Multitask Item Response Models for Response Bias Removal from Affective Ratings

Shiro Kumano and Keishi Nomura
NTT Communication Science Laboratories
Nippon Telegraph and Telephone Corporation
Kanagawa, Japan
kumano@ieee.org

***Abstract*—Response style (RS) is a tendency to choose specific categories regardless of content, e.g. extreme or midpoint categories. It degrades the validity of the analysis of subjective ratings such as correlation and variance-based analyses. However, the computational removal of RS has received little attention from the affective computing community. RS removal techniques have been proposed in areas such as marketing research. However, most of these techniques do not exploit the content-independence of RS; i.e. it should be observed consistently in various tasks, such as affective judgment tasks and standard psychological questionnaires. Therefore, this paper proposes a multitask RS removal method. An individual's responses in multiple tasks are modeled using task-independent RS parameters, and task-dependent parameters, including the item and respondent's characteristic parameters based on item response models (IRM). Through Bayesian modeling, we observed that: i) the proposed model outperformed traditional IRMs in terms of predictive accuracy; ii) our multitask framework estimated RS with higher precision than previous single-task-based RS removal methods; iii) our model replicated Japanese midpoint RS, which has been demonstrated repeatedly in previous cross-cultural studies; and iv) RS-removed predictive ratings showed higher inter-rater agreement than those including RS in valence/arousal judgment tasks.**

***Index Terms*—response style, item response theory, multitask, affect, emotion, perception**

I. Introduction

Subjective affect rating still plays an important role in the affective computing community. In fact, the development of effective rating methods is an active research topic [1], [2]. Response styles (RS) are of particular concern when using subjective rating scales. RS is defined as "a systematic tendency to respond to a range of questionnaire items on some basis other than the specific item content (that is, what the items were designed to measure)" [3]. Some of the most common RSs are acquiescent/disacquiescent RSs (ARS/DRS), in which an individual tends to use the upper/lower range of the scale (e.g. yea-saying/nay-saying), and extreme/midpoint RSs (ERS/MRS), in which a person prefers the ends/center of the scale [4]. Traditionally, such RSs were quantized in a simple manner. For example, extreme RSs are frequently measured as the proportion of extreme choices compared with the total number of items or the standard deviation of item scores within a respondent [4].

Using such simply calculated measures, researchers have demonstrated how RS degrades the validity of the analysis of subjective ratings, such as correlation (e.g. correlation between two scales) and variance-based analyses [4]. For example, shared RSs inflate interrater agreement [5] and they deflate it if the RSs differ between raters [6]. This is a serious factor in cross-cultural studies because of the cultural differences in RS. For example, American college students tend to have more extreme RSs than Japanese students [7], [8]. Traditional methods can provide measurements related to RS, but they cannot eliminate RS from ratings.

Recently, several ways have been proposed for removing RS. The first way is to obtain multiple ratings for the same target and to aggregate them while assuming random independent noise across the ratings. The target may be ratings from multiple people to a visual/auditory stimulus, or ratings from a single person to a set of items about a psychological construct (e.g. a psychological questionnaire). This type of study covers both simple aggregation methods, including averaging and majority voting, and more advanced truth discovery methods [9]–[11]. This technique is useful when the research target is perceived emotion, i.e. an emotion perceived by the general population. However, if the target is individual-level perception, such as felt emotion (what the target person is actually feeling) or how another specific individual perceives it, then such techniques are difficult to apply because of their high cost and/or poor reproducibility. The second method involves using anchors in rating to correct individual differences in the criteria for selecting a category. Example methods are ordinal rating [12], and the Anchoring Vignette method [6]. However, such anchors should be carefully designed so that they are specific to the target task; the designing per se remains a research topic.

The third way to eliminate RS is to build a rating process model with RS as a latent variable and remove the effect of RS. Several RS removal techniques have been proposed mainly in the marketing research area [6], [13]. Most use a single task. For example, when emotion judgment is the target task, RS is estimated solely by using ratings on the task. The main weakness of this approach is that it is difficult

to distinguish RS from task-dependent response tendencies; the two types are called dispositional and situational in [4]. For example, depending on the item/stimulus and/or prepared categories from which to choose, people may tend to select extreme responses in some tasks (e.g. because of the many exaggerated facial expressions that are present in an affect rating task), while choosing the middle category in other tasks (e.g. as a consequence of the ambiguity of items). In such cases, the single task framework yields different RS results, which unfortunately violates the definition of RS.

Focusing on the task independence of RS, we propose using various tasks together to allow us to extract RSs shared across tasks. In fact, in many affective computing studies, Likert-type psychological questionnaires are additionally used to examine the relationship between the results of the main task and the summary statistics (primarily the total score) of the additional tasks. Modern test theories, including the item response theory (IRT), make it possible to estimate both an individual's characteristics and each item's characteristics jointly from answers to such questionnaires. Therefore, we propose a multitask item response model for RS removal.

We believe that this paper makes two major contributions. First, this is the first attempt in the affective computing community to computationally remove RS from affective ratings. Second, it is the first IRM-based multitask framework for estimating/removing RS. This paper demonstrates how our multitask framework works in one of the most fundamental tasks in the affective computing area, namely valence and arousal judgment tasks. The proposed framework can potentially open new horizons for future affective computing studies.

II. Method

Our model is an extension of item response models (IRMs) that contain response style (RS) parameters for polytomous ratings. This section first introduces basic IRMs which have no RS terms, and more advanced IRMs with RS. After that, we describe our model.

A. Single task models

1) Basic item response models: IRMs constitute a family of multivariate generalized linear mixed models (MGLMM) [14]. An IRM consists of three elements: 1) the distribution of data, 2) a link function, which determines which transformation of the mean of the distribution should be modeled linearly, and 3) predictors.

In conventional IRMs, a multivariate Bernoulli distribution, namely a multinomial distribution with a total count equal to one, is used with an adjacent-categories logit. Such IRMs are expressed as

$$\log\left(\frac{P(y_{ij}=s|\boldsymbol{X}_\Theta)}{P(y_{ij}=s-1|\boldsymbol{X}_\Theta)}\right)=\boldsymbol{X}_\Theta \tag{1}$$

where y_{ij} denotes the response of person j to item i, and $\boldsymbol{X}_\Theta$ is a linear predictor that consists of a set of parameters Θ including a person (respondent) parameter and an item (stimulus for affective judgment or item in psychological questionnaire) parameter.

One of the most fundamental item response models is the partial credit model (PCM) [15], in which $\boldsymbol{X}_\Theta$ is defined as $\theta_j-\beta_{is}$. θ_j is the trait of person j (such as ability in the test theory domain), while β_{is} represents the characteristics of item i for category s (e.g. the difficulty of the item to obtain a score s or the selection threshold/criterion for rating s). Generalized PCM (GPCM) [16] is an extended version of PCM, where the effect of a person's ability is assumed to be different across items; namely $\boldsymbol{X}_\Theta=\alpha_i\theta_j-\beta_{is}$, where $\alpha(>0)$ is called a slope parameter (or discrimination parameter), because it determines the slope of the characteristic curve (a cumulative distribution) that represents the relationship between the probability of obtaining the category and the individual's ability. As with many other IRMs, both models ignore any temporal structure, and assume that the model does not change over time.

One of the key properties of PCM is that it inherits from the specific objectivity property of the original Rasch model; that is, the comparison of items does not depend on person parameters, and the comparison of persons does not depend on item parameters [13]. A similar assumption has also been made in the affective computing community to build computational models of social cognition, e.g. [17]. When two items i and i' are compared for the same person j, the difference of their logits has no person term:

$$\begin{aligned}\log\left(\frac{P(y_{ij}=s|\boldsymbol{X}_\Theta)}{P(y_{ij}=s-1|\boldsymbol{X}_\Theta)}\right)-\log\left(\frac{P(y_{i'j}=s|\boldsymbol{X}_\Theta)}{P(y_{i'j}=s-1|\boldsymbol{X}_\Theta)}\right)\\ =(\theta_j-\beta_{is})-(\theta_j-\beta_{i's})=\beta_{i's}-\beta_{is}.\end{aligned} \tag{2}$$

This property also holds for the difference between two persons for the same item. On the other hand, GPCM does not preserve the property due to the interaction term $\alpha_i\theta_j$.

2) Response style models for a single task: Recently, several researchers have proposed incorporating RS into traditional IRMs. They are divided into two categories depending on the definition of response styles. In terms of extreme RSs, some focus on respondent's tendency to select the extreme end points, while others exploit their greater variability of scores assigned to items [18].

One example of the former was proposed by Tutz et al. [13] who incorporated a RS term $\tilde{\gamma}$ into the threshold β_{is} as $\tilde{\beta}_{is}=\beta_{is}-\tilde{\gamma}_{js}$ (which this paper calls PCM_RSt):

$$\boldsymbol{X}_\Theta=\theta_j-(\beta_{is}-\tilde{\gamma}_{js}), \tag{3}$$

where $\tilde{\gamma}_{js}=(m-s+1)\gamma_j$, m is the midpoint category (e.g. $m=2$ when $s\in\{0,1,2,3,4\}$ and $m=2.5$ when $s\in\{0,1,2,3,4,5\}$). Positive γ represents a midpoint RS, while negative γ indicates an extreme RS. If γ is positive, the intervals of β between categories expand around the middle category m, which means that the probability of category m increases (i.e. midpoint RS). If γ is negative, it has the opposite effect, namely the intervals move toward the middle category, and consequently the probability of extreme categories increases (i.e. extreme RS).

For the latter definition of extreme RSs, researchers portrayed ERS in reference to person characteristics that reflect expanded or contracted use of the rating scale (represented by β here) [18]. For example, Jonas and Markon [6] incorporated extreme/midpoint RSs and positive/negative bias into the GPCM[1] (which this paper calls GPCM_RSj) as

$$\boldsymbol{X}_\Theta = \alpha_i\theta_j - \gamma_j(\beta_{is} - \gamma'_j). \tag{4}$$

Here, γ represents extreme/midpoint RSs, as in Tutz et al.'s model (although in Jonas & Markon's model, $\gamma > 0$ and a smaller/larger value means an extreme/midpoint RS), while γ' represents a bias toward a positive/negative category representing an acquiescent/disacquiescent RS. Tutz et al.'s model satisfies the specific objectivity property because there is no interaction between person and item parameters. On the other hand, Jonas & Markon's model does not satisfy the specific objectivity property because of the interaction term.

B. Proposed multitask models

We extend Tutz et al.'s [13] and Jonas & Markon's [6] models to a multitask framework. We incorporate a set of tasks simultaneously using task-independent parameters that describe RS. Our multitask version of Tutz et al.'s model (mtPCM_RSt) is defined as:

$$\boldsymbol{X}_\Theta = \theta_{jk} - (\beta_{iks} - \tilde{\gamma}_{js}), \tag{5}$$

where k is a task index. Note that RS parameter $\tilde{\gamma}$ excludes subscript k because of the task-independence. Our extension of Jonas & Markon's model (mtGPCM_RSj) is:

$$\boldsymbol{X}_\Theta = \alpha_{ik}\theta_{jk} - \gamma_j(\beta_{iks} - \gamma'_j). \tag{6}$$

We also a built GPCM version of mtPCM_RSt and a PCM version of mtGPCM_RSj by replacing θ_{jk} and $\alpha_{ik}\theta_{jk}$ (called mtGPCM_RSt and mtPCM_RSj, respectively). Only mtPCM_RSt satisfies the specific objectivity property, while mtGPCM_RSt, mtPCM_RSj and mtGPCM_RSj do not. Table I compares all four proposed models with the baseline models.

After estimating model parameter Θ, we can predict the ratings that are likely to be obtained from the model. Predictive rating $\hat{y}$ is estimated as:

$$\hat{y}_{ijk} \sim categorical(\boldsymbol{\pi}) \tag{7}$$

$$\pi_s = P(y = s|\boldsymbol{X}_\Theta) \tag{8}$$

In addition, the RS-removed ratings are estimated by excluding the RS term $\tilde{\gamma}$ from the predictor in Eq. 8. This can be expressed as:

$$\pi_s = P(y = s|\boldsymbol{X}_{\Theta'}) \tag{9}$$

where $\boldsymbol{X}_{\Theta'} = \theta_{jk} - \beta_{iks}$ for the PCM family and $\boldsymbol{X}_{\Theta'} = \alpha_{ik}\theta_{jk} - \beta_{iks}$ for the GPCM family.

[1]To be exact, Jonas and Markon's model [6] is based on Graded Response Model (GRM), which uses cumulative logit, rather than adjacent-categories logit.

TABLE I
LIST OF ITEM RESPONSE MODEL FAMILIES

Model	Predictors $\boldsymbol{X}_\Theta$
Models w/o response style	
Baseline models	
PCM [15]	$\theta_{jk} - \beta_{iks}$
GPCM [16]	$\alpha_{ik}\theta_{jk} - \beta_{iks}$
Models w/ response style	
Baseline models	
PCM_RSt [13]	$\theta_{jk} - (\beta_{iks} - \tilde{\gamma}_{jks})$
GPCM_RSj [6]	$\alpha_{ik}\theta_{jk} - \gamma_{jk}(\beta_{iks} - \gamma'_{jk})$
Proposed models	
mtPCM_RSt	$\theta_{jk} - (\beta_{iks} - \tilde{\gamma}_{js})$
mtGPCM_RSt	$\alpha_{ik}\theta_{jk} - (\beta_{iks} - \tilde{\gamma}_{js})$
mtPCM_RSj	$\theta_{jk} - \gamma_j(\beta_{iks} - \gamma'_j)$
mtGPCM_RSj	$\alpha_{ik}\theta_{jk} - \gamma_j(\beta_{iks} - \gamma'_j)$

All models use adjacent-categories logit as a link function. β is an item parameter subscripted with item index i (and in some cases category index s). θ is a person parameter subscripted with person index j (and in some cases category index s). α is a scale parameter subscripted with item index i. k is a task index. Note that we also include task index k in the baselines for comparison.

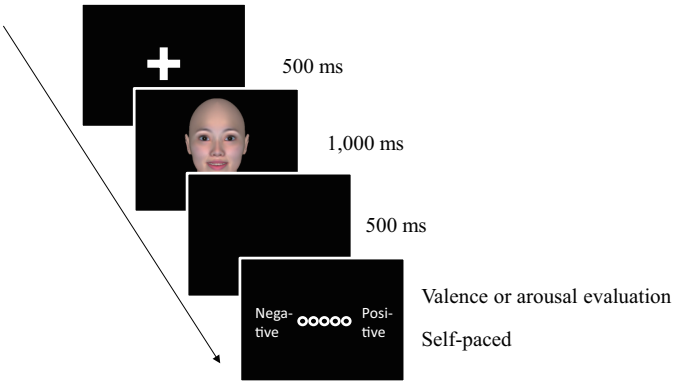

Fig. 1. Main task: valence-arousal judgment task. First, a fixation cross was displayed in the center of the screen for 500 msec. Then, a target face was shown for 1,000 msec. Next, the screen was blank for 500 msec. Finally, a valence or arousal scale was displayed until the participant selected one of the answers.

III. EXPERIMENTAL DATA

To evaluate the proposed framework, we performed valence and arousal judgment tasks using computer-generated static emotional faces as the main tasks. We also used psychological questionnaires as subtasks.

A. Observers

Fifty Japanese university students (25F) participated in the experiment. This homogeneity facilitated our verification of whether or not the estimated RS parameters really matched the Japanese midpoint RS, which has been repeatedly reported in previous studies [7], [8].

B. Main tasks: affective rating

The participants were asked to rate the valence and arousal level of artificial faces. This was a blocked design: one block for valence judgment and the other block for arousal

judgment. Each consisted of a forced choice on a 5-point scale: the extremes were labeled "Positive" and "Negative" in the valence block, and "High" and "Low" in the arousal block. Figure 1 shows the timeline of each trial. Each block consisted of 150 trials. In 120 of the 150 trials, a total of 120 original faces were displayed. The remaining 30 trials were a repetition of 30 trials that were randomly selected from the 120 trials. The aim was to calculate the test-retest reliability, i.e. the frequency with which participants gave the same rating to exactly the same face in different trials. The block order was counter-balanced, and the stimulus order and the 30 repeated faces in each block were randomized across the participants. All the labeling was performed in isolation, and all the observers successfully completed both tasks.

Various mixed facial expressions were included to make it possible to observe inter-individual differences in perceptions among respondents. The 120 stimulus faces were created using the FaceGen modeler. The faces consisted of 29 facial expressions (1 neutral and 28 non-neutral expressions) from 8 different artificial identities. Specifically, we extracted 15 expressions (1 neutral and 14 non-neutral expressions) from 4 virtual identities (called Face Set 1), and other 15 expressions (also 1 neutral and 14 non-neutral expressions) from 4 other identities (called Face Set 2). The expressions were manipulated by changing the modeler's expression-specific parameters (anger, disgust, fear, sadness, surprise, and closed- and open-mouthed smiles; a total of seven categories). Of the non-neutral expressions in Face Set 1, four were pure anger, fear, surprise and an open-mouthed smile, and the remaining 10 were combinations of the seven categories. Three of the non-neutral expressions in Face Set 2 were pure disgust, sadness and a closed-mouthed smile, and the remaining 11 were other combinations of the seven categories. The eight identities were drawn from Caucasians, Africans, Indians and Asians: each of which consisted of both masculine and feminine faces. This procedure yielded 120 (=15×4+15×4) faces.

C. Subtasks: psychological questionnaires

The participants were also asked to answer seven psychological questionnaires after the main tasks: Empathizing Quotient (EQ) [19], Systemizing Quotient (SQ) [19], Autism-Spectrum Quotient (AQ) [20], Interpersonal Reactivity Index (IRI) [21], Emotional Skills and Competence Questionnaire (ESCQ) [22], Neo-FFI or Big Five (B5) [23], and the Tokyo University Egogram (TEG) [24]. EQ, AQ, IRI and ESCQ are commonly used to measure empathy-related traits, while B5 and TEG are used for more general personality traits. They are not completely independent of each other, nor are they fully independent of valence/arousal decision tasks. However, the entire questionnaire set reasonably covers various types of traits and the number of points (ranging from a 3-point scale to a 7-point scale and including both even and odd points).

Table II summarizes the number of items and the number of points in the questionnaires. The total number of ratings was 579 items $\times$ 50 respondents $= 31,950$. There were no missing data. However, our models accept missing data in the current form, thanks to Bayesian generative modeling as described in IV-A.

TABLE II
SUMMARY OF USED TASKS

Task	#items	#points
Main tasks		
1. Valence judgment	150	5
2. Arousal judgment	150	5
Sub-tasks		
3. EQ [19]	60	4
4. SQ [19]	60	4
5. AQ [20]	50	4
6. IRI [21]	28	4
7. ESCQ [22]	28	5
8. B5 [23]	60	7
9. TEG [24]	53	3
Sum	639	

IV. EVALUATION SETTINGS

A. Bayesian parameter estimation

All the models were implemented using the Stan probabilistic programming language and its interface with the R (Stan Development Team, 2015a, b), and the edstan (v1.0.6; Furr, 2017) package. The model parameters were estimated using Stan's No-U-Turn Sampler (NUTS). As weak priors, we used zero-mean normal distributions for $\tilde{\gamma}$ (for the Tutz et al's family), γ' (for the Jonas & Markon family), β and θ, and unit-mean lognormal distributions for γ (for the Jonas & Markon family) and α. Four MCMC chains were run from random start values. The chain convergence was assessed by the $\hat{\mathrm{R}}$ statistic ($\hat{\mathrm{R}} < 1.1$). The first 7,200 iterations were used as a warm-up and discarded, and then 2,400 iterations were obtained and stored from each chain, yielding 9,600 iterations that served to empirically approximate the posterior distribution.

Predictive RS-inclusive ratings and RS-removed ratings ($\hat{y}$) were obtained for the main tasks as follows. The ratings for all respondents and items (50×300 samples) were simulated according to Eq. 7 and Eq. 8 for RS-inclusive ratings, and Eq. 9 for RS-removed ratings. This procedure was repeated 9,600 times. This yielded a posterior distribution consisting of 9,600 random samples for each $\hat{y}$ for both types of predictive ratings. Furthermore, a point estimate was determined for each $\hat{y}$ by majority voting with the 9,600 samples. The following analysis used the point estimates unless otherwise specified.

B. Performance measure

For the main valence and arousal tasks, we report the following four measures to indicate how well each model explains the observed ratings: accuracy (percent agreement, κ), Pearson's correlation coefficient (r), mean absolute error (MAE), and intra-class correlation coefficient (ICC), following [25], which recommend the joint use of multiple measures. As evaluation criteria for model comparison in term of overall generalizability to both main and sub tasks, the approximate widely applicable information criterion (WAIC) [26] and

Pareto smoothed importance sampling leave-one-out cross-validation (PSIS-LOO) [27] (an approximated LOO) were also calculated using the loo package (v.2.0.0; https://mc-stan.org/loo/). Both measures penalize model complexity, and a smaller value indicates a better model.

V. Results

This section reports various validation results, including a model comparison, and a prediction performance evaluation. All the results support the validity of our proposed framework.

A. Rating results

1) Basic statistics: The proportions of the rating categories (the marginal distribution of ratings) were (09, .32, .35, .20, .04) (from negative to positive) for valence, and (.09, .25, .31, .28, .07) (from low to high) for arousal.

The test-retest reliability (calculated in a manner similar to that used for accuracy) κ was .525 for valence and .475 for arousal. This is a percent agreement, meaning that the participants gave the same rating for the test and retest pairs at a rate of κ. Fleiss' generalized κ, κ_F, the chance-corrected agreement, were .345 and .300, respectively, for the valence and arousal ratings. Pearson's r was .556 for valence and .550 for arousal. This value is comparable to that reported in the literature, e.g. [28]. The ICC(2,1) was .48 for valence and .35 for arousal. Both are considered to be between poor and fair [29], [30]. These values provide a good demonstration of how differently people rate affective faces.

However, it is uncertain whether it is caused by an individual difference of perception or by the RS. Therefore, we investigate the impact of RS on these reliability measures in V-D.

B. Model comparison

All models successfully converged on learning. Table III summarizes their performance. For the main tasks (in terms of κ, r, MAE and ICC), our mtGPCM_RSt outperformed both the baselines (PCM and GPCM) and our remaining multitask models[2]. In terms of the overall criteria (i.e. WAIC and LOO), mtPCM_RSj was the best preferred model. However, mtPCM_RSj performed worse on the main tasks suggesting that it well fitted only on the sub tasks. Therefore, we conclude that mtGPCM_RSt is the best model.

[2]The accuracy of mtGPCM_RSt was higher than the test-retest reliability. This may sound strange, but it is possible. The *upper bound* of the prediction accuracy was estimated to be .83 for valence and .81 for arousal. The upper bounds were obtained as follows. Our data can be divided into two types and they should be considered separately. Of the 120 images (150 trials), 30 (60) were shown twice, and the remaining 90 (90) were used only once. For the 60% (=90/150) samples, perfect accuracy is possible if a very complex model is used (although this probably results in overfitting). This is because the training and test sets were identical. For the 40% (=60/150) samples, κ_F percent of samples, where the test and retest ratings are identical (not by chance), perfect accuracy is also possible. The remaining $1-\kappa_F$ percent of samples were however rated differently in a pair of trials, and thus perfect accuracy is not possible. This is because the proposed models (as well as the baselines) give the same predictive rating for each pair of trials. If the random sampling of ratings from the marginal distribution is assumed for the samples, the maximum chance levels (p_{max}) are .35 and .31 for valence and arousal tasks, respectively. Therefore, the estimated upper bound for the 40% data is $\kappa_F \times 1 + (1-\kappa_F) \times p_{max} = .574$ for valence and .517 for arousal. Taken together, the overall upper bound is expected to be $.574 \times 40\% + 1 \times 60\% = 0.83$ for valence, and $.517 \times 40\% + 1 \times 60\% = 0.81$ for arousal. The observed accuracies are within this range.

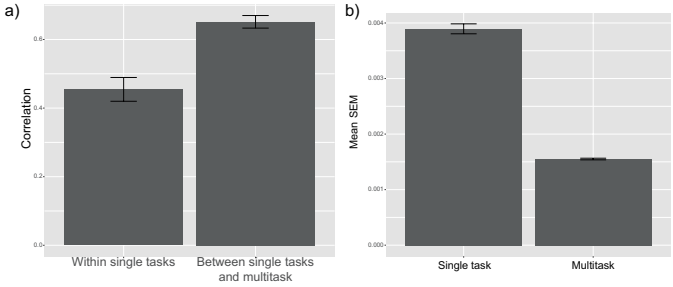

Fig. 2. (a)Correlation of estimated γ within single tasks (using GPCM_RSt) and those between single tasks (using GPCM_RSt) and multitask (using mtGPCM_RSt). (b) Mean SEM of γ's posterior distribution: GPCM_RSt vs mtGPCM_RSt. Error bar indicates SEM.

C. Single-task vs multitask

In Table III, (G)PCM_RSt outperformed our mt(G)PCM_RSt in terms of WAIC and LOO. This means that if the objective is to describe the observed ratings as accurately as possible, (G)PCM_RSt should be selected. However, as mentioned in I, the single task framework confuses task-dependent response tendencies with RS.

To further illustrate the need for the multitask framework quantitatively, Fig. 2 (a) shows the pairwise correlation of estimated γ (a 50-d vector) within 9 single tasks using GPCM_RSt (yielding a within-single-task correlation for each of ${}_9C_2 = 36$ pairs of tasks). It also includes the correlation between the γ values and those obtained using our mtGPCM_RSt. The estimated γ in the single task was closer to the estimate in the multitask than that in a different single task. This reasonably demonstrates the task-independence of RS.

In addition, Fig. 2 (b) shows another benefit of using multiple tasks; the multitask framework gave a more precise estimate. The posterior distribution of γ was narrower in the multitask scenario (mtGPCM_RSt) than in the single scenarios (GPCM_RSt). This is an important property because γ parameters are interconnected with the other parameters and thus a precise estimate of γ is expected to lead to precise estimates of the remaining parameters.

D. Estimated parameters and response style removal

Figure 3 shows a histogram of the estimated γ across 50 participants using mtGPCM_RSt. The mean value was positive (M = 0.42 ($\pm$0.07 SEM), $p < .001$, d = .81), indicating that the participants had a midpoint RS overall. The midpoint RS of Japanese people is in line with that reported in previous studies [7], [8]. Furthermore, the estimated γ values were reasonably correlated with the traditional measure of extreme RS, i.e. the proportion of extreme choices in relation to the total number of items [4] (Spearman's $\rho = -.91, p < .001$). These results validate our method. Moreover, $\tilde{\gamma}$ of GPCM_RSt and γ of PCM_RSj showed strong correlation; $\rho = .78, p < .001$.

TABLE III
PREDICTIVE PERFORMANCE OF THE PROPOSED MODELS AND BASELINES FOR ALL NINE TASKS

Model	WAIC ↓		LOO ↓		Valence task				Arousal task			
	Mean	SEM	Mean	SEM	κ ↑	r ↑	MAE ↓	ICC ↑	κ ↑	r ↑	MAE ↓	ICC ↑
Single task models												
PCM [15]	78,771	272	78,957	277	.599	.672	.457	.664	.466	.557	.665	.552
GPCM [16]	77,016	277	77,273	283	.606	.687	.446	.681	.482	.582	.638	.576
PCM_RSt [13]	73,683	284	73,870	289	.645	.699	.415	.696	.515	.590	.621	.589
GPCM_RSt	72,438	287	72,698	292	.653	.717	.401	.714	.523	.599	.606	.598
Multitask models												
mtPCM_RSt	76,589	279	76,773	284	.615	.681	.443	.676	.484	.570	.649	.567
mtGPCM_RSt	75,126	283	75,361	288	**.623**	**.700**	**.429**	**.695**	**.498**	**.596**	.624	**.594**
mtPCM_RSj	**72,117**	245	**72,345**	248	.599	.670	.456	.660	.463	.555	.661	.546
mtGPCM_RSj	74,881	240	74,949	241	.616	.680	.441	.674	.490	**.596**	**.620**	.589

"↑" and "↓" denote higher and lower levels of performance. Note that although achieving the best performance in terms of the predictive performance of *ratings*, GPCM_RSt, a single task framework, confuses task-dependent response tendencies with RS, as mentioned in I.

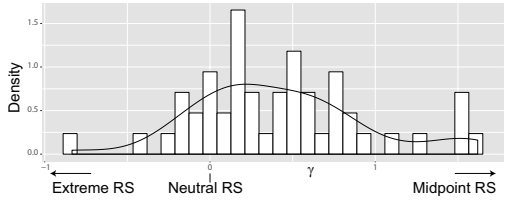

Fig. 3. Histogram of γ estimated using mtGPCM_RSt. Positive and negative values indicate midpoint and extreme response styles.

An ICC(2,1) of the estimated posterior ratings obtained with Eqs. 7 and 8 for all 9,600 samples (not their point estimates), namely a recovered ICC, was M = .49 (95% CI [.47, .51]) for valence and M = .36 (95% CI [.34, .38] for arousal. The observed ICCs (.48 and .35, respectively) were successfully replicated.

The RS-removed ratings were estimated following Eqs. 7 and 9. This slightly but statistically significantly increased the recovered ICCs; .51 (95% CI [.49, .54]) for valence and .41 (95% CI [.38, .44]) for arousal. This suggests that in our participant set, the observed ICCs were deflated because many participants had a midpoint RS while some had an extreme RS. This supports the need for RS correction.

VI. DISCUSSION

We have provided a body of evidence in support of our multitask framework. However, several issues still remain.

First, our multitask framework successfully found the Japanese midpoint RS. However, this was an indirect evaluation, and a more direct evaluation is needed. One way to achieve this is to use an anchoring vignette technique, such as [6], in which respondents are also asked to judge imaginary character(s) as *anchor* that are assumed to result in the same judgment from everyone, in order to normalize each respondent's judgment based on their judgment regarding the anchor.

Second, our model is probably not the *best model* for eliminating RS in a multitask fashion. First, although we use the same base model (PCM or GPCM) for all tasks, we can use different models for different tasks in our framework. It is reasonable to use a simple model (e.g. PCM) for psychological questionnaires, since they are basically designed to measure a single construct. However, it would be interesting to find the best, or at least a better, model for affective judgment tasks.

Thirdly, as secondary tasks this study used seven psychological questionnaires similar in size to the main task, namely 339 items compared with 300 items. This increases the cost and complexity of collecting affective ratings, especially when crowd sourcing is used. Thus, it is necessary to evaluate the effect of the number of secondary tasks.

Fourthly, this study employed a discrete annotation procedure for both time and emotion space. To apply our work to continuous annotations, as with the recent trend in the affective community, e.g. [1], [2], our model must be extended. It is also interesting to investigate whether or not the rating process is time invariant, as mentioned in [12].

Finally, this paper focused on decoder or receiver in emotional communication, i.e. the affective judgment of other people. It is also interesting to target a coder's or sender's judgment, i.e. a self-report of emotional states. This is an important step because a self-report is available only from an individual. Therefore, the impact of RS on their ratings is expected to be stronger than that of a decoder. It would be interesting to incorporate physiological signals in our framework, as used in felt-emotion studies [31]–[33].

VII. CONCLUSION

This paper proposed a multitask RS removal framework, where an individual's responses in multiple tasks are modeled using task-independent RS terms, and task-dependent terms, including item and respondent's characteristic parameters based on the item response model (IRM). Through Bayesian modeling, we observed that i) the proposed model outperformed traditional IRMs in terms of predictive accuracy; ii) our multitask framework estimated RS with higher precision than previous single-task-based RS removal methods; iii) our model replicated Japanese midpoint RS; and iv) RS-removed predictive ratings showed higher inter-rater agreement than those including RS in valence/arousal judgment tasks. The proposed RS removal technique has the potential to reveal

new/stronger results that previous methods used by the affective computing community were unable to find. Validating the potential constitutes one of the next steps.

REFERENCES

[1] R. Cowie, E. Douglas-Cowie, S. Savvidou, E. McMahon, M. Sawey, and M. Schröder, "FEELTRACE: An instrument for recording perceived emotion in real time," in *ISCA Workshop on Speech and Emotion*, 2000, pp. 19–24.

[2] G. N. Yannakakis and H. P. Martínez, "Grounding truth via ordinal annotation," in *Proc. ACII*, 2015, pp. 574–580.

[3] D. L. Paulhus, *Measures of personality and social psychological attitudes*. San Diego, CA, US: Academic Press, 1991, ch. Measurement and control of response bias, pp. 17–59.

[4] H. Baumgartner and J.-B. E. Steenkamp, "Response styles in marketing research: A cross-national investigation," *Journal of Marketing Research*, vol. 38, no. 2, pp. 143–156, 2001.

[5] S. Dolnicar and B. Grün, "Response style contamination of student evaluation data," *Journal of Marketing Education*, vol. 31, no. 2, pp. 160–172, 2009.

[6] K. G. Jonas and K. E. Markon, "Modeling response style using vignettes and person-specific item response theory," *Applied Psychological Measurement*, vol. 43, no. 1, pp. 3–17, 2019.

[7] M. Zax and S. Takahashi, "Cultural influences on response style: Comparisons of japanese and american college students," *The Journal of Social Psychology*, vol. 71, no. 1, pp. 3–10, 1967.

[8] C. Chen, S.-Y. Lee, and H. W. Stevenson, "Response style and cross-cultural comparisons of rating scales among east asian and north american students," *Psychological Science*, vol. 6, no. 3, pp. 170–175, 1995.

[9] A. P. Dawid and A. M. Skene, "Maximum likelihood estimation of observer error-rates using the EM algorithm," *Appl. Statist.*, vol. 28, no. 1, pp. 20–28, 1979.

[10] P. Welinder, S. Branson, S. Belongie, and P. Perona, "The multidimensional wisdom of crowds," in *Neural Information Processing Systems Conference (NIPS)*, 2010, pp. 2424–2432.

[11] A. Rui, O. Martinez, X. Binefa, and F. Sukno, "Fusion of valence and arousal annotations through dynamic subjective ordinal modelling," in *Proc. IEEE FG*, 2017 2017.

[12] G. N. Yannakakis, R. Cowie, and C. Busso, "The ordinal nature of emotions," in *Proc. Int'l Conf. Affective Computing and Intelligent Interaction (ACII)*, 2017, pp. 248–255.

[13] G. Tutz, G. Schauberger, and M. Berger, "Response styles in the partial credit model," *Applied Psychological Measurement*, vol. 42, no. 6, pp. 407–427, 2018.

[14] F. Tuerlinckx and W.-C. Wang, *Explanatory Item Response Models*. New York, NY: Springer, 2004, ch. Models for polytomous data, pp. 75–109.

[15] G. N. Masters, "A rasch model for partial credit scoring," *Psychometrika*, vol. 47, no. 2, pp. 149–174, Jun 1982.

[16] E. Muraki, "A generalized partial credit model: Application of an em algorithm," *ETS Research Report Series*, vol. 1992, no. 1, pp. i–30, 1992.

[17] S. Kumano, R. Ishii, and K. Otsuka, "Computational model of idiosyncratic perception of others' emotions," in *Proc. Int'l Conf. Affective Computing and Intelligent Interaction (ACII)*, Oct 2017, pp. 42–49.

[18] D. M. Bolt and J. R. Newton, "Multiscale measurement of extreme response style," *Educational and Psychological Measurement*, vol. 71, no. 5, pp. 814–833, 2011.

[19] S. Baron-Cohen, "Autism: The empathizing-systemizing (e-s) theory," *Ann. N. Y. Acad. Sci.*, vol. 1156, pp. 68–80, 2009.

[20] S. Baron-Cohen, S. Wheelwright, R. Skinner, J. Martin, and E. Clubley, "The autism-spectrum quotient (aq): Evidence from asperger syndrome/high-functioning autism, malesand females, scientists and mathematicians," *Journal of Autism and Developmental Disorders*, vol. 31, no. 1, pp. 5–17, Feb 2001.

[21] M. H. Davis, "Measuring individual differences in empathy: Evidence for a multidimensional approach," *J. Pers. Soc. Psychol.*, vol. 44, no. 1, pp. 113–126, 1983.

[22] V. Taksic, "The importance of emotional intelligence(competence) in positive psychology," in *Proc. The First International Positive Psychology Summit*, 2002.

[23] P. T. Costa and R. R. McCrae, *Revised NEO Personality Inventory (NEO-PIR) and NEO Five Factor Inventory (NEO-FFI) professional manual*. Psychological Assessment Resources, 1992.

[24] H. Suematsu, S. Nomura, and M. Wada, *Handbook of TEG. 2nd edition [in Japanese]*. Tokyo: Kaneko-shobo, 1993.

[25] H. Gunes and B. Schuller, "Categorical and dimensional affect analysis in continuous input: Current trends and future directions," *Image Vision Comput.*, vol. 31, no. 2, pp. 120–136, 2013.

[26] S. Watanabe, "Asymptotic equivalence of bayes cross validation and widely applicable information criterion in singular learning theory," *J. Mach. Learn. Res.*, vol. 11, pp. 3571–3594, 2010.

[27] A. Vehtari, A. Gelman, and J. Gabry, "Practical bayesian model evaluation using leave-one-out cross-validation and waic," *Statistics and Computing*, vol. 27, no. 5, pp. 1413–1432, 2017.

[28] T. Bänziger, D. Grandjean, and K. R. Scherer, "Emotion recognition from expressions in face, voice, and body: the multimodal emotion recognition test (MERT)," *Emotion*, vol. 9, pp. 691–704, 2009.

[29] T. K. Koo and M. Y. Li, "A guideline of selecting and reporting intraclass correlation coefficients for reliability research," *Journal of Chiropractic Medicine*, vol. 15, no. 2, pp. 155 – 163, 2016.

[30] D. V. Cicchetti, "Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology," *Psychological Assessment*, vol. 6, no. 4, pp. 284–290, 1994.

[31] M. Soleymani, J. Lichtenauer, T. Pun, and M. Pantic, "A multimodal database for affect recognition and implicit tagging," *IEEE Trans. Affect. Comput.*, vol. 3, no. 1, pp. 42–55, 2012.

[32] S. Koelstra, C. Muhl, M. Soleymani, J.-S. Lee, A. Yazdani, T. Ebrahimi, T. Pun, A. Nijholt, and I. Patras, "Deap: A database for emotion analysis ;using physiological signals," *IEEE Trans. Affect. Comput.*, vol. 3, no. 1, pp. 18–31, 2012.

[33] H. P. Martinez, Y. Bengio, and G. N. Yannakakis, "Learning deep physiological models of affect," *IEEE Comp. Intell. Magazine*, vol. 8, no. 2, pp. 20–33, 2013.