JOURNAL OF APPLIED STATISTICS
https://doi.org/10.1080/02664763.2018.1547692

Taylor & Francis
Taylor & Francis Group

Statistical considerations for crowdsourced perceptual ratings of human speech productions

Daniel Fernández [a,b], Daphna Harel [c,d], Panos Ipeirotis[e] and Tara McAllister[f]

[a]Parc Sanitari Sant Joan de Déu, Institut de Recerca Sant Joan de Déu, CIBERSAM, Spain; [b]School of Mathematics and Statistics, Victoria University of Wellington, New Zealand; [c]Department of Applied Statistics, Social Science, and Humanities, Steinhardt School of Culture, Education, and Human Develoment, New York University, New York, USA; [d]PRIISM Applied Statistics Center, New York University, New York, NY, USA; [e]Leonard N. Stern School of Business, New York University, New York, USA; [f]Department of Communicative Sciences and Disorders, Steinhardt School of Culture, Education, and Human Development, New York University, New York, USA

ABSTRACT

Crowdsourcing has become a major tool for scholarly research since its introduction to the academic sphere in 2008. However, unlike in traditional laboratory settings, it is nearly impossible to control the conditions under which workers on crowdsourcing platforms complete tasks. In the study of communication disorders, crowdsourcing has provided a novel solution to the collection of perceptual ratings of human speech production. Such ratings allow researchers to gauge whether a treatment yields meaningful change in how human listeners' perceive disordered speech. This paper will explore some statistical considerations of crowdsourced data with specific focus on collecting perceptual ratings of human speech productions. Random effects models are applied to crowdsourced perceptual ratings collected in both a continuous and binary fashion. A simulation study is conducted to test the reliability of the proposed models under differing numbers of workers and tasks. Finally, this methodology is applied to a data set from the study of communication disorders.

ARTICLE HISTORY
Received 22 August 2017
Accepted 3 November 2018

KEYWORDS
Reliability; validity; random effects models; Amazon Mechanical Turk; communication disorders; crowdsourcing

1. Introduction

Crowdsourcing enables data to be collected quickly, cheaply, and efficiently by separating projects into micro-tasks that can be completed in small amounts of time. First coined by Howe [31], crowdsourcing has become a major tool for scholarly research since its introduction to the academic sphere in 2008 [7,31]. The use of crowdsourcing in research spans many fields, including psychology [4,22,47], linguistics [63], astronomy [23], education [49], marketing [21], game theory [25], health research [38,65,67], and speech-language pathology [26,35,40,41], among many others. The Amazon Mechanical Turk (AMT) crowdsourcing platform (www.MTurk.com) and its workers have been most widely used and studied by academic researchers to date. Many studies on the utility

CONTACT Daphna Harel ✉ daphna.harel@nyu.edu PRIISM Applied Statistics Center, New York University, 246 Greene Street, 3rd Floor, New York, NY 10003, USA

of crowdsourcing have focused on the motivation of AMT workers, or comparing the population of AMT workers to those in traditional laboratory experiments. These studies have found that AMT workers are more likely to be female than male, and they have a median age of roughly 30 years [64]. Most workers do not rely on AMT as their primary source of income with most US-based workers using AMT to provide a secondary source of income [32]. AMT workers have been described as 'less naive than researchers assume' [9], and 'more representative of the U.S. population than in-person convenience samples' [5].

While AMT was developed for commercial use, researchers have exploited its interface, trusted payment system, and built-in advertising with considerable success [8]. The ease and speed of crowdsourced data collection has been described as 'revolutionary', with instances where traditional laboratory-based experiments were reproduced in roughly 2–3% of the time originally needed [13,63]. However, unlike in traditional laboratory settings, it is nearly impossible to control the conditions under which workers complete tasks. While crowdsourcing researchers do not dispute that this form of data collection may result in higher variability than traditional methods, they argue that the ability to recruit much larger samples than would ordinarily be available has the potential to overcome these drawbacks. Moreover, although an individual recruited online is unlikely to display expert-level performance in a given task, responses aggregated over numerous non-experts generally converge with responses obtained from experts. This assertion is supported by both computational modeling studies [33] and by experimental studies that validate results obtained through AMT [13,30,47].

This paper focuses on an application of crowdsourcing in the study of communication disorders, and specifically on disorders affecting speech production, where it has the potential to represent a novel solution to a longstanding problem. Communication disorders affect up to 10% of the total population and are estimated to cost the US economy $150 billion per year [27]. Because effective medical management can reduce these impacts, research investigating communication disorders represents an important public health priority. To study the efficacy of speech interventions, researchers must measure changes in speech production accuracy or intelligibility over time. From a clinical standpoint, it is most important to know whether a treatment yields a meaningful change in human listeners' perception of speech [56]. The conventional approach to obtaining perceptual ratings of speech data is to rely on certified clinicians [42] or students in speech-language pathology [39]. While it is desirable to use the expert judgment of certified clinicians, researchers may find it difficult to offer compensation in line with the typical pay rate of speech-language pathologists. In such cases, they may resort to non-optimal, and potentially biasing, methods such as using the authors or other study personnel as the source of perceptual ratings [40]. Crowdsourcing could be an important method to overcome bottlenecks in the process of obtaining valid ratings of clinical speech samples. Recent studies have validated the use of crowdsourced perceptual ratings against those obtained from expert listeners [40], compared different elicitation conditions [41], and assessed the reliability of crowdsourced ratings across repeated presentations [26]. Additionally, aggregated ratings have been found to correlate strongly with acoustic gold standard measures both when individual raters use a continuous rating scale, such as visual analog scaling [44], and when individual raters provide binary ratings [40].

Increasing the speed of data collection through crowdsourcing is only advisable if it does not compromise the quality of the data and any inferences subsequently drawn. Previous research has aimed to understand and mitigate the drawbacks of crowdsourcing that arise from the inherent variability of data obtained from multiple workers [58], specifically if some workers do not understand the requested task [71], cheat [62] or vary in quality or expertize [70,72]. Thus, previous work has developed algorithms to estimate the quality of the workers, allowing for the rejection and blocking of the low-performing workers and spammers [34,58,69,71]. Dawid and Skene [15] developed an expectation maximization (EM) algorithm to obtain maximum likelihood estimates of workers' error rates in rating tasks when the true rating (gold standard) is unknown [15]. Ipeirotis, Provost, and Wang [34] used the EM algorithm to assign quality scores to workers who answer multiple choice tasks by separating true error rates from workers' biases [34]. Bayesian extensions of the EM algorithm have been proposed to capture the skill of different workers through prior distributions that are specified from the results of similar previous experiments or pilot studies [53,71]. Others have developed algorithms to filter ratings from non-experts in settings where the quality of the workers are already known [52,72]. Recent work has combined spectral methods and EM algorithms with an optimal convergence rate up to a logarithmic factor for inferring the true ratings from the noisy ratings provided by non-expert crowdsourcing workers [73].

However, most current methods are algorithmic techniques that assign a deterministic score to classify the workers, instead of providing an underlying probabilistic model for the data and subsequent ratings obtained. Without quantifying the uncertainty inherent in such estimates, it is unclear how to assess or compare these algorithms. Probabilistic models and standard inferential tools, however, provide a principled way to assess the appropriateness of model fit for a particular dataset, the reproducibility of results to future datasets, and the generalizability of these approaches to other contexts.

This paper explores some statistical considerations of crowdsourced data with specific focus on collecting perceptual ratings of human speech productions. Following this brief introduction, Section 2 reviews random effects models, and explores their utility in modeling crowdsourced perceptual ratings collected using both continuous and binary rating scales. Section 3 presents a simulation study to determine the number of workers and tasks required to obtain robust estimates of both worker quality and task accuracy. Section 4 applies this methodology to a dataset consisting of perceptual ratings of speech produced by children receiving treatment for misarticulation of the English /r/ sound. Finally, Section 5 presents concluding remarks and discussion of future work.

2. Methodology

In crowdsourced tasks of rating human speech productions, a set of J workers are presented with I tasks, to rate on a unidimensional continuous trait, such as accuracy or intelligibility. Thus, each task has some true value, $\{\alpha_i\}$, that must be estimated from the observed responses, y_{ij}. However, each worker exhibits some degree of skill or bias in the rating task, which may be characterized by the set of parameters $\{\beta_j\}$. Therefore, the ratings obtained for each task are noisy estimates of the true, underlying value.

Despite the continuous nature of the underlying trait, perceptual ratings of human speech production, y_{ij}, may be obtained through either continuous or binary response mechanisms. In a continuous response task, workers may be presented with a visual analog scale (VAS) demarcated by two endpoints indicating fully correct and fully incorrect productions. The corresponding measurement is calculated as the proportion of the total VAS length contained between the left endpoint and the worker's click location. In a binary task, workers may be presented with two options (e.g. correct or incorrect) and forced to categorize the task as one or the other. Either of these response mechanisms may be used to derive valid measures of gradient characteristics of speech, either by averaging VAS click locations across workers, or by computing the proportion of workers who selected a particular option (e.g. correct) in a forced-choice task [41].

From a statistical perspective, this data structure falls within the framework of non-nested random effect models. We assume that the response variables y_{ij} are drawn from a probabilistic model that depends on both the worker and task random effects [20]. Models for both continuous and binary responses are described below.

2.1. Random effects model for continuous responses

When continuous rating tasks are used, a random effects model for the responses may be formulated as follows:

$$\begin{aligned} &y_{ij} = \mu + \alpha_i + \beta_j + \varepsilon_{ij}, \quad \varepsilon_{ij} \sim \mathcal{N}(0, \sigma_y^2),\\ &\alpha_i \sim \mathcal{N}(0, \sigma_\alpha^2), \quad \beta_j \sim \mathcal{N}(0, \sigma_\beta^2), \\ &y_{ij} \sim \mathcal{N}(\mu + \alpha_i + \beta_j, \sigma_y^2 + \sigma_\alpha^2 + \sigma_\beta^2) \quad i = 1, \ldots, I, \ \ j = 1, \ldots, J, \end{aligned} \tag{1}$$

where y_{ij} is the continuous measurement given by worker j to task i, μ is the intercept describing the overall mean of the response distribution, α_i and β_j are the main effects in the response y_{ij} for task i and worker j respectively, and ε_{ij} represents the residual error. This model requires I+J+3 parameters to be estimated, including the I task-level parameters, the J worker-level parameters, and three variance estimates: σ_y, σ_α and σ_β. The use of random effects for both $\{\alpha_i\}$ and $\{\beta_j\}$ rather than fixed effects allows the specific workers and tasks to be modeled as a random sample from the set of all possible workers and tasks. Furthermore, although not considered here, this framework may be extended to adjust for the effects of worker-level or task-level covariates, such as the age or gender of either the worker or the speaker producing each task.

The random effects paradigm provides a natural measure of reliability for both workers and tasks. Since workers vary in their ability to provide accurate task ratings, it is of interest to measure the reliability of the ratings obtained from a single worker. Additionally, since the reliability of a single worker's ratings may be extremely low, it is also of interest to measure the reliability of ratings averaged across the set of all workers. As defined by Shrout and Fleiss [60], these reliability measures may be formulated as $\mathrm{ICC}^w(2,1)$ and $\mathrm{ICC}^w(2,\mathrm{J})$ as follows:

$$\mathrm{ICC}^w(2,1) = \frac{\sigma_\alpha^2}{\sigma_\alpha^2 + \sigma_\beta^2 + \sigma_y^2}, \quad \text{and} \quad \mathrm{ICC}^w(2,\mathrm{J}) = \frac{\mathrm{J}\sigma_\alpha^2}{\mathrm{J}\sigma_\alpha^2 + \sigma_\beta^2 + \sigma_y^2}. \tag{2}$$

Similarly, not all tasks may lead to an equal assessment of a worker's ability. Therefore, it is also of interest to understand the reliability of a single task ($\mathrm{ICC}^t(2,1)$) or the mean of the set of all tasks ($\mathrm{ICC}^t(2,\mathrm{I})$) for predicting a worker's effect, which can be calculated as:

$$\mathrm{ICC}^t(2,1) = \frac{\sigma_\beta^2}{\sigma_\alpha^2 + \sigma_\beta^2 + \sigma_y^2}, \quad \text{and} \quad \mathrm{ICC}^t(2,\mathrm{I}) = \frac{\mathrm{I}\sigma_\beta^2}{\sigma_\alpha^2 + \mathrm{I}\sigma_\beta^2 + \sigma_y^2}. \tag{3}$$

2.2. Random effects model for binary responses

When binary rating tasks are used, a random effects model for the responses may be formulated through a logistic link function:

$$\begin{aligned} \operatorname{logit}(\Pr[y_{ij}=1]) &= \mu + \alpha_i + \beta_j, \quad \alpha_i \sim \mathcal{N}(0,\sigma_\alpha^2) \\ \beta_j &\sim \mathcal{N}(0,\sigma_\beta^2), \\ y_{ij} &\sim \text{ Bernoulli } (\Pr[y_{ij}=1]), \quad i=1,\dots,I, \; j=1,\dots,J, \end{aligned} \tag{4}$$

where y_{ij} is the binary observed response from worker j to task i, and α_i and β_j are the main effects for task i and the worker j, respectively. This model requires a total of $I+J+2$ parameters to be estimated, from the I tasks, J workers, and two variance components: σ_α and σ_β.

As above, we can obtain an estimate of a worker's reliability on a given task from model (4). Fleiss' kappa (κ) coefficient [18], which is a generalization of Scott's pi statistic [57] and Cohen's kappa [11,12] to more than two workers or tasks, calculates the inter-class reliability above and beyond that which would be expected by chance. As with the ICC for continuous responses, the κ coefficient ranges from 0 to 1, with higher values indicating higher levels of reliability. The κ coefficient for binary responses is defined as follows:

$$\kappa^w = \frac{p_a^w - p_s^w}{1 - p_s^w} \quad \text{and} \quad \kappa^t = \frac{p_a^t - p_s^t}{1 - p_s^t}, \tag{5}$$

where

$$p_a^w = \frac{1}{\mathrm{IJ}(\mathrm{J}-1)}\left(\sum_{i=1}^{\mathrm{I}}\sum_{k=0}^{1} x_{ik}^2 - \mathrm{IJ}\right), \quad p_s^w = \sum_{k=0}^{1}\left(\frac{1}{\mathrm{IJ}}\sum_{i=1}^{\mathrm{I}} x_{ik}\right)^2,$$

$$p_a^t = \frac{1}{\mathrm{IJ}(\mathrm{I}-1)}\left(\sum_{j=1}^{\mathrm{J}}\sum_{k=0}^{1} z_{jk}^2 - \mathrm{IJ}\right), \quad \text{and} \quad p_s^t = \sum_{k=0}^{1}\left(\frac{1}{\mathrm{IJ}}\sum_{j=1}^{\mathrm{J}} z_{jk}\right)^2,$$

x_{ik} is the number of workers that assign category k ($k=0,1$) to task i, and z_{jk} is the number of tasks that worker j assigns as a category k ($k=0,1$).

3. Simulation study

Random effects models provide a natural framework to partition the variability in observed responses to task-level factors and worker-level factors. In practice, researchers must collect ratings from a sufficiently large number of workers on a sufficiently large number of tasks to ensure that parameter estimates obtained from these models are robust. However, there currently do not exist guidelines for the number of workers or tasks needed to obtain robust parameter estimates, and how these numbers may vary as a function of worker and task reliability. The following simulation study presents a comparison of model fit for both continuous and binary response data, and provides guidelines for the numbers of workers and tasks required to obtain robust estimates of the model parameters.

3.1. Simulation design

The utility of random effects models are compared using a four-way factorial simulation design, resulting in a total of 24 simulation scenarios. The response mechanism was simulated to have either continuous (VAS) responses or binary responses. The values of μ, σ_α, σ_β, and for the continuous response conditions, σ_y, were chosen so that the reliability varied from low, to medium, to high, controlling separately for the reliability of tasks and the reliability of workers.

For each combination of the first three factors, either the number of workers or tasks was fixed at one of four values (10, 25, 50, 100) and the other was varied from 1 to 80. The factor that was varied, which we deem the focus of the simulation scenario, was then used to determine the minimum number of either tasks or workers required. For each combination of the simulation factors, 100 replications were generated, resulting in a total of 768,000 simulated datasets. A full list of parameters used may be found in Table 1.

3.2. Simulation statistics

Each simulated dataset consisted of an $n \times m$ matrix $Y = (y_{ij})$ where the n rows represent tasks and the m columns represent workers. For each simulated dataset, the corresponding random effects model was fit, and, given that the true value of the parameters is known, the mean squared error (MSE) for each parameter of the model was estimated by calculating the mean over all replicates of the squared difference between the estimator and the true value. However, the MSE was expected to generally decrease as the number of workers or tasks (r), depending on the focus of the scenario, increased from 1 to 80. Thus, a utility function based on cost was specified with the aim of selecting the minimum number of workers or tasks required. Specifically, this function balances the accuracy of the estimates (through the MSE) with a user-defined cost of prolonging the study, and its minimum provides the optimal number of workers or tasks required to estimate the model parameters. In paid crowdsourcing tasks, such as those through AMT, this cost may include the expected increases in fees, as well as the additional time needed to collect a larger amount of data.

As recommended in [50], we employed an exponential utility function so that a unique minimum could be identified. Specifically, we defined the utility function to be:

$$u(\mathrm{r}) = -\Delta_{\mathrm{MSE}} + (\exp(\omega \mathrm{r}) - 1), \qquad (6)$$

where $\exp(\omega \mathrm{r}) - 1$ is the penalty associated with increased costs, r is the number of workers or tasks, and ω is an imposed weight representing the cost of additional units. The quantity Δ_{MSE} is the difference in average MSE when r is increased by one unit, i.e. $\Delta_{\mathrm{MSE}} = \mathrm{MSE}_r - \mathrm{MSE}_{r-1}$ for the fit of the random effect model. Thus, the utility function has a similar structure as an information criterion measure, i.e. accuracy of the model (in this case, Δ_{MSE}) plus a penalty term (in this case, $\exp(\omega \mathrm{r}) - 1$). The utility function

Table 1. Parameters used in the simulation study for the mixed models (1) and (4). Scenarios labeled through a four letter code. The first letter indicates whether the scenario is focused on calculating the number of workers (w) or tasks (t), the second letter is the type of response—VAS (v) or binary (b), the third letter indicates if the reliability is with regard to a single worker (w), i.e. ICC^w(2,1) and κ^w, or task (t), i.e. ICC^t(2,1) and κ^t, and the last letter indicates the level of reliability—low (l), medium (m), and high (h).

			Parameters				
Scen.	Focus	Response	μ	σ_α	σ_β	σ_y	Reliability
wvtl	Worker	VAS	0.593	0.267	0.098	0.224	0.074
wvtm			0.250	0.025	0.100	0.075	0.615
wvth			0.250	0.010	0.100	0.010	0.980
wvwl			0.650	0.150	0.350	0.050	0.153
wvwm			0.593	0.267	0.098	0.224	0.544
wvwh			0.250	0.150	0.010	0.050	0.896
wbtl		Binary	−1.043	2.827	1.055		0.054
wbtm			−1.043	0.925	2.025		0.454
wbth			−1.043	0.425	3.675		0.795
wbwl			−1.043	1.025	0.825		0.131
wbwm			−1.043	2.827	1.055		0.408
wbwh			−1.043	4.325	0.150		0.811
tvtl	Task	VAS	0.593	0.267	0.098	0.224	0.074
tvtm			0.250	0.025	0.100	0.075	0.615
tvth			0.250	0.010	0.100	0.010	0.980
tvwl			0.650	0.150	0.350	0.050	0.153
tvwm			0.593	0.267	0.098	0.224	0.544
tvwh			0.250	0.150	0.010	0.050	0.896
tbtl		Binary	−1.043	2.827	1.055		0.054
tbtm			−1.043	0.925	2.025		0.454
tbth			−1.043	0.425	3.675		0.795
tbwl			−1.043	1.025	0.825		0.131
tbwm			−1.043	2.827	1.055		0.408
tbwh			−1.043	4.325	0.150		0.811

depends on the incremental change in MSE when a new rater or task is included in the study, rather than the value of the MSE itself, in order to compare the gain in goodness-of-fit with the cost of including a new rater or task.

3.3. Simulation results

The simulation study provides insight into the adequacy of model fit, as well as the optimal number of workers and/or tasks required to provide accurate parameter estimates under various conditions. For the simulated data, the random effects model was able to adequately recover the true parameter values. A summary of the MSE across all 24 simulation scenarios is provided in Appendix A.1 (Tables A1 to A4). The average MSE values over the scenarios focusing on the number of workers ranged from 0.0035 ($n = 10$) to 0.0011 ($n = 100$) for VAS response data, and from 0.3341 ($n = 10$) to 0.0684 ($n = 100$) for binary response data indicating good model fit. The average MSE values over the scenarios focusing on the number of tasks were generally higher than for the scenarios focusing on the number of workers, ranging from 0.0072 ($n = 10$) to 0.0029 ($n = 100$) for VAS response data, and from 0.4003 ($n = 10$) to 0.0573 ($n = 100$) for binary response data.

In general, the MSE was higher for binary response data than for the corresponding scenario with VAS response data. This difference was largest in the scenarios where the responses were generated to have low reliability and for small sample sizes ($n = 10$, $m = 10$). This difference may be due to a loss of information from the binary categorization process of the underlying continuous trait.

Figures 1 shows the empirical MSE for VAS response data (scenarios *wvtl*, *wvtm* and *wvth*) and binary response data (scenarios *wbtl*, *wbtm* and *wbth*) when varying the number m of workers over different reliability levels with regard to a single task. Similarly, Figure 2 shows the empirical MSE for VAS response data (scenarios *tvwl*, *tvwm* and *tvwh*) and binary response data (scenarios *tbwl*, *tbwm* and *tbwh*) when varying the number n of tasks over different reliability levels with regard to a single worker. As expected, the MSE generally decreased as the number of workers, m, the number of tasks, n, or the reliability increased.

To obtain the minimum number of workers or tasks, a utility convex function, as described above, was optimized. Figure 3 provides an illustrative example of the utility function for scenarios *wvtl* (left graph), *wbwh* (middle graph), and *wbwh* (right graph) at different number of workers and tasks. The results obtained depended on the exact specification of the weight in the utility function We tested the impact of weight in a range from $\omega \in (0, 2)$, from no influence of the cost ($\omega = 0$) to high influence ($\omega = 2$, which results in an increase of the effect of the penalty by a $\exp(2r)$ factor for each unit increase of a task or a worker). We also note here that we intentionally chose positive weights because they represent the cost, which are always positive. The impact of setting different weights did not lead to largely different results. In this simulation study and the application below,

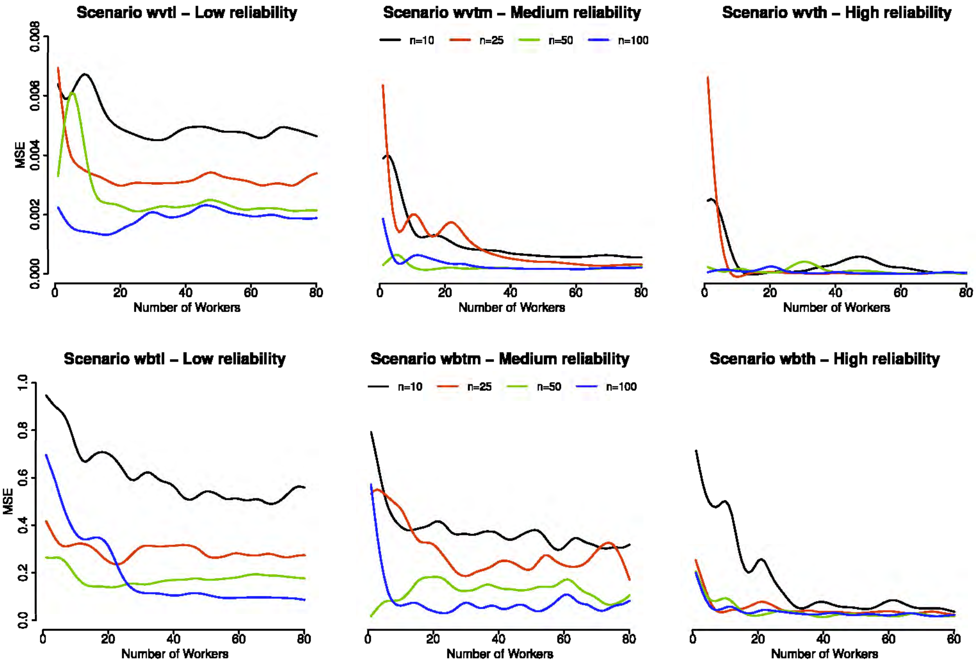

Figure 1. Simulation study. Number of workers: MSE of the worker effect estimator $\{\beta_j\}$ as a function of the number of workers at different single-task reliability levels for VAS responses (top) and binary responses (bottom) across 10, 25, 50 and 100 tasks.

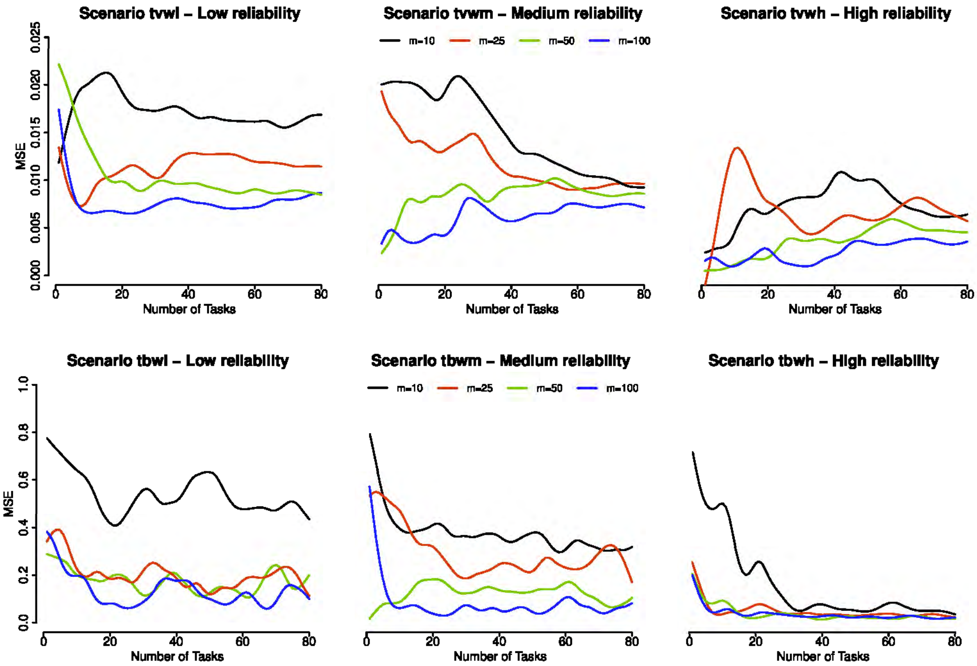

Figure 2. Simulation study. Number of tasks: MSE of the task effect estimator $\{\alpha_i\}$ as a function of the number of tasks at different single-worker reliability levels for VAS responses (top) and binary responses (bottom) across 10, 25, 50 and 100 workers.

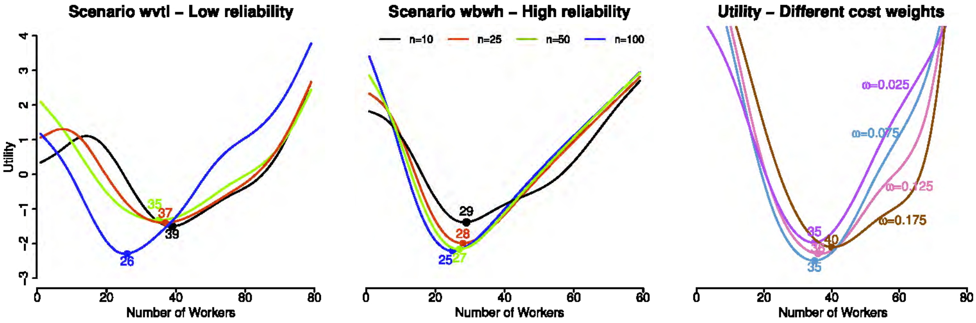

Figure 3. Utility function: Utility functions for scenarios *wvtl* (left graph) and *wbwh* (middle graph) at different number of workers *n*. Utility functions at $n = 50$ across different cost weights ω for scenario *wbwh* (right graph).

we set $\omega = 0.075$, based on this preliminary simulation analysis. However, ω (and also the utility function used) could be set depending on the requirements of any particular study.

Table 2 shows the results for the number of workers or tasks required across the 24 simulation scenarios. The number of workers required depends on the response mechanism, the number of tasks assigned, and the reliability of each worker and each task. For VAS response data, the number of workers required ranged from 8 to 44 with a median of 30.5. For binary response data, the number of workers required ranged from 11 to 43 with a

Table 2. Number of workers and tasks required for VAS and binary scales. The first half of the table shows the number of workers m at different number of tasks n. The second half shows the number of tasks n at different number of workers m.

			Number of Workers Required (m)			
Scenario	Focus	Response	$n = 10$	$n = 25$	$n = 50$	$n = 100$
wvtl	Worker	VAS	39	37	35	26
wvtm			32	30	21	13
wvth			19	15	9	8
wvwl			44	39	38	35
wvwm			34	33	32	31
wvwh			31	29	26	25
wbtl		Binary	41	39	37	30
wbtm			36	32	24	15
wbth			22	18	15	11
wbwl			43	40	39	37
wbwm			36	34	33	33
wbwh			29	28	27	25
			Number of Tasks Required (n)			
Scenario	Focus	Response	$m = 10$	$m = 25$	$m = 50$	$m = 100$
tvtl	Task	VAS	67	59	53	57
tvtm			59	53	40	33
tvth			28	23	21	18
tvwl			62	58	52	47
tvwm			58	55	52	45
tvwh			39	35	31	28
tbtl		Binary	66	55	54	52
tbtm			61	51	38	32
tbth			31	25	23	20
tbwl			65	60	57	50
tbwm			61	58	49	41
tbwh			36	32	27	25

median of 32.5. Similarly, the number of tasks required depends on the response mechanism, the number of workers hired, and the reliability of each worker and each task. For VAS response data, the number of tasks required ranged from 18 to 62 with a median of 42.5. For binary response data, the number of tasks required ranged from 20 to 66 with a median of 49.5.

4. Application: /r/ misarticulation in children

Speech sound disorders in childhood can pose a barrier to participation in social and academic activities [29], which may have negative ramifications that can persist throughout their lifespan [43]. Developmental speech errors typically resolve by the time children reach eight or nine years of age, but errors persist past this point in a subset of children [59]. One of the most common residual errors is misarticulation of the North American English rhotic /r/ [55]. These persisting errors pose a particular challenge for speech-language pathologists, who have called for novel and improved treatment methods for use with this population. In order to arrive at improved treatment methods, it is essential to be able to obtain precise measurements documenting changes in children's productions of /r/ over time.

Recent research has shown promising results from the use of crowdsourcing to collect non-expert ratings for the study of children's misarticulated /r/ sounds (see e.g. [26,40,41]), in particular when the non-expert listeners were recruited through Amazon Mechanical Turk (AMT). These studies have investigated the utility of both continuous and binary response mechanisms, and the subsequent gradient measures that may be derived from the ratings obtained. In order to illustrate the application of the model-based approaches introduced in Section 2, a crowdsourced data set was analyzed containing $n = 23{,}280$ ratings collected on $I = 40$ speech tokens (tasks) consisting of single words produced by 12 children at varying stages in the process of remediation for misarticulation of the North American English rhotic /r/. The ratings were obtained from $J = 291$ AMT non-expert raters (workers) who each rated the stimulus set twice, once using VAS and once using a binary response mechanism. The VAS mechanism consisted of a continuous line anchored on one side with the label 'correct /r/' and on the other with 'incorrect /r/'. The binary task used two buttons, with the same labels that anchored the VAS task, resulting in a forced choice of one category over another. Data collection was completed in 21.4 hours and cost $722, including Amazon fees. Originally, a total of $J = 726$ workers were recruited, but 287 participants did not meet all demographic criteria (e.g. native speaker of American English), 136 participants did not exceed chance-level performance on attentional catch trials, and 12 participants had missing or otherwise unusable data. The final set of $J = 291$ participants had a mean age of 32.4 years, with a standard deviation of 9.8 years.

This dataset was previously analyzed to compare the results of binary versus VAS rating scales when aggregating responses over a large number of non-expert listeners recruited via crowdsourcing [41]. The previous study found that both VAS and binary response mechanisms provided valid gradient measures, but showed high levels of variability in the response styles and levels of performance of the AMT workers. However, this study did not account for these differences across workers in the computation of the token estimates. Therefore, the present study extends the methodology previously used from simple means to random effect models.

The random effect model formulated in Equation (1) was fit to the 11,640 ratings obtained through the VAS response mechanism. The model was fit using restricted maximum likelihood [28,48,54,66] to avoid bias on the variance components estimates and was numerically optimized through a penalized iteratively reweighted least squares algorithm [2]. Similarly, the non-nested logistic multilevel regression model formulated in Equation (4) was fit to the 11,640 ratings obtained through the binary response mechanism via the numerical optimization of the likelihood function based on the Laplace approximation [36,61]. All statistical models were fit using the `lme4` package [3] in `R` version 3.2.3 [51]. A glossary of the commands used is provided in Appendix 2.

An acoustic measure of rhoticity, F3-F2 distance, was also collected on the 40 speech tokens. Previous research has reported that the North American English rhotic /r/ can be distinguished from other sonorant phonemes by the low height of the third formant (F3) [16,24] and a relatively high second formant (F2) [16,19], which brings F2 and F3 particularly close together [6,14]. Thus, the difference between these formants, F3-F2, is a commonly used acoustic measure of /r/ quality or rhoticity, with lower values indicating a higher degree of rhoticity [19]. Previous studies investigating the utility of crowdsourced ratings collected via AMT, including [26,40,41], have shown that both the mean VAS click location across listeners and the proportion of listeners marking a token as correct are highly correlated with F3-F2 distance. Thus, F3-F2 distance in Hz may be considered as a gold standard for the purpose of this example. Therefore, this measure was used to determine the appropriateness of the model fit by calculating the Spearman's rank correlation coefficient between the token-level estimates produced by the model and the gold standard, F3-F2 distance.

Token effect estimates derived from ratings obtained from both the VAS and binary response mechanisms are presented in Figure 4. Figure 4 shows the relationship between the token effects estimated from the VAS and binary models respectively, and the acoustic measure, F3-F2. In both cases, the Spearman's rank correlation coefficient, ρ, indicates a strong correlation between the crowdsourced and acoustic measures ($\rho = 0.81, 0.80$). The high values of these correlation coefficients indicate that the estimates produced from the random effects model are strongly related to the gold-standard, thus, confirming model fit. Furthermore, these correlations are slightly higher than those previously found in McAllister Byun *et al.* [41], indicating that the more sophisticated approach to modeling used in this study may have improved the validity of the token effects obtained. Similarly, the relationship between the token effect estimates derived from ratings obtained from the VAS and binary response mechanisms are compared in rightmost subfigure of Figure 4, indicating near perfect agreement ($\rho = 0.98$).

Results from the model fit to the VAS ratings showed high levels of variation among the speech tokens ($\hat{\sigma}_\alpha = 0.267$), and low levels of variation among the raters ($\hat{\sigma}_\beta = 0.098$) with a residual standard deviation of $\hat{\sigma}_y = 0.224$. Thus, the reliability of ratings obtained from a single worker was moderate ($\mathrm{ICC}^{\mathrm{w}}(2,1) = 0.544$), but when aggregated over the set of all workers, the reliability increased to $\mathrm{ICC}^{\mathrm{w}}(2,291) = 0.997$. Similarly, the reliability of a single task for predicting a given worker effect was extremely low ($\mathrm{ICC}^{\mathrm{t}}(2,1) = 0.074$), but increased to $\mathrm{ICC}^{\mathrm{t}}(2,40) = 0.762$ when the mean of all the tokens was used. In order to check the robustness of the model fit, we note that the parameter estimates obtained in this example were intentionally chosen as one of the simulated scenarios (*wvtl*). As shown in Table A1 in Appendix A.1, the simulation study shows proper recovery of the true parameter values.

Similarly, the model fit to the binary ratings resulted in estimated variance components of $\hat{\sigma}_\alpha^{\mathrm{logit}} = 2.827$, and $\hat{\sigma}_\beta^{\mathrm{logit}} = 1.055$ on the logit scale, with corresponding standard deviations of $\hat{\sigma}_\alpha = 0.707$ and $\hat{\sigma}_\beta = 0.264$ on the probability scale for the speech tokens and

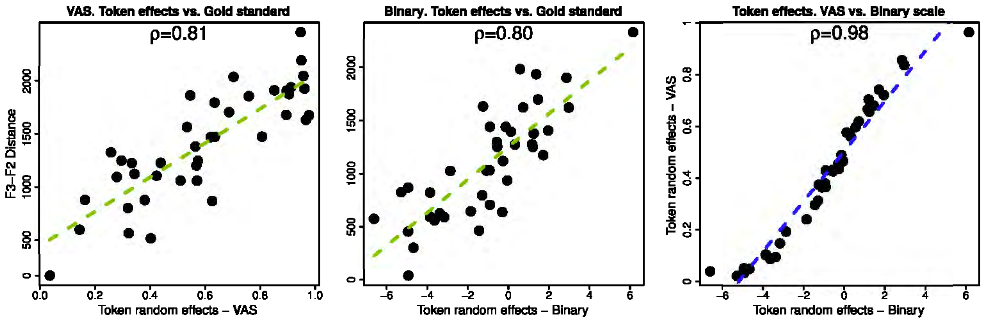

Figure 4. Misarticulation of Phoneme /r/ in Children: Correlation between between token effects in VAS model (Equation (1)) and F3-F2 distance (left graph), token effects in binary model (Equation (4)) and F3-F2 distance (middle graph), and token effects in VAS model and binary model (right graph).

raters, respectively [20, Section 5.2]. The reliability of the ratings obtained from a single worker and a single task are $\kappa^w = 0.408$ and $\kappa^t = 0.054$. As with the VAS ratings, the model fit in this example can be confirmed through the simulation study within scenario *wbtl.*

To assess the appropriateness of the models fit to the data, the conditional R^2 statistic was calculated. Although there are several approaches to calculated R^2 statistics for random effects models (see e.g. [10,17,37,46]), recently Nakagawa (2013) [45] derived an easily interpretable conditional R^2 (R^2_c) measure, which describes the proportion of variance explained by fixed and random factors. This measure is 'less susceptible to the common problems that plague alternative measures of R^2' [45]. The R^2_c statistic was calculated using the MuMIn package [1]. The values of the R^2_c for the VAS and binary models were 0.618 and 0.735, respectively, which show that the binary model fit better the data than the VAS model, and but both models adequately fit the data at hand.

Lastly, to assess the minimum number of non-expert raters required to achieve optimal performance, a bootstrap analysis was conducted varying the number of raters from $m=3$ to $m=80$. For each level of m, 1000 bootstrap resamples of the AMT raters were drawn and a random effects model was fit. From each model fit, Spearman's rank correlation was calculated between the token random effects and the acoustic measure, F3-F2 distance. Figure 5 shows the mean correlation obtained, with a 95% empirical pointwise confidence intervals, for both VAS and binary ratings, along with the results obtained from the full sample of raters. For VAS ratings, the correlation obtained from the full set of raters ($\rho = 0.81$) falls within the 95% bootstrapped CI when the number of workers exceeds 34. For binary ratings, the correlation obtained from the full set of raters ($\rho = 0.80$) falls within the 95% bootstrapped CI when the number of workers exceeds 36. These values are roughly equivalent to those obtained from the corresponding simulation scenarios when using the utility function approach (*wvtl* and *wbtl*, respectively). However, these recommendations are significantly larger than the recommendations of $m=9$ found in the literature [26,40]. This can be attributed to differing standards for equivalent results, or in other words, a different measure of how costly additional workers are. For practicality in clinical application,

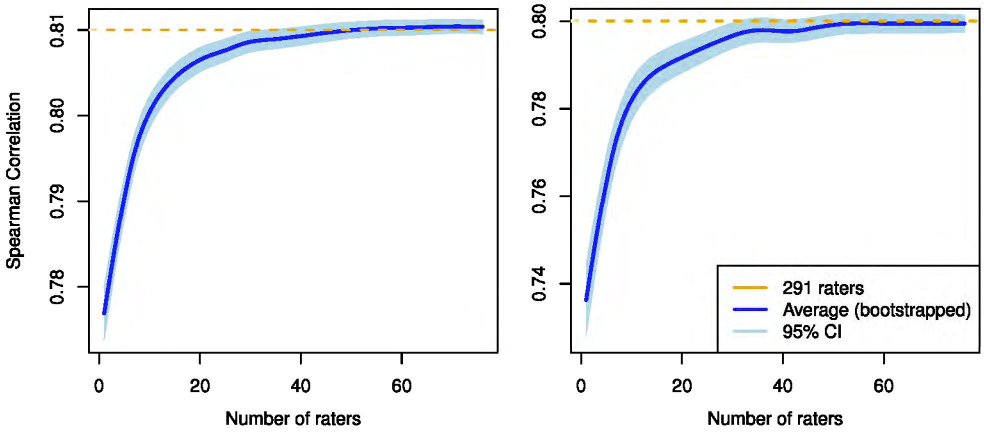

Figure 5. Bootstrap analysis: Mean Spearman's correlation and 95% CI between F3–F2 distance and token effects in VAS model (Equation (1)) (left graph), and token effects in binary model (Equation (4)) (right graph) over 1000 bootstrap resamples, for $m = 1$ to $m = 80$ workers. Horizontal line indicates value obtained from the full set of $m = 291$ workers.

these previous studies aimed to find the absolute minimum number of workers required to obtain results roughly equivalent to norms found in the existing literature. Therefore, it can be concluded that the guidelines presented in this paper can be adjusted if any expected loss in performance is within an acceptable tolerance for the application in question.

5. Discussion

Crowdsourcing allows researchers to collect data more efficiently than traditional laboratory settings. By capitalizing on crowdsourcing platforms such as AMT, researchers can draw workers from anywhere in the world to contribute ratings to micro-tasks. However, the varying quality and attentiveness of these workers, as well as the different skill sets required to complete each task, result in noisy data that must be summarized to obtain estimates of the parameters of interest. Random effects models provide a simple and natural solution that accounts for the different potential sources of variability while providing estimates of task- and worker-level parameters. This paper explored some statistical considerations when applying random effects models to crowdsourced datasets.

Section 3 presented a simulation study that assessed the performance of random effects models in estimating both worker quality and task accuracy in a variety of circumstances. These simulations yield preliminary guidelines for the numbers of workers and tasks required to ensure adequate parameter recovery under differing levels of reliability. The number of workers required to obtain optimal parameter estimates decreased as the number of tasks assigned increased, and as the reliability of the worker or task increased. Similarly, the number of tasks required to obtain optimal parameter estimates decreased as the number or reliability of the workers increased. Furthermore, in comparable scenarios, ratings obtained from binary response mechanisms showed higher levels of uncertainty than ratings obtained from VAS response mechanisms. This indicates that, in this case, binary response mechanisms require a larger number of workers and/or tasks to obtain equivalent results.

The exact results of the simulation study depended on the application of an exponential utility function. This family of utility functions is well-known and provides an easy interpretation for practitioners because it creates a trade-off between cost (penalty) and accuracy (Δ_{MSE}). However, researchers may choose different utility functions that better suit their needs, such as power utility functions [68]. The choice of utility function will depend on the specific features of a crowdsourcing experiment, e.g. whether each new worker or task is very costly. The application of alternative utility functions and an assessment of the implications of each choice is left beyond the scope of this paper and may be investigated as future work.

Section 4 showed how random effects models may be applied to real data sets, and Appendix 2 provides a description of the commands in R that can be used to reproduce our results. While other research has focused on achieving optimal efficiency, the algorithms proposed are often complicated and fall beyond the scope of what quantitative researchers who lack specific programing or statistical knowledge may be able to implement independently. Since one goal of crowdsourcing is to empower researchers to run experiments or collect data without the hurdles that limit traditional methods, random effects models may provide a framework that is both adequate and simple to use. Furthermore, the models implemented in this paper may be easily extended to answer other questions that are important to researchers by including worker-level or task-level covariates, such as the age or gender of the worker or the speaker producing each task, or allowing for heteroscedasticity in the variance of the random effects to estimate different levels of Gaussian noise for each worker or task.

Crowdsourcing provides a novel opportunity for researchers to collect data in an efficient manner. However, the statistical community has given relatively little attention to potential problems that may arise when crowdsourcing is used. This paper provides some statistical considerations regarding the numbers of workers and tasks that are required, and the types of models that may be appropriate for analyzing crowdsourced data.

Disclosure statement

No potential conflict of interest was reported by the authors.

Funding

This work was supported by the Moore-Sloan Data Sciences Environment grant from New York University, the New York University Research Challenge Fund Program, the National Institute of Health (R03DC 012883), and by the Marsden grant number E2987-3648 from the Royal Society of New Zealand.

ORCID

Daniel Fernández http://orcid.org/0000-0003-0012-2094
Daphna Harel http://orcid.org/0000-0001-7015-5989

References

[1] K. Bartoń, *Mumin: Multi-model inference. r package version 1.9.13*, The Comprehensive R Archive Network (CRAN), Vienna, Austria (2013).
[2] D. Bates, *Linear mixed model implementation in lme4*, Manuscript, University of Wisconsin 15 (2007).
[3] D. Bates, M. Mächler, B. Bolker, and S. Walker, *Fitting linear mixed-effects models using lme4*, J. Stat. Softw. 67 (2015), pp. 1–48.
[4] T.S. Behrend, D.J. Sharek, A.W. Meade, and E.N. Wiebe, *The viability of crowdsourcing for survey research*, Behav. Res. Methods 43 (2011), pp. 800–813.
[5] A.J. Berinsky, G.A. Huber, and G.S. Lenz, *Evaluating online labor markets for experimental research: Amazon.com's mechanical turk*, Polit. Anal. 20 (2012), pp. 351–368.
[6] S. Boyce and C.Y. Espy-Wilson, *Coarticulatory stability in american english/r*, J. Acoust. Soc. Am. 101 (1997), pp. 3741–3753.
[7] D.C. Brabham, *Crowdsourcing as a model for problem solving an introduction and cases*, Convergence 14 (2008), pp. 75–90.
[8] M. Buhrmester, T. Kwang, and S.D. Gosling, *Amazon's mechanical turk a new source of inexpensive, yet high-quality, data?* Perspect. Psychol. Sci. 6 (2011), pp. 3–5.
[9] J. Chandler, P. Mueller, and G. Paolacci, *Nonna"ıveté among amazon mechanical turk workers: Consequences and solutions for behavioral researchers*, Behav. Res. Methods. 46 (2014), pp. 112–130.
[10] J. Cheng, L.J. Edwards, M.M. Maldonado-Molina, K.A. Komro, and K.E. Muller, *Real longitudinal data analysis for real people: Building a good enough mixed model*, Stat. Med. 29 (2010), pp. 504–520.
[11] J. Cohen, *A coefficient of agreement for nominal scale*, Educ. Psychol. Meas. 20 (1960), pp. 37–46.

[12] J. Cohen, *Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit*, Psychol. Bull. 70 (1968), p. 213.

[13] M.J. Crump, J.V. McDonnell, and T.M. Gureckis, *Evaluating Amazon's mechanical turk as a tool for experimental behavioral research*, PLoS One 8 (2013), p. e57410. Available at http://dx.doi.org/10.1371%2Fjournal.pone.0057410.

[14] R.M. Dalston, *Acoustic characteristics of english/w, r, l/spoken correctly by young children and adults*, J. Acoust. Soc. Am. 57 (1975), pp. 462–469.

[15] A.P. Dawid and A.M. Skene, *Maximum likelihood estimation of observer error-rates using the EM algorithm*, Appl. Stat. 1 (1979), pp. 20–28.

[16] P. Delattre and D.C. Freeman, *A dialect study of american r's by x-ray motion picture*, Linguistics 6 (1968), pp. 29–68.

[17] L.J. Edwards, K.E. Muller, R.D. Wolfinger, B.F. Qaqish, and O. Schabenberger, *An r2 statistic for fixed effects in the linear mixed model*, Stat. Med. 27 (2008), pp. 6137–6157.

[18] J.L. Fleiss, *Measuring nominal scale agreement among many raters.*, Psychol. Bull. 76 (1971), p. 378.

[19] P. Flipsen, L.D. Shriberg, G. Weismer, H.B. Karlsson, and J.L. McSweeny, *Acoustic phenotypes for speech–genetics studies: Reference data for residual /ɜ/distortions*, Clin. Linguist. Phon. 15 (2001), pp. 603–630.

[20] A. Gelman and J. Hill, *Data Analysis using Regression and Multilevel/hierarchical Models*, Cambridge University Press, New York, 2006.

[21] A. Ghose, P.G. Ipeirotis, and B. Li, *Designing ranking systems for hotels on travel search engines by mining user-generated and crowdsourced content*, Marketing Sci. 31 (2012), pp. 493–520.

[22] J.K. Goodman, C.E. Cryder, and A. Cheema, *Data collection in a flat world: The strengths and weaknesses of mechanical turk samples*, J. Behav. Decis. Mak. 26 (2013), pp. 213–224. Available at http://dx.doi.org/10.1002/bdm.1753.

[23] G. Graham, J. Cox, B. Simmons, C. Lintott, K. Masters, A. Greenhill, and K. Holmes, *How is success defined and measured in online citizen science: A case study of zooniverse projects*, Comput. Sci. Eng. 99 (2015), p. 22.

[24] R. Hagiwara, *Acoustic realizations of American/r/as produced by women and men*, Vol. 90, Phonetics Laboratory, Dept. of Linguistics, UCLA, 1995.

[25] P.R. Hahn, I. Goswami, and C.F. Mela, *A bayesian hierarchical model for inferring player strategy types in a number guessing game*, Ann. Appl. Stat. 9 (2015), pp. 1459–1483.

[26] D. Harel, E.R. Hitchcock, D. Szeredi, J. Ortiz, and T. McAllister Byun, *Finding the experts in the crowd: Validity and reliability of crowdsourced measures of children's gradient speech contrasts*, Clin. Linguist. Phon. 31 (2017), pp. 104–117.

[27] J.P. Harris, J.P. Anderson, and R. Novak, *An outcomes study of cochlear implants in deaf patients. Audiologic, economic, and quality-of-life changes*, Arch. Otolaryngol. Head Neck Surg. 121 (1995), pp. 398–404.

[28] D.A. Harville, *Maximum likelihood approaches to variance component estimation and to related problems*, J. Am. Stat. Assoc. 72 (1977), pp. 320–338.

[29] E.R. Hitchcock, D. Harel, and T. McAllister Byun, *Social, emotional, and academic impact of residual speech errors in school-aged children: A survey study*, Semin. Speech. Lang. 36 (2015), pp. 283–294.

[30] J.J. Horton, D.G. Rand, and R.J. Zeckhauser, *The online laboratory: Conducting experiments in a real labor market*, Exp. Econ. 14 (2011), pp. 399–425.

[31] J. Howe, *The rise of crowdsourcing*, Wired Mag. 14 (2006), pp. 1–4.

[32] P.G. Ipeirotis, *Demographics of mechanical turk* (2010). Available at http://hdl.handle.net/2451/29585.

[33] P.G. Ipeirotis, F. Provost, V.S. Sheng, and J. Wang, *Repeated labeling using multiple noisy labelers*, Data. Min. Knowl. Discov. 28 (2014), pp. 402–441.

[34] P.G. Ipeirotis, F. Provost, and J. Wang, *Quality management on amazon mechanical turk*, Proceedings of the ACM SIGKDD Workshop on Human Computation, ACM, Washington DC, DC, USA – July 25–28, 2010, pp. 64–67.

[35] K.L. Lansford, S.A. Borrie, and L. Bystricky, *Use of crowdsourcing to assess the ecological validity of perceptual-training paradigms in dysarthria*, Am. J. Speech Lang. Pathol. 25 (2016), pp. 1–7.

[36] Y. Lee and J.A. Nelder, *Hierarchical generalised linear models: A synthesis of generalised linear models, random-effect models and structured dispersions*, Biometrika 88 (2001), pp. 987–1006.
[37] H. Liu, Y. Zheng, and J. Shen, *Goodness-of-fit measures of r 2 for repeated measures mixed effect models*, J. Appl. Stat. 35 (2008), pp. 1081–1092.
[38] M.A. Luengo-Oroz, A. Arranz, and J. Frean, *Crowdsourcing malaria parasite quantification: An online game for analyzing images of infected thick blood smears*, J. Med. Internet. Res. 14 (2012), p. e167.
[39] E. Maas and K.A. Farinella, *Random versus blocked practice in treatment for childhood apraxia of speech*, J. Speech Lang. Hear. Res. 55 (2012), pp. 561–578.
[40] T. McAllister Byun, P.F. Halpin, and D. Szeredi, *Online crowdsourcing for efficient rating of speech: A validation study*, J. Commun. Disord. 53 (2015), pp. 70–83.
[41] T. McAllister Byun, D. Harel, P.F. Halpin, and D. Szeredi, *Deriving gradient measures of child speech from crowdsourced ratings*, J. Commun. Disord. 64 (2016), pp. 91–102.
[42] T. McAllister Byun and E.R. Hitchcock, *Investigating the use of traditional and spectral biofeedback approaches to intervention for/r/misarticulation*, Am. J. Speech Lang. Pathol. 21 (2012), pp. 207–221.
[43] J. McCormack, S. McLeod, L. McAllister, and L.J. Harrison, *A systematic review of the association between childhood speech impairment and participation across the lifespan*, Int. J. Speech. Lang. Pathol. 11 (2009), pp. 155–170.
[44] B. Munson, J.M. Johnson, and J. Edwards, *The role of experience in the perception of phonetic detail in children's speech: a comparison between speech-language pathologists and clinically untrained listeners*, Am. J. Speech Lang. Pathol. 21 (2012), pp. 124–139.
[45] S. Nakagawa and H. Schielzeth, *A general and simple method for obtaining r2 from generalized linear mixed-effects models*, Methods Ecol. Evol. 4 (2013), pp. 133–142.
[46] J.G. Orelien and L.J. Edwards, *Fixed-effect variable selection in linear mixed models using r2 statistics*, Comput. Stat. Data Anal. 52 (2008), pp. 1896–1907.
[47] G. Paolacci, J. Chandler, and P.G. Ipeirotis, *Running experiments on amazon mechanical turk*, Judgm. Decis. Mak. 5 (2010), pp. 411–419.
[48] H.D. Patterson and R. Thompson, *Recovery of inter-block information when block sizes are unequal*, Biometrika 58 (1971), pp. 545–554.
[49] C. Piech, J. Huang, Z. Chen, C. Do, A. Ng, and D. Koller, *Tuned models of peer assessment in moocs*, preprint (2013). Available at arXiv:1307.2579.
[50] J.W. Pratt, *Risk aversion in the small and in the large*, Econometrica 32 (1964), pp. 122–136.
[51] R Core Team, *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria (2015). Available at https://www.R-project.org/.
[52] V.C. Raykar, S. Yu, L.H. Zhao, A. Jerebko, C. Florin, G.H. Valadez, L. Bogoni, and L. Moy, *Supervised learning from multiple experts: Whom to trust when everyone lies a bit*, Proceedings of the 26th Annual International Conference on Machine Learning, Montreal, QC, ACM, 2009, pp. 889–896.
[53] V.C. Raykar, S. Yu, L.H. Zhao, G.H. Valadez, C. Florin, L. Bogoni, and L. Moy, *Learning from crowds*, J. Mach. Learn. Res. 11 (2010), pp. 1297–1322.
[54] G.K. Robinson, *That blup is a good thing: The estimation of random effects*, Stat. Sci. 6 (1991), pp. 15–32.
[55] D.M. Ruscello, *Visual feedback in treatment of residual phonological disorders*, J. Commun. Disord. 28 (1995), pp. 279–302.
[56] S.K. Schellinger, B. Munson, and J. Edwards, *Gradient perception of children's productions of /s/ and /?/: A comparative study of rating methods*, Clin. Linguist. Phon. (2016), pp. 1–24. Available at http://dx.doi.org/10.1080/02699206.2016.1205665, PMID: 27552446.
[57] W. Scott, *Reliability of content analysis: The case of nominal scale coding*, Public Opin. Q. 19 (1955), p. 321.
[58] V.S. Sheng, F. Provost, and P.G. Ipeirotis, *Get another label? Improving data quality and data mining using multiple, noisy labelers*, Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining, ACM, 2008, pp. 614–622.

[59] L.D. Shriberg, F.A. Gruber, and J. Kwiatkowski, *Developmental phonological disorders III: Long-term speech-sound normalization*, J. Speech Lang. Hear. Res. 37 (1994), pp. 1151–1177.
[60] P. Shrout and J. Fleiss, *Intraclass correlations: Uses in assessing rater reliability.*, Psychol. Bull. 86 (1979), p. 420.
[61] Z. Shun, *Another look at the salamander mating data: A modified laplace approximation approach*, J. Am. Stat. Assoc. 92 (1997), pp. 341–349.
[62] A. Sorokin and D. Forsyth, *Utility data annotation with amazon mechanical turk*, First IEEE Workshop on Internet Vision at CVPR'08, 2008, 51 (2008), p. 820.
[63] J. Sprouse, *A validation of amazon mechanical turk for the collection of acceptability judgments in linguistic theory*, Behav. Res. Methods 43 (2011), pp. 155–167.
[64] S. Suri and D.J. Watts, *Cooperation and contagion in web-based, networked public goods experiments*, PLoS One 6 (2011), p. e16836.
[65] M. Swan, *Crowdsourced health research studies: an important emerging complement to clinical trials in the public health research ecosystem*, J. Med. Internet. Res. 14 (2012).
[66] R. Thompson, *Maximum likelihood estimation of variance components*, Statistics 11 (1980), pp. 545–561.
[67] A.M. Turner, K. Kirchhoff, and D. Capurro, *Using crowdsourcing technology for testing multilingual public health promotion materials*, J. Med. Internet. Res. 14 (2012), p. e79.
[68] P.P. Wakker, *Explaining the characteristics of the power (crra) utility family*, Health. Econ. 17 (2008), pp. 1329–1344.
[69] J. Wang, P.G. Ipeirotis, and F. Provost, *Quality-based pricing for crowdsourced workers* (2013). NYUCBA Working Paper CBA-13-06. Available at http://hdl.handle.net/2451/31833.
[70] P. Welinder, S. Branson, P. Perona, and S.J. Belongie, *The multidimensional wisdom of crowds*, Advances in Neural Information Processing Systems, Vancouver, Canada, 2010, pp. 2424–2432.
[71] P. Welinder and P. Perona, *Online crowdsourcing: Rating annotators and obtaining cost-effective labels*. In Computer Vision and Pattern Recognition Workshops (CVPRW), 2010 IEEE Computer Society Conference (pp. 25–32). IEEE.
[72] J. Whitehill, T.F. Wu, J. Bergsma, J.R. Movellan, and P.L. Ruvolo, *Whose vote should count more: Optimal integration of labels from labelers of unknown expertise*, Advances in Neural Information Processing Systems, Vancouver, Canada, 2009, pp. 2035–2043.
[73] Y. Zhang, X. Chen, D. Zhou, and M.I. Jordan, *Spectral methods meet EM: A provably optimal algorithm for crowdsourcing*, Advances in Neural Information Processing Systems, Montreal, Canada, 2014, pp. 1260–1268.

Appendix 1. Results simulation study

A.1 VAS responses

The following tables provide detailed results from the simulation study in Section 3. Tables A1 and A2 provide the estimates and MSE for the parameters of the model formulated in (1) for VAS responses focused on number of workers and tasks, respectively.

A.2 Binary responses

The following tables provide detailed results from the simulation study in Section 3. Tables A3 and A4 provide the estimates and MSEf for the parameters of the model formulated in (4) for binary responses focused on number of workers and tasks, respectively.

Appendix 2. Glossary of commands in R

The following appendix provides the commands used to fit the models in Section 4. We fit the linear mixed-effects model (1) to the VAS rating data using the command `lmer` from the R package `lme4` [3] as follows:

```
fmVAS <- lmer(VAS ~ (1|token) + (1 | rater), data=VASdata)
```

where `VASdata` is the data set consisting of 40 columns representing the speech `token` and 291 rows representing the `rater`.

Similarly, we fit the generalized linear mixed-effects model (4) to the binary rating data using the command `glmer` from the same R package `lme4` as follows:

```
fmBIN <- glmer(Binary ~ (1|token) + (1 | rater),
                    family = binomial("logit"), data=BINdata)
```

Table A1. Simulation study: VAS responses. Workers: Estimates and MSE of the intercept μ and variance components σ_α, σ_β, and σ_y of the model formulated in (1). Scenarios shown are those from Table 1 focused on calculating the number of workers at different reliability levels (low, medium, and high) with regard to a single task (wvtl, wvtm, and wvth) or worker (wvwl, wvwm, and wvwh) for VAS responses. The results are averaged over all replicates.

	True	$n = 10$		$n = 25$		$n = 50$		$n = 100$	
Scen.	param.	Estim.	MSE	Estim.	MSE	Estim.	MSE	Estim.	MSE
wvtl	$\mu = 0.593$	0.559	0.005	0.559	0.003	0.555	0.002	0.573	0.001
	$\sigma_\alpha = 0.267$	0.279	0.010	0.279	0.009	0.275	0.008	0.273	0.008
	$\sigma_\beta = 0.098$	0.090	0.002	0.093	0.001	0.094	0.001	0.099	0.002
	$\sigma_y = 0.224$	0.222	0.002	0.222	0.002	0.222	0.002	0.221	0.002
wvtm	$\mu = 0.250$	0.269	0.001	0.273	0.001	0.254	0.000	0.251	0.000
	$\sigma_\alpha = 0.025$	0.022	0.000	0.023	0.000	0.023	0.000	0.024	0.000
	$\sigma_\beta = 0.100$	0.074	0.001	0.104	0.000	0.093	0.000	0.100	0.000
	$\sigma_y = 0.075$	0.074	0.000	0.073	0.000	0.073	0.000	0.073	0.000
wvth	$\mu = 0.250$	0.245	0.000	0.254	0.000	0.245	0.000	0.251	0.000
	$\sigma_\alpha = 0.010$	0.010	0.000	0.010	0.000	0.010	0.000	0.010	0.000
	$\sigma_\beta = 0.100$	0.098	0.000	0.091	0.000	0.098	0.000	0.099	0.000
	$\sigma_y = 0.010$	0.010	0.000	0.010	0.000	0.010	0.000	0.010	0.000
wvwl	$\mu = 0.650$	0.532	0.015	0.587	0.027	0.676	0.015	0.631	0.007
	$\sigma_\alpha = 0.150$	0.086	0.005	0.098	0.004	0.143	0.012	0.147	0.006
	$\sigma_\beta = 0.350$	0.264	0.036	0.337	0.047	0.335	0.032	0.355	0.014
	$\sigma_y = 0.050$	0.063	0.002	0.064	0.002	0.054	0.002	0.054	0.002
wvwm	$\mu = 0.593$	0.527	0.005	0.525	0.005	0.558	0.006	0.589	0.004
	$\sigma_\alpha = 0.267$	0.262	0.013	0.265	0.012	0.265	0.011	0.269	0.004
	$\sigma_\beta = 0.098$	0.062	0.002	0.073	0.001	0.088	0.001	0.092	0.002
	$\sigma_y = 0.224$	0.217	0.000	0.219	0.000	0.218	0.000	0.219	0.000
wvwh	$\mu = 0.250$	0.301	0.004	0.286	0.002	0.268	0.002	0.258	0.002
	$\sigma_\alpha = 0.150$	0.139	0.002	0.146	0.001	0.148	0.001	0.149	0.001
	$\sigma_\beta = 0.010$	0.011	0.000	0.010	0.000	0.010	0.000	0.010	0.000
	$\sigma_y = 0.050$	0.053	0.000	0.051	0.000	0.048	0.000	0.049	0.000

Table A2. Simulation study: VAS responses. Tasks: Estimates and MSE of the intercept μ and variance components σ_α, σ_β, andσ_y of the model formulated in (1). Scenarios shown are those from Table 1 focused on calculating the number of tasks at different reliability levels (low, medium, and high) with regard to a single task (tvtl, tvtm, and tvth) or worker (tvwl, tvwm, and tvwh) for VAS responses. The results are averaged over all replicates.

	True	$m = 10$		$m = 25$		$m = 50$		$m = 100$	
Scen.	param.	Estim.	MSE	Estim.	MSE	Estim.	MSE	Estim.	MSE
tvtl	$\mu = 0.593$	0.511	0.008	0.609	0.001	0.571	0.003	0.586	0.001
	$\sigma_\alpha = 0.267$	0.256	0.013	0.270	0.010	0.126	0.011	0.265	0.011
	$\sigma_\beta = 0.098$	0.093	0.002	0.090	0.002	0.095	0.001	0.096	0.001
	$\sigma_y = 0.224$	0.227	0.001	0.229	0.002	0.223	0.002	0.223	0.002
tvtm	$\mu = 0.250$	0.260	0.001	0.246	0.001	0.254	0.000	0.252	0.000
	$\sigma_\alpha = 0.025$	0.021	0.000	0.022	0.000	0.023	0.000	0.025	0.000
	$\sigma_\beta = 0.100$	0.091	0.001	0.093	0.000	0.094	0.000	0.096	0.000
	$\sigma_y = 0.075$	0.073	0.000	0.073	0.000	0.073	0.000	0.074	0.000
tvth	$\mu = 0.250$	0.253	0.001	0.250	0.000	0.249	0.000	0.248	0.000
	$\sigma_\alpha = 0.010$	0.009	0.000	0.011	0.000	0.008	0.000	0.009	0.000
	$\sigma_\beta = 0.100$	0.094	0.001	0.097	0.000	0.097	0.000	0.098	0.000
	$\sigma_y = 0.010$	0.010	0.000	0.010	0.000	0.010	0.000	0.010	0.000
tvwl	$\mu = 0.650$	0.632	0.015	0.659	0.008	0.654	0.009	0.653	0.009
	$\sigma_\alpha = 0.150$	0.175	0.006	0.162	0.008	0.150	0.009	0.149	0.006
	$\sigma_\beta = 0.350$	0.362	0.016	0.312	0.015	0.336	0.007	0.344	0.008
	$\sigma_y = 0.050$	0.039	0.028	0.044	0.018	0.046	0.019	0.046	0.017
tvwm	$\mu = 0.593$	0.563	0.004	0.574	0.005	0.589	0.004	0.596	0.003
	$\sigma_\alpha = 0.267$	0.225	0.020	0.294	0.018	0.262	0.015	0.269	0.012
	$\sigma_\beta = 0.098$	0.086	0.007	0.092	0.008	0.098	0.008	0.097	0.007
	$\sigma_y = 0.224$	0.214	0.000	0.217	0.000	0.228	0.000	0.225	0.000
tvwh	$\mu = 0.250$	0.290	0.007	0.289	0.007	0.256	0.003	0.247	0.002
	$\sigma_\alpha = 0.150$	0.122	0.001	0.135	0.002	0.143	0.003	0.153	0.001
	$\sigma_\beta = 0.010$	0.008	0.000	0.010	0.000	0.010	0.000	0.010	0.000
	$\sigma_y = 0.050$	0.050	0.000	0.050	0.000	0.050	0.000	0.050	0.000

Table A3. Simulation study: Binary responses. Workers: Estimates and MSE of the intercept μ and variance components σ_α and σ_β of the model formulated in (4). Scenarios shown are those from Table 1 focused on calculating the number of workers at different reliability levels (low, medium, and high) with regard to a single task (wbtl, wbtm, and wbth) or worker (wbwl, wbwm, and wbwh) for binary responses. The results are averaged over all replicates.

	True	$n = 10$		$n = 25$		$n = 50$		$n = 100$	
Scen.	param.	Estim.	MSE	Estim.	MSE	Estim.	MSE	Estim.	MSE
wbtl	$\mu = -1.043$	−1.014	0.379	−0.789	0.206	−1.108	0.095	−1.082	0.031
	$\sigma_\alpha = 2.827$	2.489	0.348	2.728	0.139	2.762	0.047	2.833	0.034
	$\sigma_\beta = 1.055$	0.999	0.042	1.018	0.012	0.999	0.008	1.058	0.003
wbtm	$\mu = -1.043$	−1.218	0.083	−1.213	0.047	-1.109	0.029	−1.029	0.009
	$\sigma_\alpha = 0.925$	0.852	0.058	0.886	0.019	0.895	0.008	0.933	0.005
	$\sigma_\beta = 2.025$	1.792	0.093	1.861	0.042	1.963	0.014	1.996	0.007
wbth	$\mu = -1.043$	−1.313	0.108	−0.747	0.098	−1.125	0.036	−1.064	0.018
	$\sigma_\alpha = 0.425$	0.552	0.058	0.486	0.019	0.465	0.008	0.433	0.005
	$\sigma_\beta = 3.675$	3.492	0.093	3.561	0.042	3.563	0.014	3.676	0.007
wbwl	$\mu = -1.043$	−1.376	0.378	−1.088	0.191	−0.919	0.022	−1.041	0.085
	$\sigma_\alpha = 1.025$	0.921	0.155	0.989	0.029	1.014	0.052	1.022	0.004
	$\sigma_\beta = 0.825$	0.967	0.053	0.931	0.025	0.832	0.003	0.915	0.021
wbwm	$\mu = -1.043$	−0.976	0.056	−1.001	0.026	−1.073	0.089	−0.962	0.009
	$\sigma_\alpha = 2.827$	2.576	0.101	2.734	0.094	2.793	0.059	2.811	0.041
	$\sigma_\beta = 1.055$	0.868	0.028	0.943	0.014	0.991	0.011	1.018	0.003
wbwh	$\mu = -1.043$	−1.272	0.133	−0.767	0.095	−1.141	0.022	−1.038	0.015
	$\sigma_\alpha = 4.325$	4.209	0.194	4.083	0.107	4.162	0.059	4.292	0.011
	$\sigma_\beta = 0.150$	0.228	0.108	0.208	0.023	0.166	0.010	0.152	0.004

Table A4. Simulation study binary responses. Tasks: Estimates and MSE of the intercept μ and variance components σ_α and σ_β of the model formulated in (4). Scenarios shown are those from Table 1 focused on calculating the number of workers at different reliability levels (low, medium, and high) with regard to a single task (tbtl, tbtm, and tbth) or worker (tbwl, tbwm, and tbwh) for binary responses. The results are averaged over all replicates.

	True	$m = 10$		$m = 25$		$m = 50$		$m = 100$	
Scen.	param.	Estim.	MSE	Estim.	MSE	Estim.	MSE	Estim.	MSE
tbtl	$\mu = -1.043$	−1.163	0.146	−0.973	0.042	−1.073	0.011	−1.038	0.008
	$\sigma_\alpha = 2.827$	2.502	0.202	2.698	0.116	2.724	0.052	2.821	0.037
	$\sigma_\beta = 1.055$	0.819	0.038	1.002	0.027	1.041	0.012	1.050	0.006
tbtm	$\mu = -1.043$	−0.964	0.168	−0.917	0.064	−1.071	0.054	−1.012	0.016
	$\sigma_\alpha = 0.925$	0.627	0.088	1.238	0.034	0.814	0.015	0.921	0.007
	$\sigma_\beta = 2.025$	2.333	0.091	2.199	0.044	2.062	0.021	2.031	0.004
tbth	$\mu = -1.043$	−1.444	0.305	−0.817	0.128	−0.982	0.072	−1.034	0.018
	$\sigma_\alpha = 0.425$	0.689	0.098	0.500	0.031	0.435	0.014	0.425	0.004
	$\sigma_\beta = 3.675$	3.583	0.122	3.689	0.073	3.681	0.021	3.676	0.004
tbwl	$\mu = -1.043$	−0.998	0.152	−1.012	0.098	−1.036	0.021	−1.042	0.040
	$\sigma_\alpha = 1.025$	0.991	0.088	1.001	0.054	1.019	0.015	1.027	0.012
	$\sigma_\beta = 0.825$	0.815	0.072	0.821	0.023	0.835	0.011	0.834	0.004
tbwm	$\mu = -1.043$	−0.823	0.211	−0.875	0.203	−1.213	0.059	−1.109	0.036
	$\sigma_\alpha = 2.827$	3.209	0.119	2.901	0.112	2.855	0.023	2.838	0.018
	$\sigma_\beta = 1.055$	0.834	0.092	0.918	0.078	1.102	0.023	1.067	0.007
tbwh	$\mu = -1.043$	−1.289	0.166	−1.111	0.138	−1.098	0.042	−1.029	0.031
	$\sigma_\alpha = 4.325$	4.407	0.086	4.399	0.074	4.335	0.015	4.318	0.017
	$\sigma_\beta = 0.150$	0.122	0.042	0.123	0.031	0.162	0.010	0.149	0.005