RESEARCH ARTICLE

Statistics in Medicine WILEY

A Bayesian nonparametric approach to dynamic item-response modeling: An application to the GUSTO cohort study

Andrea Cremaschi[1] | Maria De Iorio[1,2,3,4] | Yap Seng Chong[1,2] | Birit Broekman[5] | Michael J. Meaney[1,6] | Michelle Z. L. Kee[1]

[1]Singapore Institute for Clinical Sciences (SICS), Agency for Science, Technology and Research (A*STAR), Brenner Centre for Molecular Medicine, Singapore

[2]Yong Loo Lin School of Medicine, National University of Singapore, Singapore, Singapore

[3]Division of Science, Yale-NUS College, Singapore, Singapore

[4]Department of Statistical Science, University College London, London, UK

[5]Department of Psychiatry, VU Medical Centre, Amsterdam, the Netherlands

[6]Department of Psychiatry, Douglas Mental Health University Research Institute, McGill University, Montreal, Quebec, Canada

Correspondence
Andrea Cremaschi, Singapore Institute for Clinical Sciences (SICS), Agency for Science, Technology and Research (A*STAR), Brenner Centre for Molecular Medicine, 30 Medical Dr, Singapore 117609, Singapore.
Email: cremaschia@sics.a-star.edu.sg

Statistical analysis of questionnaire data is often performed employing techniques from item-response theory. In this framework, it is possible to differentiate respondent profiles and characterize the questions (items) included in the questionnaire via interpretable parameters. These models are often crosssectional and aim at evaluating the performance of the respondents. The motivating application of this work is the analysis of psychometric questionnaires taken by a group of mothers at different time points and by their children at one later time point. The data are available through the GUSTO cohort study. To this end, we propose a Bayesian semiparametric model and extend the current literature by: (i) introducing temporal dependence among questionnaires taken at different time points; (ii) jointly modeling the responses to questionnaires taken from different, but related, groups of subjects (in our case mothers and children), introducing a further dependency structure and therefore sharing of information; (iii) allowing clustering of subjects based on their latent response profile. The proposed model is able to identify three main groups of mother/child pairs characterized by their response profiles. Furthermore, we report an interesting maternal reporting bias effect strongly affecting the clustering structure of the mother/child dyads.

KEYWORDS
clustering, cohort study, Dirichlet process, item-response theory, questionnaire data

1 | INTRODUCTION

Research and clinical practice in child psychopathology continues to rely heavily on the reports of adults, most notably the parents and especially mothers. Assessments of child psychopathology commonly collect information from multiple informants, including parents, most often the mother and teachers. Studies of the convergence of these measures often reveal only moderate (correlation coefficient $r = 0.3$-0.6) alignment.[1-4] Alternatively, the divergence could represent very

real effects of context, whereby the function of the child varies as a function of the setting.[5-7] The divergence might reflect reporter bias associated with the nature of the behavioral problems within a context. Teachers, for example, seem particularly sensitive to reporting of externalizing problems, likely associated with classroom disruptions. However, the major concern for parental reports is that of bias derived from the mental health of the parent. For example, depressed mothers have been found to overreport child emotional or behavioral problems.[8-10] Cognitive theories of depression predict that depressed parents and children will globally report negatively on many aspects of their lives.[11,12] Nevertheless, a number of studies[13,14] report association between maternal symptoms of depression at one stage in development and child behavioral problems at a later stage, while controlling for depressive symptoms at the time of maternal reporting. This approach should minimize the impact of the reporter mental health at the time of the assessment. Moreover, maternal symptoms of depression do predict child outcomes on a wide range of objective measures assessed directly, with the child independent of reporter bias.[15-18] Thus, reporter bias seems unlikely to account entirely for the association between mental health of the mother and that of the child.

Among studies with evidence for maternal bias, the bias appears to depend upon the nature of the psychopathological problem (eg, internalizing vs externalizing behaviors[19]) as well as the gender of the child. The nature of the maternal psychopathology is also a factor, with greater evidence for maternal bias in mothers with comorbid anxiety and depression.[19,20] The issue is clearly not as simple as a maternal bias linked to mental health status that is expressed universally. In addition to the variation associated with the target outcome and the gender of the child, there is also the possibility that a maternal bias may vary as a function of the mother and/or of the mother-child dyad.[21] A dyadic relation was reported[19] where the analysis suggested that maternal psychopathology and the reported child behavior appear to interact to determine the degree of the bias.

In this work, we analyze data on mothers reporting on their children's behavior, and on the children's self-reporting, obtained from their answers to clinical screening psychometric questionnaires. The mothers answered the Child Behavior Checklist (CBCL), which is the most commonly used parent or teacher tool for child assessment. The children answered the Child Depression Inventory and the Multidimensional Anxiety Scale for Children as self-reporting tools. The availability of both type of questionnaires allows for a deeper evaluation of the psychopathological assessment of the mother-child dyads. Furthermore, the mothers responded to questionnaires at different time points, adding a longitudinal dimension to the study. From a statistical point of view, our work builds upon techniques from *item-response theory* (IRT), which are commonly used for the analysis of questionnaire data.[22-25] However, the statistical approach used here best exploits the interesting features of the dataset, adding novel features to the existing literature. First, by setting the study in a Bayesian framework, we allow for joint modeling of both mothers' and children's answers to the respective questionnaires. By borrowing information between these two components of the dataset, the proposed approach exploits the dependencies within the dyads, which would be otherwise overlooked if analyzed in two independent marginal studies. Second, the longitudinal aspect of the maternal reports is taken into account by including an autoregressive structure in the model, capturing the temporal profile of the maternal evaluation of the children's behavior. Furthermore, in our setting there is not a complete overlap between questions answered by the mothers at different time points. We account for this feature of the data by specifying question-specific parameters which are shared over different time points when the questions are repeated over time, therefore allowing for additional sharing of information. Finally, to introduce extra flexibility, dyad-specific latent parameters capturing the respondents' profiles are modeled via Bayesian nonparametric processes, naturally inducing clustering of the mother-child pairs. As shown later, this feature of the model allows for the identification of three main clusters of interest, presenting in-depth differences. In particular, our analysis shows evidence of the existence of reporting biases, a crucial finding enabled by our novel statistical approach.

The article is structured as follows: the dataset is described in Section 2, while the proposed Bayesian semiparametric IRT model is introduced in Section 3, followed by posterior inference results in Section 4. Section 5 concludes with a discussion.

2 | DESCRIPTION OF THE DATASET

In this work, we develop a modeling approach for data collected from questionnaires answered by large heterogeneous cohorts over time. In particular, we focus on responses from a cohort of mothers to psychometric questions describing the behavior of their children at different time points (2, 3, 4, and 7 years of age of the children), available through the GUSTO cohort study.[26] Furthermore, we have access to the answers to two questionnaires given to the corresponding children at 8.5 years of age, aimed at measuring the presence of depressive/anxious traits. Due to the parental relationship between the respondents, we expect specific dependency structures in the data, relating the behaviors described by the

mothers with the well-being of the children. This dependency structure is the main object of this study and is of clinical interest. To better understand the relationship between the mother and the child's perception of behavior, we develop a Bayesian semi-parametric model able to account for covariates and time evolution.

First, we describe the data relative to the mothers' questionnaires. At each time point (2, 3, 4, and 7 years), the mothers provide information about the behavioral well-being of the children by answering the CBCL.[27] The questions in the CBCL are grouped into psychometric subscales, as reported in the second column of Supplementary Table S1, as well as syndrome subscales as reported in Supplementary Table S2. Clearly, the children's personality, attitude and abilities evolve as they grow, and therefore different questions are asked throughout time in order to properly capture the evolution of their behavioral and developmental profile. For instance, for the last time point, subscales relative to social interaction (*Social Problems*) and thinking processes (*Thought Problems*) are included, different from earlier time points. The total number of questions asked increases from 99 at the first three time points, to 119 at the last one, with only partial overlap. The number of subjects answering the questionnaires varies throughout the years, indicating that some mothers participate in the study only at specific time points due, in part, to missed visits, thus representing missing information in the dataset. In this work, we consider mothers who answered to the CBCL questionnaires at, at least, three different time points. Answers to each question in the CBCL questionnaire vary between *not true* (0), *somewhat true* (1), and *very true* (2). For each subscale, the responses are summarized in Supplementary Figures S1 to S5, by reporting the proportions of each answer category, grouped according to the subscales provided in Supplementary Table S1. The missing information (ie, unanswered questions) in the mothers' questionnaire datasets is 28%, 7%, 3%, and 11% at the four time points, respectively.

The self-reported depressive and anxiety symptoms from the children are obtained using two different questionnaires: the Children Depression Inventory 2nd edition (CDI2[28]), and the Multidimensional Anxiety Score for Children 2nd edition (MASC2[29]). The corresponding subscales are reported in the bottom part of Supplementary Table S1. The children in the study answered a total of 28 questions in CDI2 and of 50 questions in MASC2. The answers to the questions are recorded on a scale from 0 to 2 for the CDI2 system (eg, 0 = *I am sad once in a while*, 1 = *I am sad many times*, 2 = *I am sad all the time*), and from 0 to 3 for the MASC2 system (eg, 0 = *Never*, 1 = *Rarely*, 2 = *Sometimes*, 3 = *Often*). The proportions of answer categories for the CDI2 and MASC2 systems are reported in Supplementary Figures S6 and S7. The missing information from the children's questionnaires is very low, with only 1% missing information in CDI2, and none in MASC2.

Besides the answers to the above-described questionnaires, additional information is available to be used as predictors in the model. This includes gender (*Gender*) of the child and three indices measuring mothers' depression and anxiety: the Beck's Depression Inventory (BDI[30]), the Edinburgh Postnatal Depression Scale (EPDS[31]) and the State-Trait Anxiety Inventory (STAI[32]). The resulting scores are used as covariate in our model, indicated by *BDI*, *EPDS*, *STAI*. No missing information is present in the covariates for the mother/child pairs considered in the analysis.

3 | BNP FOR ITEM-RESPONSE MODELLING

The analysis of questionnaire data is usually performed using techniques from IRT. The aim of IRT is to describe the answering profiles of the respondents by quantifying their ability to answer consistently and to capture the discriminating power and difficulty of each question. All the subjects are usually evaluated on the same set of questions (items). The questions can be dichotomous (binary answers), or polytomous (answers in more than two categories). Dichotomous answers are often modeled via the Rasch model,[22] which is based on a logistic random effects model including a *latent trait* parameter for each respondent and a set of *discriminatory* and *difficulty* parameters for each question. This model can be extended to its polytomous version, allowing for answers with multiple categories.[23] In this context, different models are available such as the Rating Scale Model,[33,34] the Graded Response Model,[35] and the Partial Credit Model.[24] We opt for the latter, which provides the most general analysis framework. Specifically, the Partial Credit Model can be seen as an extension of the Rating Scale Model where the questions are allowed to have categories varying in number and structure, an appealing feature in our application. On the other hand, the Graded Response Model specifies the cumulative distribution functions of the responses directly, modeling regions of the latent space separated by boundaries representing the categories of each item. This approach does not allow the algebraic separation of latent traits and question-specific parameters, preventing the identification of a sufficient statistic for either. The Partial Credit Model describes the probability that individual *i* answers the *j*th question with category *h* as follows:

$$\mathbb{P}(Y_{ij} = h | \alpha_j, \beta_j, \theta_i) = \frac{\exp(\alpha_j(h\theta_i - \sum_{l=0}^{h}\beta_{jl}))}{\sum_{h=0}^{m_j-1}\exp(\alpha_j(h\theta_i - \sum_{l=0}^{h}\beta_{jl}))}, \quad h = 0, \dots, m_j - 1, \tag{1}$$

where m_j is the number of categories available for the jth question. The latent variable θ_i is a univariate subject-specific parameters, while α_j and β_{jh} for $h = 0, \dots, m_j - 1$ are item-specific parameters. Model (1) is able to estimate the main features involved in the answering process. In particular, the ability parameter θ_i describes the attitude of the respondent toward the features that the questionnaire intends to measure. For instance, in our application this can be the perception of the mothers toward their children's behavior, or how the children view themselves through the self-reported behavioral questionnaires. For each question $j = 1, \dots, J$, the positive parameter α_j indicates the power of each question in discriminating between subjects, acting as a question-specific steepness parameter for the probability curve as a function of the ability parameter θ_i, while the parameter β_{jh} quantifies the difficulty (offset) of reaching the hth category, expressing how likely it is that a person would answer using category h rather than $h - 1$, for $h = 0, \dots, m_j - 1$, specifying regions of the latent space for the ability parameter θ_i which correspond to each answer category.

The interpretability of the parameters is an appealing feature of this model and explains its popularity in psychology and social sciences, which has led to the development of a variety of extensions, aimed at dealing with particular applications.[25,36,37] For example, extensions of model (1) to include covariates have been proposed.[38] In a longitudinal framework, where subjects answer questionnaires at different time points, random effects can be introduced in model (1), in order to account for time variation and be able to infer changes in the behavior/condition of the respondents as compared with baseline.[39] In a Bayesian framework, flexible hierarchical models can be specified allowing for time-varying parameters and sharing of information.[40] Furthermore, still in a Bayesian framework, different covariance patterns for the latent ability parameters can be specified, depending on the modeling assumptions.[41] An important scenario sees the inclusion of fixed effects to model the subjects' group membership (eg, when additional information about the respondents is available.[42]) However, clustering of the subjects can be one of the goals of the analysis. To this aim, we mention works from the Bayesian nonparametric literature,[43,44] where the latent trait parameters θ_i for $i = 1, \dots, I$ are modeled as a sample from a Dirichlet process (DP) prior.[45] The DP describes a prior over the space of probability distributions in such a way that its samples are almost surely discrete measures, that is if a random measure $P \sim \mathrm{DP}(\kappa, P_0)$, then P is almost surely discrete. In this notation, $\kappa \in \mathbf{R}^+$ represents the mass parameter controlling the departure from the base measure P_0, around which the DP is centered. A useful and straightforward constructive definition of the DP is provided by its stick-breaking representation:[46]

$$P(\cdot) = \sum_{k=1}^{\infty} \omega_k \delta_{\phi_k}(\cdot)$$

where $\delta_x(\cdot)$ is the Dirac measure equal to 1 at the location x and 0 otherwise. The infinite sequence of location parameters $\{\phi_k\}_{k=1}^{\infty}$ is i.i.d. from the continuous base measure P_0, while the weights $\{\omega_k\}_{k=1}^{\infty}$ follow the stick-breaking construction:

$$\omega_k = v_k \prod_{l=1}^{k-1}(1 - v_l), \quad k = 1, 2, \dots$$

$$v_1, v_2, \dots \overset{\text{iid}}{\sim} \mathrm{Beta}(1, \kappa).$$

Discreteness is a key feature of the DP exploited in our modeling choice, since it allows for ties in the sample of latent trait parameters, and therefore induces a clustering structure over the set of subject indices $\{1, \dots, I\}$. In particular, in this semi-parametric setting, the partition of the subject indices is itself a random variable and object of posterior inference.

In our application, we face various challenges: (i) the subject-specific ability parameters θ_i can be of difficult interpretation, since they describe complex psychological and behavioral traits, departing from the setting where the questions have correct/wrong answers; (ii) the sets of questions from the CBCL questionnaire vary over different time points in the maternal behavioral assessments, with only partial overlap between time points and (iii) we have responses from two related sides of the same cohort, mothers and children, answering different but related questionnaires. The latter is a problem that has been explored before, for instance, by employing a Graded Response Model and linking the questionnaires answered by different respondents by sharing the same latent ability parameters,[47] therefore not differentiating between informants in the model but rather focusing on the assessment of the same topic, or by simply analyzing the questionnaires

separately and comparing the results on a common scale.[48] In our case, however, we are interested in incorporating additional information regarding the informant/respondent, therefore maintaining the latent ability traits separated, and rather introduce dependency via a common random effect. These considerations require a nontrivial generalization of model (1).

First of all, we introduce some notation. Let $\boldsymbol{Y} = (Y_{ij}^t)$ and $\boldsymbol{X} = (X_{ij}^q)$ be the data arrays of responding to the mothers and the children, respectively. In this dataset, we have $I = 112$ pairs. Each entry Y_{ij}^t represents the response at time t of mother i to question j, for $t = 1, \ldots, T^Y = 4$, $i = 1, \ldots, I$, and $j \in \mathcal{J}^{Y^t}$, with $\mathcal{J}^{Y^t}$ indicating the set of questions asked at time t. Notice that the total number of questions is allowed to change over time and therefore the global set of questions answered by the mothers is indicated as $\mathcal{J}^Y = \cup_{t=1}^{T^Y} \mathcal{J}^{Y^t}$. The number of questions answered by the mothers are $J_t^Y := |\mathcal{J}_t^Y| = 99$ for $t = 1, 2, 3$ and $J_4^Y := |\mathcal{J}_4^Y| = 119$, with $\mathcal{J}_1^Y = \mathcal{J}_2^Y = \mathcal{J}_3^Y \neq \mathcal{J}_4^Y$ and an overlap of $|\mathcal{J}_1^Y \cap \mathcal{J}_4^Y| = 50$ questions, for a total number of unique questions equal to $J^Y := |\mathcal{J}^Y| = 168$. The variable Y_{ij}^t can take value $h \in \{0, \ldots, m_j^{Y^t} - 1\}$, with $m_j^{Y^t}$ possible answer categories to question j, for $j = 1, \ldots, J^{Y^t}$. Although the number of answer categories might vary with question and time, we have here that $m_j^{Y^t} = m^Y = 3$, for $j = 1, \ldots, J^Y$ and $t = 1, \ldots, T^Y$. Similarly, the entry X_{ij}^q of the children's dataset contains the response of child i to question j of questionnaire q, for $q = 1, \ldots, Q^X$. In our application, for the children's reports we have $Q^X = 2$ different questionnaires (CDI2 and MASC2) composed of $J^{X^1} = 28$ and $J^{X^2} = 50$ questions for the CDI2 and MASC2 questionnaires, respectively, and such that $m_j^{X^1} = 3$, for $j = 1, \ldots, 28$ and $m_j^{X^2} = 4$ for $j = 1, \ldots, 50$. It is possible that some of the interviewed mothers/children do not respond to some of the questions, or only respond partially, leaving the corresponding entries of the data arrays unobserved (ie, missing). These variables will be treated as missing-at-random, and therefore imputed in a Bayesian framework under this assumption by sampling from their full-conditional distributions (more details in Section 2 of Supplementary Material). As typical in longitudinal cohort studies, covariate information is also available. Let $\boldsymbol{U}^Y$ and $\boldsymbol{U}^X$ be the covariates associated to the mothers and the children, respectively. We include the covariate *Gender* in modeling both mothers' and children's latent traits (see Equation (5) below). For the mothers, additional covariates indicating the time from the first visit (in years), that is, a variable taking values 0, 1, 2, 5, as well as the four depression/anxiety indicators introduced in Section 2, are available. Thus, $\dim(\boldsymbol{U}^Y) = (I, q^Y)$, with $q^Y = 5$ and $\dim(\boldsymbol{U}^X) = (I, q^X)$, with $q^X = 1$. We specify a joint model for $(\boldsymbol{Y}, \boldsymbol{X})$:

$$p(\boldsymbol{Y}, \boldsymbol{X}|\boldsymbol{U}^Y, \boldsymbol{U}^X) = \int p(\boldsymbol{Y}, \boldsymbol{X}|\boldsymbol{U}^Y, \boldsymbol{U}^X, \boldsymbol{\vartheta})p(d\boldsymbol{\vartheta}) = \int p(\boldsymbol{Y}|\boldsymbol{U}^Y, \boldsymbol{\vartheta}^Y)p(\boldsymbol{X}|\boldsymbol{U}^X, \boldsymbol{\vartheta}^X)p(d\boldsymbol{\vartheta}) \tag{2}$$

and introduce dependence between the two sets of data by jointly modeling the latent parameter vector $\boldsymbol{\vartheta} = (\boldsymbol{\vartheta}^Y, \boldsymbol{\vartheta}^X)$. Conditionally on $\boldsymbol{\vartheta}$ the children and mothers' responses are independent. Building on (1), we propose the following joint model:

Likelihood

$$p(\boldsymbol{Y}, \boldsymbol{X}|\boldsymbol{U}^Y, \boldsymbol{U}^X, \boldsymbol{\vartheta}) = p(\boldsymbol{Y}|\boldsymbol{U}^Y, \boldsymbol{\vartheta}^Y)p(\boldsymbol{X}|\boldsymbol{U}^X, \boldsymbol{\vartheta}^X) = \prod_{i=1}^{I} \left(\prod_{t=1}^{T^Y} \prod_{j \in \mathcal{J}^{Y^t}} p_{ij}^{Y^t} \right) \left(\prod_{q=1}^{Q^X} \prod_{j=1}^{J^{X^q}} p_{ij}^{X^q} \right), \tag{3}$$

$$p_{ij}^{Y^t} = \mathbb{P}(Y_{ij}^t = h|\boldsymbol{\vartheta}^Y) = \frac{\exp(\alpha_j^Y(h\theta_{it} - \sum_{l=0}^{h}\beta_{jl}^Y))}{\sum_{h=0}^{m^Y-1}\exp(\alpha_j^Y(h\theta_{it} - \sum_{l=0}^{h}\beta_{jl}^Y))}, \quad j \in \mathcal{J}^{Y^t}, \quad i = 1, \ldots, I$$

$$p_{ij}^{X^q} = \mathbb{P}(X_{ij}^q = h|\boldsymbol{\vartheta}^X) = \frac{\exp(\alpha_j^{X^q}(h\psi_i - \sum_{l=0}^{h}\beta_{jl}^{X^q}))}{\sum_{h=0}^{m^{X^q}-1}\exp(\alpha_j^{X^q}(h\psi_i - \sum_{l=0}^{h}\beta_{jl}^{X^q}))}, \quad j = 1, \ldots, J^{X^q}, \quad i = 1, \ldots, I.$$

Item parameters

$$\boldsymbol{\gamma}_j^Y = \left(\log\left(\alpha_j^Y\right), \beta_{j0}^Y, \ldots, \beta_{jm^Y-1}^Y \right) \sim N_{m^Y}\left(\boldsymbol{0}_{m^Y}, \mathbb{I}_{m^Y}\right), \quad \sum_{h=0}^{m^Y-1} \beta_{jh}^Y = 0, \quad j \in \mathcal{J}^{Y^t}$$

$$\boldsymbol{\gamma}_j^{X^q} = \left(\log\left(\alpha_j^{X^q}\right), \beta_{j0}^{X^q}, \ldots, \beta_{jm^{X^q}-1}^{X^q} \right) \sim N_{m^{X^q}}\left(\boldsymbol{0}_{m^{X^q}}, \mathbb{I}_{m^{X^q}}\right), \quad \sum_{h=0}^{m^{X^q}-1} \beta_{jh}^{X^q} = 0, \quad j = 1, \ldots, J^{X^q}. \tag{4}$$

Subject parameters

$$
\begin{aligned}
\theta_{it} &= \theta_i^0 + \rho_i \theta_{it-1} + \boldsymbol{U}_{it}^Y \boldsymbol{\eta}^\theta + \epsilon_{it}, \quad i = 1, \dots, I, \quad t = 2, \dots, T^Y \\
\theta_{i1} &= \theta_i^0 + \boldsymbol{U}_{i1}^Y \boldsymbol{\eta}^\theta + \epsilon_{i1}^\theta, \quad i = 1, \dots, I \\
\epsilon_{i1}^\theta, \dots, \epsilon_{iT^Y}^\theta &\overset{\text{iid}}{\sim} \mathrm{N}\left(0, \sigma_{\epsilon^\theta}^2\right), \quad i = 1, \dots, I \\
\psi_i &= \psi_i^0 + \boldsymbol{U}_i^X \boldsymbol{\eta}^\psi + \epsilon_i^\psi, \quad i = 1, \dots, I \\
\epsilon_1^\psi, \dots, \epsilon_I^\psi &\overset{\text{iid}}{\sim} \mathrm{N}\left(0, \sigma_{\epsilon^\psi}^2\right).
\end{aligned}
\tag{5}
$$

Hyperpriors

$$
\begin{aligned}
\boldsymbol{\eta}^\theta &\sim \mathrm{N}_{q^Y}\left(\boldsymbol{0}_{q^Y}, \mathbb{I}_{q^Y}\right) \\
\boldsymbol{\eta}^\psi &\sim \mathrm{N}_{q^X}\left(\boldsymbol{0}_{q^X}, \mathbb{I}_{q^X}\right) \\
\sigma_{\epsilon^\theta}^2, \sigma_{\epsilon^\nu}^2 &\sim \mathrm{IG}(3, 2).
\end{aligned}
\tag{6}
$$

Nonparametric component

$$
\begin{aligned}
\boldsymbol{\phi}_i &= \left(\theta_i^0, \psi_i^0, \rho_i\right), \quad i = 1, \dots, I \\
\boldsymbol{\phi}_1, \dots, \boldsymbol{\phi}_I | \kappa, P &\overset{\text{iid}}{\sim} P, P \sim \mathrm{DP}(\kappa, P_0) \\
P_0(\theta^0, \psi^0, \rho) &= \mathrm{N}(\theta^0 | 0, 1) \mathrm{N}(\psi^0 | 0, 1) \mathrm{Unif}(\rho | -1, 1),
\end{aligned}
\tag{7}
$$

where $\boldsymbol{0}_d$ is a d-dimensional vector of zeros and $\mathbb{I}_d$ is the d-dimensional identity matrix. The vector of latent variables $\boldsymbol{\vartheta} = (\boldsymbol{\vartheta}^Y, \boldsymbol{\vartheta}^X)$ is such that $\boldsymbol{\vartheta}^Y = (\boldsymbol{\gamma}^Y, \boldsymbol{\theta})$ and $\boldsymbol{\vartheta}^X = (\boldsymbol{\gamma}^{X^1}, \dots, \boldsymbol{\gamma}^{X^{Q^X}}, \boldsymbol{\psi})$ with $\boldsymbol{\gamma}^Y = (\boldsymbol{\gamma}_j^Y)_{j \in \mathcal{J}^Y}$ and $\boldsymbol{\gamma}^{X^q} = (\boldsymbol{\gamma}_j^{X^q})_{j=1, \dots, J^{X^q}}$, for $q = 1, \dots, Q^X$. This description emphasizes the inclusion of parameters $\boldsymbol{\gamma}^Y$ for the mothers' responses that are shared between time points, and parameters $(\boldsymbol{\gamma}^{X^1}, \dots, \boldsymbol{\gamma}^{X^{Q^X}})$ for the children's responses that are questionnaire-specific. Therefore, the parameters $\boldsymbol{\gamma}_j^Y$ are indexed over the global set of questions for $j \in \mathcal{J}^Y$ and do not change over time, allowing for sharing of information between time points. This modeling choice reflects the assumption that the questions in the CBCL questionnaires, some which are common through all time points, aim at measuring the same characteristics of the mothers' behavioral assessment. Due to the high number of free parameters in the model, in order to ensure identifiability, we impose a constraint on the difficulty parameters in (4), as discussed in previous work on the subject.[40] This implies that the prior distribution the difficulty parameters is defined over a lower dimensional space, imposing a constraint on one of the β_{jh}^Y and β_{jh}^X (in our case we choose $h = 0$). Furthermore, the prior distribution of the parameters $\boldsymbol{\gamma}^Y, \boldsymbol{\gamma}^{X^1}, \dots, \boldsymbol{\gamma}^{X^{Q^X}}$ has unitary variance.

The trait parameters for the mothers θ_{it} are modeled via a covariate-dependent autoregressive model, which is able to incorporate information about the evolution of the respondent's profile through time, as well as covariate information through the vectors $\boldsymbol{U}_{it}^Y$, for $i = 1, \dots, I$ and $t = 1, \dots, T^Y$. In particular, referring to the description of Section 2, the covariate vectors contain the variables *Gender*, *BDI*, *EPDS*, *STAI*, and a time-varying component representing the time from the first questionnaire (in years). The latter is used to provide information on how the elapsed time influences the respondent's profile,[49] besides the autoregressive term. For the children, since we only have one time point, the trait parameters ψ_i are modeled via linear regression using the covariate vectors $\boldsymbol{U}_i^X$, for $i = 1, \dots, I$, containing only *Gender*. The regression coefficients are a priori normally distributed with mean zero and variance one, as the continuous covariates are standardized so that their effect is comparable. The interpretation of the latent trait parameters is very important as, in our application, they are proxy for complex psychological behavioral assessments. In the proposed model, increasing values of the parameters θ_{it} and ψ_i correspond to respondents who are more prone to assessing the behavior described in the questions as happening more often or being more likely to be observed. Considering the nature of the specific questions posed in the CBCL, CDI2, and MASC2 questionnaires, higher values of the latent trait parameters translate into more problematic behavioral assessments. For example, for the jth question asked to the mothers characterized by the parameters $\boldsymbol{\gamma}_j^Y$, a positive value of the latent trait is associated with higher probability of answering category *very true* (ie, 2), which is associated with a more "negative/problematic" description of the child's behavior. A similar interpretation holds for the children's questionnaire, reflecting their perception of themselves. However, the latent trait parameters of the mothers and children are defined over different latent scales in our application.

Questionnaire data from mothers and children are linked through the random effect vectors $\boldsymbol{\phi}_i = (\theta_i^0, \psi_i^0, \rho_i)$, for $i = 1, \ldots, I$, which are modeled as time invariant. The vector $\boldsymbol{\phi}_i$ contains parameters influencing the mothers' behavioral perception of their children θ_i^0, the children's self-perception of behavior ψ_i^0, as well as an autoregressive coefficient ρ_i, which captures the time evolution of the maternal latent traits. These parameters can be seen as drivers of the respondents' profiles. Therefore, it is of interest to be able to clusters the mother and child pairs based on their response patterns through these parameters. To this end, we specify a joint distribution for these variables, by modeling the vectors $\boldsymbol{\phi}_1, \ldots, \boldsymbol{\phi}_I$ with a DP prior. The discreteness of this measure allows us to cluster the parameter vectors $\boldsymbol{\phi}_1, \ldots, \boldsymbol{\phi}_I$ based on ties in the sample and induces a partition of the observations a-posteriori. Specifying the base measure P_0 is straightforward if independence between the components of the vector $\boldsymbol{\phi}$ is assumed a priori, as in our case.

In Supplementary Material Section 3.2, we compare the performance of the proposed model with competitor models. In particular, we consider a parametric version of our model, in which all the observations are assumed to come from the same cluster. Furthermore, we investigate the performance of the following extensions of our modeling strategy: (i) subject-specific regression coefficients in Equation (5), that is, $\boldsymbol{\eta}_i^\theta$ and $\boldsymbol{\eta}_i^\psi$, for $i = 1, \ldots, I$ and the DP prior is defined on the augmented vector $\boldsymbol{\phi}_i = (\theta_i^0, \psi_i^0, \rho_i, \boldsymbol{\eta}_i^\theta, \boldsymbol{\eta}_i^\psi)$; (ii) a model in which the AR(1) dependency in the mothers' temporal profile is specified directly on the residuals of the latent process. In addition, the regression coefficients are time-specific, that is, $\boldsymbol{\eta}_t^\theta$ for $t = 1, \ldots, T^Y$ and the covariate *time from first questionnaire* is removed; (iii) a model where instead of modeling the temporal dependence with an autoregressive component, hence linearly, we introduce time-specific intercepts, that is, $\theta_i^0 = (\theta_{i1}^0, \ldots, \theta_{iT^Y}^0)$ and we specify the DP prior directly on the augmented vector $\boldsymbol{\phi}_i = (\theta_i^0, \psi_i^0)$. In addition, the regression coefficients are time-specific, that is, $\boldsymbol{\eta}_t^\theta$ for $t = 1, \ldots, T^Y$ and the covariate *time from first questionnaire* is removed. This latter model allows for a nonlinear temporal effect on the questionnaire temporal profiles. Our comparison shows that the proposed strategy is robust in terms of clustering results and goodness-of-fit, still maintaining interpretability. In Supplementary Material, we also highlight the properties of the alternative approaches. We note that model (iii) is the best competitor in terms of clustering properties, flexibility and regression results, although the extra flexibility deriving from not assuming a linear temporal dependence comes at the cost of increasing the dimension of the parameter space.
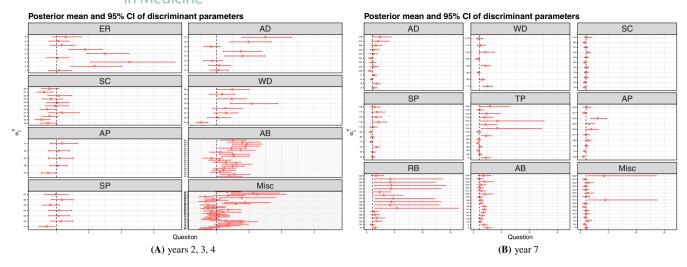
4 | POSTERIOR INFERENCE ON MOTHER/CHILD PAIRS

We fit model (3) to the $I = 112$ mother/child pairs introduced in Section 2. The prior distributions assumed for the item-specific parameters, reported in (4), are multivariate Gaussian distribution with null mean vector and identity covariance matrix. The scalar variances $\sigma_{\epsilon^\theta}^2$ and $\sigma_{\epsilon^\psi}^2$ modeling the variances of the subject-specific parameters are a priori distributed as inverse-gamma random variables with unitary mean and variance. The mass parameter κ of the DP is chosen such that the a priori expected number of clusters K_I is approximately three, yielding $\kappa = 0.4202$. In Supplementary Material Section 3.1, we perform a sensitivity analysis to the choice of the parameter κ, and show that posterior inference is sensitive (in terms of estimated number of clusters) to such choice. This is typical of nonparametric priors based on the DP and the problem could be overcome by setting a hyperprior on κ. Finally, we fix the hyperparameters of P_0 as reported in (7), using standard Gaussian distributions for the univariate parameters θ^0 and ψ^0, and a noninformative Uniform distribution defined over the interval $(-1, 1)$ for the AR(1) parameter ρ.

Inference for the proposed model is obtained by running a Markov Chain Monte Carlo (MCMC) algorithm[50] with code written in Rcpp. Most of the steps in the MCMC algorithm involve updating parameters that are nonconjugate in the proposed model. These steps are performed using an adaptive Metropolis-Hastings approach within the main Gibbs sampler.[51] The cluster allocations are updated using an algorithm for nonconjugate DP mixture models.[52,53] After an initial 100 iterations used to initialize the adaptive updates, the model is run for 50 000 iterations, of which the first 40 000 are discarded as burn-in and the last 10 000 are thinned every two iterations, yielding a final sample of size 5000 which is used for posterior inference.

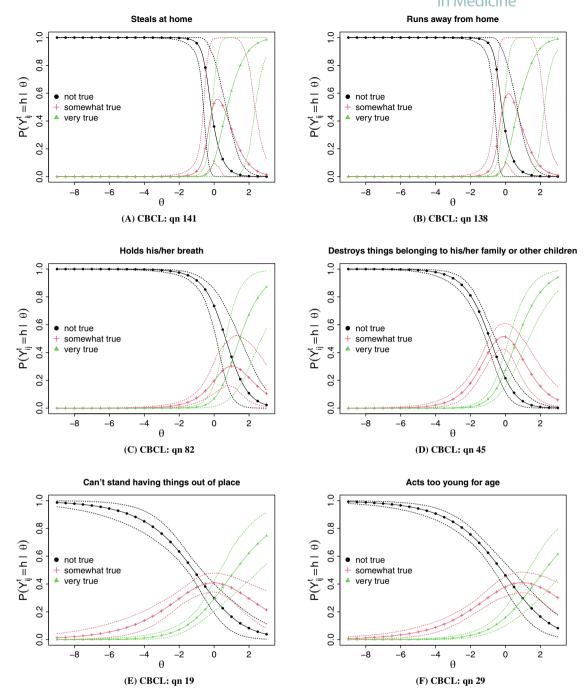## 4.1 | Analysis of question-specific parameters

The posterior distributions of the item-specific parameters γ^Y and γ^X are among the main objects of inference, since they allow identifying those questions that are able to differentiate between respondent profiles. Figure 1 shows the posterior mean and 95% credible intervals for the discriminatory parameters α_j^Y relative to the questions contained in the CBCL ($J^Y = 168$), grouped into clinical subscales for the first three and last time points, as reported in Supplementary Table S1.

FIGURE 1 Posterior means and 95% credible intervals for the discrimination parameters α_j^Y for the Child Behavior Checklist questionnaire, grouped into its clinical subscales. The abbreviations are defined in Supplementary Table S2. Vertical dashed lines indicate the reference value 1. Questions asked at years 2, 3, and 4 (A) or 7 (B) [Colour figure can be viewed at wileyonlinelibrary.com]

We observe how, within each subscale group, some questions have more discriminatory power, especially in the first time points where we have more information. Some of the subscales are characterized by higher variability in the distributions of the discriminatory parameters, probably due to the different ways mothers answered the questions contained in these subscales. For instance, while most of the questions at year 7 are characterized by narrow posterior credible intervals, indicating little uncertainty around the estimation of the corresponding discriminatory parameters, we observe high variability in the subscale *rule-breaking behavior*. An explanation of such result can be found in the nature of the questions contained in this subscale, often involving negative or extreme behaviors, which are not usually encountered by the majority of the population, therefore strongly identifying different subjects based on their answers to these questions, and requiring higher variability of such parameters. In order to better understand this behavior on a question-specific level, item characteristic curves (ICCs) are usually shown in IRT analysis. The ICCs represent the probability of answering a question with a specific category as a function of the latent trait parameter θ. By studying ICCs, one is able to locate the values of the latent traits where the propensity for a category increases with respect to the previous one, as indicated by the intersection of the two corresponding curves. Figure 2 shows some examples of ICCs for those questions in $\mathcal{J}^y$ for which the posterior medians of the corresponding discriminatory parameters α_j^Y are the highest, median and lowest (top, middle and bottom rows). The ICCs are plotted over the posterior range of values of θ spanned during the MCMC sample. The questions characterized by very high discriminatory parameters show ICCs with higher slopes, indicating clear areas of the latent space associated with a specific category. In the case of questions number 141 and 138, shown in the top row of Figure 2, larger credible intervals are also observed for category *somewhat true*, indicating an almost dichotomous characterization of the questions. On the other hand, notice the increasing flatness of the ICCs when the discriminatory power of the question decreases, allowing for wider ranges of values of θ where the middle category is preferred.

We now discuss the children's questionnaires. Figure 3 shows the posterior mean and 95% credible intervals for the discriminatory parameters $\alpha_j^{X^q}$ for the CDI2 and MASC2 questionnaires, respectively, each grouped according to the their subscales as reported in Supplementary Table S1. We observe how most of the posterior distributions of these parameters are centered around similar values for CDI2, indicating analogous discriminating power of such questions. However, this is not the case for the questions belonging to the MASC2 questionnaire, where many range below the reference unitary value (eg, *Separation anxiety/Phobias*, *Generalized Anxiety Disorder*, *Obsessions/Compulsions*, *Tense/Restless*, and *Harm avoidance* subscales). Following the same approach used for the investigation of the mothers' responses, we select the questions with highest/median/lowest posterior median of the discrimination parameters $\alpha_j^{X^q}$, for $q = 1, 2$. Their ICCs are shown in Figure 4. Once again, we identify the questions with highest posterior median (top row) to have a more clear separation between categories in association with the value of the latent trait parameter ψ, as characterized by higher slopes. By contrast, the bottom row shows how the middle category in question 15 of CDI2 and the categories *rarely* and *sometimes* in question 15 of MASC2 have overlapping credible intervals with the other categories for most values of ψ,
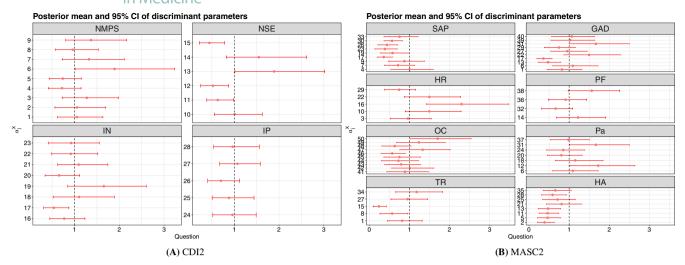
FIGURE 2 Posterior item characteristic curves and their 95% credible intervals (dashed lines) for questions presenting decreasing discriminatory power (top = highest, middle = median, bottom = lowest). The discriminatory power of each question is assessed by ranking the posterior median of the discriminatory parameters α_j^Y [Colour figure can be viewed at wileyonlinelibrary.com]

indicating lower discriminating power in terms of respondent propensity to pick a particular category. A more gradual shift between the curves for each categories is observed in the middle row.
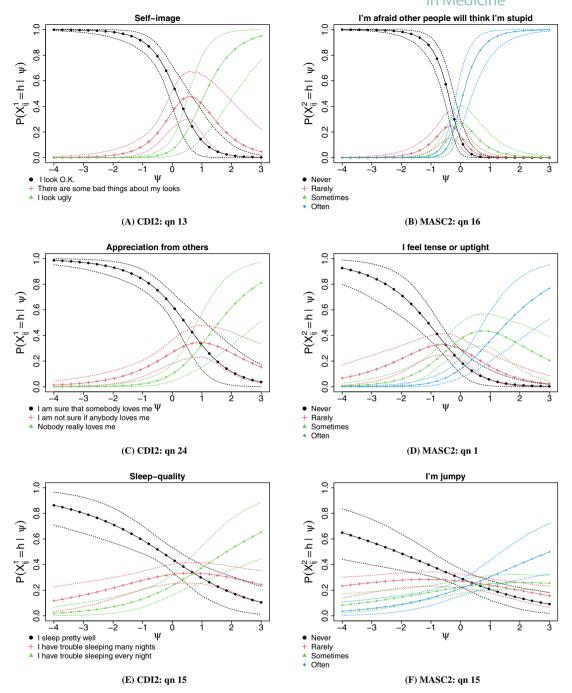
4.2 | Clustering analysis

The latent traits $\theta_{i1}, \ldots, \theta_{iT^Y}$ for the mothers and ψ_i for the children incorporate features regarding the answering profiles of the respondents pairs $i = 1, \ldots, I$. Due to the complex phenomenon under study, the interpretation of these

FIGURE 3 Posterior means and 95% credible intervals for the discrimination parameters $\alpha_j^{X^q}$ for CDI2 (A) and MASC2 (B), grouped into their respective subscales. The abbreviations are defined in Supplementary Table S2. Vertical dashed lines indicate the reference value 1 [Colour figure can be viewed at wileyonlinelibrary.com]

parameters represents a challenge. As we observed in the ICCs curves presented in the previous section, the range of parameter values associated with a specific answer category is dictated by the posterior estimates of the coefficients γ^Y and γ^X. In general, small values of the latent subject-specific parameters are related to high probability of answering the lowest category (eg, *not true*, *never*), and vice-versa large values of these parameters are associated with higher categories (eg, *very true*, *often*). In general, we can describe the behavior of the latent profiles by looking at the predictive densities of the latent variables responsible for their clustering structure, that is, of $\boldsymbol{\phi} = \left(\theta^0, \psi^0, \rho\right)$. See Figure 5. As it is easily observed from the multimodality of these predictive distributions, the driving variables for clustering are θ^0 and ρ, indicating that the mothers' latent profile contribute the most to the partition structure. This is an interesting result, suggesting that the mothers exhibit a more diverse response behavior than their children.
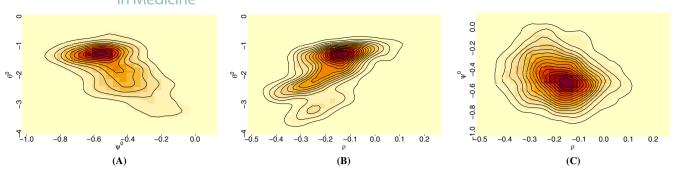
Thanks to the specification of a Bayesian nonparametric prior, the posterior distribution of the vector of parameters $\boldsymbol{\phi} = \left(\theta^0, \psi^0, \rho\right)$ allows the identification of clusters of mother/child dyads characterized by similar pairs of response profiles. First, we provide a posterior estimate of the partition of the data by minimizing the Binder's loss function with equal misclassification costs.[54] This approach minimizes a loss function assigning equal costs to each type of misclassification error, that is, when two subjects are clustered together rather than separately, and vice-versa. The Binder's loss is then obtained as the sum of such costs over all possible pairs of data points, and its expectation is minimized using the MCMC output of the algorithm.[55,56] The computation of this estimate is achieved via the R package `salso`[57] and we refer to it as *Binder partition*. The matrix of posterior coclustering probabilities used to minimize the Binder's loss criterion, as well as the posterior distribution of the number of clusters, are reported in Supplementary Figure S12. The estimated Binder partition is composed of $K_I = 6$ clusters of sizes (60, 34, 7, 6, 4, 1), respectively, and labeled according to their cluster sizes in decreasing order. Due to label switching problems,[58] estimates of the posterior distribution of the latent parameters $\boldsymbol{\phi}$ are not trivially obtained from the MCMC output. Therefore, we study the posterior distribution of the mean of the parameters $\boldsymbol{\phi}$ within each cluster.[59] To this end, we first fix the partition of the subjects to the estimated Binder partition, and compute, for each MCMC iteration, the average of the values of $\boldsymbol{\phi}$ over the subjects within each cluster and then we explore their posterior distribution. Recall that $K_I = 6$ is the number of clusters for the Binder estimated partition and Figure 6 reports such posterior distributions (one for each cluster). The cluster-specific posterior distributions of the autoregressive coefficient ρ and the children's latent trait intercept ψ^0 (two bottom rows of Figure 6) are concentrated around negative values for most of the clusters, indicating latent traits oscillating more in time (ρ), and that overall lower categories are preferred for the questions answered by children (ψ^0), indicating respondents more prone to positive characterization of behavior (eg, $0 = I$ *am sad once in a while* is better than $1 = I$ *am sad many times*, or than $2 = I$ *am sad all the time*). Interestingly, Cluster 3 presents higher variability and spans a range of values including both negative and positive terms for the parameter ρ, but it is centered around zero. On the other hand, the intercept parameter θ^0 presents strong differences between clusters, although all distributions are concentrated on negative values, with highest values
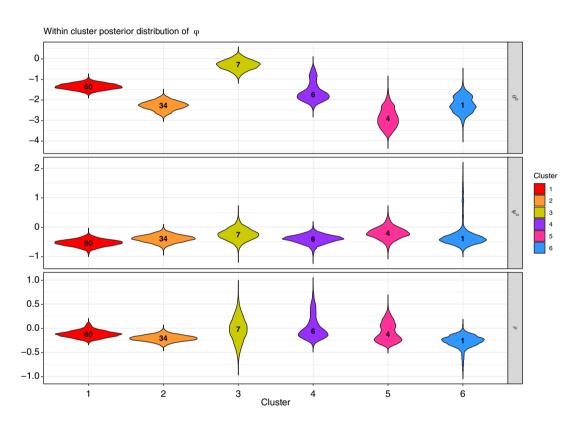
FIGURE 4 Posterior item characteristic curves and their 95% credible intervals (dashed lines) for questions presenting decreasing discriminatory power (top = highest, middle = median, bottom = lowest). The discriminatory power of each question is assessed by computing the posterior median of the discriminatory parameters $\alpha_j^{X^q}$, for $q = 1, 2$ [Colour figure can be viewed at wileyonlinelibrary.com]

for the ones relative to Cluster 4. The three largest clusters also show higher precision in the posterior distributions of the averages of θ^0, while still exhibiting opposite behavior, indicating discordance in the assessments between mothers in these clusters. For instance, the assessment of Cluster 1 is better than Cluster 3, since the average values of θ^0 are overall lower, but it is worse than that of Cluster 2.

The relationship between the clustering structure and the way the subjects within each cluster respond to the questionnaires is presented in details in Supplementary Figures S13 and S14 (for the same sets of questions previously highlighted in Figures 2 and 4), showing the posterior expected value curves (EVCs) and their 95% credible interval. EVCs represent the mean answers to each question as a function of the latent traits, easily obtainable from the MCMC output. The

FIGURE 5 Contour plots for the bivariate predictive distribution of the latent variables $\boldsymbol{\phi} = (\theta^0, \psi^0, \rho)$ [Colour figure can be viewed at wileyonlinelibrary.com]

FIGURE 6 Violin plots of the mean values of θ^0, ψ^0, and ρ within six clusters identified by the Binder partition. The numbers within the violin plots represent the cluster size [Colour figure can be viewed at wileyonlinelibrary.com]

EVCs are overlapped with the posterior means and 95% credible intervals of the average latent trait parameters θ and ψ within each cluster (and, for the mothers, for the two extreme time points), as depicted by the horizontal lines in the figures. From the mothers' EVCs, the separation between the way subjects within clusters respond is very clear, as highlighted by the different locations of the estimates of θ for each cluster. In particular, in all questions presented here, Cluster 3 is associated with the most negative assessment of the children's behavior (ie, corresponds to the highest values of θ). However, in all clusters apart from the single-dyad Cluster 6, an improvement between year 2 and year 7 is observed (ie, the values of θ at year 7 are lower than at year 2). Another important aspect highlighted by these figures, is the relationship between the clusters and the item parameters. Specifically, questions with higher discriminatory power (ie, top rows of Figures 2 and 4 and of Supplementary Figures S13 and S14) are able to isolate the most negative Cluster 3 from the other clusters more neatly. By looking at Figure Supplementary Figure S14 for the children's assessment, we observe a similar distinction between clusters in the questions from the CDI2 and MASC2 questionnaires, again

showing a strong relation between the discriminatory power of each question and the cluster separation. Interestingly, while in Supplementary Figures S13 Cluster 5 presents the lowest values for θ for the CBCL questions, Supplementary Figures S14 shows that this is true for Cluster 4, again capturing a discrepancy in assessment between mothers and children.

4.3 | Analysis of covariate effects

The model for latent trait parameters contains a regression component, including an autoregressive term for the maternal questionnaires that are observed at several time points. In a Bayesian framework, the overall effect of covariates can be established by computing the posterior 95% credible intervals (95% CI) of the corresponding coefficients, here denoted as $\boldsymbol{\eta}^\theta$ and $\boldsymbol{\eta}^\psi$ for the two groups of subjects. Whenever the credible interval contains the value zero, we regard the corresponding covariate as not having a relevant effect in explaining the subjects' latent traits. The posterior means and credible intervals for the coefficients $\boldsymbol{\eta}^\theta$ and $\boldsymbol{\eta}^\psi$ are reported in Supplementary Figure S15. In this analysis, the time from the first questionnaire has a strong effect on the maternal latent trait evolution (95% CI = $(-0.18, -0.09)$), inducing lower values of the mothers' latent traits as time passes, indicating an improvement in their behavioral assessment over time. This finding is not surprising as older children, with language development, become more responsive to parental instruction. By contrast, the maternal mental health index *BDI* shows a negative effect on the maternal rating (95% CI = $(0.10, 0.70)$). For the children's self-reporting, the only covariate included (*Gender*) has high variability although posterior mean relatively far away from zero (95% CI = $(-0.37, 0.04)$).

To further characterize the clusters in the Binder partition, we look at the empirical distribution of the time-homogeneous covariates within each cluster. To have an indication of which covariates differ the most among clusters, we perform a χ^2-test for the categorical covariate *Gender* and Kruskal-Wallis test for the continuous mental health indicators (analyses not shown). None of the covariates is associated with a significant test result, pointing toward the existence of further (unmeasured) factors driving the clustering.

4.4 | Scale-specific analysis

In this section, we combine the information provided by the available subscales, reported exhaustively in Supplementary Table S2, and the posterior clustering estimate, with the aim of understanding the behavior of the respondents from a clinical point of view.

In the analysis of questionnaire data via IRT models, a widely used summary statistics for the latent trait parameters is the *raw score*, obtained from each respondent as the sum of all the answers in the questionnaire or a specific subscale. Such statistic is sufficient for the latent trait parameters in the proposed model.[33] For each subscale of interest, conditionally on the Binder partition, we can compute the raw score statistic by sampling answers from the predictive probabilities $p_{ij}^{Y^t}$, for $j \in \mathcal{J}^{Y^t}$ and $t = 1, \ldots, T^Y$ for the mothers and $p_{ij}^{X^q}$, for $j = 1, \ldots, J^{X^q}$ and $q = 1, \ldots, Q^X$ for the children. The predictive probabilities of answering a specific category within each cluster are estimated by averaging over the posterior MCMC samples, where at each iteration the subject-specific parameters θ_{it} are replaced with their average within the clusters.[59] The raw scores so obtained are representative of the subjects' behavior within each cluster, providing clinical information regarding the behavior and well-being of the children. The distribution of such raw scores for a selection of composite subscales, for both mothers and children data in the three largest cluster, are shown in Supplementary Figures S16 to S18 for internalizing and emotional problems, anxiety and depression, and dysregulation, respectively. The latter is a composite scale of clinical interest, and is composed of scores resulting from aggregating the questions in the *aggressive behavior*, *anxious/depressed*, and *attention problems* CBCL subscales (also termed AAA).[60] The differences in the predictive distributions of the raw scores within each cluster are more evident in the mothers' subscales than in the children's ones. Furthermore, Clusters 1 and 2 appear to be similar with respect to children's self-assessment. This confirms the results discussed in Section 4.2, highlighting a discrepancy between mothers' and children's profiles. In addition, from the first two rows of Supplementary Figure S16 and the left column of Supplementary Figure S17, it is observable a general shift to smaller values in the predictive distributions of the raw scores within each cluster, indicating an improvement in the maternal assessment over time, and confirming the phenomenon already noticed in Supplementary Figure S13.

5 | DISCUSSION

Our analysis points to the existence of three main mother-child clusters based on a latent trait reflecting the nature of the maternal assessment of the child, jointly with an analogous latent trait reflecting the nature of the child self-assessment (see Figures 5 and 6). In brief, all mothers in the three largest clusters show improved child ratings over time, with mothers in Clusters 1 and 2 overall more positive in their assessments. Mothers in Cluster 3 are generally more negative (see Supplementary Figure S13). However, the same is not completely matched by the children's self-reports (see Supplementary Figure S14), who show very similar behavior in Clusters 1 and 2, supporting the evidence of bias in the mothers reporting process. These findings reflect the previously reported discrepancy between maternal reports and those of their children.[61] Furthermore, the inclusion of mental health indicators as covariates in the model show that only the maternal mental health index *BDI* is a global factor in characterizing the mothers' latent traits, while none contributes to the characterization of the clusters.

5.1 | Maternal mental health

Rater bias is a major concern in the assessment of child psychopathology (see Introduction). The greatest concern in the literature is that of a maternal bias in child reports linked to maternal mental health. Our analysis provides evidence for a global association between maternal response biases and maternal mental health (see Section 4.3). As such, our longitudinal analyses reveal evidence that the critical latent trait that reflects the response pattern of the mothers is associated with maternal mental health. In details, our findings suggest some evidence of an association between maternal symptoms of depression *BDI*, but not those of anxiety *STAI*. This finding is consistent with a detailed study comparing potential maternal reporting biases showing no evidence for symptoms of anxiety when compared directly with those of depression.[61] Furthermore, as noted above, our clustering of mothers provides a unique ability for a more nuanced analysis exploring whether a mental health-linked maternal bias might appear in a subset of mothers. In addition, there is evidence of reporting bias when clinical raw scores between mothers and children are compared (Supplementary Figures S16-S18). This is consistent with studies reporting maternal bias associated with mental health status[6,8] which examine mothers with clinical levels of depressive symptoms. Studies reporting evidence of a maternal bias compare reports of mothers with and without clinical levels of depression. An extensive analysis[62] provides compelling evidence for a maternal bias among clinically depressed mothers. Maternal reports using the CBCL of child behavior problems by mothers who are depressed and/or anxious at the time they provide the report are systematically different from those of unimpaired mothers when compared with child self-reports (using the Youth Self-Report[62]). It has been found[20] that children of mothers with a history of major depression have more internalizing problems according to both mothers and teachers. When comorbidity is taken into account, mothers with symptoms of anxiety, a combination of depression and anxiety, or substance abuse appear to overstate their children's internalizing and externalizing behavioral problems, but no evidence of overreporting is found when mothers are depressed but not anxious. The majority of the depressed mothers in a previous study[62] shows comorbid anxiety. Indeed there are other streams of evidence showing that mothers with comorbid anxiety and depression represent a significantly more affected population.[63]

5.2 | Gender

Interestingly in this analysis the gender of the child is not a major factor driving variation in the maternal latent traits across all dyads (Supplementary Figure S14, left panel). This finding is consistent with previous work[4] that finds a gender bias in reports of teachers, but not parents. However, *Gender* is a relevant factor in the modeling of the latent traits for the children's answering profiles (Supplementary Figure S14, right panel). This finding is consistent with known gender differences in the self-reports of symptoms of depression and anxiety, with greater evidence of problems in girls.[2]

5.3 | Child self-reported symptoms

We find that the children's self-report in the three main clusters are consistent with each other, with slightly higher values of the raw scores for Cluster 3, and lowest values reported in Clusters 1 and 2 (see Supplementary Figures S16-S18). This

is not completely consistent with the maternal report. For example, our analysis identifies maternal Clusters 1 and 2 to be associated with the lowest, but different, maternally reported Anxiety and Depression features, but the same is not observed among the children, where the predictive distributions of the raw scores are overlapping (Supplementary Figure S17). Recall that child self-report measures are obtained at about 9 years of age, 2 years following the final maternal CBCL reports. The same pattern is observed for internalizing and externalizing problems. Thus, the relation between the maternal response latent trait and the child's self-report is apparent across a range of distinct behavioral problems.

5.4 | Conclusions

The association between parental mental health and that of the offspring has been established for some time.[64] However, there is a long-standing methodological concern that emphasizes the weakness of the parent as both the source of risk and the reporter of child outcomes.[65] The concern, simply put, is that more depressed mothers tend to overrate the mental health problems of their children. Previous analyses of the issue reveal highly mixed results with reports of essential no effects,[66] small effects[8] or moderate effects.[6] Detailed analyses provide a more nuanced understanding. First, there is evidence for a greater maternal reporting bias associated with depressive symptoms when rating boys compared with girls.[8] Second, there is compelling evidence showing that the influence of a depression-related maternal bias depends not only on gender, but on the reported outcome: internalizing problems reveal greater evidence for a maternal reporting bias than do externalizing problems,[19] with evidence that certain emotional functions reveal greater bias than do others.[67]

A limitation of the existing literature is a reliance on crosssectional data. In this work, we analyzed maternal reports of their children's behavior in conjunction with symptom self-reports from the children obtained from clinical psychometric questionnaires commonly used as screening tools in child clinical studies. The availability of both types of reporting systems allows for a deeper evaluation of the psycho-pathological assessment of the mother-child dyads. Furthermore, the mothers responded to questionnaires at multiple time points adding a valuable longitudinal dimension to the study. A significant and novel contribution of this analysis is the ability to identify specific mother—child dyadic clusters. The implication is that of variation across clusters such that the issue of potential reporting bias must consider the nature of the mother—child dyad. Among the measures of maternal symptoms of anxiety or depression, the *BDI* influences the mothers' profiles in the direction of greater negative assessment. However, our clusters did not associate with maternal mental health indicators nor with the gender of the child.

DATA AVAILABILITY STATEMENT
The data that support the findings of this study are available from the corresponding author upon reasonable request.

ORCID
Andrea Cremaschi https://orcid.org/0000-0003-1179-0768
Maria De Iorio https://orcid.org/0000-0003-3109-0478

REFERENCES
1. Achenbach TM, McConaughy SH, Howell CT. Child/adolescent behavioral and emotional problems: implications of cross-informant correlations for situational specificity. *Psychol Bull.* 1987;101(2):213.
2. De Los RA, Kazdin AE. Informant discrepancies in the assessment of childhood psychopathology: a critical review, theoretical framework, and recommendations for further study. *Psychol Bull.* 2005;131(4):483.
3. Youngstrom E, Loeber R, Stouthamer-Loeber M. Patterns and correlates of agreement between parent, teacher, and male adolescent ratings of externalizing and internalizing problems. *J Consult Clin Psychol.* 2000;68(6):1038.
4. Collishaw S, Goodman R, Ford T, Rabe-Hesketh S, Pickles A. How far are associations between child, family and community factors and child psychopathology informant-specific and informant-general? *J Child Psychol Psychiatry.* 2009;50(5):571-580.

5. Bernal ME, Delfini LF, North JA, Kreutzer SL. Comparison of boys' behaviors in homes and classrooms. *Behavior Modification and Families*. Vol 1; New York, NY: Academic Press; 1976:204-227.

6. Fergusson DM, Lynskey MT, Horwood LJ. The effect of maternal depression on maternal ratings of child behavior. *J Abnorm Child Psychol*. 1993;21(3):245-269.

7. Biederman J, Mick E, Faraone SV. Normalized functioning in youths with persistent attention-deficit/hyperactivity disorder. *J Pediatr*. 1998;133(4):544-551.

8. Boyle MH, Pickles AR. Influence of maternal depressive symptoms on ratings of childhood behavior. *J Abnorm Child Psychol*. 1997;25(5):399-412.

9. Fitzmaurice GM, Laird NM, Zahner GE, Daskalakis C. Bivariate logistic regression analysis of childhood psychopathology ratings using multiple informants. *Am J Epidemiol*. 1995;142(11):1194-1203.

10. Offord DR, Boyle MH, Racine Y, et al. Integrating assessment data from multiple informants. *J Am Acad Child Adolesc Psychiatry*. 1996;35(8):1078-1085.

11. Birmaher B, Ryan ND, Williamson DE, et al. Childhood and adolescent depression: a review of the past 10 years. Part I. *J Am Acad Child Adolesc Psychiatry*. 1996;35(11):1427-1439.

12. Birmaher B, Ryan ND, Williamson DE, Brent DA, Kaufman J. Childhood and adolescent depression: a review of the past 10 years. Part II. *J Am Acad Child Adolesc Psychiatry*. 1996;35(12):1575-1583.

13. Bouvette-Turcot AA, Unternaehrer E, Gaudreau H, et al. The joint contribution of maternal history of early adversity and adulthood depression to socioeconomic status and potential relevance for offspring development. *J Affect Disord*. 2017;207:26-31.

14. Bouvette-Turcot AA, Fleming AS, Unternaehrer E, et al. Maternal symptoms of depression and sensitivity mediate the relation between maternal history of early adversity and her child temperament: the inheritance of circumstance. *Dev Psychopathol*. 2020;32(2):605-613.

15. Goodman SH, Rouse MH, Connell AM, Broth MR, Hall CM, Heyward D. Maternal depression and child psychopathology: a meta-analytic review. *Clin Child Fam Psychol Rev*. 2011;14(1):1-27.

16. Pearson RM, Evans J, Kounali D, et al. Maternal depression during pregnancy and the postnatal period: risks and possible mechanisms for offspring depression at age 18 years. *JAMA psychiatry*. 2013;70(12):1312-1319.

17. Shen H, Magnusson C, Rai D, et al. Associations of parental depression with child school performance at age 16 years in Sweden. *JAMA psychiatry*. 2016;73(3):239-246.

18. Evans J, Melotti R, Heron J, et al. The timing of maternal depressive symptoms and child cognitive development: a longitudinal study. *J Child Psychol Psychiatry*. 2012;53(6):632-640.

19. Kroes G, Veerman JW, De Bruyn EE. Bias in parental reports? Maternal psychopathology and the reporting of problem behavior in clinic-referred children. *Eur J Psychol Assess*. 2003;19(3):195.

20. Chilcoat HD, Breslau N. Does psychiatric history bias mothers' reports? An application of a new analytic approach. *J Am Acad Child Adolesc Psychiatry*. 1997;36(7):971-979.

21. Mick E, Santangelo SL, Wypij D, Biederman J. Impact of maternal depression on ratings of comorbid depression in adolescents with attention-deficit/hyperactivity disorder. *J Am Acad Child Adolesc Psychiatry*. 2000;39(3):314-319.

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

23. Rasch G. On general laws and the meaning of measurement in psychology. In: Neyman J, (ed). *Paper presented at: Proceedings of the 4th Berkeley Symposium on Mathematical Statistics and Probability*. Vol. 4; Berkeley, CA: University of California Press; 1961:321-333.

24. Masters GN. A Rasch model for partial credit scoring. *Psychometrika*. 1982;47(2):149-174.

25. Andersen EB. Polytomous Rasch models and their estimation. *Rasch Models*. New York, NY: Springer; 1995:271-291.

26. Soh SE, Tint MT, Gluckman PD, et al. Cohort profile: growing up in Singapore towards healthy outcomes (GUSTO) birth cohort study. *Int J Epidemiol*. 2014;43(5):1401-1409.

27. Achenbach TM. *Integrative Guide for the 1991 CBCL/4-18, YSR, and TRF Profiles*. Burlington, MA: University of Vermont Department of Psychiatry; 1994.

28. Kovacs M, Preiss M. *CDI. Children's Depression Inventory Multi-Health Systems*. New York, NY: Springer; 1992.

29. Villabø M, Gere M, Torgersen S, March JS, Kendall PC. Diagnostic efficiency of the child and parent versions of the multidimensional anxiety scale for children. *J Clin Child Adolesc Psychol*. 2012;41(1):75-85.

30. Beck AT, Ward CH, Mendelson M, Mock J, Erbaugh J. An inventory for measuring depression. *Arch Gen Psychiatry*. 1961;4(6):561-571. https://doi.org/10.1001/archpsyc.1961.01710120031004

31. Cox JL, Holden JM, Sagovsky R. Detection of postnatal depression: development of the 10-item Edinburgh postnatal depression scale. *Br J Psychiatry*. 1987;150(6):782-786.

32. Spielberger CD. State-trait anxiety inventory for adults; 1983.

33. Andersen EB. Sufficient statistics and latent trait models. *Psychometrika*. 1977;42(1):69-81.

34. Andrich D. A rating formulation for ordered response categories. *Psychometrika*. 1978;43(4):561-573.

35. Samejima F. Estimation of latent ability using a response pattern of graded scores. *Psychometrika Monogr Suppl*. 1969;34(4 (Pt. 2)):100.

36. Fischer GH, Ponocny I. An extension of the partial credit model with an application to the measurement of change. *Psychometrika*. 1994;59(2):177-192.

37. Fischer GH, Ponocny I. Extended rating scale and partial credit models for assessing change. *Rasch Models*. New York, NY: Springer; 1995:353-370.

38. Rijmen F, Tuerlinckx F, De Boeck P, Kuppens P. A nonlinear mixed model framework for item response theory. *Psychol Methods*. 2003;8(2):185.

39. van Nispen RM, Knol DL, Neve HJ, van Rens GH. A multilevel item response theory model was investigated for longitudinal vision-related quality-of-life data. *J Clin Epidemiol*. 2010;63(3):321-330.

40. Verhagen J, Fox JP. Longitudinal measurement in health-related surveys. a Bayesian joint growth model for multivariate ordinal responses. *Stat Med*. 2013;32(17):2988-3005.

41. Azevedo CL, Fox JP, Andrade DF. Bayesian longitudinal item response modeling with restricted covariance pattern structures. *Stat Comput*. 2016;26(1-2):443-460.

42. Da-Silva C, Gomes A. Bayesian inference for an item response model for modeling test anxiety. *Comput Stat Data Anal*. 2011;55(12):3165-3182.

43. Duncan KA, MacEachern SN. Nonparametric Bayesian modelling for item response. *Stat Model*. 2008;8(1):41-66.

44. Miyazaki K, Hoshino T. A Bayesian semiparametric item response model with Dirichlet process priors. *Psychometrika*. 2009;74(3):375-393.

45. Ferguson TS. A Bayesian analysis of some nonparametric problems. *Ann Stat*. 1973;1(2):209-230.

46. Sethuraman J. A constructive definition of Dirichlet priors. *Stat Sin*. 1994;4(2):639-650.

47. Iwata N, Kumagai R, Saeki I. Do mothers and fathers assess their children's behavioral problems in the same way as do their children? an IRT investigation on the strengths and difficulties questionnaire. *Jpn Psychol Res*. 2020;62(2):87-100.

48. Choi B, Shim G, Jeong B, Jo S. Data-driven analysis using multiple self-report questionnaires to identify college students at high risk of depressive disorder. *Sci Rep*. 2020;10(1):1-13.

49. Barcella W, Iorio MD, Malone-Lee J. Modelling correlated binary variables: an application to lower urinary tract symptoms. *J R Stat Soc Ser C Appl Stat*. 2018;67(4):1083-1100.

50. Patz RJ, Junker BW. A straightforward approach to Markov chain Monte Carlo methods for item response models. *J Educ Behav Stat*. 1999;24(2):146-178.

51. Haario H, Saksman E, Tamminen J. An adaptive Metropolis algorithm. *Bernoulli*. 2001;7(2):223-242.

52. Neal R. Markov Chain sampling methods for Dirichlet process mixture models. *J Comput Graph Stat*. 2000;9:249-265.

53. Favaro S, Teh Y. MCMC for normalized random measure mixture models. *Stat Sci*. 2013;28(3):335-359.

54. Binder DA. Bayesian cluster analysis. *Biometrika*. 1978;65(1):31-38.

55. Lau JW, Green PJ. Bayesian model-based clustering procedures. *J Comput Graph Stat*. 2007;16(3):526-558.

56. Medvedovic M, Sivaganesan S. Bayesian infinite mixture model based clustering of gene expression profiles. *Bioinformatics*. 2002;18(9):1194-1206.

57. Dahl DB. *salso: sequentially-allocated latent structure optimization*. R package version 0.1.16; 2020.

58. Jasra A, Holmes CC, Stephens DA. Markov chain Monte Carlo methods and the label switching problem in Bayesian mixture modeling. *Stat Sci*. 2005;20(1):50-67.

59. Molitor J, Papathomas M, Jerrett M, Richardson S. Bayesian profile regression with an application to the national survey of children's health. *Biostatistics*. 2010;11(3):484-498.

60. Biederman J, Petty CR, Day H, et al. Severity of the aggression/anxiety-depression/attention (AAA) CBCL profile discriminates between different levels of deficits in emotional regulation in youth with ADHD. *J Dev Behav Pediatr*. 2012;33(3):236.

61. Toorn V, Huizink AC, Utens EM, Verhulst FC, Ormel J, Ferdinand RF. Maternal depressive symptoms, and not anxiety symptoms, are associated with positive mother–child reporting discrepancies of internalizing problems in children: a report on the TRAILS study. *Eur Child Adolesc Psychiatry*. 2010;19(4):379-388.

62. Najman JM, Williams GM, Nikles J, et al. Mothers' mental illness and child behavior problems: cause-effect association or observation bias? *J Am Acad Child Adolesc Psychiatry*. 2000;39(5):592-602.

63. Meaney MJ. Perinatal maternal depressive symptoms as an issue for population health. *Am J Psychiatry*. 2018;175(11):1084-1093.

64. Rutter M, Quinton D. Parental psychiatric disorder: effects on children. *Psychol Med*. 1984;14(4):853-880.

65. Downey G, Coyne JC. Children of depressed parents: an integrative review. *Psychol Bull*. 1990;108(1):50.

66. Sawyer MG, Streiner DL, Baghurst P. The influence of distress on mothers' and fathers' reports of childhood emotional and behavioral problems. *J Abnorm Child Psychol*. 1998;26(6):407-414.

67. Durbin CE, Wilson S. Convergent validity of and bias in maternal reports of child emotion. *Psychol Assess*. 2012;24(3):647-660.

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of this article.

How to cite this article: Cremaschi A, De Iorio M, Seng Chong Y, Broekman B, Meaney MJ, Kee MZL. A Bayesian nonparametric approach to dynamic item-response modeling: An application to the GUSTO cohort study. *Statistics in Medicine*. 2021;1–17. https://doi.org/10.1002/sim.9167