# A Bayesian Latent Variable Model for Analysis of Empathic Accuracy


**Linh H. Nghiem, Benjamin A. Tabak, Zachary Wallmark, Talha Alvi, and Jing Cao**

**Abstract** Empathic accuracy (EA), defined as the ability to accurately understand the thoughts and emotions of others, has become a well-studied phenomenon in social and clinical psychology. A widely used computer-based EA paradigm compares perceivers' ratings of targets' feelings or affective states with the ratings of target themselves (the true ratings) and uses correlation or its monotonic transformation as a measure of EA. However, correlation has a number of notable limitations. In particular, perceivers may differ in their rating patterns, but still have similar overall correlations. To overcome the limitations, we propose a Bayesian latent variable model that decomposes EA into two separate dimensions—discrimination and variability. Discrimination measures perceivers' sensitivity in relation to the true ratings, and variability measures the variance of random error in perceiver's perceptions. Similar to the conventional correlation, the Bayesian model is able to measure the overall level of the association between perceiver and target, but more importantly, the Bayesian approach can provide insights into how perceivers differ in their EA. We demonstrate the advantages of the new EA measures in two case studies. The proposed Bayesian model has a simple specification and is easy to use in practice due to its straightforward implementation in popular software. The R code is included in the supplementary material.

L. H. Nghiem (✉)
School of Mathematics and Statistics, University of Sydney, Sydney, NSW, Australia
e-mail: linh.nghiem@sydney.edu.au

B. A. Tabak · T. Alvi
Department of Psychology, Southern Methodist University, Dallas, TX, USA
e-mail: btabak@smu.edu; talvi@smu.edu

Z. Wallmark
Department of Musicology and Ethnomusicology, University of Oregon, Eugene, OR, USA
e-mail: zwallmar@uoregon.edu

J. Cao
Department of Statistical Science, Southern Methodist University, Dallas, TX, USA
e-mail: jcao@smu.edu



H. K. T. Ng, D. F. Heitjan (eds.), *Recent Advances on Sampling Methods and Educational Statistics*, Emerging Topics in Statistics and Biostatistics,
https://doi.org/10.1007/978-3-031-14525-4_10

## 1 Introduction

Empathic accuracy (EA) is defined as the ability to correctly infer the thoughts and emotions of others (Zaki et al., 2009). In addition to the role of EA in the development and maintenance of healthy social relationships (Sened et al., 2017), clinical research has shown that performance in a standard EA video task can differentiate individuals with certain psychiatric disorders from healthy controls (Lee et al., 2011). Thus, the study of EA can help us understand general social functioning and also help identify social cognitive impairment in clinical populations.

There are a number of ways to examine EA, including matching categorical assessments (Schweinle et al., 2002) or continuous real-time assessments of the affective states of people (hereafter referred to as targets) by participants (hereafter, perceivers) (Zaki et al., 2008). Studies of EA that focus on matching categorical assessments between perceivers and targets often use signal detection theory in analyses. However, continuous EA data do not allow for this type of analysis. The focus of this study is on the analysis of EA tasks based on continuous real-time ratings. For example, EA paradigms may include a set of brief video clips in which targets discuss positive or negative events in their lives. Perceivers are asked to rate how negative or positive the target is feeling when discussing autobiographical events in real time using a 9-point scale (e.g., 1 = extremely negative; 9 = extremely positive). Responses from perceivers are captured in 2–5 s epochs throughout each video clip, and these responses are then compared to the responses of the targets, who watched the videos of themselves and completed the same ratings task in order to create a canonical index of "true" responses.

Traditionally, correlational analysis (and its monotonic transformation) is the conventional and arguably most common statistical method used for analysis of the continuous EA data. For example, based on several videos in which social targets discussed emotional events, Zaki et al. (2009) collected ratings averaged across 5-s periods and computed the Fisher transformation of the Pearson correlation coefficient to measure perceivers' EA. Also, in an fMRI validation study of a modified EA task, Mackes et al. (2018) computed the same measure and conducted paired samples t-tests to examine the neural correlates of perceived emotional intensity and mentalizing. However, this one-dimensional correlation approach, which only measures the linear association between two variables, may leave out important patterns in the data. First, unlike weight or height, EA is a latent merit that cannot be directly measured in absolute terms. For example, in a given task, the same rating may mean something different to different perceivers. In addition, although all perceivers are given the same scales (such as from 1 to 9), different perceivers may subjectively choose different ranges of their own ratings (e.g., one person may always give ratings from 4 to 7, while another person may use the whole range from 1 to 9). Second, there are at least two underlying behavioral dimensions that contribute to the discrepancy between perceivers' and targets' ratings, including different interpretations of the scale range and the random error in perceivers'

ratings. These two dimensions are distinct, so it is necessary to incorporate both of them when measuring EA. Third, correlation can only be calculated for each stimulus separately, but an EA study typically involves a number of stimuli (such as multiple videos under one condition). Due to all of the issues raised here, statistical analysis based on correlation may limit the amount of information that can be gained from EA studies.

In a broader context of modeling accuracy of human judgment, a few approaches have been proposed as an alternative for correlation, yet these approaches typically require additional data compared to what we have for our applications. For example, West and Kenny (2011) proposed the truth and bias model, in which perceivers' responses to a stimulus are assumed to be influenced by a truth force and a bias force. To use this model, each perceiver is typically asked to provide not only a response toward the target but also a self-judgment response. In our application, we only have the former but not the latter. Biesanz (2010) proposed the social accuracy model, in which accuracy of a judgment is into distinctive accuracy, the extent to which a perceiver can perceive the distinct and unique characteristics of one person, and normative accuracy, a measure of how a perceiver's perception of others corresponds to the same perceiver's perception of an average person. This social accuracy model is commonly used in modeling perception of traits, where a perceiver is asked to rate different traits of other people, and the ratings of these traits for an average person (a normative profile) are available from a larger sample or a meta-study. In our application, a perceiver is asked to provide a continuous rating over time to judge the emotion of a specific target. To the best of our knowledge, the normative profile for these continuous ratings are not available.

As pointed out by an anonymous referee, one may tend to conduct the Bland-Altman analysis (Bland & Altman, 1999) between the perceivers' and targets' ratings. In most of the applications, the Bland-Altman (BA) analysis aims to evaluate whether two different devices give the same measurements of an objective quantity. For example, in Doğan (2018), the BA method is used to evaluate whether a venous blood gas analysis and a biochemistry panel shows the same level of potassium in patients. However, the BA method is not appropriate for measuring EA. First, for one specific task, the perceivers' and target's ratings are not expected to be the same, because they can have different (subjective) interpretation of the rating scale. For example, a rating of "5" for one person is not the same as a rating of "5" for another. Furthermore, just knowing whether perceivers and targets agree with each other may be even less informative than using correlation, since the BA analysis does not quantify the extent to which a perceiver agrees with the target.

In this article, we introduce a Bayesian latent variable approach to model EA response data that is based on the previous work by Cao and colleagues (Cao et al., 2010; Cao & Stokes, 2017). The proposed Bayesian model identifies two latent dimensions of EA—discrimination and variability—that are identifiable when perceivers' ratings differ from the targets' ratings. Discrimination measures a perceiver's ability to distinguish changes in a target's emotions, while a perceiver's variability measures the variance of random error in perceivers' ratings (i.e., the difference between perceiver's and target's ratings due to inconsistency). A smaller

variance implies that the perceiver has a higher level of consistency in perceiving the target. Using the proposed Bayesian model, we are able to estimate perceivers' discrimination and variability and hence obtain more valuable information about their EA perception than correlation, which only measures the general association between perceivers' and target's ratings.

We begin by introducing the Bayesian model, including model specification and software implementation. We then describe the advantages of the Bayesian model using two case studies. In the first case study, we re-analyze the dataset in Devlin et al. (2014) that consists of perceivers' ratings of four distinct videos in which targets discuss emotional events in their lives. In this case study, we focus on explaining the underlying dimensions of EA and comparing the Bayesian estimates of discrimination and variability with the standard correlational measure. In the second case study, we analyze perceivers' ratings of 12 original music recordings expressing musician-targets' renderings of four primary emotions (three recordings per emotion), with the focus on how the underlying EA dimensions are associated with the musicality (i.e., level of musical skill and training) of the perceivers. This case study further demonstrates that the new measures can facilitate additional insights on how EA perception is related to perceivers' characteristics.

## 2 Methodology

In an EA study, suppose that there are $n$ perceivers instructed to provide ratings on $J$ stimuli, where each stimulus has $K_j$ units (i.e., there are $K_j$ points in the sequence of ratings, which can vary among stimuli). Each stimulus corresponds to a specific target. Let $x^r_{jk}$ denote the raw rating given by the corresponding target for the $k$th unit of the $j$th stimulus, $j = 1, \ldots, J$ and $k = 1, \ldots, K_j$. Note that similar to correlational analysis, the mean of target score will not affect the measurement on EA. Hence, to simplify the model specification, the raw ratings $x^r_{jk}$ are centered for each stimulus, where the centered rating is denoted as $x_{jk} = x^r_{jk} - K_j^{-1} \sum_{m=1}^{K_j} x^r_{jm}$ and is treated as the true rating. Similarly, letting $y^r_{ijk}$ denote the rating given by the $i$th perceiver for the $k$th unit of the $j$th stimulus, then the corresponding centered rating is $y_{ijk} = y^r_{ijk} - K_j^{-1} \sum_{m=1}^{K_j} y^r_{ijm}$. We specify the Bayesian latent variable model to measure EA as

$$\begin{aligned} y_{ijk} &= \beta_{ij} x_{jk} + \varepsilon_{ijk} \\ \varepsilon_{ijk} &\sim N(0, \sigma_i^2), \quad \beta_{ij} \sim N(\beta_i, \sigma_\beta^2), \end{aligned} \tag{1}$$

for $i = 1, \ldots, n$, $j = 1, \ldots, J$, $k = 1, \ldots, K_j$. In the model, $\beta_{ij}$ represents the $i$th perceiver's discrimination level on the $j$th stimulus, which is assumed to follow a normal distribution with mean $\beta_i$ and variance $\sigma^2_\beta$. Note that $\beta_i$ is the $i$th perceiver's average discrimination level over all the $J$ stimuli. We allow a perceiver to have different discrimination levels for different stimulus, but assume these discrimination levels are similar by imposing a random-effect structure on all $\beta_{ij}$'s.

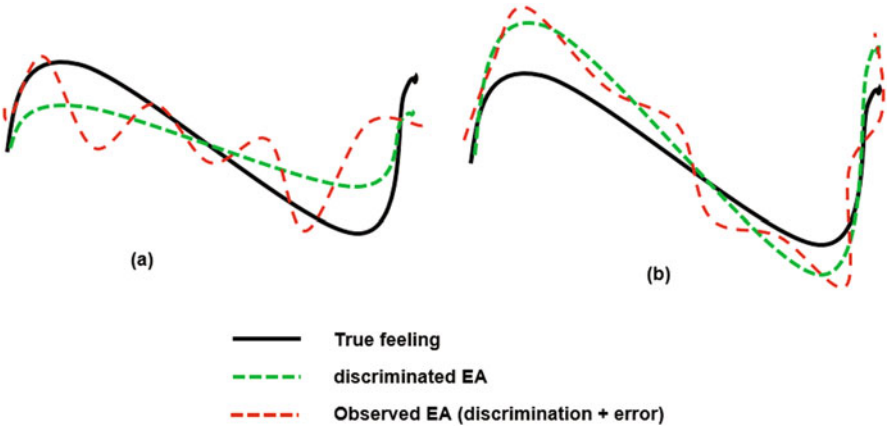


**Fig. 1** Illustration of the two latent dimensions in EA. Plot (**a**): a subject has an attenuated discrimination and relatively large variability. Plot (**b**): a subject has a magnified discrimination and relatively small variability

An empathic perceiver's discrimination parameter $\beta_i$ will be positive, indicating that on average, the perceiver's response has a congruent association with the target. A smaller value $\beta_i$ suggests that the perceiver's response signal is more attenuated compared to a perceiver with a larger $\beta_i$ value. Furthermore, a perceiver with a negative $\beta_i$ has a response that moves in opposite direction compared to the target's ratings, yet these instances are generally rare in EA studies. Additionally, Model (1) contains the random error $\varepsilon_{ijk}$, which is assumed to follow a normal distribution with mean 0 and *perceiver-specific variance* $\sigma_i^2$. The smaller the variance, the higher the consistency in the perceiver's ratings, so we refer to $\sigma_i^2$ as the measure of the variability in EA of the $i$th perceiver.

Figure 1 illustrates two examples of concept on how the two latent dimensions of EA (i.e., discrimination and variability) contributes to the actual ratings given by a perceiver. The black line depicts the (observed) target true ratings. The green dashed line represents the (unobserved) expected ratings associated with a certain discrimination level, and the red dashed line represents the (observed) actual ratings after random errors are added to the green dashed line. The plot on the left shows an example of ratings with an attenuated discrimination (i.e., a less distinctive interpretation of true signals) and relatively large variability (i.e., a large deviation between the expected ratings and the actual ratings), and the plot on the right shows an example with a magnified discrimination and relatively small variability.

To complete the Bayesian model specification, the assignment of prior distribution is listed in the following:

$$\beta_i \sim N(1, 100), \quad \sigma_i^2 \sim \text{IG}(2, 1), \quad \sigma_\beta^2 \sim \text{IG}(2, 1), \tag{2}$$

for $i = 1, \ldots, n$. Note that the ratings have a range of 9 points, so the normal prior on $\beta_i$ has a variance of 100, which is large enough to make the normal prior

a non-informative prior. The mean of $\beta_i$ is 1 because without any prior knowledge, we assume all perceivers have roughly the same interpretation as the true targets. The prior for the variance $\sigma_i^2$ is IG(2, 1), which is the inverse gamma distribution with a shape parameter of 2 and a scale parameter of 1, so the corresponding prior variance is infinite. Thus, the assigned priors are conventional conjugate non-informative priors, which facilitates data-driven inference and results in fast Bayesian computation (Sun et al., 2001). The proposed model is referred to as the BDV Model (Bayesian model with the latent dimensions on Discrimination and Variance). Finally, note that when there is only one stimulus in the EA study (i.e., $J = 1$), the BDV Model can be reduced to

$$y_{ik} = \beta_i x_k + \varepsilon_{ik}, \qquad \varepsilon_{ik} \sim N(0, \sigma_i^2), \tag{3}$$

where the priors for $\beta_i$ and $\sigma_i^2$ are the same as in (2) for $i = 1, \ldots, n$.

Note that the general BDV Model with $J \geq 2$ is a random-effect model, with a random slope (perceiver-specific discrimination) and a perceiver-specific variance. The perceiver-specific variance is a novel and indispensable part of the model because it represents a unique EA dimension. In the applications below, we demonstrate that, compared to the same models with constant variance, i.e., $\sigma_i^2 = \sigma^2$ for all $i = 1, \ldots, n$, the incorporation of perceiver-specific variance improves the model fits significantly. After fitting the BDV Model, we use the posterior mean for $\beta_i$ and $\sigma_i^2$ as the estimated discrimination and variability for the EA of the $i$th perceiver.

To facilitate the implementation of the model, we include the R code in the supplementary material. The code is based on "Just-Another Gibbs-Sampler"(JAGS) model, which is an open-source program designed to run Bayesian hierarchical model using Markov chain Monte Carlo methods (Plummer et al., 2003). With JAGS, users specify a model and its prior specification; then a Markov chain simulation is automatically implemented for the resulting posterior distribution. This frees users from manually deriving the MCMC algorithm, which is the main obstacle for the implementation of Bayesian inference in practice. JAGS is designed to work closely with the R language. Our code uses the `rjags` package (Plummer, 2019) as the interface from R to JAGS. Detailed instructions are annotated in the code.

## 3 Applications

### *3.1 Study on Social Empathic Accuracy*

In our first application, we consider a study conducted by Devlin et al. (2014) that examined the relationship between perceivers' levels of positive emotion and EA. Their study included $n = 121$ perceivers, who watched four videos

of targets discussing emotional events in their lives. These four videos vary in valence (positive or negative) and intensity (high or low), resulting in four non-homogeneous videos, including high-positive, low-positive, high-negative, and low-negative. While watching each video, perceivers provided continuous online ratings of the corresponding target's emotion using a 9-point scale (from 1 = extremely negative to 9 = extremely positive). The ratings from the perceivers were then compared with those from the targets.

To measure the EA of each perceiver, the authors calculated the Fisher transformation to the Pearson correlation between perceivers' ratings and targets' ratings for each video. In other words, each participant had four EA measures, each of which corresponds to one video. For the correlation coefficient $r$, the corresponding Fisher transformation is defined as $Z = (1/2)\log\{(1+r)/(1-r)\}$, where log denotes the natural logarithm. While $r$ ranges from $-1$ to 1, the Fisher transformed correlation can take any value from the real line, so it is more appropriate to conduct statistical analyses with the normality assumption based on $Z$ than based on $r$. In this paper, we refer to $Z$ as the "$r$-to-$Z$ EA estimate" and denote it to be $rZ$. The data from Devlin et al. (2014) are publicly available at https://doi.org/10.1371/journal.pone.0110470.

Because these four videos varied in valence and intensity, they should be treated as four distinct individual stimuli instead of multiple stimuli under one condition, and we fit the reduced BDV Model (3) to each of the four video stimulus separately. We begin with a graphical demonstration to illustrate how the two latent EA dimensions can provide more insights on EA compared to the conventional measure $rZ$. Figure 2 shows two plots, each depicting the ratings given by the target and those by three perceivers (selected for illustrative purposes) for the high-negative video. In each plot, the black line represents the true target's rating, and the other lines represent ratings of the selected perceivers watching the same video. The estimated $rZ$s between the target's and perceivers' ratings are listed in the legend, along with the estimated discrimination and variability parameters (abbreviated as D and V, respectively). In the top panel, the three perceivers demonstrated very different $rZ$, whereas in the bottom panel, the three perceivers had similar $rZ$.

In the top plot of Fig. 2, the three perceivers (denoted as P1, P2, and P3) demonstrated very different EA levels, indicated by the varying estimated $rZ$ values (1.51, $-0.22$, and 0.51). However, the correlational analysis does not explain why the three perceivers have such dramatically different EA scores. Based on the Bayesian estimates, we can see that P1's greater EA (red line, $\widehat{rZ} = 1.51$) is due to a higher level of discrimination ($\hat{D} = 0.81$) and smaller variability (i.e., higher consistency, $\hat{V} = 0.16$) when rating the target. The perceiver P2 has a negative correlation (green line, $\widehat{rZ} = -0.22$), which is due to the negative discrimination (i.e., the person perceived the target's emotion in the opposite direction, $\hat{D} = -0.14$). In addition, P2 also has the largest variability among the three perceivers reflecting the more obvious fluctuation of P2's ratings ($\hat{V} = 0.34$). Moreover, P3 (blue line, $\widehat{rZ} = 0.51$) has a moderate EA level: compared to P1, P3 has a lower discrimination (other than the initial drop, P3's ratings are quite flat, not showing the gradual decline in the target's ratings, $\hat{D} = 0.34$) and larger variability (the discrepancy between P3's ratings and the target's ratings are noticeably large in both ends of the series, $\hat{V} = 0.38$).

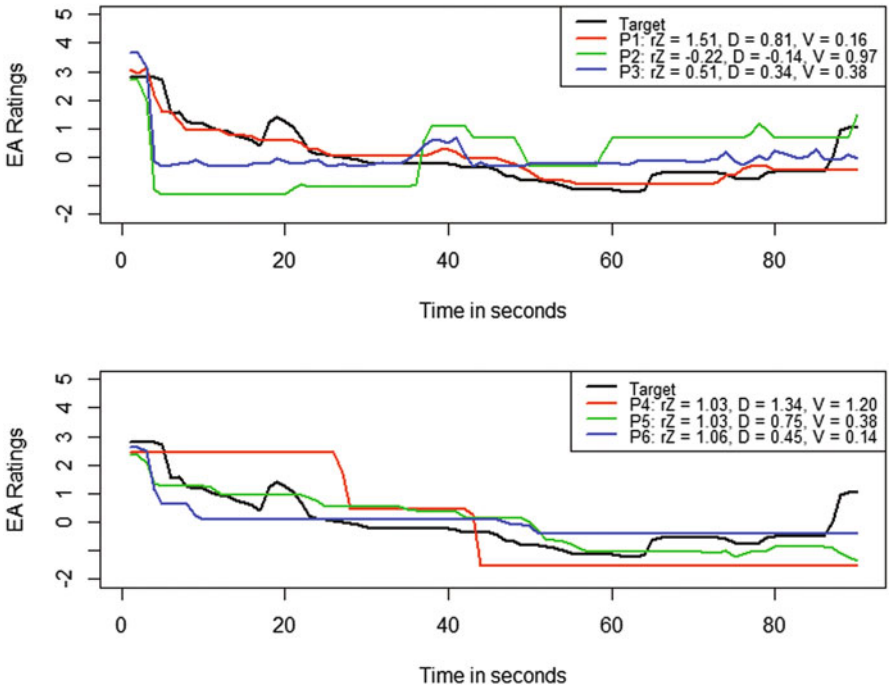


**Fig. 2** Comparison of the target's ratings and the six perceivers' ratings in the high-negative video, where $rZ$ denotes the $r$-to-$Z$ transformed correlation and D and V represent discrimination and variance in the Bayesian model, respectively

In the bottom plot of Fig. 2, we chose data from three different perceivers (P4, P5, and P6) to further demonstrate the advantage of utilizing discrimination and variability to study EA over the $rZ$ measure. In this case, the three perceivers have similar estimated $rZ$ values (1.03, 1.03, and 1.06). Hence, based on the correlational analysis, these perceivers have similar EA. However, the estimates of discrimination and variability show that their underlying EA dimensions have distinct patterns. P4 (red) has a large discrimination value ($\hat{D} = 1.34$), resulting from the fact that P4's ratings have a more dramatic decline than the target's ratings. At the same time, P4 has the largest variability among the three perceivers, reflecting P4's pronounced shift toward negative ratings at around time units 25 and 40. In addition, P6 (blue line) has both the smallest discrimination ($\hat{D} = 0.45$) and smallest variability ($\hat{V} = 0.14$) among the three perceivers. Other than the initial drop, P6's ratings are mostly flat, only spanning a narrow range of scores. Unlike the dramatic decline in P4's ratings and slow change in P6's ratings, P5's (green line) ratings follow the gradual decline in the target's ratings. Because of the inadequate drop in the beginning and opposite change in the end of the series, P5 has a larger variability ($\hat{V} = 0.38$) than P6.

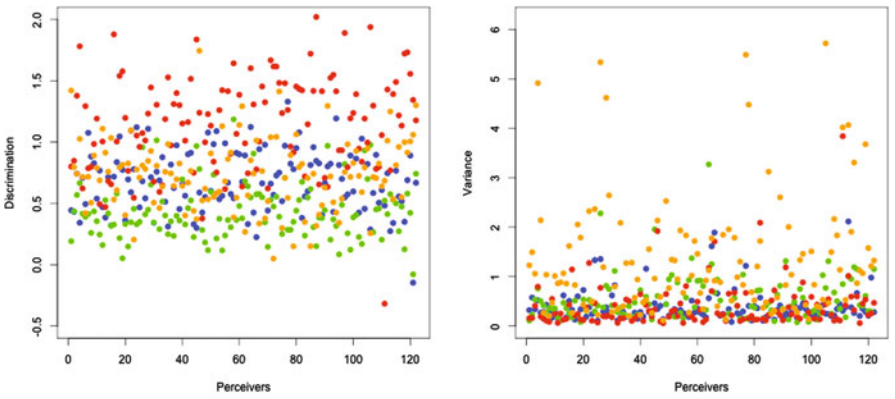


**Fig. 3** Perceivers' video-specific discrimination and variance (red = high-positive, orange = low-positive, blue = high-negative, green = low-negative)

Based on the examples included in Fig. 2, we can see that the two latent discrimination and variability dimensions specified in the BDV Model offer unique information regarding EA compared to the correlation analysis. Specifically, the proposed BDV Model is able to explain how perceivers differ in their EA and to identify possible differences in the underlying dimensions in EA when the correlation may show no differences.

Next, we compare the latent dimensions across the four videos. Figure 3 (left panel) shows that perceivers had higher discrimination ability for the high-positive video (red dots) and lower discrimination ability for the low-negative video (green dots). These findings are in agreement with previous studies of showing greater EA for positive videos compared to negative videos in both healthy and clinical samples (Lee et al., 2011). As for the variability, the largest video-specific variances are from the two low-intensity videos (orange and green dots in the right panel of Fig. 3).

As we mentioned in the last section, a novelty of the BDV Model is that it incorporates perceiver-specific variances for random errors instead of assuming a constant variance as in most of the conventional random-effect models. We demonstrate the advantage of this functionality by comparing the model fit between the BDV Model and the following random-effect model with a *constant* variance:

$$y_{ik} = \beta_i x_{ik} + \varepsilon_{ik}, \quad \varepsilon_{ik} \sim N(0, \sigma^2). \tag{4}$$

Note that Model (4) assumes that all the perceivers have the same variability, i.e., the same consistency level in EA. The model comparison is conducted using the deviance information criterion (DIC), where a small value of is preferred and a difference of more than 10 usually rules out the model with a higher DIC (Spiegelhalter et al., 2002). The results are summarized in Table 1. The evidence is clear and convincing that across all the four video groups, the BDV Model (3) provides much better model fits than Model (4). This data-driven evidence

**Table 1** Model comparison between the BDV Model (3) and Model (4) using DIC

| | High-negative | Low-negative | High-positive | Low-positive |
|---|---|---|---|---|
| BDV Model (3) | 18288.58 | 12981.28 | 9318.62 | 20196.82 |
| Model (4) | 20758.17 | 14949.47 | 12063.96 | 22649.64 |

**Table 2** Correlation between $rZ$ and the Bayesian measures on EA

| | High-negative | Low-negative | High-positive | Low-positive |
|---|---|---|---|---|
| Discrimination | **0.74 ($p < 0.01$)** | **0.57 ($p < 0.01$)** | **0.40 ($p < 0.01$)** | **0.52 ($p < 0.01$)** |
| Variance | **−0.63 ($p < 0.01$)** | −0.08 ($p = 0.41$) | **−0.72 ($p < 0.01$)** | **−0.52 ($p < 0.01$)** |
| $B_{EA}$ | **0.99 ($p < 0.01$)** | **0.98 ($p < 0.01$)** | **0.96 ($p < 0.01$)** | **0.99 ($p < 0.01$)** |

P-values are based on two-tailed tests and included in parentheses. Significant p-values (<0.05) are indicated in bold

supports the inclusion of the perceiver-specific variance, which further confirms that variability is a unique dimension aside from discrimination in EA.

Finally, we compare the results from the Bayesian model with the conventional $rZ$ estimates. For each video, we compute the Pearson correlation between perceivers' estimated $rZ$ estimates and the Bayesian estimates of discrimination and variability, respectively. Furthermore, we investigate the correlation between the $rZ$ estimates and the estimates for $\beta_i/\sigma_i$, $i = 1, \ldots, n$, which is the ratio of the discrimination and the square root of random error's variance, similar to the measure used by Cao et al. (2010). We refer to this ratio as the "Bayesian EA aggregated estimate" and denote it as $B_{EA}$. Similar to $rZ$, the measure $B_{EA}$ can take any value from the real line.A high $B_{EA}$ implies that a perceiver has a relatively large discrimination and a relatively small variability.

Table 2 shows that the association between perceivers' $rZ$ and $B_{EA}$ is consistently high, with the correlation being almost 1. However, the association between perceivers' $rZ$ and the latent dimensions on discrimination and variability are weak to moderate (though most of them are statistically significant). This indicates that the conventional correlation, as was used by Devlin et al. (2014) and most existing literature, only provides a valid aggregate measure for EA, but it does not provide much insight into the dimensions underlying the structure of EA.

## *3.2 Study on Musical Empathic Accuracy*

In a study of the association between EA and accuracy of emotion recognition in music, Tabak et al. (In press) collected data from 415 undergraduate perceivers enrolled at Southern Methodist University. Perceivers participated in a novel music EA task, in which they listened to and rated 12 brief music recordings expressing the target's (musician's) primary emotions of joy/happiness, sadness, anger, and tenderness (3 recordings per emotion). Stimuli were solo piano pieces created by

six composer-pianists. Identical to the video EA task in the previous case study, perceivers listened to the excerpts and provided continuous real-time response evaluations of how negative or positive (1 = very negative to 9 = very positive) they perceived the music to be. Samples were collected every 2 s. The same data collected from the composer-pianist targets provided the "true" target ratings to be compared with perceivers' ratings.

EA research has typically focused on cognitive empathy, i.e., perceivers' understanding of a target's thoughts, feelings, and general mental state (Zaki et al., 2009). However, recently Morrison et al. (2016) included an additional assessment of EA in which they slightly altered the instructions of the task to assess affect sharing or the extent to which a perceiver experiences the same emotion as a target (i.e., affective empathy). To examine the two different kinds of EA, perceivers in this study were randomized into an affective empathy group or a cognitive empathy group. In the affective empathy group ($n = 230$), perceivers were asked to provide their own emotional response when listening to the music, whereas in the cognitive empathy group ($n = 185$), they were instructed to try to understand the emotion being communicated or expressed by the composer-pianist in the recordings.

In this application, our goal is to use the BDV Model to investigate the association between perceivers' EA underlying dimensions and their musical training in both groups. Musical training has been shown to modulate emotion recognition of music (Di Mauro et al., 2018). Our aim here is to examine whether musical training is associated with the accurate perception of musical emotion, as operationalized according to the musician-targets' intent. Musical training is measured by the Goldsmiths Musical Sophistication Index (Müllensiefen et al., 2014), a psychometric tool for the measurement of musical attitudes, behaviors, and skills. For each group, we first compute the correlation between the estimates of each dimension in EA and the Gold-MSI among the perceivers. In addition, we examine the association of EA and Gold-MSI in three conditions: (1) across all 12 music recordings (i.e., $J = 12$), (2) among the 6 positive music recordings (i.e., $J = 6$) which consist of 3 recordings expressing happiness and 3 recordings expressing tenderness, and (3) among the 6 negative music recordings (i.e., $J = 6$) which consist of 3 recordings expressing sadness and 3 recordings expressing anger. Note that there are multiple stimuli under each condition, and it is not straightforward to compute an overall EA measure from the correlational analysis in this setting.

We fit the BDV Model (1) to multiple stimuli for each of the three above conditions. We then compute the Pearson correlation between the Gold-MSI and the Bayesian estimates of discrimination and variability, respectively. Table 3 provides the results for the affective empathy and the cognitive empathy groups. First, for the affective empathy group, none of the association between estimated discrimination nor variance with musical background is statistically significant. On the other hand, for the cognitive empathy group, we find a significant association between perceivers' estimated discrimination and their musicality across all the three conditions. However, the association between the estimated variability and perceivers' musicality is not statistically significant. In other words, higher levels of discrimination in the cognitive assessment of musician/targets' emotions are

**Table 3** Correlation between empathic accuracy latent dimensions and musicality

| | Affective group | | Cognitive group | |
|---|---|---|---|---|
| | Discrimination | Variance | Discrimination | Variance |
| All music | −0.00 ($p = 0.98$) | −0.02 ($p = 0.78$) | **0.22 (p < 0.01)** | −0.09 ($p = 0.23$) |
| Positive music | 0.07 ($p = 0.31$) | −0.01 ($p = 0.90$) | **0.19 (p = 0.01)** | −0.14 ($p = 0.06$) |
| Negative music | −0.00 ($p = 0.50$) | −0.02 ($p = 0.74$) | **0.15 (p = 0.04)** | −0.07 ($p = 0.33$) |

P-values are based on two-tailed tests and included in parentheses. Significant p-values (<.05) are indicated in bold

**Table 4** Model comparison between the BDV Model (1) and Model (5) using DIC

| | | All music | Positive music | Negative music |
|---|---|---|---|---|
| Affective group | Model (1) | 183732 | 70507 | 107962 |
| | Model (5) | 215722 | 83849 | 131688 |
| Cognitive group | Model (1) | 167756 | 63654 | 97049 |
| | Model (5) | 189105 | 72736 | 116119 |

associated with perceivers' relative degree of musical ability, while the level of consistency is not. In contrast, the congruence of one's personal emotional responses to the musician's expressive intentions (EA for affective empathy) is not related to one's training and depth of musical knowledge. Finally, in order to confirm the need for including perceiver-specific variance, similar to what was done in the previous application, we compare the model fit between the BDV Model (1) and the following random-effect model with a constant variance:

$$y_{ijk} = \beta_{ij} x_{jk} + \varepsilon_{ijk}, \quad \varepsilon_{ijk} \sim N(0, \sigma^2), \quad i = 1, \ldots, n, \ j = 1, \ldots, J, \ k = 1, \ldots, K_j. \tag{5}$$

The prior specification of model (5), other than the perceiver-specific variance, remains the same as that in (2). The model comparison using DIC is summarized in Table 4. It shows that for all the three conditions and for both the affective group and the cognitive group, DIC for the BDV Model (1) is substantially smaller than that for Model (5). The model comparison results provide strong evidence to show that the incorporation of the perceiver-specific variance improves model fit substantially. Whether examining one stimulus or multiple stimuli, variability, as measured by the perceiver-specific variance of the random error in the model, is a distinctive dimension of EA, which is inherently different from perceiver-specific discrimination. Thus, when looking at perceivers' EA patterns, including both dimensions provides more detailed information about perceivers' perceptions.

## 4 Conclusion

In this article, we have proposed a Bayesian latent variable model which serves as an alternative to the conventional correlational analysis for empathic accuracy (EA) research using continuous real-time assessments. The proposed BDV model has three main advantages over the correlational analysis. First, it is more sensitive to perceiver-level differences in EA studies, as reflected in varying response behaviors (e.g., using different ranges of the scale). Correspondingly, the BDV Model quantifies two behavioral dimensions of EA, discrimination, and variability. Similar to the correlational analysis, these two dimensions measure the overall EA level for each perceiver, but more importantly, they explain how perceivers differ in EA. Using correlation, many perceivers giving different rating patterns may have a similar EA level, but using discrimination and variability, these differences can be identified. Finally, while correlational analysis must be conducted independently for each individual target, the proposed model is capable of providing an overall EA measure where multiple stimuli are included under one condition of an EA task. Taken together, the Bayesian approach to EA can shed light on distinctions that are not detectable by simple correlational analysis.

There are many areas of research that can benefit from this approach. Broadly speaking, it could be used to increase the analytical precision of any experimental paradigm involving the comparison of sequential measurements on latent perceptual responses, such as research on social cognitive deficits in individuals with autism spectrum disorders and schizophrenia. The association of EA with social functioning in healthy and clinical populations has previously relied on the correlational approach to EA analysis (Lee et al., 2011). With the approach described here, researchers may be able to identify specific dimensions of EA that may be more or less impaired among clinical populations. For example, the discrimination parameter could be used to elucidate the extent to which the amplification of negative information and suppression of positive information that characterize individuals with depression (LeMoult & Gotlib, 2019) . The increased level of specificity could also benefit neuroscientists by examining the extent to which different dimensions of EA are correlated with real-time neural processing (Mackes et al., 2018). Furthermore, the BDV model can be improved in future research by incorporating other covariates that represent perceivers' and targets' characteristics. In general, improving the BDV model requires a consideration of both the quality of the model fit and its interpretability in the context of measuring EA.

In conclusion, the proposed Bayesian EA model is more flexible in handing perceiver-specific parameters than traditional correlational analysis. The model specification is simple, and the computation is efficient. Annotated R code is included to facilitate the implementation of the proposed model.

## References


Biesanz, J. C. (2010). The social accuracy model of interpersonal perception: Assessing individual differences in perceptive and expressive accuracy. *Multivariate Behavioral Research, 45*(5), 853–885.

Bland, J. M., & and Altman, D. G. (1999). Measuring agreement in method comparison studies. *Statistical Methods in Medical Research, 8*(2), 135–160.

Cao, J., & Stokes, L. (2017). Comparison of different ranking methods in wine tasting. *Journal of Wine Economics, 12*(2), 203–210.

Cao, J., Stokes, S. L., & 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.

Devlin, H. C., Zaki, J., Ong, D. C., & Gruber, J. (2014). Not as good as you think? trait positive emotion is associated with increased self-reported empathy but decreased empathic performance. *PloS One, 9*(10), e110470.

Di Mauro, M., Toffalini, E., Grassi, M., & Petrini, K. (2018). Effect of long-term music training on emotion perception from drumming improvisation. *Frontiers in Psychology, 9*, 2168.

Doğan, N. Ö. (2018). Bland-altman analysis: A paradigm to understand correlation and agreement. *Turkish Journal of Emergency Medicine, 18*(4), 139–141.

Lee, J., Zaki, J., Harvey, P.-O., Ochsner, K., & Green, M. F. (2011). Schizophrenia patients are impaired in empathic accuracy. *Psychological Medicine, 41*(11), 2297–2304.

LeMoult, J., & Gotlib, I. H. (2019). Depression: A cognitive perspective. *Clinical Psychology Review, 69*, 51–66.

Mackes, N. K., Golm, D., O'Daly, O. G., Sarkar, S., Sonuga-Barke, E. J., Fairchild, G., & Mehta, M. A. (2018). Tracking emotions in the brain–revisiting the empathic accuracy task. *NeuroImage, 178*, 677–686.

Morrison, A. S., Mateen, M. A., Brozovich, F. A., Zaki, J., Goldin, P. R., Heimberg, R. G., & Gross, J. J. (2016). Empathy for positive and negative emotions in social anxiety disorder. *Behaviour Research and Therapy, 87*, 232–242.

Müllensiefen, D., Gingras, B., Musil, J., & Stewart, L. (2014). Measuring the facets of musicality: The goldsmiths musical sophistication index (gold-msi). *Personality and Individual Differences, 60*, S35.

Plummer, M. (2019). *rjags: Bayesian graphical models using MCMC*. R package version 4-10.

Plummer, M. et al. (2003). Jags: A program for analysis of Bayesian graphical models using Gibbs sampling. In *Proceedings of the 3rd International Workshop on Distributed Statistical Computing* (Vol. 124, pp. 1–10). Vienna, Austria.

Schweinle, W. E., Ickes, W., & Bernstein, I. H. (2002). Emphatic inaccuracy in husband to wife aggression: The overattribution bias. *Personal Relationships, 9*(2), 141–158.

Sened, H., Lavidor, M., Lazarus, G., Bar-Kalifa, E., Rafaeli, E., & Ickes, W. (2017). Empathic accuracy and relationship satisfaction: A meta-analytic review. *Journal of Family Psychology, 31*(6), 742.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., & Van Der Linde, A. (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society: Series B (Statistical Methodology), 64*(4), 583–639.

Sun, D., Tsutakawa, R. K., & He, Z. (2001). Propriety of posteriors with improper priors in hierarchical linear mixed models. *Statistica Sinica* 77–95.

Tabak, B. A., Wallmark, Z., Nghiem, L., Alvi, T., Sunahara, C. S., Lee, J., & Cao, J. (In press). Initial evidence for a relation between behaviorally assessed empathic accuracy and affect sharing for people and music. *Emotion*.

West, T. V., & Kenny, D. A. (2011). The truth and bias model of judgment. *Psychological Review, 118*(2), 357.

Zaki, J., Bolger, N., & Ochsner, K. (2008). It takes two: The interpersonal nature of empathic accuracy. *Psychological Science, 19*(4), 399–404.

Zaki, J., Bolger, N., & Ochsner, K. (2009). Unpacking the informational bases of empathic accuracy. *Emotion, 9*(4), 478.