmentation procedure by which we simulate the underlying variables (Gu et al., 2009; Albert and Chib, 1993). The conditional distribution of Y^*_{ij}, $i = 1,\ldots,I$, $j \in \mathscr{J}$, is:

$$Y^*_{ij}|\delta_j,\beta,\mathbf{u}_i,\sigma^2_{\varepsilon,i},\gamma_0,\ldots,\gamma_K \quad \sim \quad N(\delta_j+\mathbf{x}'_{ij}\beta+\mathbf{z}'_{ij}\mathbf{u}_i,\sigma^2_{\varepsilon,i})\times I(\gamma_{k-1}<Y^*_{ij}\leq\gamma_k), \quad k=1,\ldots,K.$$

Here $I(\cdot)$ is an indicator function. As for the underlying variable approach adopted in Generalized Linear Models (Gu et al., 2009; Albert and Chib, 1993), we generate the underlying variables from a truncated normal distribution, in which the boundary is given by two consecutive cutoffs. This approach makes both the modelling and the computation straightforward. This needs to be the first step of the Gibbs since the newly generated underlying variables have to be used in the following steps as continuous observed normal variables. Accordingly, the Gibbs steps are the same as in the previous model (Chapter 2), but those below [4].

Following Albert and Chib (1993) the conditional posterior distribution of the cutoffs, $\gamma_1,\ldots,\gamma_{K-1}$ might be seen to be uniform on the respective intervals:

$$\gamma_k \quad \sim \quad U(max\{max\{Y^*_{ij}:Y_{ij}=k\},\gamma_{k-1}\},min\{min\{Y^*_{ij}:Y_{ij}=k+1\},\gamma_{k+1}\}), \quad k=3,\ldots,K-1$$

Here $U(\cdot)$ stands for Uniform distribution. Note that for identifiability purpose, γ_1 and γ_2 needs to be fixed in the HETOP models (Lockwood et al., 2018); $\gamma_0 = -\infty$ and $\gamma_K = \infty$ by the model.

The conditional posterior of the items' effects variance σ^2_δ is:

$$\sigma^2_\delta|\delta_1,\ldots,\delta_J \quad \sim \quad IG\left(a_\delta+\frac{J}{2},b_\delta+\frac{1}{2}\sum_{j=1}^{J}\delta_j^2\right)$$

Here $J = |\mathscr{J}|$ is the cardinality of the items set.

For each rater $i = 1,\ldots,I$, the error variance is updated as follows:

[4] Note that the heteroscedasticity has to be considered in the updating steps of β and $\mathbf{u}_i$.

$$\sigma^2_{\varepsilon,i}|\beta,\mathbf{u}_i,\mathbf{y}_i \quad \sim \quad IG\left(a_\varepsilon+\frac{1}{2}|\mathscr{J}_i|,b_\varepsilon+\frac{1}{2}\sum_{j\in\mathscr{J}_i}\left(Y^*_{ij}-\mathbf{x}_{ij}\beta-\mathbf{Z}_{ij}\mathbf{u}_i\right)^2\right).$$

The varying effects of both the items and the raters need to be centred at zero for identifiability purposes (Gelman et al., 2013; Heinzl et al., 2012) during the sampling scheme. Heinzl et al. (2012) use this method in a similar context.

3.4 Bayesian nonparametric ICC

The degree of association between different ratings of the same rater might be a quantity of interest. One might want to know the extent to which the ratings given by a specific rater are correlated. As introduced in the previous Chapter, this is captured by the *intra-class correlation coefficient* (ICC; see Chapter 2 for more details on this index). Under standard hierarchical models, the computation of the implied ICC is straightforward. Under the nonparametric framework, as shown in Chapter 2, it needs to be carefully considered. The nonparametric distribution of the varying effect makes the ICC computation less intuitive.

The univariate case (i.e. the raters' varying intercept specification) is here considered as it is the case in which the ICC is generally more interpretable and used.

When a stick-breaking process is involved in the DPM construction (Rigon and Durante, 2021; Heinzl et al., 2012; Sethuraman, 1994) the mixture components parameters are not marginalized out. Based on convergence and posterior consistency of the DPM in density estimation Ascolani et al. (2023); Ghosh et al. (2022); Canale and Blasi (2017); Wu and Ghosal (2010); Tokdar (2006); Amewou-Atisso et al. (2003); Ghosal et al. (1999), we can use the mixture parameters (i.e., the components' location and scale parameters) to approximate the variance of the nonparametric distribution of the varying effects. Since we approximate the distribution of the varying effects through the DPM, we can use the involved mixture parameters to estimate the first two moments of this distribution. Specifically, we are interested in the variance of the mixture distribution.

Given the following distribution of the raters' effect u_i[5]:

$$\begin{aligned}
u_i|\mu,\mathbf{Q}, &\overset{iid}{\sim} \sum_{c=1}^{R} \pi_c N_q(\mu_c, Q_c), \quad i=1,\ldots,I \\
\mu_c, Q_c &\overset{iid}{\sim} G_0 \\
\pi_c &= v_c \prod_{l<c}(1-v_l), \text{ where} \\
v_c &\overset{iid}{\sim} Be(1,\alpha), \ c=1\ldots,R,
\end{aligned}$$

the variance of u_i (i.e., the variance of the mixture distribution), $i=1,\ldots,I$ is:

$$\sigma_u^2 = \sum_{c=1}^{R} \pi_c(Q_c+\mu_c^2)-\mu_u^2,$$

where μ_u is the expected value $E[u]=\sum_{c=1}^{R}\pi_c\mu_c$.

At each iteration t, the ICC, given the other parameters, might be computed as follows:

$$ICC_{a_i} = = \frac{\sigma_u^2}{\sigma_\delta^2+\sigma_u^2+\sigma_{\varepsilon,i}^2}$$

where σ_u^2 is the variance of rater i effect u, σ_δ^2 is the variance of the effect of the item δ and σ_ε^2 is the rater i's error variance. It indicates the correlation between ratings of the same rater across different items. Smaller values indicate a low correlation, whereas larger values indicate a higher correlation.

To assess inter-rater reliability, the following *IRR* index might be computed:

$$IRR_i = \frac{\sigma_\delta^2}{\sigma_\delta^2+\sigma_u^2+\sigma_{\varepsilon,i}^2}.$$

It is the ratio between the variance due to the item σ_δ^2, also referred to as the *true* variance, and the total variance. As for the *ICC*, the *IRR* values are on the unit interval $IRR \in (0,1)$; smaller values indicate poor reliability of the rater, as a large portion of the total variance of the ratings is due to both the rater effect and the error. On the contrary, values of *IRR* near one indicate a highly

[5] See Chapter 2 for more details on this stick-breaking construction.

reliable rating behaviour, as most of the total variance is due to the item's effect. [6] At any rate, the parameter $\sigma^2_{\varepsilon,i}$ is sufficiently informative about the reliability of rater i. See Martinková et al. (2023) for more details on inter-rater reliability.

3.5 Further remarks and future directions

The proposed statistical framework due to its flexibility might be applied in several contexts. The three-fold generalization makes the model able to capture novel interesting information about the rating process. It may inform the researcher about the heterogeneity among different raters and might provide a more flexible tool.

We discussed the HETOP as a more general model, but many features might be used also in other types of data. For instance, the heteroscedasticity might be easily addressed in the same manner in the continuous data case. The non-ordinal case might be a new compelling research line.

Future simulation studies and motivation applications might highlight the advantages of the model. At least in an asymptotic regime, the credible interval of the ICC_a computed with the proposed method is expected to be tighter than those based on normal assumptions. This is because the estimate of σ^2_u is more accurate. It is not affected by the strong assumptions.

[6] It might be interpreted as the average correlation between different scores given *ideally* to the same item by the same rater.

Chapter 4

Hierarchical Bayesian Modelling of Peer Grading

This Chapter is based on a joint work with Professors Irini Moustaki and Yunxiao Chen.

4.1 Introduction

In the previous chapters, the set of the raters and that of the items were assumed to be different, that is the items (which might be subjects or objects) are evaluated by different independent raters. This data structure, discussed in the previous chapters, might be seen as a ”unidirectional” rating scheme. In some cases, when the objects of the evaluation are people it might be possible to have a ”bidirectional” rating scheme. More specifically, under this scheme people rate each other, a person evaluates other people and, in turn, he/she is evaluated by others as well. People might have a *twofold* role, one as a rater and another as an object of rating, that is they are evaluated. It is a valuable rating solution in situations of peers, for instance in the educational contexts in which each student is evaluated by the other students. As a consequence, they might be regarded both as examinees and as graders (i.e., raters). Based on the motivating application from the educational setting and considering that the evaluation that we are referring to grades/scores, we use the more specific term *grader* instead of rater; with this, we aim to make the discussion more contextualized and clearer.

This specific rating scheme is referred to as peer grading or peer assessment and it is a system of formative assessment in education whereby students assess and give feedback on one another's work. It not only substantially reduces teachers' burden for grading coursework but may also improve students' understanding of the subject and critical thinking (Shengkai Yin and Chang, 2022; Panadero and Alqassab, 2019). Consequently, it is widely used in many educational settings, including massive open online courses (MOOCs) (Gamage et al., 2021), large university courses (Double et al., 2016) and small classroom settings (Sanchez et al., 2017).

In a peer grading system, one student's work is assigned, often randomly, to several other students to grade, and the resulting grades are then aggregated through a simple scheme – often by taking the mean or the median – to give the final grade for his/her work (Sajjadi et al., 2015; Reily et al., 2009). Such a naïve system may be problematic because there are reliability concerns due to the grader bias and variability, a more accurate score is needed. It does not provide a mechanism to assess the grader's performance, which is neglected.

Existing work on peer grading data Several methods have been proposed during the last three decades to both mitigate graders' personal bias and improve peer assessment reliability (Alqassab et al., 2023). The two major approaches that have been proposed to overcome the naïve system are machine learning, e.g., Han et al. (2020); Li et al. (2019); Sajjadi et al. (2015) and latent variable models, e.g., Xiong and Suen (2016); Piech et al. (2013). Both provide methods to aggregate the peer grades to come up with a more accurate final score. The former aims to predict the true grades using those given by an instructor to a training set. The second one provides some interpretable estimates which highlight interesting features of the grading process. Specifically, Piech et al. (2013) formulated three similar peer grading models to estimate the student's true grades as well as grader biases and reliability. They aimed to recover the true grade of each student controlling for the tendency of the grader to give higher (vs. lower) grades (i.e., the grader bias) and his/her consistency across different scores (i.e., grader reliability). The observed grade is assumed to be Normal distributed with the location parameter given by the sum of the true grade of the student and the grader bias, and the scale parameter being the grader reliability. In particular, they introduced a dependency between the reliability of a student as a grader and his/her ability as a student to do the assignments. Their approach overcomes the naïve one and offers a better solution to the reliability concerns. Nonetheless, a more informative and less restrictive modelling might be adopted.

In this chapter, we propose a more general and flexible modelling for peer grading data within the latent variable approach. We model each student's grade for each coursework to be a function of four key latent features: the ability of the student, the bias and the reliability of the grader, and the difficulty level of each assignment. We decompose each grade accordingly and we extend the model to accommodate for longitudinal trajectories of student-specific latent ability through a growth curve modelling. The peer grading data structure implies that each student is also a grader, which might result in some sort of dependency between certain parameters of the model. For this reason, we adopt a joint modelling of student-specific latent variables.

The proposed model is closely related to two lines of psychometric models – models for measuring students' ability (van der Linden, 2016; Reckase, 1997; Birnbaum, 1969) and rater models for modelling rater effect (Martinková et al., 2023; Casabianca et al., 2015; DeCarlo, 2008). As for the first line, we model the observed grade to be a function of both the latent student's ability and the difficulty level of the assignment, which might be conceived as the item difficulty parameter in the Item Response Theory models (van der Linden, 2016). As with the raters' models, student's performance is assessed by graders. We model the rater effect by introducing some grader-specific parameters that account for their bias and reliability.

The main difference between these two psychometric lines (i.e., models for measuring students' ability and models for raters' effects) and the one proposed here is that in the peer grading system, each student has two roles – the examinee being assessed and the grader that assesses peers. This requires us to introduce multiple latent variables to capture the features of both the examinee and grader of a student. In this regard, there are some similarities between our model and those proposed by Piech et al. (2013). On one hand, we both consider the dependency of these different features (i.e., those related to the role of the examinee and the others to that of the grader) within the same individual. On the other hand, we made different modelling choices. For instance, they specified an autoregressive path between the same grader's biases across different times; they addressed separately by two different models the correlation due to the double role of each student and the longitudinal dependencies. Moreover, we modelled each assignment to have a different difficulty level, this aspect was not taken into account within their modelling. We propose a general statistical framework which might be applied both to a cross-sectional and a longitudinal context. It may enable a more detailed analysis of the aspects involved in peer grading and provide students with highly accurate grades.

The rest of the chapter is organised as follows. In Section 4.2, a general statistical framework is proposed, under which several specific models are described. In Section 4.3, an extension of the model to the latent growth curve is presented. In Section 4.4 results from simulations studies are reported. Real data examples are provided in Section 4.5. Advantages, limits and future direction of the present work are discussed in Section 4.6.

4.2 Proposed Model

4.2.1 Problem Setup

Consider N students who receive T coursework assignments over a period of time. Generally, each assignment has to be done at different times. Therefore, each time point $t = 1, \ldots, T$ implies a different assignment hereafter [1]. Student i's work for assignment t is randomly assigned to a small subset of other students. We denote this subset by $S_{it} \subset S \setminus i$, where S is the set of all the students and has cardinality $|S| = N$. Each grader $g \in S_{it}$ comes up with a grade Y_{igt}, following some scoring rubrics. We focus on the setting where Y_{igt} is a continuous random variable as motivated by a real-data example to be considered in Section 4.5. The model framework below can be easily extended to ordinal data through an underlying variable formulation (Bartholomew et al., 2011, see e.g., Chapter 5.9). Typically but not necessarily, the number of grades $|S_{it}|$ is the same for different students and coursework assignments. A naïve estimate of student i's performance on the tth assignment is the average score, defined as $(\sum_{g \in S_{it}} Y_{igt})/|S_{it}|$, or other summary statistics that measure the centrality of Y_{igt}, $g \in S_{it}$, for example, the median. However, as pointed out earlier,

[1] The models presented below are easily extended to the case in which multiple assignments are done at the same time point.

such a score does not take into account the heterogeneity of the graders and, thus, may be biased and suffer from a high variance. In addition, this score also fails to adjust for the difficulty levels of the assignments. Consequently, the scores are not comparable across different assignments.

4.2.2 Proposed Model

Modelling Peer Grade Y_{igt}. We assume the following decomposition for the peer grade Y_{igt}

$$Y_{igt} = \theta_{it} + \tau_{igt} - \delta_t, \quad i = 1,...,N, \quad t = 1,\ldots, \quad T, \quad g \in S_{it}. \tag{4.1}$$

Here, $\delta_t \in \mathbb{R}$ captures the difficulty level of assignment t, with δ_1 set to zero as the reference. A larger value of δ_t implies a more difficult assignment, which results in a lower observed grade. In addition, $\theta_{it} \in \mathbb{R}$ represents the student i's true score for assignment t. Thus, $\theta_{it} - \delta_t$ is the student i's expected score for assignment t, which takes into account the assignment's difficulty. Finally, $\tau_{igt} \in \mathbb{R}$ represents the difference between the observed peer grade Y_{igt} and the expected score $\theta_{it} - \delta_t$. We assume that θ_{it}, τ_{igt} and δ_t are independent.

Modelling True Score θ_{it}. For each student i, we assume that his/her true scores for different assignments θ_{it}, $t = 1,\ldots,T$, are independent and identically distributed, following a Normal distribution $N(\mu_{\theta,i}, \sigma^2_{\theta,i})$. Here, $\mu_{\theta,i} \in \mathbb{R}$ and $\sigma^2_{\theta,i} \in \mathbb{R}_+$ are two student-specific latent variables. They capture two different features of a student which are related to the role of the examinee.

The first latent variable, $\mu_{\theta,i}$, captures a student's overall performance across different assignments. It might be considered as the average ability of the student in doing the assignments. Larger values of this variable imply better student performance and, on average, higher grades. In contrast, smaller values of $\mu_{\theta,i}$ indicate an overall poor performance of a student across assignments.

The other latent variable, $\sigma^2_{\theta,i}$, captures the stability of a student's proficiency. Namely, the extent to which the student consistently does well on the assignments. Lower values of $\sigma^2_{\theta,i}$ imply more stable performances of i, i.e. a small variance between true scores across assignments $\theta_{it}, t = 1,\ldots,T$. On the contrary, higher values imply less stable performances across assignments, namely a larger variance between his/her true scores.

Within this modelling, we assume that a student's overall performance does not change across assignments. This assumption might be relaxed if it is of interest to assess students' growth throughout a course. We present this case in Section 4.3.

Modelling Grader Effect τ_{igt}. Each student g grades multiple pieces of work. We denote this set by $H_g = \{(i,t) : g \in S_{it}\}$. For each student g, we assume that τ_{igt}, $(i,t) \in H_g$, are independent and identically distributed, following a Normal distribution $N(\mu_{\tau,g}, \sigma^2_{\tau,g})$. Here, $\mu_{\tau,g} \in \mathbb{R}$ and $\sigma^2_{\tau,g} \in \mathbb{R}_+$ are two student-specific latent variables. They capture two different features of a student which are related to the role of grader.

The first latent variable, $\mu_{\tau,g}$, captures the bias of student g as a grader. A positive value of $\mu_{\tau,g}$ implies that the student tends to have a lower standard and gives grades that are, on average, above the expected scores. In contrast, a student with a negative value of $\mu_{\tau,g}$ tends to follow a high standard when grading and gives grades that are, on average, below the expected score.

The variance $\sigma^2_{\tau,g}$ is another student-specific latent variable capturing the reliability of student g as a grader. Smaller values imply, on average, a more consistent grading and more reliable grades. On the contrary, larger values of $\sigma^2_{\tau,g}$ indicate a student g's poorer grading quality and, as a result, less reliable grades.

Here it is assumed that a student's grading quality does not change over time. Indeed, both the grader role-related latent variables do not change across the assignment (i.e., over time). This assumption might be relaxed letting both $\mu_{\tau,g}$ and $\sigma^2_{\tau,g}$ depend on time t.

Joint Modelling of Student Specific Latent Variables. The above model specification introduces four student-specific latent variables $(\mu_{\theta,i}, \sigma^2_{\theta,i}, \mu_{\tau,i}, \sigma^2_{\tau,i})$ for each student i. It is reasonable to assume some dependency between these quantities since they are all features of the same student i. We model these variables to be independent and identically distributed across students, following a multivariate Normal distribution $\left(\mu_{\theta,i}, log(\sigma^2_{\theta,i}), \mu_{\tau,i}, log(\sigma^2_{\tau,i})\right) \overset{iid}{\sim} MVN(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, $i = 1, \dots, N$. Here, $\boldsymbol{\mu} \in \mathbb{R}^4$ is the vector of expected values of the latent variables and $\boldsymbol{\Sigma}$ is a 4×4 symmetric positive definite variance-covariance matrix. This *iid* assumption might be relaxed allowing the students to be nested within clusters, this point is discussed in the conclusion section of this chapter. For identifiability purposes we need to fix the expected value of the true score $\mathbb{E}[\mu_{\theta,i}] = 0$ and that of the bias $\mathbb{E}[\mu_{\tau,i}] = 0$. We do not impose any constrain over $\mathbb{E}[log(\sigma^2_{\theta,i})]$ and $\mathbb{E}[log(\sigma^2_{\tau,i})]$. This allows us to estimate the mean stability of students' ability as an examinee and their mean reliability as a grader, respectively. Higher values of $\mathbb{E}[log(\sigma^2_{\theta,i})]$ imply that students perform as examinees, on average, consistently across different assignments. Higher values of $\mathbb{E}[log(\sigma^2_{\tau,i})]$ imply that students are, on average, consistent graders. Lower values of this quantity raise serious reliability concerns since grades might be highly unreliable.

This joint modelling allows us to borrow information from a student's grading behaviour when predicting his/her true score. It allows us to address some substantive questions, such as whether a student who performs better in the coursework also tends to be a more reliable or severe grader. All this information is provided by the variance-covariance matrix:

$$\mathbf{\Sigma} = \begin{pmatrix} \sigma_1^2 & \sigma_{1,2} & \sigma_{1,3} & \sigma_{1,4} \\ - & \sigma_2^2 & \sigma_{2,3} & \sigma_{2,4} \\ - & - & \sigma_3^2 & \sigma_{3,4} \\ - & - & - & \sigma_4^2 \end{pmatrix}$$

The variance parameters are informative concerning the variability of the latent variables across students. The parameter σ_1^2 is the variance of the overall student ability. Larger values imply a higher heterogeneity in the coursework among students. On the contrary, smaller values suggest that the student's proficiency level is quite similar. The variance of the student stability (i.e., how consistently they perform over time) is captured by σ_2^2. Larger values imply that, on average, the students have a discontinuous performance over time and show big jumps in their proficiency. In contrast, smaller values indicate that the students show a quite constant proficiency level over time. The parameter σ_3^2 is the variance of students' bias as a grader. Higher values suggest that there is more variability among the students in their grading standards. This results in more biased grades. Whereas, smaller values of σ_3^2 imply more accurate grades. The variability of students' reliability as a grader is captured by σ_4^2. Larger values indicate that students are very different in terms of reliability. On the contrary, smaller values imply that the students show a more similar degree of reliability in grading.

The off-diagonal elements (i.e., $\sigma_{1,2},\ldots,\sigma_{3,4}$) capture the covariance between the student-specific latent variables; the main diagonal elements (i.e., $\sigma_1^2,\ldots,\sigma_4^2$) are the respective variances. The parameter $\sigma_{1,2}$ captures the covariance between the student's features related to the role of examinee: the overall ability and the consistency of the performance over time; $\sigma_{1,3}$ is the covariance between the overall ability of a student and his/her severity as a grader (i.e., grader bias); $\sigma_{1,4}$ captures the covariance between a student's overall ability as examinee and his/her reliability as a grader. The covariance between the consistency of a student's proficiency and his/her bias and reliability are captured by $\sigma_{2,3}$ and $\sigma_{2,4}$, respectively. The parameter $\sigma_{3,4}$ is the covariance between the two features related to the role of the grader: the bias and the reliability.

Some Remarks. This statistical modelling allows us to decompose Y_{igt} to have useful information about the peer grading process. Since we have repeated measures of the students both as an examinee and as a grader, we can recover the different features of each student in both roles. This implies several advantages. First, we can correct the grader's bias and use his/her reliability to reduce the variance and make better predictions. Second, it is possible to estimate the difficulty level of each assignment once the examinee's and grader's features are considered. Third, we can estimate the dependency patterns among the student-specific latent variables.

A naïve system might be a detrimental choice. As a simple summary statistics of the centrality of Y_{igt}, it doesn't allow us to correct for the bias and the reliability of the graders. Unreliable or rogue grading behaviours might dramatically affect the examinee's final grade. On the contrary, the proposed statistical framework overcomes these issues and provides the students with fairer and more accurate grades.

Based on the proposed model the variance of the grade Y_{igt} is $Var(Y_{igt})|\delta_t = \sigma_1^2 + \sigma_{\theta,i}^2 + \sigma_3^2 + \sigma_{\tau,g}^2$. We can express the covariance between two grades, Y_{igt} and $Y_{ig't}$, given by g and g', with $g \neq g'$, to the same student i for the same assignment t as $Cov(Y_{igt}, Y_{ig't})|\delta_t = \sigma_1^2 + \sigma_{\theta,i}^2$. The covariance between the grades of the same student across different assignments is $Cov(Y_{igt}, Y_{ig't'})|\delta_t, \delta_{t'} = \sigma_1^2$.

The covariance between two grades Y_{igt}, $Y_{i'gt}$ given by the same grader g to two different students $i \neq i'$ for the same assignment t is $Cov(Y_{igt},\, Y_{i'gt})|\delta_t = \sigma_3^2$. Since we assume that the quality of the grader g doesn't change over time, also the covariance between Y_{igt}, $Y_{i'gt'}$ is $Cov(Y_{igt},\, Y_{i'gt'})|\delta_t, \delta_{t'} = \sigma_3^2$. Since the grades of different students given different grades are assumed to be independent, $Cov(Y_{igt},\, Y_{i'g't})|\delta_t = 0$ and $Cov(Y_{igt},\, Y_{i'g't'})|\delta_t, \delta_{t'} = 0$.

The reduced model: the single assignment case

A reduced cross-sectional model might be specified if the interest is on a single assignment, i.e., when $T = 1$. With continuous data, we can estimate the assignment difficulty parameter δ. In this setting, we can not estimate the parameters $\sigma^2_{\theta,i}$, student's proficiency consistency over time. Since students' stability parameters $\sigma_{\theta,i}$ are here fixed to zero, the student–specific latent variables are here three $(\mu_{\theta,i}, \mu_{\tau,i}, \sigma^2_{\tau,i})$. As in the main model, they are jointly modelled to follow a multivariate Normal distribution, $\big(\mu_{\theta,i}, \mu_{\tau,i}, log(\sigma^2_{\tau,i})\big) \overset{iid}{\sim} MVN(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, $i = 1, \dots, N$. The same identifiability constraints of the main model are still needed, i.e. $\mathbb{E}[\mu_{\theta,i}] = 0$ and $\mathbb{E}[\mu_{\tau,i}] = 0$. The notation for the parameters of the multivariate normal, as in the multiple assignment case, is consistent with the order of the variables. For instance, the parameter σ_2^2 here refers to the variance of the second student-specific latent variable $\mu_{\tau,i}$, i.e. the bias; accordingly, $\sigma_{1,2}$ is the covariance between $\mu_{\theta,i}$ and $\mu_{\tau,i}$.

This cross-sectional model implies the following equivalences. The variance of each grade Y_{ig} is $Var(Y_{ig}) = \sigma_1^2 + \sigma_2^2 + \sigma_{\tau,g}$. The covariance between the grades given to the same examinee i by two different graders, $g \neq g'$, Y_{ig} and $Y_{ig'}$, is $Cov(Y_{ig}, Y_{ig'}) = \sigma_1^2$. The covariance between two grades, Y_{ig} and $Y_{i'g}$, given by the same grader g to two different students $i \neq i'$ is $Cov(Y_{ig}, Y_{i'g}) = \sigma_2^2 + \sigma^2_{\tau,i}$. The grades given to different students $i \neq i'$ by different graders $g \neq g'$ are assumed to be independent, $Cov(Y_{ig}, Y_{i'g'}) = 0$.

4.2.3 Bayesian Inference

The following model parameters are unknown and have to be estimated: $\delta_2, \dots, \delta_T$, i.e. the assignment difficulty levels; $\boldsymbol{\mu}, \boldsymbol{\Sigma}$, i.e . the parameters of the multivariate Normal distribution of the latent variables. We adopt a full Bayesian procedure to do inference. To this regard, we need to place some prior over the unknown parameters.

Prior specification. We have to specify a prior for the assignment difficulty parameters $\delta_1, \dots, \delta_T$. From the frequentist standpoint we need to put some constrain on these parameters for identifiability purposes. In this regard, we might fix the first parameter to be zero $\delta_1 = 0$ as a reference point; alternatively, we may constrain these parameters to sum to zero $\sum_{t=1}^{T} \delta_t = 0$. These constraints are commonly adopted to estimate the item difficulty parameters in Item Response Theory models (van der Linden, 2016). Nevertheless, within Bayesian multilevel modelling, these constraints are unnecessary as suggested by Gelman and Hill (2006, p.314-320). To compute the posterior we have to multiply the likelihood by the prior. Since different values of the same parameter might result in the same likelihood, we can use the prior to inform the posterior; as a consequence, it results in a unique solution (Gelman et al., 2013). this allows us to estimate all the set of assignment difficulty parameters without fixing any of them. A standard Normal distribution might be a valuable prior for these parameters, $\delta_1, \dots, \delta_T \overset{iid}{\sim} Normal(0,1)$ [2]. This is the prior we elicit in both the simulation and the real case application.

A prior have to be specified for the parameters of the multivariate Normal distribution of the student-specific latent variables $(\mu_{\theta,i}, log(\sigma^2_{\theta,i}), \mu_{\tau,i}, log(\sigma^2_{\tau,i}))$, i.e. $\boldsymbol{\mu}, \boldsymbol{\Sigma}$. For identifiability purposes the expected value of two latent variables are fixed, $\mathbb{E}[\mu_{\theta,i}] = 0$ and $\mathbb{E}[\mu_{\tau,i}] = 0$. We place the same Normal distribution $Normal(0,5)$ as a prior for both the marginal location parameters of $log(\sigma^2_{\theta,i})$ and $log(\sigma^2_{\tau,i})$.

We express the variance-covariance matrix Σ in terms of standard deviation and correlation:

$$\boldsymbol{\Sigma} = \mathbf{S}\boldsymbol{\Omega}\mathbf{S}.$$

Here $\mathbf{S} = diag(\sigma_1, \dots, \sigma_4)$ is the 4×4 diagonal matrix of the latent variables marginal standard deviations; $\boldsymbol{\Omega}$ is the 4×4 correlation matrix, with the main diagonal elements equal to one, e.g. $\omega_{1,1} = 1, \omega_{2,2} = 1$, and the off-diagonal elements given by $\omega_{m,n} = \sigma_{m,n}/\sqrt{\sigma^2_m \sigma^2_n}$, $m \neq n$,

[2] Fixing only the location parameter might be sufficient to identify the parameters.

e.g. $\omega_{1,2} = \sigma_{1,2}/\sqrt{\sigma_1^2\sigma_2^2}$. We specify a Half-Cauchy prior distribution for each element of $\mathbf{S}$, $\sigma_1, \ldots, \sigma_4 \overset{iid}{\sim} Half-Cauchy(0,5)$. This distribution is supported on the set of all real numbers that are greater than or equal to zero, $[0,\infty)$ and put most of the mass toward zero. We specify a LKJ correlation distribution over $\boldsymbol{\Omega} \sim LkjCorr(1)$. This places a uniform density over the correlation matrices.

Posterior computation. Taking into consideration the peer grading data structure, the proposed model includes crossed (i.e., non-nested or cross-classified) varying effects. As widely argued in literature Goplerud (2022); Gin et al. (2020); Jeon et al. (2017), the large number of parameters raises challenging computational issues. The marginal likelihood involves high-dimensional integrals, numerical integration or approximate methods might be prohibitive or not satisfactory in these cases. Bayesian estimation is a flexible solution for scalable and accurate inference Cho and Rabe-Hesketh (2011). For the present model, a ”No-U-Turn” Hamiltonian Monte Carlo sampler Hoffman and Gelman (2014) is implemented through the Stan programming language .

To resolve the convergence issue and make the MCMC mix well we used a non-centred reparametrization for the multivariate Normal distribution (Gelman et al., 2013; Betancourt and Girolami, 2013; Papaspiliopoulos et al., 2007). We express the distribution of the vector of student-specific latent variables through an affine transformation, such that:

$$\left(\mu_{\theta,i}, log(\sigma^2_{\theta,i}), \mu_{\tau,i}, log(\sigma^2_{\tau,i})\right) = \boldsymbol{\mu} + \mathbf{S}\left(\mathbf{L}\boldsymbol{\alpha}_i\right).$$

Here $\mathbf{L}$ is the Cholesky factor[3] of the correlation matrix $\boldsymbol{\Omega} = \mathbf{L}\mathbf{L}'$; the element of the four-dimensional vector $\boldsymbol{\alpha}_i \in \mathbb{R}^4$ are independent and identically distributed following a standard Normal distribution, $\alpha_{i,1}, \ldots, \alpha_{i,4} \overset{iid}{\sim} Normal(0,1)$, $i = 1, \ldots, N$. Notice that, as stated above, $\mu_1 = 0$ and $\mu_3 = 0$ for identifiability purposes.

Under this inference procedure, each parameter and student-specific latent variables might be provided with a posterior point estimate, e.g. the posterior mean, and an interval estimate, e.g. a

[3] This is a method to decompose a specific type of matrix into the product of a lower triangular matrix (i.e., a square matrix in which all the entries above the main diagonal are zero) and its transpose.

95% quantile-based credible interval. The latter might be seen as an uncertainty measure of the estimated quantity; broader intervals suggest more uncertainty about the values of the parameters, whereas narrower intervals reflect less uncertainty about their values.

Model comparison. Different nested models might be compared under the current statistical framework. Some parameters of the main model might be fixed and some dependencies removed. For instance, the single-assignment model might be considered as a reduced version of the main model when only one coursework is considered. Another reduced model results from fixing some student-specific latent variables to be equal across students. For instance, we might specify $\sigma^2_{\theta,i} = \sigma^2_{\theta,i+1}$, $i = 1,\ldots,N-1$, assuming that students are equally consistent in doing the assignments. An equal reliability might be assumed across graders, which implies $\sigma^2_{\tau,i} = \sigma^2_{\tau,i+1}$, $i = 1,\ldots,N-1$. These models might be compared in terms of predictive performance. A Bayesian leave-one-out cross-validation procedure might be used. We use the expected log point-wise predictive density (elpd) as a measure of predictive accuracy of each data point (i.e., each grade) taken one at a time (Vehtari et al., 2017). The Bayesian LOO estimate of out-of-sample predictive fit is defined as:

$$elpd_{loo} \quad = \quad \sum_{t=1}^{T}\sum_{i=1}^{N}\sum_{g\in S_{it}} logp(Y_{igt}|\mathbf{Y}_{-igt})$$

where $\mathbf{Y}_{-igt}$ is the vector of all the grades without the grade Y_{igt}; and the conditional probability $p(Y_{igt}|\mathbf{Y}_{-igt}) = \int p(Y_{igt}|\boldsymbol{\eta})p(\boldsymbol{\eta}|\mathbf{Y}_{-igt})d\boldsymbol{\eta}$, with $\boldsymbol{\eta}$ being the vector of the model parameters. Models that correspond to higher *elpd* are preferred since they are more accurate in their predictions. Throughout this paper, we use the R software rstan (Stan Development Team, 2023) for the analysis and the CmdStan interface for posterior sampling. The loo R package (Vehtari et al., 2017) was used for model comparisons.

4.3 Extension to Latent Growth Curves

The main model might be extended to account for students' growth throughout a course and to capture their specific longitudinal trends. When the number of time points T is greater than 2, we might explicitly model the change over time of the student-specific ability by letting the mean of the True Score $\mu_{\theta,i}$ depend on the time t. To capture non-linear growth trends a linear basis expansion in t might be adopted (James et al., 2013; Hastie et al., 2009). Denote by $h_k(t) : \mathbb{N} \rightarrow \mathbb{R}$ the k-th transformation of t, $k = 1, \ldots, K$, we model the student specific ability [4]:

$$\mu_{\theta,i,t} \quad = \quad \beta_{0,i} + \sum_{k=1}^{K} \beta_{k,i} h_k(t), \quad i = 1, \ldots, N; \quad t = 1, \ldots, T. \tag{4.2}$$

This implies that a student's average ability changes over time and the True Score distribution at each assignment is $\theta_{it} \sim Normal(\mu_{\theta,i,t}, \sigma^2_{\theta,i})$. Here $\beta_{0,i}, \beta_{1,i}, \beta_{2,i}$ are student-specific latent variables and t is a variable coding for the time. According to the main model, we specify a joint distribution for each student's latent variables. More specifically, for $K = 2$, we model all the student-specific latent variables to be independent and identically distributed across students, following a multivariate normal distribution $(\beta_{0,i}, \beta_{1,i}, \beta_{2,i}, \mu_{\tau,i}, log(\sigma^2_{\tau,i})) \overset{iid}{\sim} MVN(\boldsymbol{\mu}, \boldsymbol{\Sigma}), i = 1, \ldots, N$. As stated above for the main model, this assumption might be relaxed, for instance allowing the students to be grouped into independent clusters. Notwithstanding, in the absence of any particular information, it is reasonable to assume that the students are from the same population and that they are independent. Here, $\boldsymbol{\mu} \in \mathbb{R}^5$ is a five-dimensional vector and $\boldsymbol{\Sigma}$ is a symmetric positive definite 5×5 variance-covariance matrix. This latent growth curves (LGC) modelling might be considered as an extension of the main model introduced in Section 4.2, therefore the properties and the interpretation of $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ are consistent with those introduced above. As for the previous model, the following expected values are fixed $\mathbb{E}[\mu_1], ..., \mathbb{E}[\mu_4] = 0$.
The model 4.2 reduces to the polynomial quadratic approximation for $h_k(t) = t^k$ and $K = 2$:

$$\mu_{\theta,i,t} \quad = \quad \beta_{0,i} + \beta_{1,i} t + \beta_{1,i} t^2, \quad i = 1, \ldots, N; \quad t = 1, \ldots, T. \tag{4.3}$$

This is an approximation which is used to capture non-linear relations. Through the LGC modelling the general ability of each student varies over time and this growth trend is captured. This is a flexible and realistic solution when students are provided with assignments throughout a course

[4] Note that the basis function $h_1(\cdot), \ldots, h_K(\cdot)$ are fixed and known.

since their abilities might considerably change over time.

The prior specifications discussed in Section 4.2 for the unknown parameters, $\delta_1, \ldots, \delta_T$ and $\boldsymbol{\mu}, \boldsymbol{\Sigma}$, might be consistently adopted for the current model. The same procedures and reparametrization used for the posterior computation introduced above might be freely applied here for the multivariate Normal distribution. The Bayesian model comparison procedure discussed in Section 4.2 might be used to compare the models under both the main and LGC framework. As for the previous models, each parameter and student-specific latent variables might be provided with both a posterior point estimate, e.g. the posterior mean, and an interval estimate, e.g. a 95% quantile-based credible interval. Indeed it might be seen as an uncertainty measure of the relative estimated quantity.

4.4 Simulation Studies

The current statistical framework is proposed as a better alternative to the naïve average score system discussed in Section 4.1. For this purpose, we compare the True Score recovery performance of these two approaches under different settings. Two simulation studies are conducted considering both the single and the multiple assignment case. For each of them, different graders reliability and sample size scenarios were designed.

4.4.1 Longitudinal Setting

In the first simulation study the longitudinal setting is considered, i.e. when students receive different coursework assignments over a period of time. The interest here is the accuracy of the True Score estimate under the proposed framework. The objectives of the current study are manyfold.

1. It aims to compare the naïve average score system with the model-based approach introduced above.

2. We want to analyze the role of grader reliability in our estimates. More specifically, whether the accuracy decreases when students are unreliable as graders.

3. We want to know whether a larger number of graders per student might mitigate reliability concerns. Namely, to which extent a larger number of repeated measures of the same student coursework might result in better accuracy.

Manipulated factors and data generation procedure. To address these research questions, we generate the data from the same model, but we assign different values to some quantities across different scenarios. We define a $2 \times 2 \times 2$ experimental design that results from the combination of two levels of the following variables: the average grader reliability level (high vs. low), the number of graders per student and the total number of students.

Data are generated according to the LGC framework presented in Section 4.3. Orthogonal polynomials with a degree of $K = 2$ are used to model the student-specific mean ability at each time. The following quantities are fixed across scenarios. The assignment difficulty parameters are sampled independently from a standard Normal distribution, $\delta_1, \ldots, \delta_T \overset{iid}{\sim} Normal(0, 1)$. Accordingly to the discussion on the identifiability of these parameters in Section 4.2.3. Note that the first 4 elements of the $\boldsymbol{\mu}$, i.e. the expected values of the latent variables, are constrained to be fixed $\mu_1, \ldots, \mu_4 = 0$ for identifiability purposes. The values assigned to the last parameter are discussed below. The latent variables correlation matrix $\boldsymbol{\Omega} = \boldsymbol{I}$ is a 5-dimensional identity matrix and the diagonal matrix of their variances is $\boldsymbol{S} = diag(1, 1, 1, 1, 0.2)$. This implies a unit variance for each latent variable, i.e. $\beta_{0,i}, \beta_{1,i}, \beta_{2,i}, \mu_{\tau,i}$, but $\sigma_5^2 = 0.2$ for $log(\sigma_{\tau,i})$. This is a reasonable choice considering the logarithmic scale of these quantities[5]. The total number of time points (i.e., different assignments) is $T = 8$.

To make different levels of average grader reliability we manipulate the parameter μ_5 of the multivariate Normal distribution of the latent variables to be either $\mu_5 = 1$, low-reliability scenario, or $\mu_5 = -1$, high-reliability scenario. It is the marginal location parameter of the student-specific reliability $\sigma_{\tau,i}$. Higher values imply, on average, less reliable graders (i.e., larger variance between their grades); on the contrary, smaller values imply more reliable graders (i.e., smaller variance between their grades). The other two quantities that vary across different scenarios are the number

[5] Any increase or decrease in the logarithmic scale is expressed in exponential units and it makes this scale non-linear. In other words, a large difference on the original scale might correspond to a small one on the logarithmic scale

of graders per coursework, $|S_{it}| = \{3,6\}$ and the sample size, $N = \{50,200\}$. In sum, for two different sample sizes, $N = 50$ and $N = 200$, four different simulations are performed: LR3 and LR6, low-reliability scenarios where $\mu_5 = 1$, with $|S_{it}| = 3$ and $|S_{it}| = 6$, respectively; HR3 and HR6, high-reliability scenarios where $\mu_5 = -1$, with $|S_{it}| = 3$ and $|S_{it}| = 6$, respectively.

Estimation procedure and model predictive assessment. We fit the LGC model on the several generated data sets. To this regard, the previous prior specifications and posterior computations are employed, see Section 4.2.3. Four independent chains of 5000 iterations were used for the posterior sampling; the first 1000 iterations were used as warm-ups. For each fitted model, we check MCMC mix and convergence through the trace-plot and the $\hat{R}$ index Gelman et al. (2013). The comparison between the naïve average score system and the model-based peer grading is made using the Mean Square Error (MSE). For the naïve approach it is computed as the mean square difference between the average grade given by different graders to the same coursework and the actual True Score. The recovery performance of the model is assessed using the MSE, computed as follows:

$$MSE = \frac{1}{T \cdot N} \sum_{t=1}^{T} \sum_{i=1}^{N} \left[(\hat{\theta}_{it} - \hat{\delta}_t) - (\theta_{it} - \delta_t) \right]^2 .$$

Here θ_{it} is the True Score and $\hat{\theta}_{it}$ is the posterior mean of the Trues Score. The Pearson correlation coefficient r between these two quantities is computed for each scenario.

Results. The model-based peer grading is, on average, ≈ 26.7 times more accurate than the naïve procedure (see Figure 4.1). Our method outperformed the naïve alternative in each scenario. This is supported by an average decrease of MSE values of $\approx -96\%$ from the naïve estimation to the model-based solution. Both methods produce better estimates in high-reliability scenarios.

The True Score recovery accuracy of the naïve system is highly affected by graders' reliability, that is, in a context in which graders are poorly reliable the resulting estimated score is largely different from the actual True Score. As a consequence, in high-reliability scenarios (HR3 and HR6) the estimates are considerably more accurate than the low-reliability ones (LR3 and LR6).

It emerges that when graders are poorly reliable, a larger number of graders per coursework might improve the accuracy of the estimated score. Under the low-reliability scenarios (LR3 and LR6) a larger number of graders per coursework result in substantially better estimates. This is not true under the high-reliability scenarios, a larger number of graders per coursework doesn't improve the accuracy.

A similar trend is shown by our method (see Tables 4.1 and 4.2). Model's estimates are more accurate when graders are more reliable (HR3, HR6) than when they are poorly reliable (LR3 and LR6). On average it implies almost 20 times lower MSE values. Marginally to the reliability level (i.e. under the same reliability scenarios) a larger number of graders per coursework results in better estimates. This implies, on average, a 40% increase in accuracy in terms of MSE for the low-reliability scenarios; but not a notable difference for the high-reliability scenarios. These results are consistent, as expected, with the values of the Pearson correlation r reported in Table 4.1 as well. Similar results are shown across different sample sizes.

4.4.2 Cross-sectional Setting

The second simulation study is devoted to the same research questions as the previous one, but under a cross-sectional setting, i.e. when only one assignment is considered, $T = 1$. We aim to compare the naïve average score system with the model-based approach introduced in Section 4.2.2. We also focus on the consequences of highly unreliable graders on the True Score recovery performance in cross-sectional settings. In this regard, we investigate the marginal role of the number of graders per coursework on estimation accuracy.

Manipulated factors and data generation procedure. The data-generating process and the experimental design are consistent with the previous simulation and adapted here to the cross-sectional setting. Accordingly, the graders' reliability level, μ_3, the number of graders per coursework, $|S_i|$ and the sample size, N, are manipulated.

We generate the data from the one assignment model specified in Section 4.2.2. In this case, we only have three student-specific latent variables because we don't have any time trend. The assignment difficulty parameter is not identifiable, as discussed above, thus it is fixed to zero in

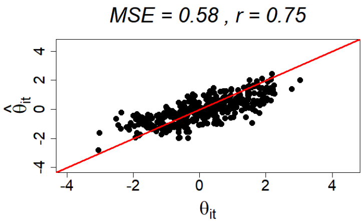
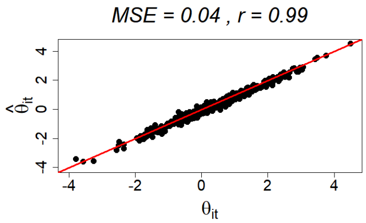
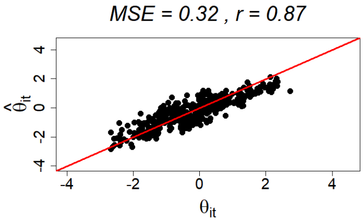
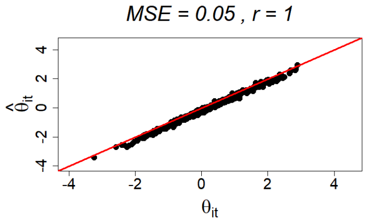
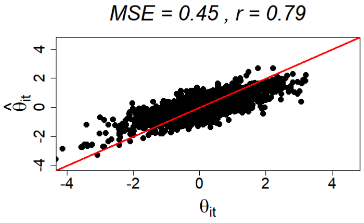
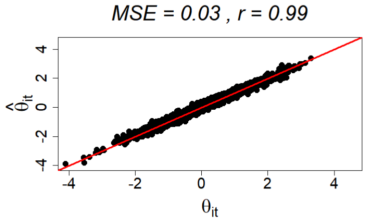
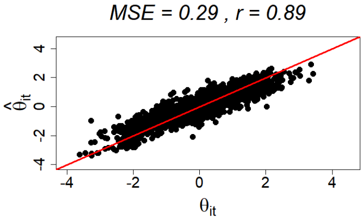
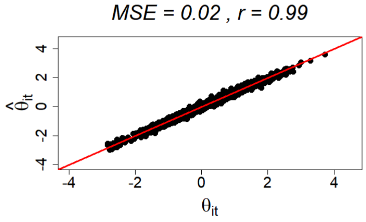
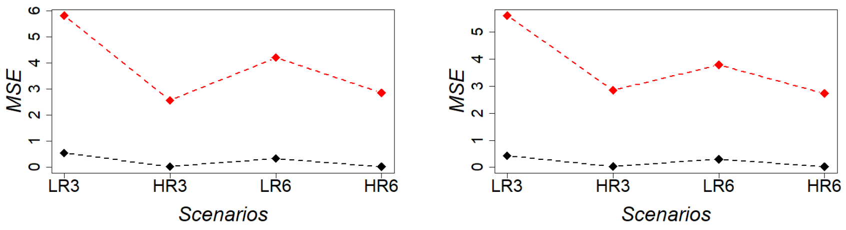
	$\mu_5 = 1$	$\mu_5 = -1$
N = 50		
$\lvert S_{it}\rvert = 3$	 	
$\lvert S_{it}\rvert = 6$	 	
N = 200		
$\lvert S_{it}\rvert = 3$	 	
$\lvert S_{it}\rvert = 6$	 	

Table 4.1: For each simulation scenario the true value of θ_{it} against the posterior mean estimate $\hat{\theta}_{it}$ is plotted. A 45-degree line is plotted to highlight possible under- or over-estimate trends. Here μ_5 is the mean grader reliability, smaller values imply higher reliable graders; $|S_{it}|$ is the number of graders per coursework. The sample size, i.e. the total number of students, is indicated by N.

	Reliability	Graders	Naïve system	Model based
$N = 50$				
			MSE	MSE
	Low	3	5.83	0.58
		6	4.21	0.32
	High	3	2.56	0.04
		6	2.85	0.05
$N = 200$				
			MSE	MSE
	Low	3	5.62	0.45
		6	3.79	0.29
	High	3	2.85	0.03
		6	2.73	0.02

Table 4.2: The Mean Square Error (MSE) values of both the naïve and the model-based grading systems for each simulation are reported. They are divided into low ($\mu_5 = 1$) vs. high ($\mu_5 = -1$) reliability scenarios. In turn, different cardinality values are chosen for the graders' subsets, $|S_{it}| = 3$ vs. $|S_{it}| = 6$.



(a)

(b)

Figure 4.1: The MSE of both the naïve (red line) and the model-based systems (black line) are plotted. LR3 and LR6 are the low-reliability scenarios, i.e. $\mu_5 = 1$, with $|S_{it}| = 3$ and $|S_{it}| = 6$ graders per coursework, respectively. HR3 and HR6 stand for High-reliability scenarios, i.e. $\mu_5 = -1$, with $|S_{it}| = 3$ and $|S_{it}| = 6$ graders per coursework, respectively. The values of MSEs for different sample sizes, $N = 50$ and $N = 200$, are shown in, respectively, in sub-figure (a) and (b).

the data-generating process. The following quantities are fixed across scenarios. The assignment difficulty level is fixed to zero $\delta = 0$. The expected values of the first two latent variables are fixed, $\mu_1, \mu_2 = 0$, for identifiability purposes. The latent variables correlation matrix $\boldsymbol{\Omega} = \boldsymbol{I}$ is a 3-dimensional identity matrix, and the diagonal matrix of their variances is $\boldsymbol{S} = diag(1, 1, 0.2)$. This implies a unit variance for the first two latent variables, i.e. $\mu_{\theta,i}, \mu_{\tau,i}$. A variance of $\sigma_3^2 = 0.2$ is assigned for the third student-specific latent variable $log(\sigma_{\tau,i})$. This is consistent with the choice made in the previous simulation.

The $2 \times 2 \times 2$ design results from the combination of the values we assign to three different quantities. We manipulate the overall graders' reliability level through the parameter μ_3. This is the mean student-specific reliability, higher values imply a lower average grader reliability, and smaller values imply higher grader reliability. We fix it to be $\mu_3 = 1$, for the lower reliability scenarios, and $\mu_3 = -1$ for the higher ones. The other two quantities that vary across different scenarios are the number of graders per coursework, $S_{it} = \{3, 6\}$ and the sample size, $N = \{50, 200\}$. Summarize, for two different sample sizes, $N = 50$ and $N = 200$, four different simulations are performed: LR3 and LR6, low reliability scenarios where $\mu_3 = 1$, with $|S_i| = 3$ and $|S_i| = 6$, respectively; HR3 and HR6, high reliability scenarios where $\mu_3 = -1$, with $|S_i| = 3$ and $|S_i| = 6$, respectively.

Estimation procedure and model predictive assessment. Accordingly, in the previous study, we fit the cross-sectional model to the data sets generated under the different scenarios. The full Bayesian procedure introduced above is adopted here. The prior specification and the posterior computations are considered consistently in Section 4.2.3. For the posterior sampling, four independent chains of 3000 iterations are used; the first 1000 iterations are used as a warm-up. We check the MCMC mix and convergence through the trace plot and the $\hat{R}$ index Gelman et al. (2013).

The comparison between the naïve average score system and the proposed model-based grading is made using the MSE. For the naïve system it is computed as described in Section 4.4.1 for the longitudinal setting. Here we assess the accuracy of the True Score estimation computing the

MSE between the posterior mean $\hat{\theta}_i$ and the true value θ_i of this quantity as follows:

$$MSE \quad = \quad \frac{1}{N}\sum_{i=1}^{N}\left(\hat{\theta}_i - \theta_i\right)^2.$$

The Pearson correlation coefficient r between these two quantities is computed for each scenario.

Results. The proposed single assignment peer grading is, on average, more than three times more accurate than the naïve score system (see Figure 4.2). The current model provides better True Score estimates than the latter in each simulation scenario. Our method decreases by $\approx -70\%$ the MSE values of the naïve procedure. Both methods produce better estimates in high-reliability scenarios. The True Score recovery accuracy of the naïve system is highly affected by graders' reliability. In high-reliability scenarios (HR3 and HR6) the estimates are considerably more accurate and closer to those of our model. Under the low and the high-reliability scenarios (LR3, LR6 and HR3, HR6, respectively) an increasing number of graders per coursework results in higher accuracy levels. It means that using the naïve method a larger number of graders per coursework results in a considerably better estimation of the True Score.

A similar trend is shown by our method (see Tables 4.3 and 4.4). Model's estimates are more accurate when graders are more reliable (HR3, HR6) than when they are poorly reliable (LR3 and LR6). On average, it implies approximately 10 times lower MSE values. Marginally to the reliability level (i.e. under the same reliability scenarios) a larger number of graders per coursework results in better estimates. This implies, on average, a $\approx 50\%$ increase in accuracy in terms of MSE. These results are consistent with the values of the Pearson correlation r reported in Table 4.3 as well. Similar results are shown across different sample sizes.

4.5 Real Data Examples

A real-world application is presented for illustrative purposes. In this section, the longitudinal setting is considered. The data set is analyzed as a case of the multiple assignment context. Different models are compared and each student is provided with a Score estimate for each assignment as

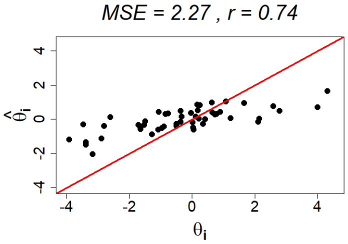
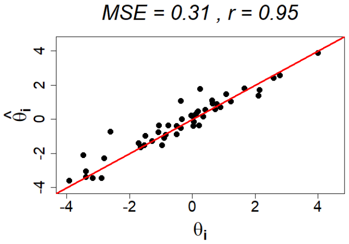
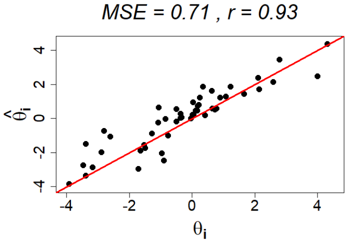
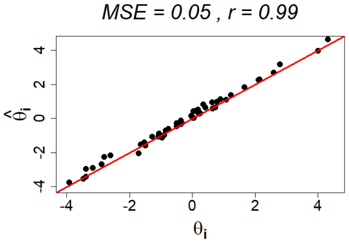
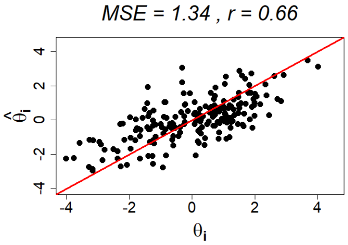
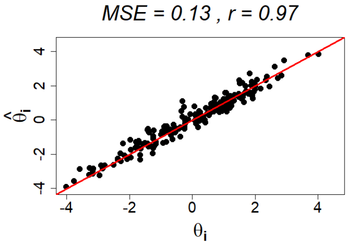
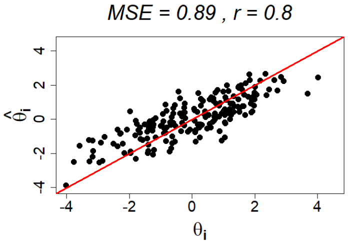
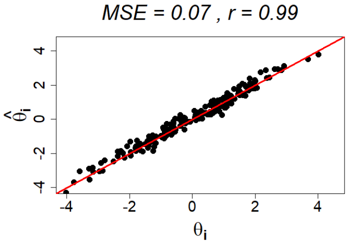
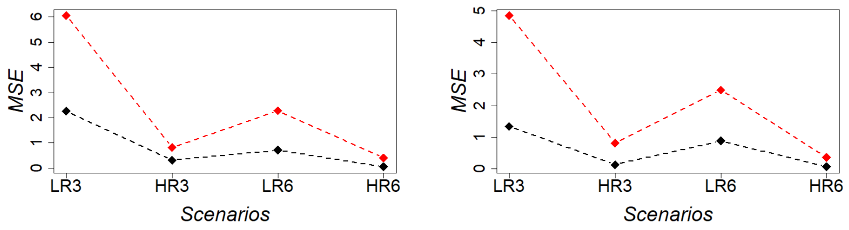
	$\mu_5 = 1$	$\mu_5 = -1$
N = 50		
$\lvert S_i \rvert = 3$	 *MSE = 2.27 , r = 0.74* 	 *MSE = 0.31 , r = 0.95*
$\lvert S_i \rvert = 6$	 *MSE = 0.71 , r = 0.93* 	 *MSE = 0.05 , r = 0.99*
N = 200		
$\lvert S_i \rvert = 3$	 *MSE = 1.34 , r = 0.66* 	 *MSE = 0.13 , r = 0.97*
$\lvert S_i \rvert = 6$	 *MSE = 0.89 , r = 0.8* 	 *MSE = 0.07 , r = 0.99*

Table 4.3: For each simulation scenario, the true value of θ_i against the posterior mean estimate $\hat{\theta}_{it}$ is plotted. A 45-degree line is plotted to highlight possible under- or over-estimate trends. Here μ_5 is the mean grader reliability, smaller values imply higher reliable graders; $|S_i|$ is the number of grader per coursework. The sample size, i.e. the total number of students, is indicated by N.

	Reliability	Graders	Naïve system	Model based
$N = 50$				
			MSE	MSE
	Low	3	6.06	2.27
		6	2.27	0.71
	High	3	0.82	0.31
		6	0.40	0.05
$N = 200$				
			MSE	MSE
	Low	3	4.84	1.34
		6	2.49	0.89
	High	3	0.82	0.13
		6	0.36	0.07

Table 4.4: The Mean Square Error (MSE) values of both the naïve and the model-based grading systems for each simulation are reported. They are divided into low ($\mu_5 = 1$) vs. high ($\mu_5 = -1$) reliability scenarios. In turn, different cardinality values are chosen for the graders' subsets, $|S_i| = 3$ vs. $|S_i| = 6$.



(a) (b)

Figure 4.2: The MSE of both the naïve and the model-based systems are plotted. LR3 and LR6 are the low-reliability scenarios, i.e. $\mu_3 = 1$, with $|S_{it}| = 3$ and $|S_{it}| = 6$ graders per coursework, respectively. HR3 and HR6 stand for High-reliability scenarios, i.e. $\mu_3 = -1$, with $|S_i| = 3$ and $|S_i| = 6$ graders per coursework, respectively. The values of MSEs for different sample sizes, $N = 50$ and $N = 200$, are shown in, respectively, sub-figure (a) and (b).

an examinee, and other quantities are used to assess the student as a grader.

4.5.1 Multiple Assignment Setting

These peer grading data are from Zong et al. (2021). Participants are $N = 274$ American undergraduate students attending a Biology course. A double-blinded individual peer assessment was implemented for four different assignments, $T = 4$ throughout the course. Each student was graded, on average, by a random set of 5 other students for each assignment. The coursework was rated on a $1-7$ Likert scale with instructor-provided anchor descriptions for each rating level. For the current analysis, only the students who completed at least three assignments were considered. This allows us to fit the LGC model. It results in a sample size of $N = 212$ students.

Model comparison. Five different models of increasing complexity are compared. In the first model, we only specify one student-specific latent variable and the assignment difficulty level (M1). This implies an equal ability of the student across assignments. We model also the rater effect in the second model specifying other two latent variables for each student, i.e. their bias and the reliability (M2). The student-specific latent variables are jointly modelled specifying a multivariate Normal distribution. This makes this model close to the cross-sectional one. In the third model, we let the student's ability vary across different assignments, and a fourth latent variable is introduced. This is the main model presented in Section 4.2. The LGC model presented in Section 4.3 is specified as the most complex one (M4). Two degrees orthogonal polynomials are used for the time coding to fit M4. We compared these models with that proposed by Piech et al. (2013)(hereafter called MP). The model comparison is based on the predictive performance criteria discussed in Section 4.2.3.

The prior specifications and the posterior procedure discussed in Section 4.2.3 are employed to fit the models. Given that grading is on a $1-7$ Likert scale, we specify as location parameter for the prior of the assignment difficulties the centre of this scale. Each student is then provided with an estimate of the True Score for each assignment and a reliability index as a grader.

Results. From the model comparison procedure, it emerges that the LGC model (M4) is the best in terms of predictive performance. The value of the leave-one-out expected log point-wise density $elpd_{loo}$ is reported in Table 4.5 for each considered model, coupled with the relative standard error. The pairwise difference between M4 and each other model in terms of $elpd_{loo}$ is shown. Considering the small number of time points it is somehow expected that M4 over-performed M3. They might be too few to capture each student's latent growth. The estimates from M4 are here considered to make inferences and presented below.

The assignment difficulty levels seem to be in increasing order (see Figure 4.3 and Table 4.6). The first assignment is the easiest one, the fourth is the most difficult one and the others have a medium level of difficulty. The 95% quantile-based credible intervals of the assignment difficulty parameters are moderately narrow, suggesting a low level of uncertainty for these parameters. These values are reported in Table 4.6.

Students' latent variables parameters suggest that, on average, each student's True Scores varies considerably across different assignments. This information is borrowed by the parameter μ_4 (see Table 4.6). Note that the fourth latent variable of the multivariate Normal is the variance of the error from the latent growth curve approximation. Considering the parameter μ_6, it emerges that they are, on average, moderately reliable graders. The low values of σ_1^2 and σ_5^2 suggest that students are very homogeneous both in their overall ability as an examinee and in their bias as a grader. They are slightly different in terms of consistency across assignment σ_4^2 and reliability σ_6^2. Note that these values are on a logarithmic scale, this must be taken into account in their interpretation.

The large negative values of ω_{14}, the correlation between $\beta_{0,i}$ and $log(\sigma^2_{\theta,i})$, suggests that students with higher overall ability have smaller variance as examinees over time. A strong correlation between the bias, $\mu_{\tau,i}$, and the reliability, $log(\sigma_{\tau,i})$, of the same grade is indicated by ω_{56}. More severe students tend to have a smaller variance as a grader. Both the results seem to be reasonable from a substantive point of view. Concerning the other correlation parameters, their $95\%CI$ includes zero, suggesting a less clear relation between the respective quantities. The only exception is ω_{13} whose values indicate a moderate negative correlation between the mean of the ability of a student $\beta_{0,i}$ and its quadratic growth $\beta_{2,i}$.

Going to the student-specific level, a latent growth curve might be estimated. As reported in Figure 4.4, for each student a non-linear trend is captured over time. It indicates the latent growth of the student in the overall ability (as examinee) during the course.

Moreover, each student might be provided with a Score estimate for each assignment. It might be regarded as an *official grade*. For explanatory purposes, some examples of them are reported in Figure 4.5. As a measure of uncertainty, the 95% quantile-based credible intervals are reported. The posterior mean of $\hat{\theta}_{it} - \delta_t$ might be a point estimate for students' Scores. It might be used for grading purposes.

The posterior distributions of both the average bias and the reliability of each grader might be useful information to assess their grading behaviour. An accurate and reliable grader should have both $\mu_{\tau,i}$ and $\sigma_{\tau,i}$ close to zero. On the contrary, values far from zero indicate respectively, a biased and unreliable grading behavior. In Figures 4.6 and 4.7 the posterior distribution of these two quantities (the average bias and the reliability) of four different grades are reported. As a measure of uncertainty for both quantities, the 95% quantile-based credible intervals are reported.

Model	$elpd_{loo}$	SE	$\Delta elpd_{loo}$	$SE\Delta$
M4	-3745.2	49.9	–	–
M3	-3794.9	50.3	-2.0	4.4
M2	-4070.3	55.8	-325.1	35.6
M1	-4470.4	54.2	-725.2	47.0
MP	-4794.8	46.1	-1049.6	45.5

Table 4.5: Five different model specifications are compared through a leave-one-out cross-validation approach. For each model, the expected log point-wise density value ($elpd_{loo}$) and the respective standard error (SE) are reported. Models are in decreasing order concerning the $elpd_{loo}$. In the last two columns, the pairwise comparisons between each model and the model with the largest $elpd_{loo}$ (M4) are reported; $\Delta elpd_{loo}$ is the difference between M3 and each of the other models, $SE\Delta$ is the standard error of the difference (which should not be expected to equal the difference of the standard errors).

	Parameter	95% CI	Parameter	95% CI
Assignments	δ_1	$(-6.42, -6.27)$	δ_3	$(-5.41, -5.26)$
	δ_2	$(-5.45, -5.30)$	δ_4	$(-5.05, -4.92)$
Students	μ_4	$(3.01, 4.21)$	μ_6	$(1.51, 1.65)$
	σ_1^2	$(0.20, 0.28)$	σ_4^2	$(0.61, 0.99)$
	σ_2^2	$(0.02, 0.22)$	σ_5^2	$(0.32, 0.39)$
	σ_3^2	$(0.14, 0.31)$	σ_6^2	$(0.29, 0.37)$
	ω_{12}	$(-0.02, 0.73)$	ω_{26}	$(-0.24, 0.61)$
	ω_{13}	$(-0.86, -0.38)$	ω_{34}	$(-0.06, 0.69)$
	ω_{14}	$(-0.92, -0.59)$	ω_{35}	$(-0.41, 0.12)$
	ω_{15}	$(-0.26, 0.08)$	ω_{36}	$(-0.28, 0.29)$
	ω_{16}	$(-0.03, 0.32)$	ω_{45}	$(-0.23, 0.18)$
	ω_{23}	$(-0.74, 0.27)$	ω_{46}	$(-0.12, 0.33)$
	ω_{24}	$(-0.61, 0.41)$	ω_{56}	$(-0.83, -0.63)$
	ω_{25}	$(-0.64, 0.17)$	-	-

Table 4.6: Model M4 parameters. For each Assignment difficulty parameter and each student's latent variables distribution parameter the 95% quantile-based credible interval (CI) is reported. The parameter δ_t is the difficulty level of the assignment t; μ_2 and μ_4 are the location parameters of the second and the fourth latent variables. Here, $\sigma_1^2, \ldots, \sigma_6^2$ are their variances; ω_{mn} is the correlation parameter between the latent variable m and n.

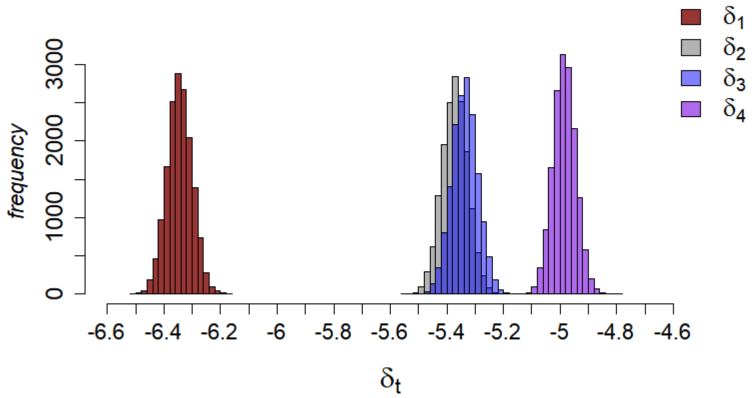
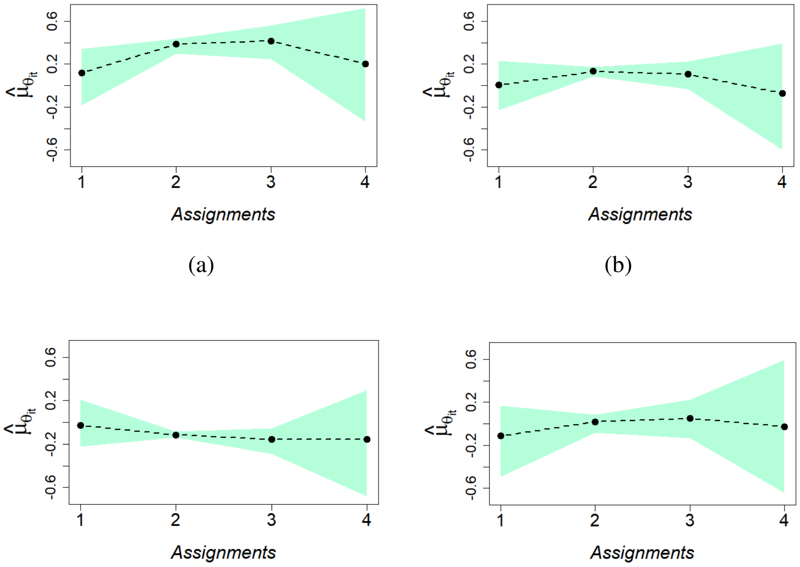


Figure 4.3: The posterior distributions of the difficulty parameters, $\delta_1 \ldots, \delta_4$, are reported.



(a)

(b)

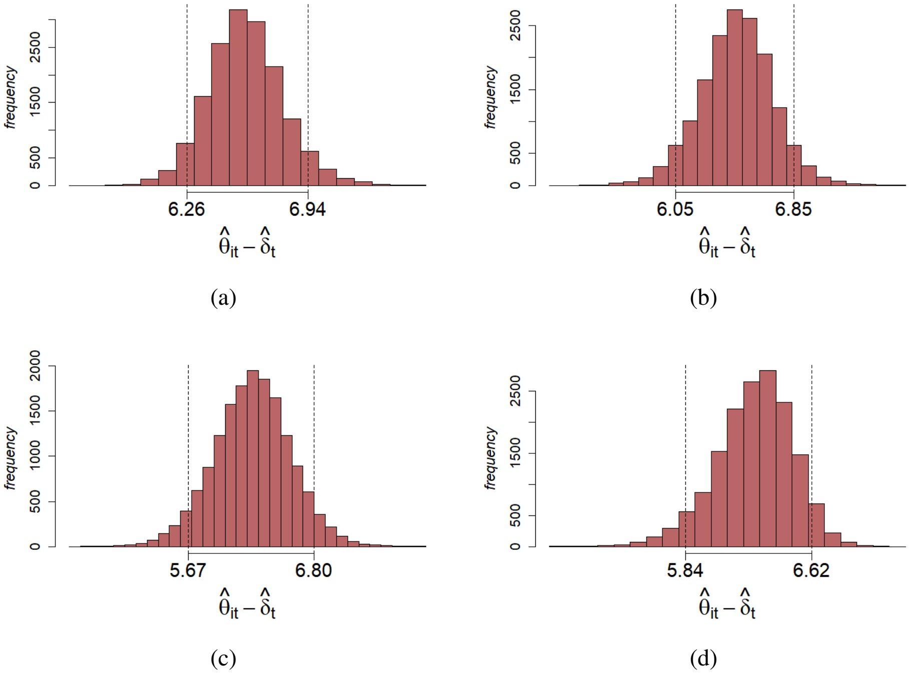
(c)

(d)

Figure 4.4: The latent growth of the overall ability $\hat{\mu}_{\theta_{it}}$ of four different students is reported for illustrative purposes. Here $\hat{\mu}_{\theta_{it}}$ is the estimated average ability of i in doing the assignment t. The posterior mean is indicated by a black dot and the 95% credible interval is indicated by the light green shaded area around each black dot.

4.6 Discussions

We introduced a broad statistical framework for peer grading data. It relates two different psychometric model classes: the models for measuring students' latent ability (van der Linden, 2016; Reckase, 1997; Birnbaum, 1969) and those for modelling rater effect (Martinková et al., 2023; Casabianca et al., 2015; DeCarlo, 2008). It comes as a consequence of the peer grading data structure since each student is both an examinee and as a grader. This twofold role of the students makes each observation (i.e. each grade) related to both the features of the examinee and those of the grader. For this reason, the current statistical modelling might be considered as a cross-classified or crossed random effect model (Goplerud, 2022; Jeon et al., 2017; Cho and Rabe-Hesketh, 2011). This poses our framework close to the lines of Social Relations Models (Nestler et al., 2020, 2017) and Dyadic IRT (Murphy, 2021; Gin et al., 2020). In contrast to these models, we jointly model the student-specific latent variables and the graders' reliability (i.e., we let the residual variance depend on the grader). Furthermore, we don't have a Round-Robin structure, since each student's



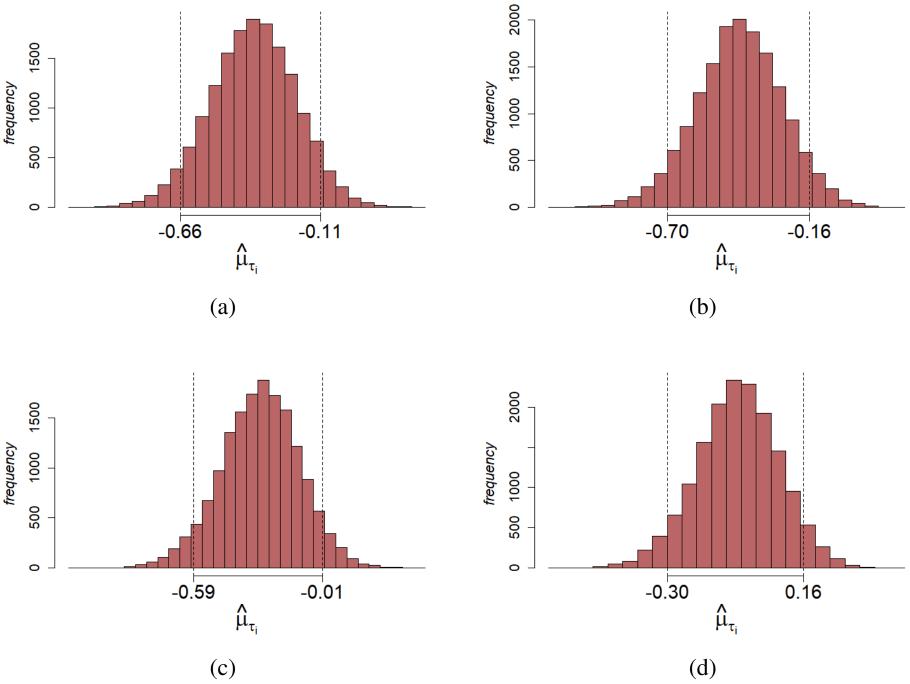
(a)

(b)

(c)

(d)

Figure 4.5: The posterior distribution of the score $\theta_{i,t} - \delta_t$ is reported for four different examinees. These quantities are referred to the first assignment, i.e. $t = 1$. The vertical dotted lines indicate the 95% quantile-based credible interval (CI).



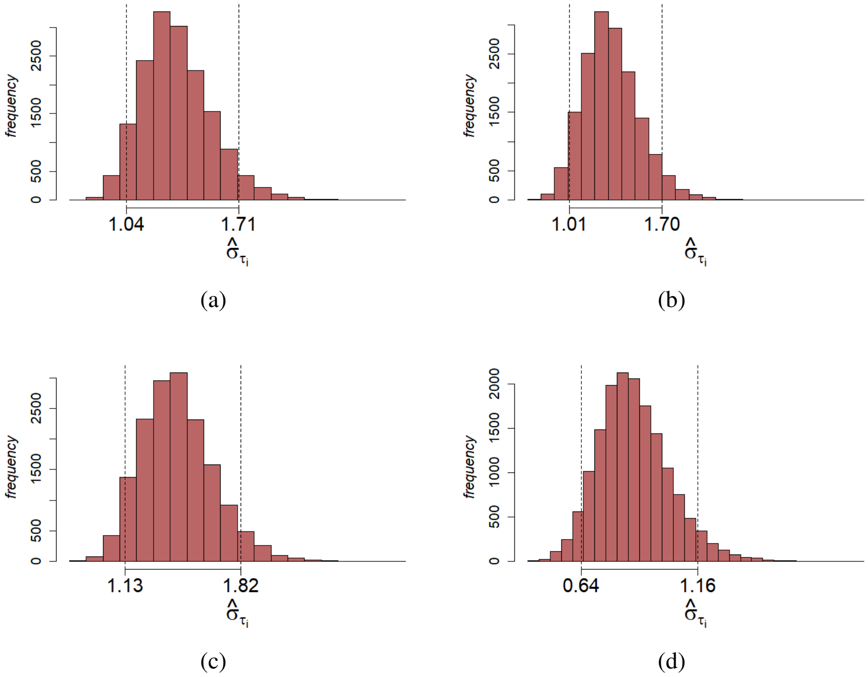
(a)

(b)

(c)

(d)

Figure 4.6: The posterior distribution of the average bias $\mu_{\tau,i}$ is reported for four different graders. The vertical dotted lines indicate the 95% quantile-based credible interval (CI).



(a)

(b)

(c)

(d)

Figure 4.7: The posterior distribution of the standard deviation of the bias $\sigma_{\tau,i}$ is reported for four different graders. The vertical dotted lines indicate the 95% quantile-based credible interval (CI).

coursework is generally not graded by all the other students.

Different settings might be considered under the proposed modelling. It might be applied both to single and multiple assignments data. In both settings, it enables to provide each student with a more accurate score as an examinee and a reliability index as a grader. Here the focus is on the continuous case (i.e., when grades are on a continuous scale), but the extension to the ordinal data is straightforward. The effect of explanatory variables might be considered in the model specification. It might be useful for research purposes in educational contexts. This notwithstanding, when the model is used as a grading system it might raise ethical concerns. For this reason, we didn't consider any covariates in this chapter.

The simulations highlight that the current framework considerably outperforms the naïve system in both the longitudinal and the cross-sectional settings. The degree of accuracy of our True Score estimates might depend on the peer grading procedure (e.g. the number of graders per coursework, the size of the course class) or the reliability of the students as graders. It might be appreciably informative in many educational contexts.

Future work might aim to model student heterogeneity in both roles. Possible latent classes of students might be found both in terms of examinee and grader. It might be a viable solution to find malevolent grading behaviours and exclude the effect of some graders from the grading system. Another important point to be addressed might be model scalability. More efficient posterior sampling schemes for crossed-random effects might be used (Papaspiliopoulos et al., 2020). Possible extensions of these models might regard multidimensional grades, i.e. when each coursework is scored on different rubrics or aspects. The presence of items (i.e. the rubrics) might be accounted for under a latent variables approach (Bartholomew et al., 2011).

Chapter 5

Conclusion

The Bayesian nonparametric (Chapters 2 and 3) and the parametric (Chapter 4) frameworks introduced in the present work aim to be useful statistical tools in the field of ratings. Their features, broadly discussed in the previous chapters, make them very flexible and applicable to a wide rage of contexts. The mixture polarization index might be a viable solution in assessing raters heterogeneity. It quantifies the *degree of polarization* between different groups of raters. These clusters naturally arises from the DPM priors, which accommodates for this sort of heterogeneity. The λ polarization index is based on the density distribution of the raters effect and it doesn't require any parametric assumption. This makes it suitable to a very broad class of statistical models and widely applicable in many contexts. For instance, it might be used as a latent bias polarization index in political psychology (in analysing differences between right- and left-wing people), in educational psychology (to assess any possible bias differences among teachers) or in experimental psychology (to quantify differences in stimuli sensitivity)

Further work may aim to clearly link the framework presented in Chapters 2 and 3 to that discussed in Chapter 4. The peer grading modeling might be extended and a nonparametric modelling of some effects might be possible. Nonetheless, this extension needs to be carefully considered both from a statistical and substantive point of view. The nonparametric distribution might refer to the rating behaviors (i.e., the rater/grader effects). Since some dependency between the two-fold role of each student is modelled (as an examinee and as a grader), it might be not clear if the clustering might concerns the two features together or separately. If a nonparametric distribution is specified over all the student-specific latent variables the clustering is among students (in both roles). On the contrary, if it is placed only over a single role (e.g., the rating/grading behavior) it might require a different modelling of the dependency between the examinee and the grader's features. In the present peer grading model, this features (i.e., the student-specific latent variables) are assumed to be independent and identically distributed across students and following a multivariate normal distribution. In case of a nonparametric modelling, this needs to be modified accordingly.

From a substantive perspective these frameworks are very informative. They capture different specific aspects of the hetero-evaluation which are meaningful from a psychological point of view. The interpretation of the shape of the nonparametric distribution of raters effect and the consideration of the polarization index are crucial information. As discussed in Chapter 1, the rating process might be dramatically affected by several aspects. Not always they are meant to be spurious or something to get rid of. They might concern something involved in the rating process itself and need to be carefully considered both from a statistical and a methodological perspective. This work focuses on the first aim. We attempted to address all the points raised in Chapter 1. The elements of heterogeneity among different raters or observers might hide substantive discrepancies and divergences. For better and fairer rating processes, they need to be considered and modelled. In this way we aim to make this heterogeneity something beyond mere noise, but a tool to deepen substantive topics.

The hetero-evaluation is gaining much attention in several psychological fields (Lee et al., 2022; Hou et al., 2020; Levinson et al., 2017; Cho and Rabe-Hesketh, 2011). The demand for fairer and more fruitful rating systems is increasing across psychological fields. Educational psychologists are looking for more accurate and formative grading systems (Shengkai Yin and Chang, 2022; Panadero and Alqassab, 2019). Work psychologists are focusing on more informative method of performance rating (Salgado and Moscoso, 2019; Schneider et al., 2019; Strahl et al., 2019). Test developers are focusing on new rating methods to assess test content validity (Spoto et al., 2023). The methods presented in the previous Chapters might be viable statistical solutions for most of these demands. The estimated quantities and the relative indexes (λ, ICC_a) might be interpreted and, in some cases, removed from the final score.

The present work aims to be a further step in deepen the crucial importance of rating process and it has proved to be an effective way to address the complexity of the latter and shed light on the whole evaluation process. We believe the methods presented might borrow precious information for a deep discussion and consideration of rater heterogeneity in several scientific fields. When some observers are asked to evaluate the same set or different sets of items (e.g., subjects, objects) the proposed framework might be a valuable solution to get some precious information about both the raters and the items. For instance, it might be applied to the educational setting to get some estimates of teachers' bias and reliability; it might be helpful to identify too much strict or lenient grading behaviors and eventually to remove their effect from the observed score (Nucci et al., 2021). The indices λ and ICC might be used to quantify the polarization and the relevance of the teachers' bias, respectively. The very same procedure might be applied to any rating context with a proper interpretation of the proposed model and indices.

Bibliography

Agresti, A. (2015). *Foundations of Linear and Generalized Linear Models*. John Wiley Sons, INC.

Alavi, M., Biros, E., and Cleary, M. (2022). A primer of inter-rater reliability in clinical measurement studies: Pros and pitfalls. *Journal of Clinical Nursing*, 31:e39–e42.

Albert, J. H. and Chib, S. (1993). Bayesian analysis of binary and polychotomous response data. *Journal of the American Statistical Association*, 88(422):669–679.

Alonderiene, R. and Majauskaite, M. (2016). Leadership style and job satisfaction in higher education institutions. *International Journal of Educational Management*, 30(1):140–164.

Alotaibi, A., Alghamdi, A., Reynard, C., and Body, R. (2021). Accuracy of emergency medical services (ems) telephone triage in identifying acute coronary syndrome (acs) for patients with chest pain: A systematic literature review. *BMJ Open*, 11.

Alqassab, M., Strijbos, J.-W., Panadero, E., Ruiz, J. F., Warrens, M., and To, J. (2023). A systematic review of peer assessment design elements. *Educational Psychology Review*, 35:18.

Amewou-Atisso, M., Ghosal, S., Ghosh, J. K., and Ramamoorthi, R. V. (2003). Posterior consistency for semi-parametric regression problems. *Bernoulli*, 9.

Anderson, J. R. and Crawford, J. (1980). *Cognitive psychology and its implications*. W.H. Freeman, San Francisco.

Antoniak, C. E. (1974). Mixtures of Dirichlet Processes with Applications to Bayesian Nonparametric Problems. *The Annals of Statistics*, 2(6):1152 – 1174.

Araújo, P., Kirkwood, R., and Figueiredo, E. (2009). Validity and intra-and inter-rater reliability of the observational gait scale for children with spastic cerebral palsy. *Brazilian Journal of Physical Therapy*, 13:267–273.

Ascolani, F., Lijoi, A., Rebaudo, G., and Zanella, G. (2023). Clustering consistency with dirichlet process mixtures. *Biometrika*, 110.

Barbot, B., Hein, S., Luthar, S. S., and Grigorenko, E. L. (2014). Capturing age-group differences and developmental change with the basc parent rating scales. *Journal of applied developmental psychology*, 35(4):294–303.

Barneron, M., Allalouf, A., and Yaniv, I. (2019). Rate it again: Using the wisdom of many to improve performance evaluations. *Journal of Behavioral Decision Making*, 32(4):485–492.

Bartholomew, D., Knott, M., and Moustaki, I. (2011). *Latent Variable Models and Factor Analysis: A Unified Approach*.

Bartoš, F., Martinková, P., and Brabec, M. (2019). Testing heterogeneity in inter-rater reliability. In *The Annual Meeting of the Psychometric Society*, pages 347–364. Springer.

Beach, C., Boyce, W., Peat, M., and Malakar, S. (1995). Inter-rater reliability of a paediatric outcome measure in nepal. *International Journal of Rehabilitation Research*, 18.

Betancourt, M. and Girolami, M. (2013). Hamiltonian monte carlo for hierarchical models.

Birnbaum, A. (1969). Statistical theory for logistic mental test models with a prior distribution of ability. *Journal of Mathematical Psychology*, 6(2):258–276.

Blackwell, D. and MacQueen, J. B. (1973). Ferguson Distributions Via Polya Urn Schemes. *The Annals of Statistics*, 1(2):353 – 355.

Bohm, K. and Kurland, L. (2018). The accuracy of medical dispatch - a systematic review.

Bonefeld, M. and Dickhäuser, O. (2018). (biased) grading of students' performance: Students' names, performance level, and implicit attitudes. *Frontiers in Psychology*, 9.

Bouchard-Côté, A., Doucet, A., and Roth, A. (2017). Particle gibbs split-merge sampling for bayesian inference in mixture models. *Journal of Machine Learning Research*, 18:1–39.

Briesch, A., Hemphill, E., Volpe, R., and Daniels, B. (2014). An evaluation of observational methods for measuring response to classwide intervention. *School psychology quarterly : the official journal of the Division of School Psychology, American Psychological Association*, 30.

Brookhart, S. M. (1993). Teachers' grading practices: Meaning and values. *Journal of Educational Measurement*, 30.

Bygren, M. (2020). Biased grades? changes in grading after a blinding of examinations reform. *Assessment & Evaluation in Higher Education*, 45(2):292–303.

Canale, A. and Blasi, P. D. (2017). Posterior asymptotics of nonparametric location-scale mixtures for multivariate density estimation. *Bernoulli*, 23.

Canale, A. and Dunson, D. B. (2011). Bayesian kernel mixtures for counts. *Journal of the American Statistical Association*, 106(496):1528–1539. PMID: 22523437.

Canale, A. and Prünster, I. (2017). Robustifying bayesian nonparametric mixtures for count data. *Biometrics*, 73(1):174–184.

Cao, J., Stokes, S. L., and Zhang, S. (2010). A bayesian approach to ranking and rater evaluation: An application to grant reviews. *Journal of Educational and Behavioral Statistics*, 35(2):194–214.

Casabianca, J. M., Lockwood, J. R., and Mccaffrey, D. F. (2015). Trends in classroom observation scores. *Educational and Psychological Measurement*, 75:311–337.

Cassata, R., Gianini, G., Anisetti, M., Bellandi, V., Damiani, E., and Cavaciuti, A. (2022). Interrater agreement based risk assessment scheme for ict corporates. volume 309.

Childs, T. M. and Wooten, N. R. (2023). Teacher bias matters: an integrative review of correlates, mechanisms, and consequences. *Race Ethnicity and Education*, 26(3):368–397.

Chin, M. J., Quinn, D. M., Dhaliwal, T. K., and Lovison, V. S. (2020). Bias in the air: A nationwide exploration of teachers' implicit racial attitudes, aggregate bias, and student outcomes. *Educational Researcher*, 49(8):566–578.

Cho, S.-J. and Rabe-Hesketh, S. (2011). Computational statistics and data analysis alternating imputation posterior estimation of models with crossed random effects. *Computational Statistics and Data Analysis*, 55:12–25.

Cicchetti, D. V. (1976). Assessing inter-rater reliability for rating scales: Resolving some basic issues. *British Journal of Psychiatry*, 129(5):452–456.

Cooper, C. W. (2003). The detrimental impact of teacher bias: Lessons learned from the standpoint of african american mothers. *Teacher Education Quarterly*, 30(2):101–116.

Crimmins, G., Nash, G., Oprescu, F., Alla, K., Brock, G., Hickson-Jamieson, B., and Noakes, C. (2016). Can a systematic assessment moderation process assure the quality and integrity of assessment practice while supporting the professional development of casual academics? *Assessment & Evaluation in Higher Education*, 41(3):427–441.

Dahlin, J., Kohn, R., and Schön, T. B. (2016). Bayesian inference for mixed effects models with heterogeneity.

Dalwai, M., Tayler-Smith, K., Twomey, M., Nasim, M., Popal, A. Q., Haqdost, W. H., Gayraud, O., Cheréstal, S., Wallis, L., and Valles, P. (2018). Inter-rater and intrarater reliability of the south african triage scale in low-resource settings of haiti and afghanistan. *Emergency Medicine Journal*, 35(6):379–383.

De la Cruz-Mesia, R. and Marshall, G. (2006). Non-linear random effects models with continuous time autoregressive errors: a bayesian approach. *Statistics in Medicine*, 25(9):1471–1484.

Deakin, C. D., England, S., and Diffey, D. (2017). Ambulance telephone triage using 'nhs pathways' to identify adult cardiac arrest. *Heart*, 103(10):738–744.

DeCarlo, L. T. (2008). Studies of a latent-class signal-detection model for constructed-response scoring. *ETS Research Report Series*, 2008(2):i–55.

Dee, T. S. (2005). A teacher like me: Does race, ethnicity, or gender matter? *American Economic Review*, 95(2):158–165.

DiMaggio, P., Evans, J., and Bryson, B. (1996). Have american's social attitudes become more polarized? *American Journal of Sociology*, 102(3):690–755.

Dorazio, R. M. (2009). On selecting a prior for the precision parameter of dirichlet process mixture models. *Journal of Statistical Planning and Inference*, 139(9):3384–3390.

Double, K. S., Mcgrane, J. A., and Hopfenbeck, T. N. (2016). The impact of peer assessment on academic performance: A meta-analysis of control group studies. *Hattie and Timperley*.

Drake, S., Auletto, A., and Cowen, J. M. (2019). Grading teachers: Race and gender differences in low evaluation ratings and teacher employment outcomes. *American Educational Research Journal*, 56.

Dressler, W. W., Balieiro, M. C., and dos Santos, J. E. (2015). Finding culture change in the second factor: Stability and change in cultural consensus and residual agreement. *Field Methods*, 27(1):22–38.

Elliot, S. N., Busse, R., and Gresham, F. M. (1993). Behavior rating scales: Issues of use and development. *School psychology review*, 22(2):313–321.

Esteban, J.-M. and Ray, D. (1994). On the measurement of polarization. *Econometrica*, 62(4):819–851.

Ferguson, T. S. (1973). A Bayesian Analysis of Some Nonparametric Problems. *The Annals of Statistics*, 1(2):209 – 230.

Forchheimer, D., Forchheimer, R., and Haviland, D. (2015). Improving image contrast and material discrimination with nonlinear response in bimodal atomic force microscopy. *Nature communications*, 6:6270.

Gamage, D., Staubitz, T., and Whiting, M. (2021). Peer assessment in moocs: Systematic literature review. *Distance Education*, 40:1–22.

Gelman, A., Carlin, J., Stern, H., Dunson, D., and Vehtari, A.and Rubin, D. (2013). *Bayesian Data Analysis*. Chapman and Hall/CRC.

Gelman, A. and Hill, J. (2006). *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Analytical Methods for Social Research. Cambridge University Press.

Gere, A. R., Limlamai, N., Wilson, E., Saylor, K. M., and Pugh, R. (2019). Writing and conceptual learning in science: An analysis of assignments. *Written Communication*, 36(1):99–135.

Gerke, O., Vilstrup, M. H., Segtnan, E. A., Halekoh, U., and Høilund-Carlsen, P. F. (2016). How to assess intra- and inter-observer agreement with quantitative pet using variance component analysis: A proposal for standardisation. *BMC Medical Imaging*, 16.

Ghosal, S., Ghosh, J. K., and Ramamoorthi, R. V. (1999). Posterior consistency of dirichlet mixtures in density estimation. *Annals of Statistics*, 27.

Ghosh, S., Hastie, T., and Owen, A. B. (2022). Backfitting for large scale crossed random effects regressions. *Annals of Statistics*, 50.

Giammarino, M., Mattiello, S., Battini, M., Quatto, P., Battaglini, L. M., Vieira, A. C., Stilwell, G., and Renna, M. (2021). Evaluation of inter-observer reliability of animal welfare indicators: Which is the best index to use? *Animals*, 11.

Gill, J. and Casella, G. (2009). Nonparametric priors for ordinal bayesian social science models: Specification and estimation. *Journal of the American Statistical Association*, 104(486):453–454.

Gin, B., Sim, N., Skrondal, A., and Rabe-Hesketh, S. (2020). A dyadic irt model. 85:815–836.

Gisev, N., Bell, J. S., and Chen, T. F. (2013). Interrater agreement and interrater reliability: Key concepts, approaches, and applications. *Research in Social and Administrative Pharmacy*, 9(3):330–338.

Goplerud, M. (2022). Fast and Accurate Estimation of Non-Nested Binomial Hierarchical Models Using Variational Inference. *Bayesian Analysis*, 17(2):623 – 650.

Gu, Y., Fiebig, D. G., Cripps, E., and Kohn, R. (2009). Bayesian estimation of a random effects heteroscedastic probit model. *Econometrics Journal*, 12.

Gwet, K. L. (2008). Computing inter-rater reliability and its variance in the presence of high agreement. *British Journal of Mathematical and Statistical Psychology*, 61(1):29–48.

Han, Y., Wu, W., Yan, Y., and Zhang, L. (2020). Human-machine hybrid peer grading in spocs. *IEEE Access*, 8:220922–220934.

Hastie, T., Tibshirani, R., and Friedman, J. (2009). *The elements of statistical learning: data mining, inference and prediction*. Springer, 2 edition.

Heimann, A. L., Ingold, P. V., and Kleinmann, M. (2020). Tell us about your leadership style: A structured interview approach for assessing leadership behavior constructs. *The Leadership Quarterly*, 31(4):101364.

Heinzl, F., Kneib, T., and Fahrmeir, L. (2012). Additive mixed models with dirichlet process mixture and p-spline priors. *AStA Advances in Statistical Analysis*, 96.

Heinzl, F. and Tutz, G. (2013). Clustering in linear mixed models with approximate dirichlet process mixtures using em algorithm. *Statistical Modelling*, 13(1):41–67.

Hoffman, M. D. and Gelman, A. (2014). The no-u-turn sampler. *The Journal of Machine Learning Research*, 15:1593–1623.

Hou, Y., Benner, A. D., Kim, S. Y., Chen, S., Spitz, S., Shi, Y., and Beretvas, T. (2020). Discordance in parents' and adolescents' reports of parenting: A meta-analysis and qualitative review. *American Psychologist*, 75(3):329.

Hsiao, C. K., Chen, P.-C., and Kao, W.-H. (2011). Bayesian random effects for interrater and test–retest reliability with nested clinical observations. *Journal of Clinical Epidemiology*, 64(7):808–814.

Hörlin, E., Ehrlington, S. M., Henricson, J., John, R. T., and Wilhelms, D. (2022). Inter-rater reliability of the clinical frailty scale by staff members in a swedish emergency department setting. *Academic Emergency Medicine*, 29.

Ishwaran, H. and James, L. (2001). Gibbs sampling methods for stick-breaking priors. *Journal of the American Statistical Association*, 96:161–173.

James, G., Witten, D., Hastie, T., and Tibshirani, R. (2013). *An Introduction to Statistical Learning: with Applications in R*. Springer.

James, G. M. and Sugar, C. A. (2003). Clustering for sparsely sampled functional data. *Journal of the American Statistical Association*, 98(462):397–408.

Jang, J. H., Manatunga, A. K., Taylor, A. T., and Long, Q. (2018). Overall indices for assessing agreement among multiple raters. *Statistics in Medicine*, 37(28):4200–4215.

Jeon, M., Rijmen, F., and Rabe-Hesketh, S. (2017). A variational maximization-maximization algorithm for generalized linear mixed models with crossed random effects. *Psychometrika*, 82:693–716.

Jia, W. and Zhang, P. (2023). Open access rater cognitive processes in integrated writing tasks: from the perspective of problem-solving. *Asia*, 13:50.

Kahrari, F., Ferreira, C. S., and Arellano-Valle, R. B. (2019). Skew-Normal-Cauchy Linear Mixed Models. *Sankhya B: The Indian Journal of Statistics*, 81(2):185–202.

Kim, S., Tadesse, M. G., and Vannucci, M. (2006). Variable selection in clustering via Dirichlet process mixture models. *Biometrika*, 93(4):877–893.

Komárek, A. and Komárková, L. (2013). Clustering for multivariate continuous and discrete longitudinal data. *The Annals of Applied Statistics*, 7(1):177 – 200.

Komárek, A., Hansen, B. E., Kuiper, E. M. M., van Buuren, H. R., and Lesaffre, E. (2010). Discriminant analysis using a multivariate linear mixed model with a normal mixture in the random effects distribution. *Statistics in Medicine*, 29(30):3267–3283.

Koudenburg, N. and Kashima, Y. (2022). A polarized discourse: Effects of opinion differentiation and structural differentiation on communication. *Personality and Social Psychology Bulletin*, 48(7):1068–1086. PMID: 34292094.

Koudenburg, N., Kiers, H. A. L., and Kashima, Y. (2021). A new opinion polarization index developed by integrating expert judgments. *Frontiers in Psychology*, 12.

Kyung, M., Gill, J., and Casella, G. (2010). Estimation in dirichlet random effects models. *The Annals of Statistics*, 38(2):979–1009.

Kyung, M., Gill, J., Casella, G., Kyung, M., Gill, J., and Casella, G. (2011). Sampling schemes for generalized linear dirichlet process random effects models. *Stat Methods Appl*, 20:259–290.

Lee, J. P., Binger, C., Harrington, N., Evelyn, S., Kent-Walsh, J., Gevarter, C., Richardson, J., and Hahs-Vaughn, D. (2022). Aided language measures: Establishing observer agreement for communicators in early language phases. *American Journal of Speech-Language Pathology*, 31.

Levin, I. P., Rouwenhorst, R. M., and Trisko, H. M. (2005). Separating gender biases in screening and selecting candidates for hiring and firing. *Social Behavior and Personality: an international journal*, 33(8):793–804.

Levinson, D., Potash, J., Mostafavi, S., Battle, A., Zhu, X., and Weissman, M. (2017). Brief assessment of major depression for genetic studies: validation of cidi-sf screening with scid interviews. *European Neuropsychopharmacology*, 27:S448.

Li, H., Xiong, Y., Hunter, C. V., Guo, X., and Tywoniw, R. (2019). Does peer assessment promote student learning? a meta-analysis. *Assessment and Evaluation in Higher Education*.

Liljequist, D., Elfving, B., and Skavberg Roaldsen, K. (2019). Intraclass correlation – a discussion and demonstration of basic features. *PLOS ONE*, 14(7):1–35.

Lin, T. I. and Lee, J. C. (2008). Estimation and prediction in linear mixed models with skew-normal random effects for longitudinal data. *Statistics in Medicine*, 27(9):1490–1507.

Lobbestael, J., Leurgans, M., and Arntz, A. (2011). Inter-rater reliability of the structured clinical interview for dsm-iv axis i disorders (scid i) and axis ii disorders (scid ii). *Clinical Psychology and Psychotherapy*, 18.

Lockwood, J. R., Castellano, K. E., and Shear, B. R. (2018). Flexible bayesian models for inferences from coarsened, group-level achievement data. *Journal of Educational and Behavioral Statistics*, 43.

Makransky, G., Terkildsen, T., and Mayer, R. (2019). Role of subjective and objective measures of cognitive processing during learning in explaining the spatial contiguity effect. *Learning and Instruction*.

Martinková, P., Bartoš, F., and Brabec, M. (2023). Assessing inter-rater reliability with heterogeneous variance components models: Flexible approach accounting for contextual variables. *Journal of Educational and Behavioral Statistics*, 48(3):349–383.

McCulloch, C. E. and Neuhaus, J. M. (2021). Improving predictions when interest focuses on extreme random effects. *Journal of the American Statistical Association*, 0(0):1–10.

McHugh, M. (2012). Interrater reliability: The kappa statistic. *Biochemia medica : časopis Hrvatskoga društva medicinskih biokemičara / HDMB*, 22:276–82.

Mokkink, L. B., Terwee, C. B., Gibbons, E., Stratford, P. W., Alonso, J., Patrick, D. L., Knol, D. L., Bouter, L. M., and De Vet, H. C. (2010). Inter-rater agreement and reliability of the cosmin (consensus-based standards for the selection of health status measurement instruments) checklist. *BMC Medical Research Methodology*, 10:1–11.

Müller, P., Quintana, F. A., Jara, A., and Hanson, T. (2015). *Bayesian nonparametric data analysis*, volume 1. Springer.

Murphy, J. P. (2021). Explanatory item response models for dyadic data from multiple groups. *Original Article Sociological Methodology*, 51:112–145.

Möller, H. J. (2000). Rating depressed patients: Observer- vs self-assessment.

Navarro, D. J., Griffiths, T. L., Steyvers, M., and Lee, M. D. (2006). Modeling individual differences using dirichlet processes. *Journal of Mathematical Psychology*, 50(2):101–122. Special Issue on Model Selection: Theoretical Developments and Applications.

Nelson, K. and Edwards, D. (2015). Measures of agreement between many raters for ordinal classifications. *Statistics in medicine*, 34.

Nelson, K. P. and Edwards, D. (2008). On population-based measures of agreement for binary classifications. *Canadian Journal of Statistics*, 36.

Nelson, K. P., Zhou, T. J., and Edwards, D. (2020). Measuring intrarater association between correlated ordinal ratings. *Biometrical Journal*, 62(7):1687–1701.

Nestler, S., Geukes, K., Hutteman, R., and Back, M. D. (2017). Tackling longitudinal round-robin data: A social relations growth model. *psychometrika*, 82.

Nestler, S., Lüdtke, O., and Robitzsch, A. (2020). Maximum likelihood estimation of a social relations structural equation model. *Psychometrika*, 85:870–889.

Noveanu, J., Amsler, F., Ummenhofer, W., von Wyl, T., and Zuercher, M. (2017). Assessment of simulated emergency scenarios: are trained observers necessary? *Prehospital Emergency Care*, 21(4):511–524.

Nucci, M., Spoto, A., Altoè, G., and Pastore, M. (2021). The role of raters threshold in estimating interrater agreement. *Psychological Methods*, 26(5):622.

Oravecz, Z., Vandekerckhove, J., and Batchelder, W. H. (2014). Bayesian cultural consensus theory. *Field Methods*, 26(3):207–222.

Ormrod, J. E. (1999). *Human learning*. Merrill Upper Saddle River, NJ.

Panadero, E. and Alqassab, M. (2019). An empirical review of anonymity effects in peer assessment, peer feedback, peer review, peer evaluation and peer grading. *Assessment & Evaluation in Higher Education*, 44(8):1253–1278.

Papaspiliopoulos, O., Roberts, G., and Sköld, M. (2007). A general framework for the parametrization of hierarchical models. *Statist Sci*, 22.

Papaspiliopoulos, O., Roberts, G. O., and Zanella, G. (2020). Scalable inference for crossed random effects models. *Biometrika*, 107.

Paredes, V. (2014). A teacher like me or a student like me? role model versus teacher bias effect. *Economics of Education Review*, 39:38–49.

Pfeifer, M., Eggemann, L., Kransmann, J., Schmitt, A. O., and Hessel, E. F. (2019). Inter- and intra-observer reliability of animal welfare indicators for the on-farm self-assessment of fattening pigs. *Animal*, 13.

Piech, C., Huang, J., Chen, Z., Do, C., Ng, A., and Koller, D. (2013). Tuned models of peer assessment in moocs. *Proceedings of the 6th International Conference on Educational Data Mining (EDM 2013)*.

Quinn, D. M. (2020). Experimental evidence on teachers' racial bias in student evaluation: The role of grading scales. *Educational Evaluation and Policy Analysis*, 42.

Randall, J. and Engelhard, G. (2009). Differences between teachers' grading practices in elementary and middle schools. *Journal of Educational Research - J EDUC RES*, 102:175–186.

Reardon, S. F., Shear, B. R., Castellano, K. E., and Ho, A. D. (2017). Using heteroskedastic ordered probit models to recover moments of continuous test score distributions from coarsened data. *Journal of Educational and Behavioral Statistics*, 42.

Reckase, M. D. (1997). The past and future of multidimensional item response theory. *Applied Psychological Measurement*, 21(1):25–36.

Reily, K., Finnerty, P., and Terveen, L. (2009). Two peers are better than one: Aggregating peer reviews for computing assignments is surprisingly accurate. pages 115–124.

Rigon, T. and Durante, D. (2021). Tractable bayesian density regression via logit stick-breaking priors. *Journal of Statistical Planning and Inference*, 211:131–142.

Rodriguez, A. and Dunson, D. (2011). Nonparametric bayesian models through probit stick-breaking processes. *Bayesian Analysis*, 6:145–178.

Rothstein, H. R. (1990). Interrater reliability of job performance ratings: Growth to asymptote level with increasing opportunity to observe. *Journal of Applied Psychology*, 75:322–327.

Sajjadi, M. S. M., Alamgir, M., and Luxburg, U. (2015). Peer grading in a course on algorithms and data structures: Machine learning algorithms do not improve over simple baselines.

Salgado, J. F. and Moscoso, S. (1996). Meta-analysis of interrater reliability of job performance ratings in validity studies of personnel selection. *Perceptual and Motor skills*, 83(3_suppl):1195–1201.

Salgado, J. F. and Moscoso, S. (2019). Meta-analysis of interrater reliability of supervisory performance ratings: Effects of appraisal purpose, scale type, and range restriction. *Frontiers in Psychology*, 10.

Sanchez, C., Atkinson, K., Koenka, A., Moshontz, H., and Cooper, H. (2017). Self-grading and peer-grading for formative and summative assessments in 3rd through 12th grade classrooms: A meta-analysis. *Journal of Educational Psychology*, 109.

Schielzeth, H., Dingemanse, N. J., Nakagawa, S., Westneat, D. F., Allegue, H., Teplitsky, C., Réale, D., Dochtermann, N. A., Garamszegi, L. Z., and Araya-Ajoy, Y. G. (2020). Robustness of linear mixed-effects models to violations of distributional assumptions. *Methods in Ecology and Evolution*, 11(9):1141–1152.

Schneider, I., Mädler, M., and Lang, J. (2019). Comparability of self-ratings and observer ratings in occupational psychosocial risk assessments: Is there agreement? *BioMed Research International*, 2019.

Sethuraman, J. (1994). A constructive definition of dirichlet priors. *Statistica Sinica*, 4(2):639–650.

Sheaffer, A. W., Majeika, C. E., Gilmour, A. F., and Wehby, J. H. (2021). Classroom behavior of students with or at risk of ebd: Student gender affects teacher ratings but not direct observations. *Behavioral Disorders*, 46.

Shear, B. R. and Reardon, S. F. (2021). Using pooled heteroskedastic ordered probit models to improve small-sample estimates of latent test score distributions. *Journal of Educational and Behavioral Statistics*, 46.

Shengkai Yin, F. C. and Chang, H. (2022). Assessment as learning: How does peer assessment function in students' learning? *Frontiers in Psychology*, 13:912568.

Shirazi, M. A. (2019). For a greater good: Bias analysis in writing assessment. *SAGE Open*, 9(1):2158244018822377.

Simpson, D., Rue, H., Riebler, A., Martins, T. G., and Sørbye, S. H. (2017). Penalising Model Component Complexity: A Principled, Practical Approach to Constructing Priors. *Statistical Science*, 32(1):1 – 28.

Snooks, H., Evans, A., Wells, B., Peconi, J., Thomas, M., Woollard, M., Guly, H., Jenkinson, E., Turner, J., and Hartley-Sharpe, C. (2009). What are the highest priorities for research in emergency prehospital care? *Emergency Medicine Journal*, 26(8):549–550.

Spoto, A., Nucci, M., Prunetti, E., and Vicovaro, M. (2023). Improving content validity evaluation of assessment instruments through formal content validity analysis. *Psychological Methods*.

Stan Development Team (2022). RStan: the R interface to Stan. R package version 2.21.7.

Stan Development Team (2023). RStan: the R interface to Stan. R package version 2.32.3.

Stefanucci, M. and Canale, A. (2021). Multiscale stick-breaking mixture models. *Statistics and Computing*, 31:13.

Strahl, A., Gerlich, C., Alpers, G. W., Gehrke, J., Müller-Garnn, A., and Vogel, H. (2019). An instrument for quality assurance in work capacity evaluation: Development, evaluation, and inter-rater reliability. *BMC Health Services Research*, 19.

Sun, Y. and Cheng, L. (2014). Teachers' grading practices: Meaning and values assigned. *Assessment in Education: Principles, Policy and Practice*, 21.

Swanson, E. A. and Vaughn, S. (2010). An observation study of reading instruction provided to elementary students with learning disabilities in the resource room. *Psychology in the Schools*, 47(5):481–492.

Tang, T., Ghorbani, A., Squazzoni, F., and Chorus, C. G. (2022). Together alone: a group-based polarization measurement. 56:3587–3619.

Tippins, N., Sackett, P., and Oswald, F. (2018). Principles for the validation and use of personnel selection procedures.

Tokdar, S. T. (2006). Posterior consistency of dirichlet location-scale mixture of normals in density estimation and regression. *Sankhya: The Indian Journal of Statistics*, 68.

Tutz, G. and Oelker, M.-R. (2017). Modelling clustered heterogeneity: Fixed effects, random effects and mixtures. *International Statistical Review*, 85(2):204–227.

Ulker, Y., Günsel, B., and Cemgil, T. (2010). Sequential monte carlo samplers for dirichlet process mixtures. In Teh, Y. W. and Titterington, M., editors, *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, volume 9 of *Proceedings of Machine Learning Research*, pages 876–883, Chia Laguna Resort, Sardinia, Italy. PMLR.

Uto, M. (2022). A bayesian many-facet rasch model with markov modeling for rater severity drift. *Behavior Research Methods*.

van der Linden, W. J. (2016). *Handbook of item response theory*, volume 1. CRC Press.

Van Trappen, S. (2022). Candidate selection and ethnic minority aspirants: Exploring the effect of party selectors' biases in a pr system. *Party Politics*, 28(6):1123–1135.

Vehtari, A., Gelman, A., and Gabry, J. (2017). Practical bayesian model evaluation using leave-one-out cross-validation and waic. *Statistics and Computing*, 27:1413–1432.

Verbeke, G. and Lesaffre, E. (1996). A linear mixed-effects model with heterogeneity in the random-effects population. *Journal of the American Statistical Association*, 91(433):217–221.

Vieira, A., Battini, M., Can, E., Mattiello, S., and Stilwell, G. (2018). Inter-observer reliability of animal-based welfare indicators included in the animal welfare indicators welfare assessment protocol for dairy goats. *Animal*, 12.

Villarroel, L., Marshall, G., and Barón, A. E. (2009). Cluster analysis using multivariate mixed effects models. *Statistics in Medicine*, 28(20):2552–2565.

Walker, S. G. (2007). Sampling the dirichlet mixture model with slices. *Communications in Statistics - Simulation and Computation*, 36(1):45–54.

Walter, S. R., Dunsmuir, W. T., and Westbrook, J. I. (2019). Inter-observer agreement and reliability assessment for observational studies of clinical work.

Wang, W.-L. and Lin, T.-I. (2014). Multivariate t nonlinear mixed-effects models for multi-outcome longitudinal data with missing values. *Statistics in Medicine*, 33(17):3029–3046.

Wirtz, M. A. (2020). *Interrater Reliability*, pages 2396–2399. Springer International Publishing, Cham.

Wu, Y. and Ghosal, S. (2010). The l1-consistency of dirichlet mixtures in multivariate bayesian density estimation. *Journal of Multivariate Analysis*, 101.

Xiong, Y. and Suen, H. (2016). Hierarchical bayesian modeling for peer assessment in a mooc.

Zenk, S. N., Schulz, A. J., Mentz, G., House, J. S., Gravlee, C. C., Miranda, P. Y., Miller, P., and Kannan, S. (2007). Inter-rater and test–retest reliability: methods and results for the neighborhood observational checklist. *Health & place*, 13(2):452–465.

Zhang, C., Mapes, B. E., and Soden, B. J. (2003). Part a no. 594 q. *J. R. Meteorol. Soc*, 129:2847–2866.

Zhu, Y., Fung, A. S.-L., and Yang, L. (2021). A methodologically improved study on raters' personality and rating severity in writing assessment. *SAGE Open*, 11(2):21582440211009476.

Zong, Z., Schunn, C. D., and Wang, Y. (2021). What aspects of online peer feedback robustly predict growth in students' task performance? *Computers in Human Behavior*, 124:106924.

Zupanc, K. and Štrumbelj, E. (2018). A bayesian hierarchical latent trait model for estimating rater bias and reliability in large-scale performance assessment. *PLOS ONE*, 13(4):1–16.

Appendix A

Nonparametrics hierarchical models

A.0.1 Remarks for multiple ratings

When raters rate the same set of items $\mathscr{J}_i = \mathscr{J}$, $i = 1, \ldots, I$ a varying intercept can be identified for each item (Bartoš et al., 2019; Agresti, 2015; Nelson and Edwards, 2015). This term might be added to equation 2.1 (which is the same in both the standard and nonparametric formulation):

$$y_{ij} \quad = \quad \mathbf{x}_{ij}'\beta + \mathbf{z}_i'\mathbf{u}_i + \delta_j + \varepsilon_{ij}, \;\; i = 1, .., I, \; j \in \mathscr{J}. \tag{A.1}$$

In both the standard HLM (i.e., assuming a multivariate normal distributed hierarchical rater effect) and the nonparametric HLM (i.e., specifying a DPM over the rater effect) the following distribution might be specified:

$$\delta_j \quad \sim \quad N(0, \sigma_\delta^2) \quad j = 1, .., J.$$

where $\sigma_\delta > 0$ is the scale parameter of δ and $J = |\mathscr{J}|$. See Section 2.2 and 2.3 for the other quantities and their distribution assumption. Specifying a conjugate prior for σ_δ additional steps might be added to the Gibbs sampling for the nonparametric HLM.

The main results of the present work and the interpretation of λ (see Section 2.5) still hold for this model specification.

A.0.2 Details on Dirichlet Process Mixture

As noticed in Chapter 2.2, α is proportional to the *concentration* of the realizations of G in point masses. Indeed, considering the partition (A, A^c) of Ω, the variance of G(A) is defined as

$$Var[G(A)] = \frac{G_0(A)(1-G_0(A))}{\alpha+1}$$

Thus, larger values of α, conditioning on the number of raters I, reduce the variability of the DP, i.e. the process samples most of the time from G_0, G tends to be an infinite number of point masses: the empirical distribution of G tends to become a discrete approximation of the parametric G_0. In this case, there is not a strong clustering since the probability of ties is very low. On the contrary, smaller values of α induce a strong clustering, the random weights distribution concentrates the probability mass to a few points of the support of G and the probability of ties is higher. That is, in the present model means that several u_i will be independent and identically distributed from a normal distribution indexed by the same parameters. Moreover, Antoniak 1974 showed that for large I:

$$\mathbb{E}[C|I,\alpha] \approx \alpha \ln\left(\frac{I+\alpha}{\alpha}\right)$$

where C is the number of clusters. Thus, the expected number of point masses of G is proportional to both the α and the number of raters I. Every consideration regarding the role of the precision parameter on the distribution of G should be conditioned to I.

A.0.3 Details on the Gibbs sampling

Further details regarding some parameters of the posterior sampling are shown as follows.
The following matrix notation is here adopted: $\mathbf{X}_i = (\mathbf{x}'_{i1}, \ldots, \mathbf{x}'_{i|\mathscr{J}_i|})$, $\mathbf{Z}_i = (\mathbf{z}'_{i1}, \ldots, \mathbf{z}'_{i|\mathscr{J}_i|})$, are the design matrices for each rater $i = 1, \ldots, I$; and $\mathbf{X} = (\mathbf{X}_1, \ldots, \mathbf{X}_I)$ and $\mathbf{Z} = diag(\mathbf{Z}_1, \ldots, \mathbf{Z}_I)$ are the full design matrices.

1. Referring to the non varying effects:

$$\begin{aligned}
\mathbf{b}_{\beta}^{*} &= \left(\mathbf{B}_{\beta}^{-1}+\frac{1}{\sigma_{\varepsilon}^{2}}\mathbf{X}'\mathbf{X}\right)^{-1}\left(\mathbf{B}_{\beta}^{-1}\mathbf{b}_{\beta}+\frac{1}{\sigma_{\varepsilon}^{2}}\mathbf{X}'\left(\mathbf{y}-\mathbf{Z}\mathbf{u}\right)\right)\\
\mathbf{B}_{\beta}^{*} &= \left(\mathbf{B}_{\beta}^{-1}+\frac{1}{\sigma_{\varepsilon}^{2}}\mathbf{X}'\mathbf{X}\right)^{-1}
\end{aligned}$$

2. Referring to hierarchical effects:

- For each rater $i=1,...,I$:

$$\begin{aligned}
\mu_{c_i}^{*} &= \left(\mathbf{D}_0^{-1}+\frac{1}{\sigma_{\varepsilon}^{2}}\mathbf{Z}_i'\mathbf{Z}_i\right)^{-1}\left(\mathbf{D}_0^{-1}\mu_{c_i}+\frac{1}{\sigma_{\varepsilon}^{2}}\mathbf{Z}_i'\left(\mathbf{y}_i-\mathbf{X}_i\beta\right)\right)\\
\mathbf{Q}_{c_i}^{*} &= \left(\mathbf{D}_0^{-1}+\frac{1}{\sigma_{\varepsilon}^{2}}\mathbf{Z}_i'\mathbf{Z}_i\right)^{-1}
\end{aligned}$$

Here μ_{c_i} is the location parameter vector of the cluster where the rater i is allocated.

- For each component $r=1,...,R$ and each variable $d=1,...,q$, associated with an hierarchical effect:

$$\begin{aligned}
\mu_{0_r}^{*} &= \left(\frac{c_r}{\sigma_{Q_d}^{2}}+\frac{1}{\sigma_{D_{0_d}}^{2}}\right)^{-1}\left(\frac{c_r}{\sigma_{Q_d}^{2}}\overline{u}_{d,r}+\frac{\mu_{0_r}}{\sigma_{D_{0_d}}^{2}}\right)\\
\sigma_{D_{0_d}}^{2*} &= \left(\frac{c_r}{\sigma_{Q_d}^{2}}+\frac{1}{\sigma_{D_{0_d}}^{2}}\right)^{-1}\\
\sigma_{Q_{dr}}^{2}|\mu,\mathbf{u} &\sim IG\left(a_{Q_0}+\frac{r_c}{2},b_{Q_0}+\frac{1}{2}\sum_{i=1}^{r_c}(u_{i_d}-\mu_{dr})^2\right)
\end{aligned}$$

Here $\overline{u}_{dr}$ is the mean of the d-th hierarchical effect in the cluster r.

- For rater $i = 1, ..., I$ and each component $r = 1, ..., R$:

$$\omega_{i_r}^{*} = \frac{\pi_r N_q(\mathbf{u}_i|\mu_r, \mathbf{Q}_r)}{\sum_{r=1}^{R} \pi_r N_q(\mathbf{u}_i|\mu_r, \mathbf{Q}_r)}.$$