Citation: Jameson, K. A., Deshpande, P.S., Tauber, S., Chang, S. M.& Gago, S. (2016). Using individual differences to better determine normative responses from crowdsourced transcription tasks: An application to the R. E. MacLaury Color Categorization Archive. Invited Proceedings paper for the 2016 IS&T International Symposium on Electronic Imaging (EI 2016). Technical Session on Human Vision and Electronic Imaging.

Using individual differences to better determine normative responses from crowdsourced transcription tasks: An application to the R. E. MacLaury Color Categorization Archive

Kimberly A. Jameson; Institute for Mathematical Behavioral Sciences, University of California, Irvine; Irvine, California, USA
Prutha S. Deshpande; Cognitive Sciences, University of California, Irvine; Irvine, California, USA
Sean Tauber; Institute for Mathematical Behavioral Sciences, University of California, Irvine; Irvine, California, USA
Stephanie M. Chang; Calit2/Computer Science, University of California, Irvine; Irvine, California, USA
Sergio Gago; Calit2, University of California, Irvine; Irvine, California, USA

Abstract

Individual differences inherent in human perceptual and behavioral data pose challenges for researchers who aim to develop standardized models of phenomena and procedures for normative assessment. A common approach used when modeling individual variation is to adopt criteria for identifying and excluding the individual data of outliers. We present investigations that use an alternative approach for analyzing response variation, which makes use of individual differences in data, to define a robust process model of both response variation and the information shared by individuals in a group. Crowdsourced perceptual identification tasks and formal analysis methods – Cultural Consensus Theory (CCT) – are employed to evaluate participants' responses to transcription tasks, towards the aim of digitizing approximately 23,000 handwritten pages of an irreplaceable cross-cultural color categorization survey by Robert E. MacLaury. Preliminary results show (1) utility of several original crowdsourced tasks for database transcription, (2) the appropriateness of CCT as a formal model for aggregating transcription data, (3) novel ways of addressing "expertise" using CCT analyses, and (4) the accurate derivation of correct transcription "answer keys", suggesting the potential for CCT methods to contribute to accurate transcription results even in the presence of large individual differences in participants responses. Research presented suggests that crowdsourcing in conjunction with CCT considerably reduces, without loss of accuracy, the number of participants needed for expeditious transcription of large, handwritten, corpora.

Introduction

Identifying patterns of results across individuals, especially in cases where participants perceptual or behavioral data varies greatly, requires principled procedures be used to analyze group data and to develop normative models of behavior. One challenge such procedures must address is defining what constitutes individual variation (and individual "expertise") for particular phenomena, when very little may be known concerning what is typical for specific kinds tasks or judgments, or specific sets of questions. This paper describes novel use of data aggregation procedures originally developed for evaluating general knowledge and information domains across groups of individuals where the underlying correct answers were unknown [1]. We generalized "Cultural Consensus Theory" procedures to aggregate and present preliminary analyses of data from several perceptual identification tasks. New modeling and empirical results presented illustrate that individual response variation found in tested perceptual tasks is well-accommodated by the procedures, that individual differences producing response variation in the tasks enhances estimation of underlying patterns in group responses and are useful for deriving a normative response model for the domain, and parameters inherent in the analyses provide useful indices of (a) observer "expertise" and individual variation from an underlying shared "truth", (b) "correct answer" estimation, and (c) shared group response coherence or consensus.

Our investigations address a specific challenge related to archival handwritten-document processing, namely, how to analyze data collected from crowdsourced tasks which aim to convert a large corpus of handwritten material, recorded using different writing styles, into a digitally addressable database. The corpus at issue here is the 23,000 page Robert E. MacLaury color categorization archive – an anthropological color cognition survey that includes data from 116 indigenous languages from mesoamerica, plus 70 languages sampled from major continents worldwide ([2]; http://colcat.calit2.uci.edu).

A description of the archive's extensive features is provided elsewhere [3], but based on similar examples (see [4]) an archive of this size and complexity could require at least 12 years to fully transcribe (if daily an individual expert converted 5 data pages). Also, because the archive's data is (i) recorded using at least 142 different cursive/printing styles, often using different alphabet scripts, and (ii) is largely of degraded optical quality, the development of a general machine-learning approach to transcribe the archive remains a challenge (although preliminary advances [5] suggest optical character recognition might provide a portion of a transcription solution when combined with methods reported here).

For these reasons, alternative methods were sought for quickly transcribing the archive across individuals and for intelligently aggregating transcription products given the expectation that individual response variation would likely exist. Here we report partial results from (1) novel crowdsourced designs aiming to collect the data to digitize the entire MacLaury archive; and (2) efficient aggregation methods customized to combine perceptual-identification transcription data across observers in a principled manner when (a) variations in stimuli are expected to contribute both uncertainty and systematic bias to observers' data, and (b) individual differences departing from correct identifica-

tion of graphemic stimuli are likely to be predictable given what is known regarding perceptual confusability among alpha-numeric symbols (cf. [6]).

Crowdsourced transcription designs

To facilitate large-scale, rapid, collection of transcribed information derived from image scans of the archive's 23,000 pages, a series of empirical investigations were designed involving "simulated crowdsourcing" done in the laboratory and actual internet-based crowdsourced data collected via Amazon's Mechanical Turk platform (M-turk). Crowdsourcing is frequently used when very large databases or corpora require evaluation or input from multiple participants. Existing online examples range from projects to transcribe the collected works of philosopher Jeremy Bentham (*blogs.ucl.ac.uk/transcribe−bentham/*), to great historical projects such as transcribing Jeremiah White Graves's farm journal (*beta.fromthepage.com/collection/show?collection*$_id = 5$) and the Civil War Diaries & Letters Transcription Project (*digital.lib.uiowa.edu/cwd/*), or the USGS/PWRC North American Bird Phenology Program which transcribes handwritten records of migratory bird activities in North America (*www.pwrc.usgs.gov/BPP/v4/index.php*). Approaches we use permit unpacking large transcription challenges – converting pages of data – into smaller, more manageable, problems – such as a portion of a data page – for which piecewise solutions can be sought, and can be implemented, in part, as automated procedures, all of which make it appropriate and useful for converting the MacLaury archive.

Table 1: Crowdsourcing designs and task types investigated

Empirical Design	Task Format
1: OCR verification (pattern recognition)	2-AFC yes/no
2: Crowdsource verification	2-AFC match/no-match
3: OCR transcription (training data)	reCAPTCHA task
4: Naming ranges 1	reCAPTCHA task & confidence
5: Naming ranges 2	N-AFC + confidence
6: Focus transcription 1	reCAPTCHA task & confidence
7: Focus transcription 2	reCAPTCHA task

"Crowdsourcing" in the laboratory

Several transcription task designs were developed for the general aim of assessing whether crowdsourcing methods were (a) feasible as an approach for transcribing the kind of data in the MacLaury corpus, and (b) if, of the several designs tested, the tasks provided analyzable and useful transcription data. Major task design consideration was given to accuracy of the transcribed product, speed of transcription, and expense. Designs were based on the idea that transcription of hand-coded data in the MacLaury archive were reducible to a set of perceptual identification type tasks. Table 1 summarizes 7 tasks that were designed for investigating several different archive data formats. Only Tasks 4 and 7 which use reCAPTCHA procedures are described here. Details of all data formats are reported in [7].

Subjects and Design

Design 4 Participants (N=30) were recruited from the University of California, Irvine, School of Social Sciences Human Subjects Pool to collect human transcription responses for a small portion of the archive. The participants received course extra-credit for completing seven perceptual transcription tasks in one-hour. Investigations were performed with participants' informed consent. All procedures used adhered to protocols based upon the world medical association declaration of Helsinki ethical principles for research involving human subjects, and were approved by the ethical review board of the University of California, Irvine.

Figure 2 illustrates the Design 4 entry screen for which one of the four reCAPTCHA-type tasks used (see Table 1) to transcribe freelisted color naming data in the survey. A "reCAPTCHA" is essentially a free-response, fill-in-the-blank, type task in which a distorted or degraded image of letters and numbers is shown and the observer enters the keystrokes that reproduce the alpha-numeric string in the displayed image, usually to verify they are not an internet-trawling robot (see a general example in Figure 1, also see [8]). Scanned images of data-sheets used in Design

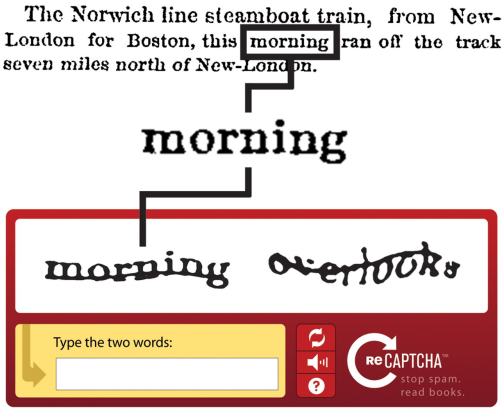

Figure 1. *An example of a standard reCAPTCHA method used to digitize words from written text. reCAPTCHA image courtesy of CyLab at Carnegie Mellon University, Pittsburgh, PA. www.cylab.cmu.edu/partners/success-stories/recaptcha.html. Reproduced with permission.*

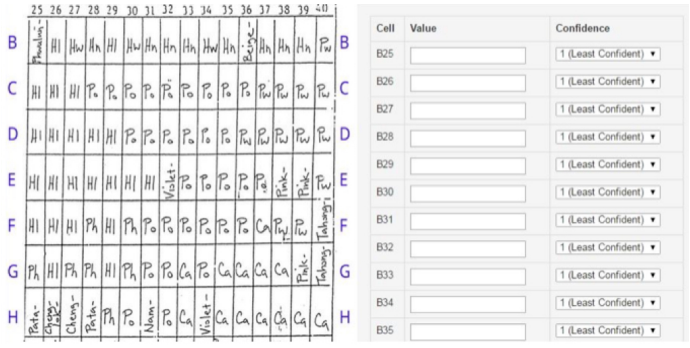

Figure 2. *An example of a Design 4 stimulus and response input fields used in laboratory-based investigations.*

4 consisted of handwritten abbreviated codes for freelisted color category names, organized in a tabular format corresponding to a color array widely used in cross-linguistic color category research (see [4], [2]). Thus, the format of input data for this piloted "crowdsourced" transcription task resembled a checkerboard of cells containing handwritten strings of 2 or more letters. Participants' task was to identify the coded string at each Alpha-row and Numeric-column position, and to enter it using keystrokes in the corresponding cell in the response field (see [7] for further details). Figure 2 shows that confidence judgments were also collected in Design 4 for each reCAPTCHA string provided, but those results are not reported here. The specific aim of Design 4 was to evaluate the use of reCAPTCHA transcription methods for the MacLaury archive using a simulated crowdsourcing format. Analyses of these data are discussed in Results below.

Crowdsourcing on Mechanical Turk

Efficient and accurate transcription of the MacLaury archive using internet-based crowdsourcing requires implementing transcription tasks on a platform like Amazon's Mechanical Turk. Below we discuss a M-turk implementation variant of Table 1's Design 7. In general, implementing designs across both lab-based and M-turk platforms required slight formatting and task display variations. However, resulting minor variations were not to a degree that altered the difficulty or the basic design of tasks assessed. For example, Figure 3 illustrates reCAPTCHA Design 7 datasheet stimulus only (the response area - not shown - is similar to the response fields shown in Figure 2). Compared to Figure 2's naming range stimuli, Design 7 images of category focus stimuli are sparse, needing far fewer items transcribed (see [3]) for definition of archive data types). Nevertheless, the task for our transcription goal remains the same: namely, reproduce with keystrokes the contents of checkerboard cells that contain handwritten letter strings.

Another example of a between-implementation difference is that Design 7 on M-turk shows an entire stimulus page in Figure 3 whereas the preliminary lab-based implementation presented only the portion of the stimulus (outlined here in black to depict this cross-platform difference). Despite such differences in the specific number of judgments on an M-turk screen compared to that in the lab-based studies, this did not effect the quality of responses (indeed, as reported below, the M-turk participants providing a greater number of Design 7 responses actually provided equally or more reliable and consistent data).

Subjects and Design

Design 7's participants or M-turk *workers* were recruited using Amazon Services M-turk platform (www.mturk.com) to complete a Human Intelligence Tasks (or "HITs") providing transcription responses. Twenty-two individuals participated in HITs (some completing more than one HIT) evaluating 3 different archive data-sheets. Here we report on only 10 workers who contributed to transcription results for Figure 3's stimuli. Participants were US citizens above the age of 18 and native English speakers, compensated at $1.50 per hour. Similar to Design 4, informed consent was obtained, via the M-turk platform, as approved by the University of California Irvine Institutional Review Board.

As with Design 4, participants' task was to identify the coded string at each Alpha-row and Numeric-column position, and to enter it using keystrokes in the corresponding cell in the response field. Unlike Design 4, Design 7 did not assess confidence judgments for each reCAPTCHA string provided. The specific aim of Design 7 was to evaluate the M-turk crowdsourcing platform for obtaining a perceptually-based transcription of the MacLaury archive, and to serve as a comparison for subject performance in parallel lab-based investigations. Analyses of Design 7 data are discussed in Results section below.

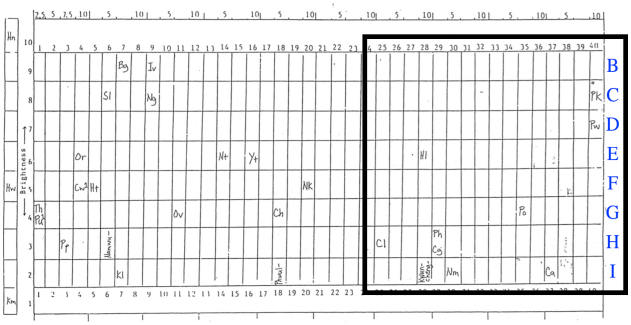

Figure 3. *Design 7 stimulus mock-up illustrating both laboratory-based study stimuli (items within the black frame) and the M-turk crowdsourced study stimuli (items within and outside the frame) for one of the stimulus images assessed.*

Implementing a Multiple-Choice Bayesian Consensus Theory

Here we greatly summarize the Multiple-Choice Bayesian Consensus Theory approach, used in the present investigations, which is developed and discussed in detail elsewhere by Deshpande and colleagues.[7]

For the purposes of aggregating, analyzing and evaluating the transcription results collected via the above mentioned designs, we [7] generalized procedures that were originally developed for evaluating general knowledge and information domains across groups of individuals where the underlying correct answers were unknown.[1] The original model called "Cultural Consensus Theory," or CCT, is a formal process model whose analyses permit inference and estimation of (i) shared latent knowledge for assessed domains (or, in this case, "true" transcriptions), (ii) individual ability, or participant "competence," and (iii) measures of item difficulty and response bias based upon patterns inherent to participants' response data. The original form of CCT [1] has been recently modified to incorporate a Bayesian informant model approach, providing extensive model developments for dichotomous, or "true/false," choice data.[9]

In order to extend the original CCT model, which uses a subjectwise guessing bias appropriate for general information type questions, we chose to enrich the model underlying Dichtomous Bayesian form of CCT [9] to handle N-alternative forced-choice and free response data formats. To do this we modified the model to better address response-option perceptual confusability by eliminating individual subject guessing bias parameters and adding guessing bias parameters for each stimulus item. The revised model assumes that each item has its own bias (linked to an item's distinctive feature confusability with other response options) that is shared across all individuals to the extent that individuals perceptually process the stimuli in the same ways. We

consider this novel itemwise bias model to be appropriate for this kind of transcription task because perceptual confusability is the basis for correctly identifying letter strings for transcription – as opposed to general semantic knowledge or general information question type tasks. In analyses below we refer to the developed multiple-alternative CCT models with different guessing bias configurations as the "subjectwise CCT model" and the "itemwise CCT model".

Although other Bayesian CCT formats have been explored, to our knowledge no Bayesian CCT model has been developed for multiple-choice / free response formats where the number of response alternatives is greater than two. Moreover, with few exceptions (e.g., [11]) CCT it has not been applied to the aggregation of data collected from perceptually-based tasks.

Since CCT was designed for general knowledge domains, how do we know it will work for aggregating data from the kind of perceptual-based judgments that are inherent to transcribing handwritten corpora? Some existing cognitive-perceptual investigations previously used human subject judgments for alphabetic symbols and aggregated responses using CCT.[11, 10] That is, in those investigations CCT measures showed Consensus Theory was appropriate for the domain, and CCT analyses were found to: (a) objectively identify individuals known to be "experts" from within a sample of participants with heterogenous expertise; (b) provide "correct answer" estimates for perceptually-based questions which were in agreement with an underlying cognitive model of perceptual confusability of items; and (c) be capable of objectively estimating item reliability for the domain, thereby identifying "well-formed" and "uncertain" question/stimulus items in a manner compatible with an underlying model of perceptual confusability among items. Such results suggest Bayesian CCT may be useful for the present investigations.

In brief, the present approach aims to use internet-based transcription tasks to collect transcription responses across many individuals, for small portions of a handwritten corpus, which subsequently will be aggregated and analyzed by CCT towards producing an accurate crowdsourced transcription product of the entire corpus. The approach breaks up the large transcription challenge into smaller tasks that are distributed across participants. The expectation is that the piecewise transcription design and its analyses will effectively handle individual differences in transcription data to derive the correct transcription. Finally, although crowdsourcing typically uses large subject samples, as discussed below the present approach employs smarter analyses of smaller samples, using CCT's formal process model, aiming to produce solutions that are as robust as those from large amounts of "averaged" data, and is thus likely to provide a quick, cost effective approach, which – due to it's way of handling potential response bias – may yield superior transcription results. Further description of specifics of the modeling of the crowdsourced tasks, their implementation, extensions of CCT using a Bayesian framework and the results of several other tasks are described elsewhere.[7]

Results

Findings are presented for Designs 4 and 7 investigations only. Results for the other Table 1 task results in that support findings presented here, as well as extensive model assessment and analysis of each design in Table 1 is describe elsewhere.[7] Results summarized below aim to demonstrate whether (a) crowdsourcing is a viable approach for collecting transcription data for the MacLaury archive stimuli, and (b) whether our novel use of CCT proves useable as a data aggregation approach given the perceptual-identification nature of the transcription tasks at investigated.

Variable response patterns versus response uniformity across individuals

Table 2: Comparing CCT analyses from 30 lab-based participants with 10 M-turk workers

CCT method	n	μ(comp)	μ(itemdiff.)
1: Itemwise	30	0.601	0.216
2: Itemwise+Ex	31	0.601	0.208
3: Ss_wise	30	0.603	0.213
4: Ss_wise+Ex	31	0.605	0.208
5: Itemwise	10	0.763	0.328
6: Itemwise+Ex	11	0.757	0.302
7: Ss_wise	10	0.763	0.323
8: Ss_wise+Ex	11	0.761	0.314

Initial analyses examined whether quality of responses collected across two platforms were comparable, if CCT aggregation procedures were appropriate for these data, and how individual differences in the data impacted the fit of the CCT data aggregation tool and model for these data.

Table 2 summarizes CCT analyses comparing 30 lab-based participants and 10 M-Turk participants based on the nine stimuli both investigations assessed (see Figure 3 caption). Rows 1-4 provide the lab-based CCT results, and rows 5-8 provide M-turk platform results, for two different bias models ("Ss_wise" and "Itemwise", and for cases where data from a single expert responder "Ex" was additionally analysed.

Important Table 2 results, for all 8 cases considered, are: (1) Mean participant competence (μ(comp)) obtained from CCT analysis indicates high levels of participant agreement, suggesting a latent "truth" underlying knowledge shared within the groups for this perceptual classification task, and the appropriateness of the CCT approach employed for this task. (2) Mean CCT item difficulty (μ(itemdiff.) is uniformly low, conveying that the nine questions assessed were well-formed and were responded to by transcribers in a manner that correlated with individual competence. And, (3) for all cases, estimated correct "answers" obtained from CCT-aggregated transcription data, corresponded 100% with the known correct answers for all 9 items evaluated (not shown in Table 2). In addition, results concerning CCT's appropriateness suggest that M-turk results (based on 10 participants) are on-par, or trending better, than those observed from the lab-based investigation (based on 30 participants).

These three results indicate that the crowdsourcing designs and CCT modeling are appropriate and useful as an aggregation method for the present transcription data. Similar results were obtained from analyses on different sets of the archive's images, for larger numbers of items, and different participants. In all cases considered, the appropriateness of CCT for aggregating transcription data was robustly demonstrated.

To expand on the comparison just described, additional anal-

yses comparing the lab-based and M-turk based investigations reveal how individual differences in response data were accommodated by the CCT analyses. Figure 4 shows the distributions of participant Design 7 responses for nine items assessed in both the laboratory (shown at left in panel a) and the M-Turk platforms (at right in panel b).

In Figure 4 nine radial wedges of color convey response homogeneity observed for nine transcription items assessed, which are outlined by a black frame in Figure 3. Figure 4(a)'s concentric circles depict 30 participants' responses for nine sequentially numbered items in the lab-based study. Whereas panel (b)'s concentric circles display 10 different participants' responses, independently assessed via M-Turk, for the same sequence of 9 items. Where shown, a wedge of uniform color (excepting lines demarking participants) indicates an item that was uniformly responded to by all participants, and wedges with concentric stripes of varying color show items for which participants provided transcription variations. Response variants observed for the nine items (1 to 9, respectively, left to right) appear grouped in the legend underneath the plots shown, and provide the key for plot color-codes used. Finally, the outermost concentric circle in both panels shows estimated "correct answers" for each of the nine items as derived from CCT analyses of the respective group data. To the degree that the color code of the outermost ring of a given wedge duplicates that seen in the remaining portion of that wedge, the estimated "answer" for that item is tracking a majority rule choice. Conversely, if the color of a given wedge's outermost ring differs from more than 50% of it's remaining area, then CCT's estimated "answer" for that wedge deviates from a majority rule choice.

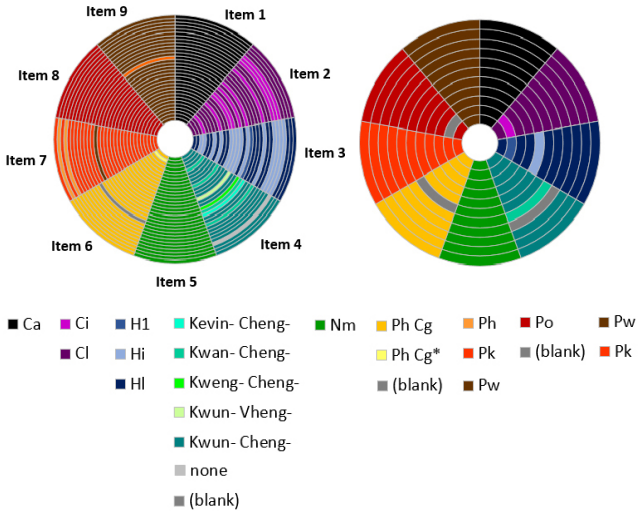

Figure 4. *Transcription response data for 9 questions assessed using both (a) Laboratory based (shown at left) and (b) Mechanical Turk based investigations (shown at right). See narrative for explanation.*

Three observations can be made of Figure 4. First, both investigations exhibit individual differences in participant's response patterns as shown by variegated color striping within a given wedge. Specifically, in the lab-based investigation (30 participants)14% of responses deviated from the known correct response, whereas in the M-Turk investigation (10 participants) 7% or responses deviated from the known correct response.

Moreover, for 3 items responded to by both groups with the greatest variability – that is, items 2, 3, and 4 in the respective Figure 4 diagrams – there is considerable individual variation in response patterns which approximates near-equal splits in preferences for response options. This kind of individual variation is captured by quantitative indices arising from CCT analyses, but is potentially problematic for majority-rule based aggregation approaches, as it is a response data feature that may be overlooked if responses are merely averaged and the underlying distributions of participants responses are not carefully examined.

Second, in both Figure 4 investigations, invariably when an incorrect transcription response was provided, the alternate responses were either (i) omissions, or, more commonly, (ii) were perceptually similar to – or shared distinctive features with – the correct answers. An example of this is found in the second legend entry in Figure 4 shows "Ci" (uppercase "C" with lowercase "I") and "Cl" (uppercase "C" with lowercase "L") were alternate responses observed for stimulus item 2 shown in cell H-25 of Figure 3. This supports the earlier suggestion that this transcription task resembles a perceptual-identification task, as opposed to a general information type task, and supports development of the present form of CCT that employs an item-bias model (before any forms of subject-bias are incorporated) for similar kinds of transcription data aggregation jobs.

Third, and most importantly, even in the presence of individual response differences both investigations find that the estimated correct answers produced by both CCT analyses arrive at the same correct-answer key for the 9 items considered here, and that answer-key reproduced the actual correct answer for all 9 items assessed. In this case, these items were robustly estimated regardless of whether 10 or 30 participants were assessed, regardless of whether an item was responded to homogeneously or not, and regardless of whether a "majority vote," "modal" answer, or even a near equal split in responses might suggest an alternative "correct" answer.

While the findings summarized above are preliminary, they are additionally supported by analyses done on similar investigations carried out using many more items, and obtained using other transcription task designs (Table 1), using varying numbers of participants. See [7] for further investigations.

Robustness under sample size variation

In addition to analyses discussed above, further analyses were conducted to determine if CCT results remained robust and produced comparable findings when based on sample sizes smaller than those typically seen in cognitive and psychological science investigations ($n = \sim 30$), and, in particular, would robustness be found in cases where two equally popular response strategies existed for some items. It should be noted that CCT was designed for scenarios where small-sized samples (6-10 participants) are typical, as is common in anthropological investigations, and it is been suggested by the developers of CCT that, majority choice rule and CCT typically lead to the same results when estimating correct answers from a large number of informants (pp. 79-81 of [1], pp. 327-8 [12]). By comparison, CCT developers assert that for small numbers of informants and heterogeneous competencies (*a.k.a.* individual differences or individual variations in expertise), CCT's answer-key estimation procedures easily permit a minority of informants with higher competencies to

outweigh a majority of informants with lower competencies. This is the mechanism by which CCT's underlying formal model permits accurate estimation of correct answers using smaller groups of participants than would typically be needed when using a majority choice aggregation rule. The question is whether this feature of the data aggregation procedures can be use to benefit of the present large-scale transcription problem.

To investigate whether CCT would perform satisfactorily when aggregating data from smaller participant samples, we computed CCT analyses on subsets of data from 30 individuals, participating in Design 4, for which two strategies for some items were observed (e.g., the above mentioned response alternatives "Ci" and "Cl" for perceptually ambiguous stimuli). The subsets were constructed by randomly drawing samples of size 8, and included 4 individuals randomly drawn from the pool of participants who adopted one strategy and 4 additional individuals who adopted an alternative strategy.

Five such subset samples were drawn, one after another, with replacement. CCT analyses were carried out on all five subsets, the results of which are presented in Table 3, where they are compared with CCT analyses using Design 4's total sample of size 30. Table 3's trends for these 8 participant subgroups suggest, for this example, that 8 participants are as informative and useful for correct transcription estimation as that seen with 30 participants (Table 3, row 6). Thus, while this analysis is preliminary, its results accord with the original aims of CCT[1], and suggest that additional bootstrapping or empirical investigations addressing this issue may confirm that CCT is appropriate for modeling data from crowdsourced sample sizes much smaller than 30 participants.

Table 3: Task 4 CCT analyses on randomly-sampled subsets of 8 participants compared to the total group ($n = 30$)

Sample size	answer-key correct %	μ(Ss.comp)	μ(itemdiff.)
1: 8 Ss	100%	0.929	0.466
2: 8 Ss	100%	0.937	0.460
3: 8 Ss	100%	0.914	0.459
4: 8 Ss	100%	0.942	0.464
5: 8 Ss	100%	0.935	0.464
6: 30 Ss	100%	0.917	0.366

Objectively identifying "expertise"

One valuable feature of CCT over other standard aggregation approaches is that CCT uses principled model-based procedures to provide "competence" estimates for each participant when analyzing a dataset. One advantage of such competence estimates is that they provide an objective index of individual expertise even when there is large variability across participants' response patterns. The utility of this was previously demonstrated in earlier investigations using CCT to assess the graphemic well-formedness of alphabetic items [10, 11]. For example, Figure 5 uses existing data [11] to show that, based solely on response data, participants who are known to be experts in typeface design (i.e., employees of Adobe Systems, Palo Alto, CA) can be objectively identified by CCT "competence" indices as typeface design "experts" when compared to college undergraduate participants with no training in symbol system design. The research also showed that in addition to Adobe "experts" having higher average competence measures, even specializations in Adobe participants' expertise could be differentiated by CCT competence indices as shown by the plotted competence-based differences between of typeface designers of *kanji*/*kana* symbol systems (Figure 5's 2 outliers scoring the lowest competence in the "Expert" column of the graph), relative typeface designers expert in Latin-based symbol systems, receiving higher competence measures in the analyses.

The relevance of this to the present investigations is that when differences in individual response patterns are found in crowdsourced transcription data, CCT analyses of such data may permit objective identification of both experts and non-experts in a sample, and response patterns of identified experts in a sample can be used to resolve transcription ambiguity, similar to the example illustrated described below.

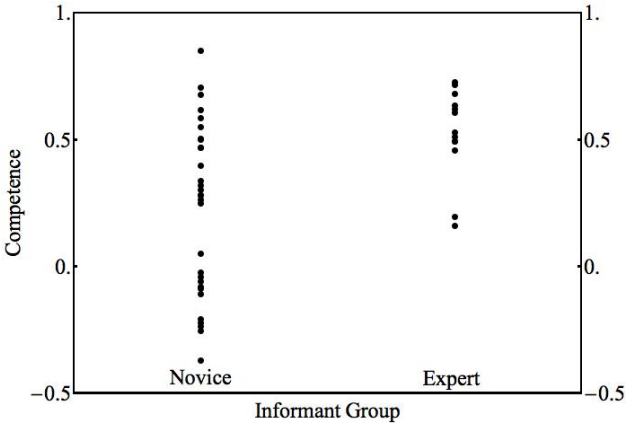

Figure 5. *CCT Competence measures objectively identify novice and expert participants. Based on data from [?].*

Can individual response variation help disambiguate competing transcription solutions?

Both CCT and majority rule-like approaches work very well as aggregation methods when response patterns across individuals are mostly homogeneous. However, one advantage of CCT over standard methods of data aggregation is that it performs equally well under circumstances when there is substantial variability in the response patterns of participants in a sample.

To illustrate the usefulness CCT indices that can identify expertise in transcription data, we can first examine analyses of all 29 M-turk responses from Design 7 which nicely illustrate how CCT analyses out-perform majority-choice aggregation methods when individual response variation does not easily disambiguate among two equally likely response options. That is, for all 29 M-turk items in Figure 3 analyses using both itemwise and subjectwise bias models correctly inferred transcription solutions for 28 of the 29 M-turk items. The one item that both models inferred incorrectly was "namwu-" (row H, col. 6 of Figure 3). Examination of the actual participant response data for these 29 items shows that four of the ten M-turk participants provided the known-correct transcription for the "namwu-" item, while the remaining 6 participant majority provided incorrect transcriptions

(4 responding “narnwu-”, 1 responding “namwy”, and 1 non-responder).

However, more to the point, when the transcription responses of an “expert” participant – that is, someone with advance knowledge of the correct answers – were added to the group’s data analyses, both itemwise bias and subjectwise bias models produced the correct transcription solution for the previously incorrectly estimated ”namwu-” item. Thus, even when the majority of participant responses were still incorrect (5 correct responses of ”namwu-”, 6 incorrect responses), the addition of an expert respondent produced a weighting scenario in CCT procedures that correctly allowed a minority choice-option to be identified as the best option for the estimated correct-answer for that item.

This preliminary result is consistent with findings from similar investigations on symbol system processing[11], and suggests that for this perceptual-identification task CCT methods are (i) robust against the uncertainty that contributes to individual variation in participants’ responses, and (ii) incorporating the data of a single ”expert” in CCT analyses may result in more reliable transcription solutions.

When individual differences make transcription data equivocal, how do the two bias models compare?

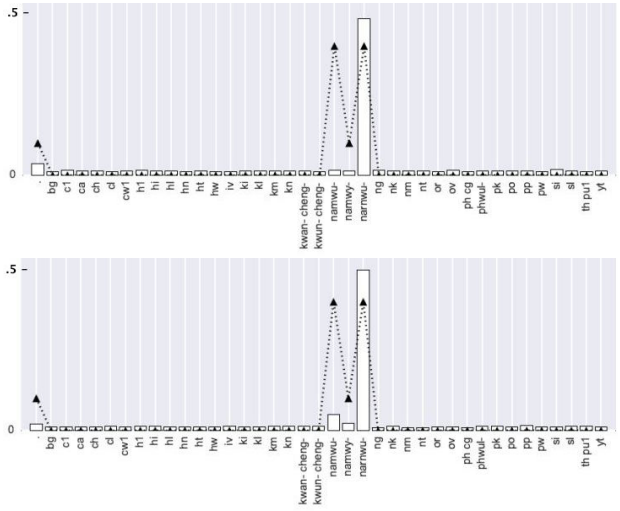

Figure 6. *Model predictions of subject response distributions using subjectwise bias CCT (at top) and itemwise bias CCT (bottom) for the “namwu-” item in the M-turk study. In both, response distribution of participant data is shown by filled triangles and dashed lines; and CCT’s modeled response distribution predictions are shown by the vertical white bars.*

Given the foregoing example of how individual response variation can give rise to equivocal transcription items (e.g.,“namwu-”), it is useful to detail the differences between our itemwise bias CCT model compared to a standard subjectwise bias CCT model, and to examine how the two versions of the model vary with respect to the observed response uncertainty associated with equivocal items like “namwu-”. Figure 6 shows two forms of model predictions for “namwu-” depicted as histograms, as well as a curve depicting the observed response distributions for the alternatives answers for items. The comparison shows that, in both cases, histogram bars (model predictions) generally coincide with the connected line of the data’s distribution of responses (dashed line), illustrating that both item-wise and subject-wise models produced good results and correct answer keys. However, in the itemwise graph (bottom panel of Figure 6) there is a trend showing less of a separation between the model prediction histograms and the dashed line showing the distribution of observed responses. This suggests that, for this item, the itemwise model appears to be better at reproducing the response patterns in the human data than does the subjectwise model. This kind of result has been previously observed for several additional equivocal items, using different data and separate analyses, and is reported in detail elsewhere.[7] Thus, even though our preliminary results show that the standard CCT form of subjectwise modeling can infer sensible answer keys much of the time, for this task subjectwise bias tends to model the actual response data less well than the itemwise modeling does. This result, in conjunction others [7], provides further confidence in the notion that data arising from perceptual confusability among transcription choice options is likely to involve some form of item-bias, in addition to forms of systematic subject bias that may be possible in this task.

Summary

For the purpose of investigating whether crowdsourcing methods are feasible as an approach for transcribing perceptually ambiguous information of the kind found in the MacLaury archive, and for illustrating the appropriateness of CCT as a data aggregation approach under two different data collection platforms, we have shown:

- reCAPTCH-based task designs were found useful for providing reliably transcribed portions of the archive, and were robust across both lab and internet-based platforms.
- Transcription products obtained from both platforms (i) could be aggregated across individuals with varying expertise in ways that robustly recovered the know latent truth, and (ii) were modeled appropriately using the principled theory on which CCT analyses are based.
- CCT remains robust even when individual differences give rise to response pattern variation across sampled individuals (whereas individual differences are often discounted or handled as “outliers” by standard aggregation methods).
- CCT was shown to be robust for both small and large sample sizes (where standard aggregation methods are typically problematic with small sample sizes).
- CCT is able to robustly estimate “correct” answers to perceptually-based questions for which the input stimulus alone may limit disambiguation of confusable options (e.g., for degraded/confusable letter-string stimuli with distinctive feature similarities, such as“ci,” “cl” or “c1”).
- CCT can disambiguate response uncertainty arising within a sample, especially for cases where computing a standard mean or modal response from a survey may be problematic (e.g., when informant groups exhibit $\sim 50/50$ split response patterns), or where majority-rule aggregation procedures produce unrepresentative results.
- CCT’s principled aggregation approach is designed to handle variable response patterns arising from individual differences within a sample, and is able to objectively identify differences arising from “expertise.” And,

- CCT succeeds in aggregating group data by using relative weights (based on individual's estimated competence) to intelligently determine the impact each individual's response data should have on estimating the correct"answers" for the questions assessed.

In accord with our previous work [7], the findings for the present perceptual classification task suggest that CCT procedures provide a viable alternative for aggregating perceptual data which makes principled use of – rather than deemphasizing – individual differences in participants' response patterns, and permits identification of domain-specific individual expertise from the individual variation in the data. The findings reported here are also supported by investigations carried out using additional participants, different designs, and additional stimuli. More further discussion, as well as supporting results, are detailed elsewhere.[7]

It is noteworthy that features of the present transcription problem and approach may apply elsewhere. The perceptual nature of our tasks differ from general information surveys or opinion-poll data, as suggested earlier, in that response bias is likely to be item-based rather than associated with individual informants, accommodating the likely event of more than one valid decision strategy when targets are ambiguous.

Our initial work in this area [7] is, to our knowledge, the first systematic application of CCT analyses to data from perceptual processing tasks that emphasize detection of graphemic targets from a generative cognitive symbol space. Such tasks differ from those that assess domains invovling general knowledge type questions, to which CCT is typically applied. CCT has previously been successfully used to assess whether color deficient participants could achieve color naming consensus in the absence of a perceptual comparison strategy for confusable color stimuli ([14]), and the present results additionally suggest that CCT is useful for data aggregation in investigations that assess perceptual stimuli in visual processing domains when distinctive feature properties are used as a basis for disambiguating stimulus targets from distractor stimuli.

This investigation proposes an intelligent model of data aggregation that may permit trading off smarter data for bigger data, and give a more economical approach to accurately deriving robust results using internet-based crowdsourcing methods. As such, this present data analysis procedure and modeling approach shows promise for aggregating data from other perceptual-like tasks that have recently used crowdsourcing, such as visual search campaigns for locating downed airliners using satellite data (www.tomnod.com/), assisting disaster relief efforts on the ground through photos taken in disaster-affected areas (geotagx.org/), mapping features of ocean floors (exploretheseafloor.net.au/), or global environmental monitoring (/www.geowiki.org/), to name a few.

Acknowledgments

Many thanks to Maru MacLaury. The authors also thank Robert E. MacLaury *ColCat* Archive Team members, Nathan Benjamin and Yang Jiao. We greatly appreciate support for this project provided by *Calit2* at UCI. Portions of this work were funded by a University of California Pacific Rim Research Program grant (2010-2015) to K.A. Jameson, PI. National Science Foundation 2014-2017 (#SMA-1416907, K.A. Jameson, PI). UC Irvine's Undergraduate Research Opportunity Program Awards. All components of this research were approved by UCI IRB protocols: HS#2013-9921, HS#2015-1976, and HS#2015-9606. The views and opinions expressed in this work are those of the authors and do not necessarily reflect the official policy or position of any agency of the University of California or The National Science Foundation.

References

[1] William H. Batchelder & A. Kimball Romney, Test theory without an answer key, Psychometrika, 53, 71 (1988).

[2] Robert E. MacLaury, Color and Cognition in Mesoamerica: Constructing Categories as Vantages. Austin, TX: University of Texas Press. (1997).

[3] Kimberly A. Jameson, Nathan A. Benjamin, Stephanie M. Chang, Prutha S. Deshpande, Sergio Gago, Ian G. Harris, Yang Jiao and Sean Tauber. Mesoamerican Color Survey Digital Archive. In Encyclopedia of Color Science and Technology, (Ronnier Luo, Ed.). Springer Berlin Heidelberg. DOI 10.1007/978-3-642-27851-8. (2015).

[4] Paul Kay, Brent Berlin, Louisa Maffi, William R. Merrifield & Richard Cook, The World Color Survey (1st edition). Center for the Study of Language and Inf. Stanford, CA, 2011.

[5] Yang Jiao, Sergio Gago, Ian Harris & Kimberly A. Jameson, Optical Character Recognition of Handwritten Tabular Data, Poster presentation. The 22nd Annual UCI Undergraduate Research Symposium, Irvine, CA. (2015).

[6] James T. Townsend, Gary G. Hu, & R. J. Evans, Modeling feature perception in brief displays with evidence for positive interdependencies, Perception & Psychophysics, 36, 35 (1984).

[7] Prutha S. Deshpande, Sean Tauber, Stephanie M. Chang, Sergio Gago & Kimberly A. Jameson, Aggregation of Crowdsourced Transcription Data: A Cultural Consensus Theory Approach. Technical Report Series. #MBS 16-01, Institute for Mathematical Behavioral Sciences, University of California at Irvine. Irvine, CA, USA. (2016).

[8] Luis Maurer, Benjamin McMillen, Colin Abraham, David Blum, Manuel, reCAPTCHA: Human-Based Character Recognition via Web Security Measures. Science 321 (5895): 1465. (2008).

[9] Zita Oravecz, Jaochim Vandekerckhove & William H. Batchelder, Bayesian Cultural Consensus Theory. Field Methods. 26, 207, (2014).

[10] Kimberly A. Jameson, Empirical Methods for Evaluating Generative Semiotic Models: An Application to the Roman Majuscules. In Writing Systems and Cognition, The Neuropsychology and Cognition series. (W. C. Watt, Ed.) Kluwer Publishers, pg. 248. (1994).

[11] Kimberly A. Jameson & A. Kimball Romney, Consensus on Semiotic Models of Alphabetic Systems, Quantitative Anthropology, 2, 289 (1990).

[12] Batchelder, William H., & Royce Anders. Cultural consensus theory: Comparing different concepts of cultural truth. Journal of Mathematical Psychology 56.5, 316, (2012).

[13] Kimberly Jameson, An Empirical Investigation of Semiotic Characterizations of Alphabetic Systems. Doctoral Dissertation, University of California, Irvine. (1989).

[14] Valerie Bonnardel, Color naming and categorization in inherited color vision deficiencies Vis. Neurosci., 23(3-4), 637, (2006).

Author Biography

Kimberly A. Jameson received her PhD in psychology from the University of California, Irvine (1989). Her empirical and theoretical work includes research on color perception and photopigment opsin genetics; the mathematical modeling of color category evolution among communicating artificial agents; individual variation and universals in human color

cognition and perception; and investigating the ways individuals name and conceptualize color. She is Associate Project Scientist at the Institute for Mathematical Behavioral Sciences, UC Irvine.

Prutha Deshpande received her BS in Cognitive Science from the University of California, Irvine (2015). She is currently a Junior Specialist in the UCI's Cognitive Science Department, and Lab Manager of the Color Cognition Laboratory at UC Irvine. Her on-going research focuses on empirically investigating the influence of bilingualism on cognitive color representation.

Sean Tauber received his PhD in Psychology from the University of California, Irvine (2013) and is currently an Assistant Project Scientist in the Institute for Mathematical Behavioral Sciences at UC Irvine. His work focuses on understanding human behavior and cognition using mathematical and computational models.

Stephanie Chang, BS in Computer Science (expected 2016), University of California, Irvine. Research assistant (2014-2016) at the California Institute for Telecommunications and Information Technology Irvine, and application developer for UC Irvine Student Housing (2014-2016). She enjoys working with both full stack web and mobile application development and aspires to be a software engineer.

Sergio Gago earned a master's degree in industrial engineering and a master's degree in product and system design from Barcelona's Polytechnic University of Catalonia (UPC). He earned his Ph.D., also from UPC in engineering for emotion in human-device interaction (2012). Following, he joined Calit2 in UC Irvine as an Associate Specialist researching user-product interaction and experience. He manages projects and mentors students in research areas such as products and systems engineering, informatics and computer sciences.