Approval Voting and Incentives in Crowdsourcing

NIHAR B. SHAH, Machine Learning Department and Computer Science Department, Carnegie Mellon University
DENGYONG ZHOU, Google Brain

The growing need for labeled training data has made crowdsourcing a vital tool for developing machine learning applications. Here, workers on a crowdsourcing platform are typically shown a list of unlabeled items, and for each of these items, are asked to choose a label from one of the provided options. The workers in crowdsourcing platforms are not experts, thereby making it essential to judiciously elicit the information known to the workers. With respect to this goal, there are two key shortcomings of current systems: (i) the incentives of the workers are not aligned with those of the requesters; and (ii) the interface does not allow workers to convey their knowledge accurately by forcing them to make a single choice among a set of options. In this article, we address these issues by introducing approval voting to utilize the expertise of workers who have partial knowledge of the true answer and coupling it with two strictly proper scoring rules. We additionally establish attractive properties of optimality and uniqueness of our scoring rules. We also conduct preliminary empirical studies on Amazon Mechanical Turk, and the results of these experiments validate our approach.

CCS Concepts: • **Information systems** → **Crowdsourcing**;

Additional Key Words and Phrases: Proper scoring rules, incentives, labeling

ACM Reference format:
Nihar B. Shah and Dengyong Zhou. 2020. Approval Voting and Incentives in Crowdsourcing. *ACM Trans. Econ. Comput.* 8, 3, Article 13 (June 2020), 40 pages.
https://doi.org/10.1145/3396863

1 INTRODUCTION

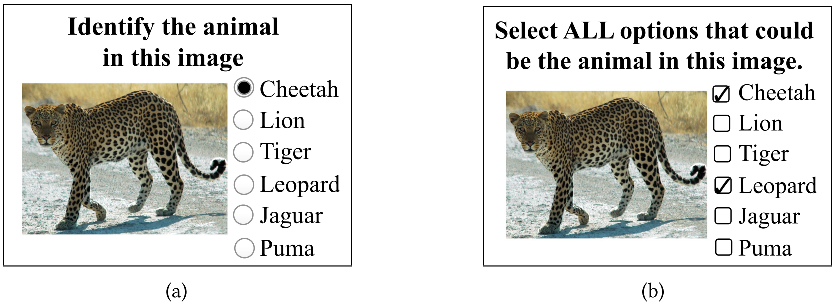
In the big data era, with the ever-increasing complexity of machine learning models such as deep learning, the demand for large amounts of labeled data is growing at an unprecedented scale. These labeling tasks used to be done by domain experts. However, the limited pool of experts would limit the size of the datasets. In the modern day, these massive labeling tasks are performed through commercial web services such as Amazon Mechanical Turk, where millions of crowdsourcing workers or annotators perform tasks in exchange for monetary payments [49].

A crowdsourcing labeling task consists of a set of items such as images to be labeled, and each item is associated with a set of exclusive categories (or options). Typically, each worker is required

This article was presented in part at the International Conference on Machine Learning (ICML) 2015. The work of Nihar B. Shah was supported in part by a Microsoft Research PhD fellowship and in part by NSF grant 1755656

Authors' addresses: N. B. Shah, Machine Learning Department and Computer Science Department, Carnegie Mellon University, 5000 Forbes Avenue, Pittsburgh PA 15213, USA; email: nihars@cs.cmu.edu, D. Zhou, Google Brain, 747 6th St South, Kirkland, WA 98033; email: dennyzhou@google.com.

2167-8375/2020/06-ART13
https://doi.org/10.1145/3396863



Identify the animal in this image

● Cheetah
○ Lion
○ Tiger
○ Leopard
○ Jaguar
○ Puma

(a)

Select ALL options that could be the animal in this image.

☑ Cheetah
□ Lion
□ Tiger
☑ Leopard
□ Jaguar
□ Puma

(b)

Fig. 1. Illustration of a task with (a) the standard single-selection interface and (b) an approval-voting interface.

to select the category (i.e., a single option) that she believes is most likely to be correct. More formally, it involves eliciting the mode of the worker's belief. Such a "single-selection" crowdsourcing setting has been studied extensively both empirically and theoretically.

In practice, the data obtained via crowdsourcing is typically quite erroneous [31, 59, 60] due to the lack of expertise of workers, lack of appropriate incentives, and often the lack of an appropriate interface for the workers to express their knowledge.

In this article, we consider an "approval-voting" means of eliciting labels from the workers. Approval voting [2, 32, 44, 61] is a form of voting in which each voter can "approve of" (that is, select) multiple candidates. No further preferences among these candidates are specified by the voter. In our context of crowdsourcing, the approval voting interface allows workers to select multiple options for every question.[1] See Figure 1 for an example. Approval voting is known to have many advantages over single-selection systems in psychology and social choice theory [2, 8–10, 20, 24, 36]. First, an approval voting interface is easy to understand [36]. Approval voting provides workers more flexibility to express their beliefs; for instance, allowing the worker to express any confusion between multiple options, instead of being forced to select one of the options. Approval voting also utilizes the expertise of workers with partial knowledge more effectively. For instance, Coombs [9] posits that *"It seems to be a common experience of individuals taking objective tests to feel confident about eliminating some of the wrong alternatives and then guess from among the remaining ones."* Further, Coombs [9] argues that *"Individuals taking the test should be instructed to cross out all the alternatives which they consider wrong."* This instruction forms the motivation for the requirements we pursue in this article.

Under this approval-voting interface, we require the worker to select the options that she feels are "quite likely" to be correct (we formalize this in subsequent sections). In the setting of crowdsourcing, as compared to single-selection, selecting multiple options would allow for obtaining more information about the partial knowledge of these non-expert workers. This additional information is particularly valuable for difficult labeling questions, allowing for the identification of the sources of difficulty.

Let us illustrate the utility of this setting in crowdsourcing by means of an example illustrated in Figure 1. The question requires the worker to identify the animal in the image—a leopard in this case. Suppose there are two workers. The first worker believes the true label to be either "Cheetah" or "Leopard," but certainly not any other option; the second worker is confused about some other aspect of the image and believes the true label to be either "Jaguar" or "Leopard," but certainly none of the others. If each worker is allowed to select only a single answer (Figure 1(a)), and if

[1] The literature on psychology often refers to approval voting as "subset selection."

the workers choose one of the two options they are confused about uniformly at random, there is a 25% chance that the first worker selects "Cheetah" and the second worker selects "Jaguar." Moreover, there is a 50% chance that one worker chooses "Leopard" and the other worker chooses a different option. In each of these cases, their responses will thus not provide any definitive answer about the true label. In contrast, under the approval voting interface (Figure 1(b)), we allow the worker to select both the options that they are confused about, that is, ("Cheetah," "Leopard") from the first worker and ("Jaguar," "Leopard") from the other worker, then "Leopard" becomes a clear winner. Indeed, "Leopard" is the language in Figure 1. For our second example, we continue to consider the question in Figure 1. Now suppose that one worker knows the correct answer to be "Leopard" for sure, while the second worker is completely confused between "Cheetah," "Jaguar," and "Leopard." In a single-selection setting, the second worker may select one of the other three options at random, and in the case "Leopard" is not selected, it provides an inconclusive set of answers from the two workers comprising two different options. On the other hand, in the approval voting setting, the second worker is allowed to communicate her confusion by selecting all three options under consideration; that allows for inference of "Leopard" as the correct answer.

Despite the flexibility it offers in eliciting partial knowledge, approval voting alone may not suffice for high-quality crowdsourcing due to lack of incentives. We additionally need to incentivize the workers to report their answers appropriately. To this end, we need to couple approval voting with a payment mechanism such that a worker receives her maximum expected payment if and only if she truthfully discloses her partial knowledge on the crowdsourcing question.

In crowdsourcing tasks comprising objective questions, it is a standard practice [19] to include "gold standard questions," that is, questions to which the system designer (or, the principal in game-theoretic terms) already knows the answers. The gold standard questions are mixed at random within the actual questions, and the worker is unaware of which questions are the gold standard. These gold standard questions are employed to verify the answers provided by the workers and form the basis of the payments made to the workers. The gold standard questions are typically generated by experts (who are often much more expensive than crowdsourcing workers) or are obtained as an aggregate of the answers of a large number of crowdsourcing workers. In this article, we will not concern ourselves with the source of these gold standard questions, but only assume that we have access to a set of gold standard questions to which we know the correct answers.

The framework of scoring rules [4, 21, 35, 51] considers the design of payment mechanisms (or scoring rules) to elicit predictions about an event whose actual outcome will be observed in the future. The payment is a function of the agent's response and the outcome of the event. The payment is called a "strictly proper scoring rule" if its expectation, with respect to the belief of the agent about the event, is strictly maximized when the agent reports her true belief. Our setting can be considered as that of designing strictly proper scoring rules, since the gold standard questions provide the actual outcome.

With this context, there are two main goals we pursue in this article. The *first goal*, in informal terms, is to design strictly proper scoring rules that incentivize the worker to select only those options that she thinks are quite likely to be correct. We consider two settings based on the precise meaning of "quite likely"—one where the meaning is in absolute terms and one relative—and these settings are formally specified later in the article. Now, for the *second goal*, we note that while proper scoring rules have previously been studied under quite generic settings, the general theory provides a very broad class of scoring rules and does not specify any particular scoring rule for use. However, in our application of crowdsourcing, we must choose one specific scoring rule for deployment. Thus, in addition to strict properness, we strive to establish additional properties of uniqueness and/or optimality of our proposed scoring rules to additionally motivate their use.

Related work

In this work, we design proper scoring rules that rely on the presence of some "gold standard" questions to validate the workers' responses. However, there are many settings where gold standard questions may not be available; for instance, if the questions are subjective or if the experts who generate these gold standard questions are too expensive. There is a parallel line of literature [11, 16, 29, 34, 41, 45] that explores the design of payment mechanisms that operate in the absence of any gold standard questions. The payment mechanisms designed therein can, however, generally provide only weaker guarantees (e.g., multiple Nash equilibria) and/or require elicitation of additional information (e.g., prediction of other workers' responses) due to the absence of a gold standard answer to compare with.

Of particular interest is the paper by Lambert and Shoham [35], which considers a general framework for eliciting truthful answers for multiple choice questions. The paper restricts attention to the study of individual multiple-choice questions, since, quoting Reference [35], "we can consider a payoff for the full questionnaire by summation of the payoffs of each question." This approach is orthogonal to a key focus of our work, which is precisely on how to combine the payoffs across questions:

- Consider a single approval voting question. This question may be considered as a set of sub-questions with each sub-question in this set corresponding to asking whether one particular option is correct or not. The results of Lambert and Shoham [35] yield a (strictly) proper scoring rule for each sub-question, which can then be summed up to form a (strictly) proper scoring rule for one question. In Section 3 of the present article, we in fact show that there is exactly one way of combining these sub-questions.
- Moving on to a general questionnaire with multiple questions, Lambert and Shoham [35] suggest summing up the payments across all questions. In contrast, in Section 4 of the present article, we show that, surprisingly, there is a different method of combining payoffs across multiple questions (*multiplying* the payoffs of each question) and this method uniquely satisfies certain desirable properties.

In principle, one could alternatively consider the entire questionnaire as a single question, but as noted by Lambert and Shoham [35], this can be computationally prohibitive.

A few prior works study approval voting in the context of crowdsourcing applications for specific settings, such as for question-and-answer forums [27] and Doodle polls [68]. The focus of the present article is on the design of incentive payment mechanisms with properties that fundamentally hold irrespective of the nature of the setting.

Past work [55] by a subset of the authors of the present article considers a crowdsourcing setup with the traditional single-selection setting, also eliciting the workers' confidences for each response. A strictly proper scoring rule is proposed for that setting, which is then shown to be the only one to satisfy a "no-free-lunch" axiom proposed therein. While the setting of the present article is different from that of Shah and Zhou [55], interestingly, our scoring rule derived for a different interface and under a different set of assumptions turns out to be the only one that can satisfy the no-free-lunch axiom (adapted to our setting).

An important complementary problem is to design methods to aggregate responses of multiple workers to estimate the true answers to the various questions. This is required, since the data obtained from workers is typically noisy. To this end, many statistical aggregation methods [7, 12, 26, 30, 33, 37, 47, 53, 62, 66, 67] have been proposed in the literature; however, they primarily consider the single-selection setting where any worker can select only one option for each question. These works consider certain statistical models for the errors in the responses and propose algorithms to estimate the correct answers to the questions often with theoretical guarantees under the assumed model. Our approach complements these techniques in that we endeavor to obtain higher-quality labels directly from crowdsourcing platforms via novel interface and incentive mechanisms. That said, the literature on statistical aggregation largely focuses on the single-selection setting, and it is an important open problem to design statistical aggregation algorithms for the approval voting setting.

A related problem is that of choosing good workers or assigning appropriate tasks to workers, and these problems are studied in References [1, 22, 58, 64]. In our work, we assume that the workers are already chosen and the specific task is already assigned to that worker. Some recent works focus on eliciting data from multiple workers with the overall goal of performing certain specific estimation tasks [5, 13, 17, 18]. In contrast, our goal is to ensure that workers censor their own low-quality (raw) data without restricting our attention to any specific downstream algorithm or task. Non-financial incentives that rely on reputation or social norms are designed in References [23, 65]. On the other hand, we consider paid crowdsourcing platforms such as Amazon Mechanical Turk as our setting.

Summary of results

We consider two settings in the context of incentivizing the worker, for each question, to select options that she thinks are "quite likely" to be correct. The two settings differ in the precise meaning of the term "quite likely." We now briefly describe these settings and our associated results.

The first setting we consider involves eliciting the options for which the probability of being correct, according to the worker's beliefs, is greater than some pre-defined threshold. We term this setting as *absolute thresholding*. We design a scoring rule for this setting and show that it is strictly proper. In addition, we also prove that in a certain regime, our proposed rule is the only possible strictly proper scoring rule.

The second setting involves *relative thresholding,* where the worker must select options whose likelihood of correctness is above a certain threshold relative to the other selected options. We design a strictly proper scoring rule for this setting and prove several additional appealing properties for this rule: (1) it is the only strictly proper scoring rule that can satisfy a simple condition, which we term "no-free-lunch"; (2) it is also strictly proper for eliciting the support of workers' beliefs; (3) it achieves the fundamental limits of performance-based payments.

We finally report results from preliminary experiments that we conduct to verify certain basic hypotheses underlying our approach and to catch any possible impediments that may arise in a practical implementation and evaluation. In the data obtained from the workers, we observe that this set of preliminary experiments indeed support our hypotheses. Moreover, the experimental deployment did not raise any red flags towards practical use of our setting and scoring rule.

Organization of the article

The rest of the article is organized as follows: We begin with a description of the formal problem setting and the goals of the article in Section 2. In Section 3, we present theoretical results on the setting of absolute thresholding. Then, in Section 4, we present theoretical results on relative thresholding, which includes the problem of eliciting the support of the beliefs as a special case. We present experimental results in Section 5. We conclude the main text of the article with a discussion on future work in Section 6.

The article also contains three appendices. Appendix A contains certain auxiliary results related to the contents of the article. Appendix B presents the proofs of the theoretical results that are not included in the main text. Appendix C provides additional details regarding the experiments.

2 PROBLEM SETUP

Consider $N \geq 1$ questions, each of which has B options ($2 \leq B < \infty$) to choose from. For each option, exactly one of the B options is correct. We assume that these N questions contain $G (1 \leq G \leq N)$ "gold standard" questions, that is, questions to which the principal knows the answers *a priori*. These gold standard questions are assumed to be mixed uniformly at random among the N questions, and the worker is evaluated based on her performance on these G questions. For any integer K, we use the standard notation of $[K]$ as a shorthand for the set $\{1, \ldots, K\}$. We let $\mathbf{1} : \{\text{True}, \text{False}\} \rightarrow \{0, 1\}$ denote the indicator function defined as $\mathbf{1}\{x\} = 1$ if x is true and $\mathbf{1}\{x\} = 0$ otherwise.

In this setting of strictly proper scoring rules—where we use the gold standard questions to evaluate workers' answers and compute the payments—we can consider the scoring rule for any worker independent of all other workers. Thus, in the problem setting, we consider only one worker, with the understanding that the scoring rule can be applied independently to all workers.

For each of the N questions, we assume that the worker has, in her mind, a probability distribution over the B options representing her beliefs of the probabilities of the respective options being correct. Formally, consider any worker and any question $i \in [N]$. For any option $b \in [B]$, the worker believes the probability of option $b \in [B]$ being correct is $p_{ib} \in [0, 1]$ for some latent values $p_{i1}, \ldots, p_{iB}$ that sum to one. We assume that these belief-distributions of a worker are mutually independent across questions [20].

2.1 Payment Function (Scoring Rule)

As mentioned earlier, the worker's performance is evaluated based on her responses to the gold standard questions. For any question in the gold standard, we denote the evaluation of the worker's performance on this question by a value in the set $\{-(B-1), \ldots, B\}$: The magnitude of this value represents the number of options selected by the worker, and the sign is positive if the correct answer is in the set of selected options and negative otherwise. For instance, if the worker selected four options for a certain gold standard question but none of them was correct, then the evaluation of this response is denoted as "-4"; if the worker selects two options for a gold standard question and one of them is the correct option, then the evaluation of this response is denoted as "$+2$." Note that we do not distinguish between the incorrect options; that is, the payment scheme depends on selection of the correct option and the number of selected options, but is assumed to be independent of the identities of the incorrect options. The uniqueness results presented subsequently operate under this assumption.

We assume that the payments are bounded, that is, any payment must lie in the interval $[\alpha_{\min}, \alpha_{\max}]$ for some values $\alpha_{\min}$ and $\alpha_{\max} > \alpha_{\min}$. The choice of the two parameters $\alpha_{\min}$ and $\alpha_{\max}$ may be made keeping various factors in mind, such as guidelines of the crowdsourcing platform used, the budget constraints, and the minimum wage. We will assume that the values of the two parameters are given to us.

Let

$$f : \{-(B-1), \ldots, B\}^G \rightarrow [\alpha_{\min}, \alpha_{\max}]$$

denote the payment function. We will use the terms "payment function" and "scoring rule" interchangeably throughout the article. It is this function f that must be designed to incentivize the worker. To bring all possible scoring rules on an equal footing, we fix $\alpha_{\max}$ as the payment for the best possible outcome, which is when the worker selects exactly the correct option (and nothing else) for each question in the gold standard:

$$f(1, \ldots, 1) = \alpha_{\max}. \tag{1}$$

Throughout the article, we assume all scoring rules to satisfy Equation (1), along with the requirement to lie in the interval $[\alpha_{\min}, \alpha_{\max}]$.

In the sequel, we use the notation $f^{\#}$ and f^{*} to denote the two scoring rules proposed in this article and f to denote any general scoring rule.

2.2 Expected Payment

A quantity central to our analysis is the *expected payment*, where the expectation is from the point of view of the worker and is taken over the randomness in the choice of the G gold standard questions among the N questions and over the N probability distributions representing her beliefs for the N questions. Let us formalize this notion. Suppose that for question $i \in [N]$, let $y_i \in [B]$ denote the number of options selected by the worker. Further, let $s_i \in [0,1]$ denote the probability, under the worker's beliefs, that the correct answer to question i lies in this set of y_i selected options. In other words, s_i denotes the sum of the beliefs for the y_i options selected by the worker: If the worker selects options $\ell_1, \ldots, \ell_{y_i}$, then $s_i = (p_{i\ell_1} + \cdots + p_{i\ell_{y_i}})$. Then, from the worker's point of view, her expected payment for this selection equals

$$\frac{1}{\binom{N}{G}} \sum_{(j_1,\ldots,j_G)\subseteq[N]} \sum_{(\epsilon_1,\ldots,\epsilon_G)\in\{-1,1\}^G} \left(f(\epsilon_1 y_{j_1}, \ldots, \epsilon_G y_{j_G}) \prod_{i=1}^{G} (1 - s_{j_i})^{\mathbf{1}\{\epsilon_i=-1\}} s_{j_i}^{\mathbf{1}\{\epsilon_i=1\}} \right). \tag{2}$$

The outer summation in Equation (2) corresponds to the expectation with respect to the random distribution of the G gold standard questions in the N total questions. The inner summation in Equation (2) corresponds to the expectation with respect to the worker's beliefs of her choices being correct. In more detail, the values $\epsilon_1, \ldots, \epsilon_G$ iterate over all possible events regarding whether or not the correct answer lies in the selected set of options for each of the G gold standard questions, and the term $\prod_{i=1}^{G}(1 - s_{j_i})^{\mathbf{1}\{\epsilon_i=-1\}} s_{j_i}^{\mathbf{1}\{\epsilon_i=1\}}$ represents the probability (according to the worker's beliefs) of the occurrence of each such event. For instance, if the worker selects all options ($y_i = B$) for every question $i \in [N]$, then the correct answer must necessarily lie in the set of selected options (that is, $s_i = 1$ for every $i \in [G]$), and then the payment evaluates to exactly $f(B, \ldots, B)$.

Given the presence of gold standard questions, the performance of any worker can be verified based on only her own answers (without depending on the answers of other workers). The payments made to different workers thus do not depend on each other, and hence, we consider only one worker without loss of generality.

In this article, we assume that the worker aims to maximize her expected payment, where the expectation is taken with respect to the randomness in the choice of the G standard questions, and also with respect to the beliefs of the worker regarding the correctness of various options for each question.

2.3 Goal

At a high level, our goal is to incentivize the worker, for each question, to select all options that she believes are quite likely to be correct. We consider two goals based on whether one wishes to elicit options whose beliefs are above a pre-defined threshold in an *absolute* or a *relative* sense.

2.3.1 *Absolute Thresholding.*

In this setting, we assume to be given the value of a parameter $\sigma \in (0,1)$. Then, for every question $i \in [N]$, we want the worker to select precisely the set of options

$$\{b \in [B] \mid p_{ib} > \sigma\}, \tag{3a}$$

while not selecting the options

$$\{b \in [B] \mid p_{ib} < \sigma\}. \tag{3b}$$

Then, the goal is to design payment mechanisms that incentivize this action.

Definition 1 (Strictly Proper Scoring Rule for Absolute Thresholding). A payment function is a strictly proper scoring rule for absolute thresholding if the expected payment (2) from the worker's point of view is strictly maximized when she selects all options in the set (3a) and no option in the set (3b) for every question $i \in [N]$.

The worker is allowed to act either way for options for which her belief is exactly σ. We do not impose a requirement from the scoring rule when the worker's belief equals σ for some option, since this is a boundary case that is impossible to incentivize (see Appendix A.2 for a formal proof of this impossibility).

We address the setting of absolute thresholding in Section 3.

2.3.2 Relative Thresholding. Consider the following two scenarios on the beliefs of a worker for a question with $B = 10$ options: Scenario I: the beliefs are $(1/4, 1/4, 1/16, \ldots, 1/16)$ for the B options; Scenario II: the beliefs are $(2/3, 1/3, 0, \ldots, 0)$. Then, in Scenario I, one may wish to have the worker select the first two options, since the other options have quite low probabilities. In Scenario II, one may wish to have the worker select the first option but not the second due to the significant difference between the beliefs pertaining to the first and the second options. However, no fixed absolute threshold σ for the setting of Section 2.3.1 can meet both these requirements simultaneously. Hence, in this section, we consider an alternate requirement based on relative thresholding: The worker is incentivized to select options one-by-one, selecting an option if and only if it contributes sufficiently to the belief relative to the already selected options.

Formally, the requirement is associated to a pre-specified value $\rho \in (0, 1)$. Consider any question $i \in [N]$. Let $\{(1), \ldots, (B)\}$ denote an ordering of the B options such that the worker's beliefs for question i follow this order, that is, $p_{i(1)} \geq \cdots \geq p_{i(B)}$ with ties broken arbitrarily by the worker. Then, we want the worker to select precisely the set of options

$$\left\{ b \in [B] \;\middle|\; \frac{p_{i(b)}}{\sum_{\ell=1}^{b} p_{i(\ell)}} > \rho \right\}, \tag{4a}$$

while not selecting the options

$$\left\{ b \in [B] \;\middle|\; \frac{p_{i(b)}}{\sum_{\ell=1}^{b} p_{i(\ell)}} < \rho \right\}. \tag{4b}$$

In words, after selection of the most likely option, the remaining options must be selected one-by-one in decreasing order of the beliefs as long as the selected option contributes a fraction more than ρ to the total belief of the selected options. For instance, both situations described earlier in this subsection can be accommodated simultaneously with the choice $\rho = 0.4$. Note that as in the setting of absolute thresholding (Section 2.3.1), we do not impose any requirement if the relative value of the belief in Equation (4) is at the boundary equaling exactly ρ.

Definition 2 (Strictly Proper Scoring Rule for Relative Thresholding). A payment function is a strictly proper scoring rule for relative thresholding if the expected payment (2) from the worker's point of view is strictly maximized when she selects all options in the set (4a) and no option in the set (4b) for every question $i \in [N]$.

We address the setting of relative thresholding in Section 4.

3 ABSOLUTE THRESHOLDING

In this section, we consider the setting of incentivizing the worker to select all options for which her belief is strictly greater σ, for some fixed parameter $\sigma \in (0,1)$, as detailed in Section 2.3.1. Before proceeding, we must specify certain pedantic details of the problem setting. Let us define two integers $s_{\min}$ and $s_{\max}$ as $s_{\min} = \mathbf{1}\{\sigma < \frac{1}{B}\}$ and $s_{\max} = \min\{\lceil \frac{1}{\sigma} \rceil - 1, B\}$. Consider any question. Observe that if if $\sigma < \frac{1}{B}$, then it is meaningless to let the worker select zero options, since the belief for at least one option must be $\frac{1}{B}$ or higher. Also observe that for any value of $\sigma \in (0,1)$, it is meaningless to allow the worker to select $\lceil \frac{1}{\sigma} \rceil$ or more options, since it is mathematically impossible for those many options to have probabilities more than σ. As a result, we will mandate the worker to select at least $s_{\min}$ and at most $s_{\max}$ options for any question. Letting $x_1, \ldots, x_G$ denote the evaluations of the workers' answers to the G gold standard questions (recall from Section 2.1), the goal thus is to design the payment function $f(x_1, \ldots, x_G)$ when $|x_i| \in \{s_{\min}, \ldots, s_{\max}\}$ for every $i \in [G]$. Finally, we note that if $B = 2$ or if $\sigma \geq \frac{1}{2}$, then the setting degenerates to the "skip-based" single-selection setting studied in Shah and Zhou [55]. Hence, we consider the regime $B \geq 3$ and $\sigma \in (0, \frac{1}{2})$ throughout this section.

3.1 Proposed Scoring Rule

Our proposed scoring rule for the setting of this section is provided as Scoring rule 1. For convenience of notation, we denote this scoring rule as $f^{\#}$.

Scoring rule 1: Strictly proper scoring rule for absolute thresholding

- **Input:** Evaluations of the answers to the G gold standard questions $(x_1, \ldots, x_G)$
- **Output:** The payment

$$f^{\#}(x_1, \ldots, x_G) = \kappa^{\#} \prod_{i=1}^{G} \Big((B - |x_i| - 1)\sigma + \mathbf{1}\{x_i \geq 1\} \Big) + \alpha_{\min},$$

where $\kappa^{\#} = \frac{\alpha_{\max} - \alpha_{\min}}{((B-2)\sigma + 1)^G}$

Let us interpret this scoring rule. For any question $i \in [G]$, the component $\Big((B - |x_i| - 1)\sigma + \mathbf{1}\{x_i \geq 1\}\Big)$ of the scoring rule $f^{\#}$ penalizes the selection of an incorrect option by σ and rewards the selection of the correct option by 1. The overall payment is then a product of these components over all gold standard questions. The constant $\kappa^{\#}$ simply serves to scale the payment to accommodate the $(\alpha_{\min}, \alpha_{\max})$-requirements.

The following theorem now proves guarantees associated to our scoring rule:

Theorem 3.1. *Consider any $\sigma \in (0, \frac{1}{2})$, $N \geq G \geq 1$ and $B \geq 3$. Then, Scoring rule 1 is strictly proper for absolute thresholding.*

The proof of this result first computes the expected payment in case of honest responses and then via some algebraic arguments shows that every other response must lead to a strictly smaller payment. The remainder of this subsection is devoted to this proof.

Proof of Theorem 3.1. Without loss of generality, assume that $\alpha_{\min} = 0$, since the property of a scoring rule being strictly proper is invariant to any constant shift and positive scale of the payment. We adopt the succinct notation of $\alpha := \alpha_{\max} - \alpha_{\min}$. Also recall that the term "expected payment" always refers to the expectation with respect to the worker's beliefs regarding the correctness of various options and the randomness in the choice of the G gold standard questions in the N questions.

First consider the case of $N = G = 1$. Suppose that the worker's beliefs for the B options are $p_1, \ldots, p_B$. It is easy to verify that the expected payment when the worker selects the options $\{o_1, \ldots, o_m\}$, for some m, equals

$$(B-1)\sigma + \sum_{\ell=1}^{m} (p_{o_\ell} - \sigma).$$

Consequently, the selection of any option o_i such that $p_{o_i} < \sigma$ contributes a term $p_{o_i} - \sigma < 0$ to the expected payment, whereas the selection of any option o_j such that $p_{o_j} > \sigma$ contributes a positive amount $p_{o_i} - \sigma > 0$. It follows that the payment is strictly maximized when the worker selects all options whose beliefs are greater than σ and does not select any option whose belief is lower than σ. This completes the proof for the case $N = G = 1$.

Let us now consider the case of $N = G \geq 1$. Suppose that for any question $i \in [G]$, the worker selects the m_i options $\{o_{i1}, \ldots, o_{im_i}\}$. By the independence of the beliefs of the worker across the questions, the expected payment equals

$$\prod_{i=1}^{G} \left((B-1)\sigma + \sum_{\ell=1}^{m_i} (p_{o_{i\ell}} - \sigma) \right). \tag{5}$$

Note that each term $\left((B-1)\sigma + \sum_{\ell=1}^{m_i} (p_{o_{i\ell}} - \sigma)\right)$ in the product is non-negative. Moreover, for any question $i \in [G]$, there is an action by the worker—that of selecting all options—that will make the corresponding term $\left((B-1)\sigma + \sum_{\ell=1}^{m_i} (p_{o_{i\ell}} - \sigma)\right)$ strictly positive. Since the worker aims to maximize the expected payment, the actions that maximize Equation (5) with respect to the worker's beliefs will ensure that every term in Equation (5) is strictly positive. Given this fact, if each individual component in the product Equation (5) is maximized, then the product is also necessarily maximized. Each individual component simply corresponds to the setting of $N = G = 1$ discussed earlier. Thus, calling upon our earlier result, we get that the expected payment for the case $N = G \geq 1$ is maximized when the worker selects options as desired for every question.

Let us finally consider the general case of $N \geq G \geq 1$. Recall from Equation (2) that the expected payment for the general case is a cascade of two expectations: the outer expectation is with respect to the uniformly random distribution of the G gold standard questions among the N total questions, while the inner expectation is taken over the worker's beliefs of the different questions conditioned on the choice of the gold standard questions and restricts attention to only these G questions. The arguments above for the case $N = G$ prove that every individual term in the inner expectation is maximized when the worker acts as desired. The outer expectation does not affect this argument. The expected payment is thus maximized when the worker selects options in the desired manner. □

3.2 Uniqueness

In this section, we address our second goal of choosing a strictly proper scoring rule among many possible options. In particular, we show that the core structure of Scoring rule 1 must necessarily be a part of any strictly proper scoring rule.

Theorem 3.2. *Consider any $\sigma \in (0, \frac{1}{2})$, $N \geq 1$, and any $B \geq 3$. When $G = 1$, our proposed scoring rule, Scoring rule 1, is the only possible strictly proper scoring rule up to a constant shift and positive scaling.*

The proof of this result shows that any strictly proper scoring rule f must necessarily satisfy the following four sets of equations (when $G = 1$):

$$\begin{aligned} f(m+1) &= (1-\sigma)f(m) + \sigma f(-m) \qquad \text{for all } m \in \{1,\dots,s_{\max}-1\},\\ f(m+2) &= (1-2\sigma)f(m) + 2\sigma f(-m) \qquad \text{for all } m \in \{1,\dots,s_{\max}-2\},\\ f(-s_{\max}) &= f(s_{\max}) - f(s_{\max}-1) + f(-(s_{\max}-1)), \qquad \text{and}\\ f(0) &= \sigma f(1) + (1-\sigma)f(-1). \end{aligned}$$

These four sets of conditions together leave only two degrees of freedom for the choice of the payment function f and hence uniquely characterize the scoring rule up to a constant shift and scale. The complete proof is provided in Appendix B.1.

While we do not have a complete answer as to what the "best" or "unique" scoring rule is for general values of N and G, but (based on analogous results in Section 4) we conjecture that Scoring rule 1 may possess more attractive uniqueness and/or optimality properties.

4 RELATIVE THRESHOLDING

In what follows, we first present our proposed strictly proper scoring rule for this problem and subsequently derive several appealing properties for our scoring rule.

4.1 Proposed Strictly Proper Scoring Rule

We begin by presenting our proposed scoring rule, denoted by f^*, as Scoring rule 2.

Scoring rule 2: Strictly proper scoring rule for relative thresholding

- **Input:** Evaluations of the answers to the G gold standard questions $(x_1,\dots,x_G)$
- **Output:** The payment

$$f^*(x_1,\dots,x_G) = \kappa^* \prod_{i=1}^{G} \left((1-\rho)^{|x_i|}\, \mathbf{1}\{x_i \ge 1\}\right) + \alpha_{\min},$$

where $\kappa^* = \frac{\alpha_{\max}-\alpha_{\min}}{(1-\rho)^G}$

The payment is based only on the evaluation of the worker's responses to the gold standard questions. It is easy to describe the scoring rule in words: The payment is $\alpha_{\min}$ plus

(i) 0 if the correct answer is not selected for one or more questions, otherwise
(ii) κ^* reduced by $(100\rho)\%$ for each option selected.

Part (i) of the payment rule is a result of the indicator function $\mathbf{1}\{x_i \ge 1\}$ in the description of Scoring rule 2. Part (ii) arises due to the term $(1-\rho)^{|x_i|}$. The term κ^* is only used to ensure that the $(\alpha_{\max}, \alpha_{\min})$-conditions are satisfied. The following theorem shows that this multiplicative scoring rule is indeed strictly proper:

Theorem 4.1. *Consider any $\rho \in (0,1)$, $N \ge G \ge 1$, and $B \ge 2$. Then, Scoring rule 2 is strictly proper for relative thresholding.*

The proof of Theorem 4.1 is provided in Appendix B.2. The proof first computes the expected payment when the worker selects options as desired and then, by means of some careful algebraic arguments, shows that any other selection of options will lead to a strictly lower payment.

4.2 Additional Properties

We now address towards our second goal by providing several additional appealing properties of our proposed Scoring rule 2.

4.2.1 An Axiomatic "Uniqueness" Derivation.

We first characterize our proposed scoring rule with an axiomatic derivation that proves the "uniqueness" of the scoring rule.

The derivation involves a "no-free-lunch axiom" of Shah and Zhou [55], which we adapt to our approval-voting–based setting. We say that the response of a worker to a question is "wrong" if the correct option does not lie in the set of options that the worker selected for that question. Moreover, we say that a worker has "attempted" a question if the worker selects at least one option and also leaves out at least one option—that is, when the worker provides some distinguishing information across the options. With these preliminaries in place, the no-free-lunch axiom is defined as follows:

Definition 3 (No-free-lunch axiom; adapted from [55]). If the response to every attempted question in the gold standard turns out to be wrong, then the worker gets the minimum payment, namely,

$$f(x_1,\ldots,x_G) = \alpha_{\min} \forall (x_1,\ldots,x_G) \in \{-(B-1),\ldots,-1,B\}^G \backslash \{B\}^G.$$

First, observe that the no-free-lunch axiom is a very mild condition. For instance, even if a worker selects $\frac{B}{2}$ options uniformly at random for each question, there is only a $\frac{1}{2^G}$ chance that the no-free-lunch axiom will come into play. Second, observe that we do not impose any restriction under the event $(x_1,\ldots,x_G) = \{B\}^G$. Imposing the no-free-lunch condition on this event would only make the no-free-lunch requirement stronger, and it follows from Theorem 4.2 below that it is impossible for any strictly proper scoring rule to satisfy this stronger requirement. That said, we show later in Theorem 4.5(a) that our scoring rule is optimal for this event as well.

We now come to the main result of this section.

Theorem 4.2. *Scoring rule 2 is the only strictly proper scoring rule for relative thresholding that satisfies the no-free-lunch axiom.*

The proof of Theorem 4.2 is provided in Appendix B.3. The proof relies on the following lemma, which provides a necessary condition that must necessarily be satisfied by any strictly proper scoring rule (which may or may not satisfy no-free-lunch):

Lemma 4.3. *Any strictly proper scoring rule f for relative thresholding must satisfy*

$$\begin{aligned} &f(x_1,\ldots,x_{i-1},x_i+1,x_{i+1},\ldots,x_G)\\ &\quad = (1-\rho)f(x_1,\ldots,x_{i-1},x_i,x_{i+1},\ldots,x_G) + \rho f(x_1,\ldots,x_{i-1},-x_i,x_{i+1},\ldots,x_G), \end{aligned}$$

for every $i \in [G]$ and $(x_1,\ldots,x_{i-1},x_{i+1},\ldots,x_G) \in \{-(B-1),\ldots,-1,1,\ldots,B\}^{G-1}$, $x_i \in [B-1]$.

Note that the lemma does *not* use the no-free-lunch condition. In words, the lemma considers any arbitrary evaluations of the answers to the gold standard questions $\{1,\ldots,i-1,i+1,G\}$. Fixing the evaluations of these questions, it says that the payment under any strictly proper scoring rule for the evaluation $(x_{i+1}+1)$ of question i must be a convex combination of the payments under the evaluations x_i and $-x_i$ for question i. Moreover, the coefficient in that convex combination must precisely equal the parameter ρ.

The proof of this lemma is provided in Appendix B.3.1. A repeated application of this lemma, along with the no-free-lunch axiom, leads to the result of Theorem 4.2.

4.2.2 Eliciting Support. This section is motivated by a proposition by Coombs [9] that *"Individuals taking the test should be instructed to cross out all the alternatives which they consider wrong."* With this motivation, we consider incentivizing workers to not select options that they consider wrong and select the other options. More formally, for each question, we wish to incentivize the worker to select the *support* of the belief distribution over the B options:

$$\{b \in [B] \mid p_{ib} \neq 0\}, \tag{6}$$

for every question $i \in [N]$.[2] Another motivation for this requirement is its simplicity of description of the required task to the workers.

Since the support of the belief distribution must necessarily contain at least one item, we will mandate the worker to select at least one option for each question. Consequently, the scoring rules considered in this section are functions mapping the set $\{-(B-1), \ldots, -1, 1, \ldots, B\}$ to the interval $[\alpha_{\min}, \alpha_{\max}]$.

In Appendix A.1 (Proposition A.1), we show that there is no strictly proper scoring rule that can achieve this exact goal. However, we make an appeal to literature in psychology to make an additional assumption that will allow us to address this problem. There is an extensive literature in psychology establishing the coarseness of processing and perception in humans. For instance, Miller's celebrated paper [40] establishes the information and storage capacity of humans: that an average human being can typically distinguish at most about seven states. This granularity of human computation is verified in many subsequent experiments [50, 56]. Jones and Loe [28] establish the ineffectiveness of finer-granularity response elicitation. Mullainathan et al. [42] hypothesize that humans often group things into categories; this hypothesis is experimentally verified by Siddiqi [57] in a specific setting. In the same spirit of coarseness of human processing, we make the following "coarse beliefs" assumption:

Consider some (fixed and known) value $\rho > 0$, and assume that the probability of any option for any question, according to the worker's belief, is either zero or greater than ρ.[3] Since one must necessarily take into account situations when a worker is totally clueless about a question—that is, when her belief is distributed uniformly over all options—we restrict $\rho < \frac{1}{B}$. To summarize, we make the following "coarse belief" assumption.

Definition 4 (Coarse Belief Assumption). The worker's belief for any option for any question lies in the set $\{0\} \cup (\rho, 1]$ for some (fixed and known) value $\rho \in (0, \frac{1}{B})$.

Our goal is to design strictly proper scoring rules for support elicitation under the coarse beliefs assumption. One choice is to use Scoring rule 1 by setting $\sigma = \rho/2$; this will be strictly proper. However, we will pursue a more interesting direction here. We show that our Scoring rule 2 also fits the bill—the proof of strict properness follows as a corollary of Theorem 4.1, where we show that eliciting the support under the coarse beliefs assumption is a special case of the setting of relative thresholding. We then show in subsequent sections that this scoring rule possesses a number of additional attractive properties in terms of this goal.

Corollary 4.4. *Any strictly proper scoring rule for relative thresholding is also strictly proper for eliciting support of the worker's beliefs for every question under the coarse beliefs assumptions. Consequently, Scoring rule 2 is strictly proper with respect to eliciting the support of the worker's belief under the coarse beliefs assumption.*

The proof of this corollary is provided in Appendix B.4.

[2] This is of course a stylized requirement where, by a zero belief, we are considering extremely low probabilities that are treated practically as zero by the worker.

[3] It will be clear in a moment that the choice of notation "ρ" here is not a coincidence.

4.2.3 Performance-based Payments. Recall that we mandate a payment of $\alpha_{\max}$ to a worker who answers perfectly ($f(1,\ldots,1) = \alpha_{\max}$), and we also mandate a payment of at least $\alpha_{\min}$ to every worker ($f(x) \in [\alpha_{\min}, \alpha_{\max}]$ for every x). With the amounts $(\alpha_{\max}, \alpha_{\min})$ fixed, the requester may wish to ensure a performance-based payment, which ensures a higher pay for a good work relative to others. Given a certain monetary budget, such a performance-based payment will ensure that most payment is made for better performance, thereby being able to support good work and deterring spammers. These implications would in turn benefit the good workers in the long run, since they would receive a greater share of the requester's budget. In this section, we explore the limits of this notion of a performance-based payment.

Consider the problem of eliciting the support of the worker's beliefs for each question under the coarse beliefs assumption. We now establish certain properties of our scoring rule f^* (Scoring rule 2) as compared to any other strictly proper scoring rule f.

Recall that for every set of answers $(x_1,\ldots,x_G) \notin \{1,\ldots,B\}^G$ that contain at least one incorrect answer, we have $f(x_1,\ldots,x_G) \geq \alpha_{\min} = f^*(x_1,\ldots,x_G)$. When a question has the correct answer selected, it is intuitive to see that the information obtained decreases as more options are selected. With this intuition, the first part of the following theorem considers the event where all options are selected for all questions—this event provides no distinguishing information between the options.[4] The remaining three parts of the theorem are discussed after its statement.

Theorem 4.5. *Consider either of the following two settings: (i) relative thresholding or (ii) eliciting support under the coarse beliefs assumption. Let f denote any arbitrary strictly proper scoring rule and f^* denote Scoring rule 2.*

(a) *For any $N \geq G$, it must be that*

$$f(B,\ldots,B) > f^*(B,\ldots,B).$$

(b) *When $N = G$, for any $(x_1,\ldots,x_G) \in \{-(B-1),\ldots,-1,1,\ldots,B\}^G$, it must be that*

$$f(x_1,\ldots,x_G) \geq f^*(x_1,\ldots,x_G).$$

(c) *For any $N \geq G$, and any $(x_1,\ldots,x_G) \in \{-(B-1),\ldots,-1,1,\ldots,B\}^G \backslash \{1,\ldots,B\}^G$, it must be that*

$$f(x_1,\ldots,x_G) \geq f^*(x_1,\ldots,x_G).$$

(d) *For any $N \geq G$, consider any value $\gamma \in [G]$. Suppose that f satisfies*

$$f(x_1,\ldots,x_G) = f^*(x_1,\ldots,x_G)$$

for every $(x_1,\ldots,x_G) \in \{-(B-1),\ldots,-1,1,\ldots,B\}^G$ such that $\sum_{i=1}^G \mathbf{1}\{x_i = 1\} \geq \gamma$. Then, it must be that

$$f(x_1',\ldots,x_G') \geq f^*(x_1',\ldots,x_G')$$

for every $(x_1',\ldots,x_G') \in \{-(B-1),\ldots,-1,1,\ldots,B\}^G$ such that $\sum_{i=1}^G \mathbf{1}\{x_i' = 1\} \geq \gamma - 1$.

Let us now parse parts (b), (c), and (d) of Theorem 4.5. Part (b) demonstrates that our scoring rule is pointwise the most frugal among all strictly proper scoring rules under the setting $N = G$. Although the setting of $N = G$ is not directly useful for crowdsourcing in practice, it is important, since it forms the basis for all the theoretical guarantees. Part (c) is simple—it simply says that our scoring rule pays minimum when one or more answers are incorrect. Finally, part (d) then goes

[4] In fact, in our initial presentations on this work, the most frequently asked question pertained to the performance of the scoring rule under this event.

on to show that any other strictly proper scoring rule that pays the same amount as our scoring rule for a certain quality of work must pay at least as much as our scoring rule for a worse quality. The notion of quality in the statement of this part is determined by the number of questions to which the worker chose only the correct answer $\sum_{i=1}^{G} \mathbf{1}\{x_i = 1\}$. The complete proof Theorem 4.5 is provided in Appendix B.5.

The performance-based payments we studied here are quantitatively different from the no-free-lunch axiom from Section 4.2.1. Despite this difference, somewhat surprisingly, both these notions lead to the same strictly proper scoring rule—Scoring rule 2.

5 PRELIMINARY EXPERIMENTS

In this section, we present preliminary experiments on the Amazon Mechanical Turk crowdsourcing platform (mturk.com). We consider the problem of eliciting the support of the beliefs (studied in Section 4.2.2) due to its motivation in literature [9] and the simplicity of description of the task to the workers. The goal of this preliminary experimental exercise is to perform a basic check on whether our proposed strictly proper scoring rule (Scoring rule 2) has the potential to work in practice, and identify any possible red flags. Specifically, our goal is to evaluate the primary hypotheses underlying the theory in terms of the following three questions:

(Q1) Are workers are able to make judicious use of the approval voting setup?
(Q2) Does the presence of the strictly proper scoring rule make any difference as compared to having a proper scoring rule that is not *strictly* proper?
(Q3) Is there an opposition from the workers to the approval voting interface or to our proposed scoring rule for any reason?

It is important to keep in mind that conclusive experiments for mechanism design are, in general, quite expensive with respect to time (workers may need months to understand a new mechanism) and budget. They are unlike typical machine learning experiments that require only existing benchmark datasets. Moreover, the wordings or the interface may exert a significant influence on the workers' behavior. We position our work primarily as a theoretical study. We hope that more detailed experiments will follow the publication of our work; indeed, it is best if experiments on such incentive schemes are conducted by multiple groups.

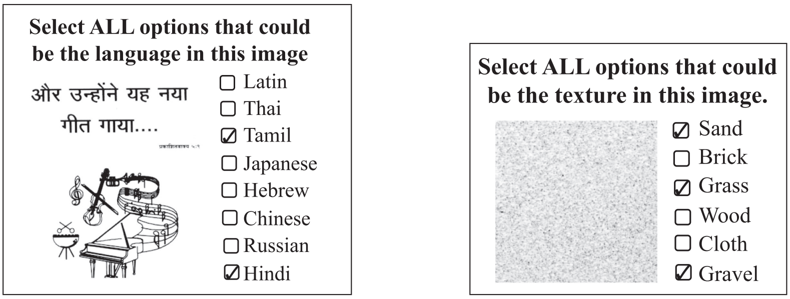
5.1 Methods

Experiments: We conducted three separate sets of experiments, with over 200 workers in each experiment:

- Identifying languages from displayed text (Figure 1)
- Identifying animals in displayed images (Figure 2(a))
- Identifying textures in displayed images (Figure 2(b)).

The experiments on identifying languages, animals, and textures contained ($N =$) 25, 16, and 16 questions, respectively, which included ($G =$) 4, 3, and 3 gold standard questions, respectively, and each question had ($B =$) 8, 6, and 6 options, respectively. We had a total of 213, 203, and 203 workers, respectively, in the three experiments.

Interfaces and scoring rules: In each experiment, we compared four different settings as described below. Each of the four settings is associated to a certain interface and a certain scoring rule. In each experiment, every worker was assigned one of the four settings uniformly at random. Each worker was made a fixed payment of $\alpha_{\min} = 10$ cents, and the scoring rule was conducted in terms of a bonus payment. Specifics of the four interfaces and scoring rules are as follows:


Select ALL options that could be the language in this image
और उन्होंने यह नया गीत गाया....
□ Latin
□ Thai
☑ Tamil
□ Japanese
□ Hebrew
□ Chinese
□ Russian
☑ Hindi
Select ALL options that could be the texture in this image.
☑ Sand
□ Brick
☑ Grass
□ Wood
□ Cloth
☑ Gravel

(a) (b)

Fig. 2. Illustration of two of the experiments we conducted on Amazon Mechanical Turk.

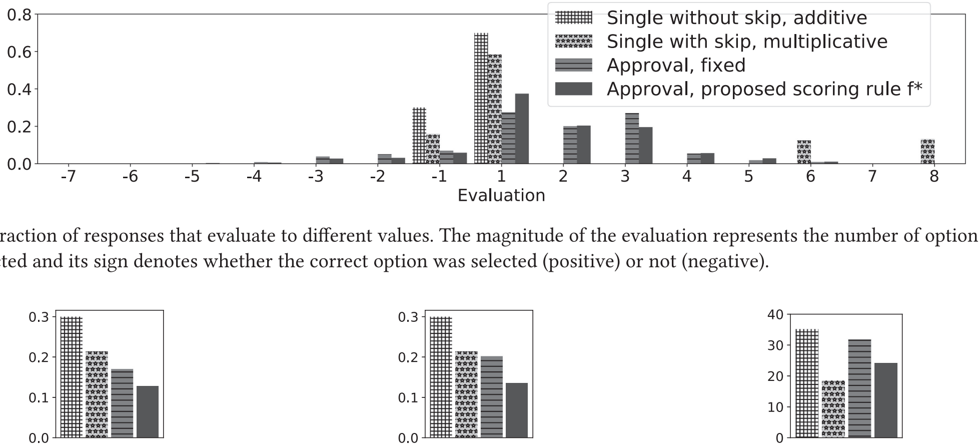
- Single-selection interface with additive payments: The worker must select exactly one option for each question. The bonus starts at zero and is increased (additively) by a fixed amount X for every correct answer in the gold standard, where X equals 12, 15, and 15 for the experiments on identifying languages, animals, and textures, respectively. This is a strictly proper scoring rule for the single-selection interface.
- Skip-based single-selection interface with multiplicative payments [55]: For each question, the worker can select either one option or can skip the question. The bonus starts at 70, 50, and 50 cents for the experiments on identifying languages, animals, and textures, respectively, is multiplied by 0.5, 0.6, and 0.6, respectively, for each correct answer in the gold standard, and becomes zero if one or more questions in the gold standard (that are not skipped) are answered incorrectly. As shown by Shah and Zhou [55], this is a strictly proper scoring rule for the skip-based single-selection interface.
- Approval-voting interface with a fixed payment: The worker can choose any number of options for each question. The bonus is fixed at 35, 30, and 30 cents for the experiments on identifying languages, animals, and textures, respectively. This is a proper scoring rule, but is not strictly proper.
- Approval-voting interface with Scoring rule 2: The worker can choose any number of options for each question. We set $\rho = 0.1$ for each experiment and set $\alpha_{\max}$ as 80, 60, and 60 cents for the experiments on identifying languages, animals, and textures, respectively. As shown in Corollary 4.4 (and Theorem 4.1), this is a strictly proper scoring rule.

Interface: Given the caveats associated to experiments on mechanism design as mentioned earlier, we provided detailed instructions about the task and the scoring rule to each worker and also made them work through multiple examples. In each experiment and for each scoring rule, the workers' interface contained the following details (in the order listed below):

- A higher level instruction for the task (Figure 4(a)).
- Detailed instructions and an example for the task that the worker must work out (Figure 4(b)).
- Description of the scoring rule—that is, instructions for the bonus payment (Figure 4(c)).
- Three examples illustrating the scoring rule (Figure 4(d)).

These instructions were followed by the actual set of questions that the worker had to answer.

Additional details and illustrations on the interface presented to the worker are provided in Figure 4 in Appendix C. The entire data related to the experiments, including the interfaces used,



(a) Fraction of responses that evaluate to different values. The magnitude of the evaluation represents the number of options selected and its sign denotes whether the correct option was selected (positive) or not (negative).

(b) Fraction wrong among attempted questions

(c) Fraction wrong when only one option was selected

(d) Mean bonus per worker (cents)

Fig. 3. Raw data from the three experiments conducted on Amazon Mechanical Turk.

specifics about the scoring rule, and the responses of the workers, is available on the website of the first author:

https://cs.cmu.edu/~nihars/data/data_approval.zip

5.2 Results

Let us begin by eyeballing the raw data obtained from the experiments. Figure 3 presents combined results from the three experiments. Figure 3(a) shows the breakup of the evaluations of all the collected responses. The magnitude of the evaluation represents the number of options selected and its sign denotes whether the correct option was selected (positive) or not (negative). Figure 3(b) depicts the fraction of responses to attempted questions (where the worker did not skip or choose all options) that turned out to be wrong (the correct option was not selected). Figure 3(c) depicts the fraction of responses that were correct when only one option was selected. Figure 3(d) depicts the mean payment per worker. Using this data, let us now investigate the three questions posed at the beginning of the experiments:

(Q1) Are workers able to make judicious use of the approval voting setup? One can observe from Figure 3(a) more than 40% responses comprised a selection of two or three options, suggesting that the workers did use approval voting.

(Q2) Does the presence of the strictly proper scoring rule make any difference as compared to having a proper scoring rule that is not *strictly* proper? We compare the data from the approval voting setup under the fixed payment setting (which is a proper, but not a strictly proper scoring rule) with the data from the approval voting setup under Scoring rule 2. To test for a possible difference, we employ Hotelling's T-squared test [25] that is used to test whether two multivariate data samples have the same mean (the null hypothesis) or have different means (the alternative hypothesis). In applying this test, we treat the response by any worker to any question as a two-dimensional data point, with the number options selected and the correctness of the answer as the two dimensions. The results of this test are listed in the following table, which lists values of the parameters T^2, F, and p associated to the test.

Experiment	T^2	F	p
Languages	15.7	7.8	0.0004
Textures	21.3	10.7	0.000025
Animals	10.2	5.1	0.0062

We could reject the null hypothesis (that the two sets of data are drawn from distributions with identical means) with $p < 0.01$ for each of the three experiments, thereby concluding a statistically significant difference between the data between the proper scoring rule (fixed payment) and the strictly proper scoring rule (Scoring rule 2).

(Q3) Is there is an opposition from the workers to the approval voting interface or our proposed scoring rule for any reason? We also elicited feedback about the task from every worker. In more detail, we presented the worker with a text box at the end of the task, asking the worker if they had any feedback for us. We also clearly stated that the feedback will not affect their payment in any way. We received mostly neutral feedback, some positive feedback, and no negative feedback about either the approval voting interface or our scoring rule.

All in all, these preliminary experiments did not present any "red flag" towards our scoring rule, while indicating a potential to be useful for practical applications such as collecting labeled data for machine learning.

A concluding remark: A standard means of denoising data from crowdsourcing is to ask every question to multiple workers and employ a statistical aggregation algorithm to aggregate the data so obtained. A future goal is to evaluate the performance of our proposed interface and scoring rule on such aggregated data. This can be accomplished upon the resolution of the open problem of designing algorithms for statistical aggregation of data collected through this interface and the proposed scoring rules.

6 DISCUSSION AND OPEN PROBLEMS

Our goal is to deliver high-quality labels for machine learning applications, at low costs, by means of incentive mechanisms or aggregation algorithms or both. In this article, we pursue the former approach. We take an approval-voting–based means of gathering labeled data from crowdsourcing. We design scoring rules via a principled theoretical approach and prove appealing properties of optimality and uniqueness of our proposed scoring rules. Preliminary experiments conducted on Amazon Mechanical Turk corroborate the usefulness of our scoring rules for practical scenarios. Our scoring rules may also draw more experts to the crowdsourcing platform, since their compensation will be significantly higher than that of mediocre workers, unlike most compensation mechanisms in current use.

We conclude with a discussion on closely related topics that merit investigation in the future.

Aggregation of labels. For the traditional single-selection setting, there is a long, existing line of work on statistical methods to aggregate redundant noisy data from multiple workers [7, 12, 26, 30, 37, 47, 62, 67]. An open problem is the design of aggregation algorithms for approval-voting–based data: algorithms that can exploit the specific structure of the responses that arise as a result of the approval voting interface and the proposed scoring rule. There is indeed work on aggregation algorithms [3, 6, 39, 46] and probabilistic models [14, 15, 38, 48] for approval-voting in the context of social choice theory; their objective, however, is primarily of fairness and strategyproofing of the voting procedure, as opposed to our goal of denoising data obtained from multiple heterogeneous workers as required for labeling tasks in crowdsourcing.

Choosing the right interface. There are tradeoffs between various interfaces for crowdsourcing. For instance, the approval voting interface elicits one or more options with high enough values of the belief (e.g., the support of the belief), whereas the single selection interface elicits the mode. Choosing among these two interfaces would depend on the application under consideration, and moreover, one may adaptively switch between the two depending on the data obtained. A natural question that one may further ask is, why not elicit the entire belief distribution itself? While the entire belief distribution seems to supersede the support and the mode, providing the distribution will also require considerably more time and effort from the workers and often also suffer from a higher noise (e.g., see Reference Shah et al. [52] for empirical evaluations revealing greater noise in cardinal scores as compared to choosing among options). These tradeoffs must be taken into account when choosing the interface for the application at hand.

Choice of parameters σ and ρ. We assumed that the parameters σ and ρ are given to us. In practice, one would have to make a principled choice of these parameters and a problem of interest is to design an adaptive strategy to tune these parameters across tasks to obtain the best data for a given budget.

Other applications beyond crowdsourcing. More generally, questions on choosing the right interfaces and incentive schemes are vital in various applications that rely on evaluations performed by people, such as peer review [43, 54, 63]. It will be fruitful to explore how ideas from crowdsourcing could potentially lead to theoretically and/or practically impactful solutions for such a broader variety of applications.

APPENDICES

A AUXILIARY NEGATIVE RESULTS

In this section, we present a pair of auxiliary results that were referred to in the main text of the article.

A.1 Impossibility of Support Elicitation without Additional Assumptions

In the setting of eliciting support of beliefs (Section 4.2.2), we made a coarse beliefs assumption that the probability of correctness of any option, according to the worker's belief, must either be zero or exceed a certain threshold ρ. The following proposition shows that there exists no strictly proper scoring rule in the absence of this assumption:

Proposition A.1. *For any $N \geq G \geq 1$ and $B \geq 2$, there is no strictly proper scoring rule towards incentivizing the worker to select precisely the support of her distribution for each question in the absence of any additional assumption.*

The proof of this result is provided in Appendix B.6. To put this negative result in perspective with the positive results of Section 4.2.2, observe that $\rho = 0$ reduces Scoring rule 2 to $f^*(x_1, \ldots, x_G) = \kappa^* \prod_{i=1}^{G} \mathbf{1}\{x_i \geq 1\} + \alpha_{\min}$. One can see that it no longer remains a strictly proper scoring rule: The worker is incentivized to simply select all options for every question. The impossibility result of Proposition A.1 proves that every other scoring rule must necessarily suffer this fate.

A.2 Impossibility of Absolute Thresholding When a Belief Exactly Equals the Threshold

Recall that when defining a strictly proper scoring rule for the setting of eliciting options with beliefs above a certain threshold σ (Section 2.3.1), we did not restrict the scoring rule to any specific choice when the probability of the correctness of an option equaled exactly σ. This is because, as one would intuitively expect, incentivizing a certain action at the boundary value of σ may not be possible. The following proposition provides a formal proof for this claim:

Proposition A.2. *For any $N \geq G \geq 1$, in the setting of absolute thresholding, there is no strictly proper scoring rule to incentivize the worker to select or reject an option when the worker's belief exactly equals the threshold.*

The proof of this result is provided in Appendix B.7.

B PROOFS

In this section, we present proofs of the various theoretical results presented in the article.

B.1 Proof of Theorem 3.2: Uniqueness of Scoring Rule 1

Let f denote any strictly proper scoring rule for absolute thresholding. Consider any $m \in \{1, \ldots, s_{\max} - 1\}$. Consider the set of beliefs $p_1 = \sigma + \delta, p_2 = \cdots = p_{m+1} = \frac{1-\sigma-\delta}{m}$ and $p_{m+2} = \cdots = p_B = 0$, for some value of δ in the neighborhood of 0. For the values of m under consideration, one can verify that $\sigma < \frac{1-\sigma}{m} < 1$. Consequently, there exists some value $\delta_{\max} > 0$ such that for every $\delta \in [-\delta_{\max}, \delta_{\max}]$, we have $0 \leq \sigma + \delta \leq 1$ and $\sigma < \frac{1-\sigma-\delta}{m} \leq 1$. To achieve the stated goal, we would thus require to incentivize the worker to select options 1 through $(m+1)$ if $\delta > 0$ and select options 2 through $(m+1)$ if $\delta < 0$. The scoring rule f therefore must satisfy the pair of inequalities

$$f(m+1) \underset{\delta>0}{\overset{\delta<0}{\lessgtr}} (1-\sigma-\delta)f(m) + (\sigma+\delta)f(-m).$$

Since the right-hand side of the expression above is linear in δ but the left-hand side is a constant, we must have

$$f(m+1) = (1-\sigma)f(m) + \sigma f(-m) \qquad \text{for all } m \in \{1, \ldots, s_{\max} - 1\}. \tag{7}$$

We will return to this set of equations later.

Next consider any $m \in \{1, \ldots, s_{\max} - 2\}$. Consider the set of beliefs $p_1 = \sigma + \delta$, $p_2 = \sigma + \delta$, $p_3 = \cdots = p_{m+2} = \frac{1-2\sigma-2\delta}{m}$, and $p_{m+3} = \cdots = p_B = 0$, for some value of δ in the neighborhood of 0. For the values of m under consideration, one can verify that $\sigma < \frac{1-2\sigma}{m} < 1$. Consequently, there exists some value $\delta_{\max} > 0$ such that for every $\delta \in [-\delta_{\max}, \delta_{\max}]$, we have $0 \leq \sigma + \delta \leq 1$ and $\sigma < \frac{1-2\sigma-2\delta}{m} \leq 1$. To achieve the stated goal, we would thus require to incentivize the worker to select options 1 through $(m+2)$ if $\delta > 0$ and select options 3 through $(m+2)$ if $\delta < 0$. The scoring rule f thus must satisfy

$$f(m+2) \underset{\delta>0}{\overset{\delta<0}{\lessgtr}} (1-2\sigma-2\delta)f(m) + (2\sigma+2\delta)f(-m).$$

Since the right-hand side of the expression above is linear in δ but the left-hand side is a constant, we must have

$$f(m+2) = (1-2\sigma)f(m) + 2\sigma f(-m) \qquad \text{for all } m \in \{1, \ldots, s_{\max} - 2\}. \tag{8}$$

It follows from Equations (7) and (8) that the values of $f(m)$ for every $m \in \{-(s_{\max} - 1), \ldots, -1, 1, \ldots, s_{\max} - 2\}$ can be expressed in terms of a linear combination of $f(s_{\max})$ and $f(s_{\max} - 1)$. We will now prove that the same holds true for $f(-s_{\max})$ and $f(0)$ as well, whenever these quantities are defined.

The quantity $f(-s_{\max})$ is defined only when $s_{\max} < B$. The reason is that when $s_{\max} = B$, $f(-s_{\max}) = f(-B)$ corresponds to a scenario where all the options are selected and the correct option is not, which is impossible. Now consider the set of beliefs $p_1 = \sigma + \delta$, $p_2 = \cdots = p_{s_{\max}} =$

$\frac{1-\sigma-\delta-\epsilon}{s_{\max}-1}$, $p_{s_{\max}+1} = \epsilon$, and $p_{s_{\max}+2} = \cdots = p_B = 0$, for some values of $\epsilon \geq 0$ and δ in the neighborhood of 0. From the definition of $s_{\max}$, one can easily verify that $\sigma < \frac{1-\sigma-\epsilon}{s_{\max}-1} < 1$ whenever $s_{\max} > 1$. Consequently, there exist some values $\delta_{\max} > 0$ and $\epsilon_{\max} \in (0, \sigma)$ such that for every $\delta \in [-\delta_{\max}, \delta_{\max}]$ and for every $\epsilon \in [0, \epsilon_{\max}]$, we have $0 \leq \sigma + \delta \leq 1$ and when $s_{\max} > 1$, we also have $\sigma < \frac{1-\sigma-\delta-\epsilon}{s_{\max}-1} \leq 1$. To achieve the stated goal, we would thus require to incentivize the worker to select options 1 through $s_{\max}$ if $\delta > 0$ and select options 2 through $s_{\max}$ if $\delta < 0$. The scoring rule f therefore must satisfy

$$(1-\epsilon)f(s_{\max}) + \epsilon f(-s_{\max}) \underset{\delta>0}{\overset{\delta<0}{\lessgtr}} (1-\sigma-\delta-\epsilon)f(s_{\max}-1) + (\sigma+\delta+\epsilon)f(-(s_{\max}-1)).$$

Since the right-hand side of the expression above is linear in δ but the left-hand side does not depend on δ, we must have

$$(1-\epsilon)f(s_{\max}) + \epsilon f(-s_{\max}) = (1-\sigma-\epsilon)f(s_{\max}-1) + (\sigma+\epsilon)f(-(s_{\max}-1)).$$

Since this equation must be true for every $\epsilon \in [0, \epsilon_{\max}]$, we must have

$$-f(s_{\max}) + f(-s_{\max}) = -f(s_{\max}-1) + f(-(s_{\max}-1)).$$

Thus, the term $f(-s_{\max})$, whenever applicable, can also be written as a linear combination of $f(s_{\max})$ and $f(s_{\max}-1)$.

The quantity $f(0)$ is defined only when $\sigma > \frac{1}{B}$. The reason is that when $\sigma \leq \frac{1}{B}$, it is mathematically impossible for the beliefs for all the B options to be less than or equal to σ (recall our assumption that no belief equals exactly σ). Now consider the set of beliefs $p_1 = \sigma + \delta$, $p_2 = \cdots = p_B = \frac{1-\sigma-\delta}{B-1}$, for some value of δ in the neighborhood of 0. One can verify that in this case of $\sigma > \frac{1}{B}$, it must be that $0 < \frac{1-\sigma}{B-1} < \sigma$. Consequently, there exists some value $\delta_{\max} > 0$ such that for every $\delta \in [-\delta_{\max}, \delta_{\max}]$, we have $0 \leq \sigma + \delta \leq 1$ and $0 \leq \frac{1-\sigma-\delta}{B-1} < \sigma$. To achieve the stated goal, we would thus require to incentivize the worker to select option 1 if $\delta > 0$ and select no options if $\delta < 0$. The mechanism f therefore must satisfy

$$(\sigma+\delta)f(1) + (1-\sigma-\delta)f(-1) \underset{\delta>0}{\overset{\delta<0}{\lessgtr}} f(0).$$

Since the left-hand side of the expression above is linear in δ but the right-hand side is a constant, we must have

$$\sigma f(1) + (1-\sigma)f(-1) = f(0).$$

Thus, the term $f(0)$, whenever applicable, can also be written as a linear combination of $f(s_{\max})$ and $f(s_{\max}-1)$.

From the arguments above, we get that the design of f has only two degrees of freedom. Given that our claim is only up to some shift and scale, the claim is proved.

B.2 Proof of Theorem 4.1: Scoring Rule 2 Is Strictly Proper for Relative Thresholding

Without loss of generality, assume that $\alpha_{\min} = 0$, since in our setting, the property of being strictly proper is invariant to any constant shift and positive scale of the payment. We adopt the succinct notation of $\alpha := \alpha_{\max} - \alpha_{\min}$.

First consider the case of $N = G = 1$. For this case, Scoring rule 2 reduces to $f^*(x) = \alpha(1-\rho)^{(x_1-1)}\mathbf{1}\{x_1 \geq 0\}$. Suppose without loss of generality that the worker's beliefs for the B options are $p_1 \geq \cdots \geq p_B$ with ties broken arbitrarily by the worker.

Suppose a worker decides to select some ℓ of the B options, say, options $\{o_1, \ldots, o_\ell\} \subseteq [B]$. Then, it is easy to see that her expected payment,

$$\alpha \sum_{i=1}^{\ell} p_{o_i} (1-\rho)^{\ell-1},$$

is maximized when she selects options $\{1, \ldots, \ell\}$, i.e., the ℓ options that are most likely to be correct. Under the monotonicity $p_1 \geq \cdots \geq p_B$, it is easy to see that for any fixed choice of $\ell \in [B]$, the expected payment is maximized when the worker selects options $\{1, \ldots, \ell\}$ (and where ties can be broken arbitrarily by the worker).

Let Ψ_ℓ denote the expected payment when the worker selects the first ℓ options:

$$\Psi_\ell = \alpha \sum_{i=1}^{\ell} p_i (1-\rho)^{\ell-1}.$$

Hence, for any $\ell \in \{2, \ldots, B\}$, we have

$$\frac{\Psi_{\ell-1}}{\Psi_\ell} = \frac{\alpha \sum_{i=1}^{\ell-1} p_i (1-\rho)^{\ell-2}}{\alpha \sum_{i=1}^{\ell} p_i (1-\rho)^{\ell-1}} = \frac{1}{1-\rho}\left(1 - \frac{p_\ell}{\sum_{i=1}^{\ell} p_i}\right). \tag{9}$$

Now consider any option $b \in [B]$ such that $p_b = 0$. Then, one can see that selection option b makes the expected payment strictly smaller as compared to not selecting that option. Hence, a worker is never incentivized to select any option in the set $\{b \in [B] \mid p_b = 0\}$. Let B' denote the number of options with non-zero beliefs—that is, we have $p_1 \geq \cdots \geq p_{B'} > 0 = p_{B'+1} = \cdots = p_B$. Then, it must be that the ratio

$$\frac{p_\ell}{\sum_{i=1}^{\ell} p_i}$$

strictly decreases with an increase in $\ell \in [B']$.

Now, for the moment, let us suppose that $\frac{p_\ell}{\sum_{i=1}^{\ell} p_i} \neq \rho$ for every $\ell \in [B]$. Then, it follows that there is a value $m \in [B']$ such that $\frac{p_\ell}{\sum_{i=1}^{\ell} p_i} > \rho$ for every $\ell \in \{1, \ldots, m\}$ and $\frac{p_\ell}{\sum_{i=1}^{\ell} p_i} < \rho$ for every $\ell \in \{m+1, \ldots, B\}$. It follows from Equation (9) that $\frac{\Psi_\ell}{\Psi_{\ell-1}} > 1$ for all $\ell \leq m$ and $\frac{\Psi_\ell}{\Psi_{\ell-1}} < 1$ for all $\ell > m$, or in other words, we have

$$\cdots < \Psi_{m-2} < \Psi_{m-1} < \Psi_m > \Psi_{m+1} > \Psi_{m+2} > \cdots.$$

It follows that the worker is incentivized to choose the set of options $\{1, \ldots, m\}$.

One can see that the argument above continues to hold even when $\frac{p_\ell}{\sum_{i=1}^{\ell} p_i} = \rho$ for some option $\ell \in [B']$. The only difference in this case is the incentive for this option ℓ; however, that is irrelevant, as we have not imposed any condition on this boundary case.

Let us now consider the case of $N = G \geq 1$. By our assumption of the independence of the beliefs of the worker across the questions, the expected payment equals

$$\prod_{i=1}^{G} \left[\alpha (1-\rho)^{(x_i - 1)} \mathbf{1}\{x_i \geq 0\}\right].$$

Since the payments are non-negative, if each individual component in the product is maximized, then the product is also necessarily maximized. Each individual component simply corresponds to the setting of $N = G = 1$ discussed earlier. Thus, calling upon our earlier result, we get that the expected payment for the case $N = G \geq 1$ is maximized when the worker acts as desired for every question.

Let us finally consider the general case of $N \geq G \geq 1$. Recall from Equation (2) that the expected payment for the general case is a cascade of two expectations: The outer expectation is with respect to the uniformly random distribution of the G gold standard questions among the N total questions, while the inner expectation is taken over the worker's beliefs of the different questions conditioned on the choice of the gold standard questions and restricts attention to only these G questions. The arguments above for the case $N = G$ prove that every individual term in the inner expectation is maximized when the worker acts as desired. The outer expectation does not affect this argument. The expected payment is thus maximized when the worker acts as desired.

B.3 Proof of Theorem 4.2: Uniqueness Under No-free-lunch

Without loss of generality, assume that $\alpha_{\min} = 0$, since the property of a scoring rule being a strictly proper is invariant to any constant shift and positive scale of the payment. We adopt the succinct notation of $\alpha := \alpha_{\max} - \alpha_{\min}$. The proof of this theorem employs some of the tools developed in Shah and Zhou [55] and also relies on Lemma 4.3 stated earlier in Section 4.2.1. We reproduce the statement of Lemma 4.3 here for convenience: Any strictly proper scoring rule f for relative thresholding must satisfy

$$\begin{aligned}&f(x_1, \ldots, x_{i-1}, x_i + 1, x_{i+1}, \ldots, x_G)\\&\qquad = (1-\rho) f(x_1, \ldots, x_{i-1}, x_i, x_{i+1}, \ldots, x_G) + \rho f(x_1, \ldots, x_{i-1}, -x_i, x_{i+1}, \ldots, x_G),\end{aligned}$$

for every $i \in [G]$ and $(x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_G) \in \{-(B-1), \ldots, -1, 1, \ldots, B\}^{G-1}$, $x_i \in [B-1]$. The proof of this lemma is provided in Section B.3.1.

Consider any strictly proper scoring rule f that satisfies the no-free-lunch condition. We first show that the scoring rule must necessarily make a zero payment when one or more questions in the gold standard are attempted incorrectly. To this end, observe that, since $f \geq 0$ and $\rho \in (0, 1)$, the statement of Lemma 4.3 necessitates that for every $i \in [G]$ and $(x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_G) \in \{-(B-1), \ldots, B\}^{G-1}$, $x_i \in [B-1]$:

$$\begin{aligned}&\text{If } f(x_1, \ldots, x_{i-1}, x_i + 1, x_{i+1}, \ldots, x_G) = 0,\\&\qquad \text{then } f(x_1, \ldots, x_{i-1}, x_i, x_{i+1}, \ldots, x_G) = f(x_1, \ldots, x_{i-1}, -x_i, x_{i+1}, \ldots, x_G) = 0.\end{aligned}$$

A repeated application of this argument implies:

$$\text{If } f(x_1, \ldots, x_{i-1}, B, x_{i+1}, \ldots, x_G) = 0, \quad \text{then } f(x_1, \ldots, x_{i-1}, x_i, x_{i+1}, \ldots, x_G) = 0,$$

for all $x_i \in \{-(B-1), \ldots, -1, 1, \ldots, B-1\}$.

Now consider any evaluation $(x_1, \ldots, x_G)$ that has at least one incorrect answer. Suppose without loss of generality that the first question is the one answered incorrectly, i.e., $x_1 \leq -1$. The no-free-lunch condition then makes $f(x_1, B, \ldots, B) = 0$. Applying our arguments from above, we get that $f(x_1, x_2, \ldots, x_G) = 0$ for every value of $(x_2, \ldots, x_G) \in \{-(B-1), \ldots, -1, 1, \ldots, B\}$.

Substituting this necessary condition in Lemma 4.3, we get that for every question $i \in \{1, \ldots, G\}$ and every $(x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_G) \in [B]^{G-1}$, $x_i \in [B-1]$, it must be that

$$f(x_1, \ldots, x_{i-1}, x_i + 1, x_{i+1}, \ldots, x_G) = (1-\rho) f(x_1, \ldots, x_{i-1}, x_i, x_{i+1}, \ldots, x_G).$$

Substituting $f(1, \ldots, 1) = \alpha$, we obtain the claimed result $f = f^*$.

B.3.1 Proof of Lemma 4.3: Necessary Condition for Relative Thresholding. Consider any strictly proper scoring rule f for relative thresholding. The proof does not make any assumptions such as no-free-lunch on this scoring rule.

First consider the case of $G = N$. Consider some $\eta, \gamma \in \{0, \ldots, G-1\}$ with $\eta + \gamma < G$. Suppose $i = \eta + \gamma + 1$, $x_1, \ldots, x_\eta \in [B-1]$, $x_{\eta+1}, \ldots, x_{\eta+\gamma} \in -[B-1]$ and $x_{\eta+\gamma+2}, \ldots, x_N = B$.

For every question $j \in [\eta + \gamma]$, suppose the worker's belief is $\delta_j \in (0, \rho)$ for the last option, is $\frac{1-\delta_j}{|x_j|}$ each for the first $|x_j|$ options, and is 0 for the remaining options. One can verify that, since $\delta_j < \rho < \frac{1}{B}$ and $|x_j| \leq B - 1$, it must be that $\frac{1-\delta_j}{|x_j|} > \delta_j$, and that the requirement of being a strictly proper scoring rule for relative thresholding requires incentivizing the worker to select the first $|x_j|$ options. Suppose the worker does so. Now, for every question $j' \in \{\eta + \gamma + 2, \ldots, N\}$, suppose the belief of the worker is uniform across all B options. The worker should be incentivized to select all B options in this case; suppose the worker does so. Finally, for question i, suppose the worker's belief is $\delta \in (\frac{\rho}{2}, \frac{3\rho}{2})$ for the last option, is $\frac{1-\delta}{|x_i|}$ each for the first $|x_i|$ options, and is 0 for the remaining options. Then, the worker must be incentivized to select the first $|x_i|$ options alone if $\delta < \rho$ and select the last option along with the first $|x_i|$ options if $\delta > \rho$.

Define $\{r_j\}_{j\in[\eta+\gamma]}$ as $r_j = \delta_j$ for $j \in [\eta]$, and $r_j = 1 - \delta_j$ for $j \in \{\eta + 1, \eta + \gamma\}$. Let $\boldsymbol{\epsilon} := \{\epsilon_1, \ldots, \epsilon_{\eta+\gamma}\} \in \{-1, 1\}^{\eta+\gamma}$. The requirement of incentivizing for question i necessitates

$$\begin{aligned}
&(1-\delta) \sum_{\boldsymbol{\epsilon}\in\{-1,1\}^{\eta+\gamma}} \left(f(\epsilon_1 x_1, \ldots, \epsilon_\eta x_\eta, \epsilon_{\eta+1} x_{\eta+1}, \ldots, \epsilon_{\eta+\gamma} x_{\eta+\gamma}, x_i, B, \ldots, B) \prod_{j\in[\eta+\gamma]} r_j^{\frac{1-\epsilon_j}{2}} (1-r_j)^{\frac{1+\epsilon_j}{2}} \right) \\
&\quad + \delta \sum_{\boldsymbol{\epsilon}\in\{-1,1\}^{\eta+\gamma}} \left(f(\epsilon_1 x_1, \ldots, \epsilon_\eta x_\eta, \epsilon_{\eta+1} x_{\eta+1}, \ldots, \epsilon_{\eta+\gamma} x_{\eta+\gamma}, -x_i, B, \ldots, B) \prod_{j\in[\eta+\gamma]} r_j^{\frac{1-\epsilon_j}{2}} (1-r_j)^{\frac{1+\epsilon_j}{2}} \right) \\
&\overset{\delta>\rho}{\underset{\delta<\rho}{\lesseqgtr}} \sum_{\boldsymbol{\epsilon}\in\{-1,1\}^{\eta+\gamma}} \left(f(\epsilon_1 x_1, \ldots, \epsilon_\eta x_\eta, \epsilon_{\eta+1} x_{\eta+1}, \ldots, \epsilon_{\eta+\gamma} x_{\eta+\gamma}, x_i + 1, B, \ldots, B) \prod_{j\in[\eta+\gamma]} r_j^{\frac{1-\epsilon_j}{2}} (1-r_j)^{\frac{1+\epsilon_j}{2}} \right).
\end{aligned}$$

The left-hand side of this expression is the expected payment if the worker chooses the first $|x_i|$ options for question $(\eta + \gamma + 1)$, while the right-hand side is the expected payment if she chooses the first $|x_i|$ options as well as the last option. For any real-valued variable q and for any real-valued constants a, b, and c,

$$aq \overset{q<c}{\underset{q>c}{\lesseqgtr}} b \quad \Rightarrow \quad ac = b.$$

With $q = 1 - \delta$ in this argument, we get

$$\begin{aligned}
&(1-\rho) \sum_{\boldsymbol{\epsilon}\in\{-1,1\}^{\eta+\gamma}} \left(f(\epsilon_1 x_1, \ldots, \epsilon_\eta x_\eta, \epsilon_{\eta+1} x_{\eta+1}, \ldots, \epsilon_{\eta+\gamma} x_{\eta+\gamma}, x_i, B, \ldots, B) \prod_{j\in[\eta+\gamma]} r_j^{\frac{1-\epsilon_j}{2}} (1-r_j)^{\frac{1+\epsilon_j}{2}} \right) \\
&\quad + \rho \sum_{\boldsymbol{\epsilon}\in\{-1,1\}^{\eta+\gamma}} \left(f(\epsilon_1 x_1, \ldots, \epsilon_\eta x_\eta, \epsilon_{\eta+1} x_{\eta+1}, \ldots, \epsilon_{\eta+\gamma} x_{\eta+\gamma}, -x_i, B, \ldots, B) \prod_{j\in[\eta+\gamma]} r_j^{\frac{1-\epsilon_j}{2}} (1-r_j)^{\frac{1+\epsilon_j}{2}} \right) \\
&- \sum_{\boldsymbol{\epsilon}\in\{-1,1\}^{\eta+\gamma}} \left(f(\epsilon_1 x_1, \ldots, \epsilon_\eta x_\eta, \epsilon_{\eta+1} x_{\eta+1}, \ldots, \epsilon_{\eta+\gamma} x_{\eta+\gamma}, x_i + 1, B, \ldots, B) \prod_{j\in[\eta+\gamma]} r_j^{\frac{1-\epsilon_j}{2}} (1-r_j)^{\frac{1+\epsilon_j}{2}} \right) = 0.
\end{aligned} \tag{10}$$

The left-hand side of Equation (10) represents a polynomial in $(\eta + \gamma)$ variables $\{r_j\}_{j=1}^{\eta+\gamma}$, which evaluates to zero for all values of the variables within an $(\eta + \gamma)$-dimensional solid ball. Thus, the coefficients of the monomials in this polynomial must be zero. In particular, the constant term must be zero. The constant term appears when $\epsilon_j = 1 \, \forall \, j$ in the summations in Equation (10). Setting the constant term to zero gives

$$\begin{aligned}
(1-\rho) f(x_1, \ldots, x_{\eta+\gamma}, x_{\eta+\gamma+1}, B, \ldots, B) + \rho f(x_1, \ldots, x_{\eta+\gamma}, -x_{\eta+\gamma+1}, B, \ldots, B)& \\
- f(x_1, \ldots, x_{\eta+\gamma}, x_{\eta+\gamma+1} + 1, B, \ldots, B) = 0,&
\end{aligned}$$

as desired. Since the arguments above hold for any permutation of the N questions, this completes the proof for the case of $G = N$.

Now consider the case $G < N$. Let $g : \{-(B-1), \ldots, -1, 1, \ldots, B\}^N \rightarrow \mathbb{R}_+$ represent the expected payment given an evaluation of all the N answers when the identities of the gold standard questions are unknown. Here, the expectation is with respect to the (uniformly random) choice of the G gold standard questions. If $(x_1, \ldots, x_N) \in \{-(B-1), \ldots, -1, 1, \ldots, B\}^N$ are the evaluations of the worker's answers to the N questions, then the expected payment is

$$g(x_1, \ldots, x_N) = \frac{1}{\binom{N}{G}} \sum_{(i_1, \ldots, i_G) \subseteq \{1, \ldots, N\}} f(x_{i_1}, \ldots, x_{i_G}). \tag{11}$$

Applying the same arguments to g as done to f above gives

$$\begin{aligned}(1-\rho)g(x_1, \ldots, x_{\eta+\gamma}, x_{\eta+\gamma+1}, B, \ldots, B) + \rho g(x_1, \ldots, x_{\eta+\gamma}, -x_{\eta+\gamma+1}, B, \ldots, B)\\ - g(x_1, \ldots, x_{\eta+\gamma}, x_{\eta+\gamma+1} + 1, B, \ldots, B) = 0.\end{aligned} \tag{12}$$

The proof now proceeds via an induction on the quantity $(G - \eta - \gamma - 1)$. We begin with the case of $(G - \eta - \gamma - 1) = G - 1$, which implies $\eta = \gamma = 0$. In this case Equation (10) simplifies to

$$(1-\rho)g(x_1, B, \ldots, B) + \rho g(-x_1, B, \ldots, B) = g(x_1 + 1, B, \ldots, B).$$

Applying the expansion of function g in terms of function f from Equation (11) for some $x_1 \in [B-1]$ gives

$$\begin{aligned}(1-\rho)\left(c_1 f(x_1, B, \ldots, B) + c_2 f(B, B, \ldots, B)\right) + \rho\left(c_1 f(-x_1, B, \ldots, B) + c_2 f(B, B, \ldots, B)\right)\\ = c_1 f(x_1 + 1, B, \ldots, B) + c_2 f(B, B, \ldots, B)\end{aligned}$$

for constants $c_1 > 0$ and $c_2 > 0$ that, respectively, represent the probabilities that the first question is picked and not picked in the set of G gold standard questions. Canceling out the common terms on both sides of the equation, we get the desired result

$$(1-\rho)f(x_1, B, \ldots, B) + \rho f(-x_1, B, \ldots, B) = f(x_1 + 1, B, \ldots, B).$$

Next, we consider the case when $(G - \eta - \gamma - 1)$ questions are skipped in the gold standard and assume that the result is true when more than $(G - \eta - \gamma - 1)$ questions are skipped in the gold standard. In Equation (12), the functions g decompose into a sum of the constituent f functions. These constituent functions f are of two types: the first, where all of the first $(\eta + \gamma + 1)$ questions are included in the gold standard; and the second, where one or more of the first $(\eta + \gamma + 1)$ questions are not included in the gold standard. The second case corresponds to situations where there are more than $(G - \eta - \gamma - 1)$ questions skipped in the gold standard and hence satisfies our induction hypothesis. The terms corresponding to these functions thus cancel out in the expansion of Equation (12). The remainder comprises only evaluations of function f for arguments in which the first $(\eta + \gamma + 1)$ questions are included in the gold standard. Since the last $(N - \eta - \gamma - 1)$ questions are skipped by the worker, the remainder evaluates to

$$\begin{aligned}(1-\rho)c_3 f(x_1, \ldots, x_{\eta+\gamma}, x_i, B, \ldots, B) + \rho c_3 f(x_1, \ldots, x_{\eta+\gamma}, -x_i, B, \ldots, B)\\ = c_3 f(x_1, \ldots, x_{\eta+\gamma}, x_i + 1, B, \ldots, B)\end{aligned} \tag{13}$$

for some constant $c_3 > 0$. Dividing throughout by c_3 gives the desired result.

Finally, the arguments above hold for any permutation of the first G questions, thus completing the proof.

B.4 Proof of Corollary 4.4: Eliciting Support

Suppose the beliefs of the worker for any particular question are $p_1 \geq \cdots \geq p_m > \rho > p_{m+1} = \cdots = p_B = 0$ for some $m \in [B]$. Notice that we have used the coarse beliefs assumptions in the condition "$p_m > \rho > p_{m+1}$." Under these beliefs, we have

$$\frac{p_b}{\sum_{i=1}^{b} p_i} = \frac{0}{\sum_{i=1}^{b} p_i} = 0 < \rho \qquad \text{for all } b \geq m+1,$$

and

$$\frac{p_b}{\sum_{i=1}^{b} p_i} \geq \frac{p_b}{1} > \rho \qquad \text{for all } b \leq m.$$

This satisfies the conditions for relative thresholding. Hence, any scoring rule that is strictly proper for relative thresholding is also strictly proper for eliciting support under the coarse beliefs assumption. From Theorem 4.1, we know that Scoring rule 2 is strictly proper for relative thresholding and hence is also strictly proper for eliciting support.

B.5 Proof of Theorem 4.5: Performance-based Payments

Without loss of generality, assume that $\alpha_{\min} = 0$, since in our setting, the property of being strictly proper is invariant to any constant shift and positive scale of the payment. We adopt the succinct notation of $\alpha := \alpha_{\max} - \alpha_{\min}$. Consider any strictly proper scoring rule f such that $f(1, \ldots, 1) = \alpha$. The proof employs the following lemma:

Lemma B.1. *Consider some $y, y' \in [B]^N$ and some $\mathcal{I} \subseteq [N]$ such that $y_i = y'_i + 1$ for all $i \in \mathcal{I}$, and $y_i = y'_i$ for all $i \notin \mathcal{I}$. Then, any strictly proper scoring rule f must necessarily satisfy*

$$\frac{1}{\binom{N}{G}} \sum_{(j_1,\ldots,j_G)\subseteq[N]} f(y_{j_1}, \ldots, y_{j_G}) \geq \frac{1}{\binom{N}{G}} \sum_{(j_1,\ldots,j_G)\subseteq[N]} (1-\rho)^{|\mathcal{I}\cap\{j_1,\ldots,j_G\}|} f(y'_{j_1}, \ldots, y'_{j_G}). \tag{14a}$$

Furthermore, a necessary condition for the above inequality to be satisfied with equality is

$$f(\epsilon_1 y'_{j_1}, \ldots, \epsilon_G y'_{j_G}) = \alpha_{\min} \tag{14b}$$

for all $(j_1, \ldots, j_G) \subseteq [N]$, and all $\{(\epsilon_1, \ldots, \epsilon_G) \in \{-1, 1\}^G \backslash \{1\}^G \mid \epsilon_i = 1$ whenever $j_i \notin \mathcal{I}\}$.

The lemma derives lower bounds on the payment made by any strictly proper scoring rule under any evaluation y as compared to another evaluation y' that differs from y only in the questions in some set $\mathcal{I}$. In particular, the left-hand side of Equation (14a) is simply the expected payment under a strictly proper scoring rule f under an evaluation y. The right-hand side is a rescaling of the expected payment under the evaluation y', where the payment is rescaled by a factor $(1-\rho)$ for every additional option selected in y' as compared to y. The second part (14b) of the lemma then shows that any strictly proper scoring rule that achieves (14a) with equality must make a minimum payment whenever one or more questions in the set $\mathcal{I}$ does not have the correct answer selected. The proof of Lemma B.1 is provided later in Appendix B.5.5.

B.5.1 Part (a). First, a non-strict inequality: Consider any $x_0 \in [B-1]$. Applying Lemma B.1 with $y = (x_0+1, \ldots, x_0+1)$, $y' = (x_0, \ldots, x_0)$ and $\mathcal{I} = [G]$ gives

$$f(x_0+1, \ldots, x_0+1) \geq (1-\rho)^G f(x_0, \ldots, x_0).$$

A repeated application of this inequality for every $x_0 \in [B-1]$ gives

$$\begin{aligned} f(B, \ldots, B) &\geq (1-\rho)^G f(B-1, \ldots, B-1) \geq \cdots \geq (1-\rho)^{(B-1)G} f(1, \ldots, 1) \\ &= (1-\rho)^{(B-1)G}\alpha. \end{aligned}$$

Scoring rule 2 achieves this lower bound on $f(B, \ldots, B)$ with equality, thereby completing the proof.

Strict inequality: In the remainder of the proof, we show that any strictly proper scoring rule that achieves

$$f(B, \ldots, B) = (1 - \rho)^{G(B-1)}\alpha$$

must be identical to our Scoring rule 2. In other words, we show that such a scoring rule f must satisfy $f(x) = f^*(x)$ for every evaluation x. We partition the rest of the proof into two cases, depending on whether $x_G > 0$ or $x_G < 0$.

In what follows, we only consider the set of evaluations x whose elements are non-decreasing, that is, $x_1 \geq x_2 \geq \cdots \geq x_G$. The proof for any other ordering of the elements follows in an identical manner.

Case of $x_G > 0$: We first consider any x such that $x_G > 0$. Then, due to the monotonicity of the entries of x, we must have that $x_1 \geq \cdots \geq x_G > 0$. We define the following notation that we will subsequently use for our induction arguments:

- Let $\gamma(x)$ denote the number of distinct entries in x:
$$\gamma(x) := 1 + \sum_{i=1}^{G-1} \mathbf{1}\{x_i \neq x_{i+1}\}.$$
- Let $\sigma(x)$ denote the size of the last jump in x:
$$\sigma(x) := x_j - x_{j+1} \qquad \text{where } j = \operatorname*{arg\,max}_{i \in [G-1]} x_i \neq x_{i+1}.$$
- Let $\beta(x)$ denote the numeric value of x in a B-ary number system:
$$\beta(x) := \sum_{i=1}^{G} B^{G-i}(x_i - 1).$$

For example, if $B = 5$, $G = 5$, and $x = (5, 5, 4, 1, 1)$, then $\gamma(x) = |\{5, 4, 1\}| = 3$, $\sigma(x) = 4 - 1 = 3$ (where $j = 3$), and $\beta(x) = 4 \cdot 5^4 + 4 \cdot 5^3 + 3 \cdot 5^2 + 0 \cdot 5^1 + 0 \cdot 5^0 = 3{,}075$. The proof involves three nested levels of induction: on γ, on σ, and then on β.

Base case for induction on γ: We first induct on γ. The base case is the set $\{x | \gamma(x) = 1\}$, that is, the set of vectors that have the same value for all its components. We now show that $f(x) = f^*(x)$ for every x such that $\gamma(x) = 1$. To this end, observe that from the definition of γ, the only values of x that have $\gamma(x) = 1$ are those that have all elements identical. Consider any $x_0 \in [B - 1]$. Applying Lemma B.1 with $y = (x_0 + 1, \ldots, x_0 + 1)$ and $y' = (x_0, \ldots, x_0)$ gives

$$f(x_0 + 1, \ldots, x_0 + 1) \geq (1 - \rho)^G f(x_0, \ldots, x_0).$$

Since this inequality is true for every $x_0 \in [B - 1]$, we have the sandwich inequalities

$$f(B, \ldots, B) \geq (1 - \rho)^{(B-x_0)G} f(x_0, \ldots, x_0) \geq (1 - \rho)^{(B-1)G} f(1, \ldots, 1).$$

Setting $f(B, \ldots, B) = (1 - \rho)^{(B-1)G}\alpha$ and $f(1, \ldots, 1) = \alpha$ implies that each of the above inequalities is in fact an equality, thereby proving the base case

$$f(x_0, \ldots, x_0) = (1 - \rho)^{x_0 G} f(1, \ldots, 1) = f^*(x_0, \ldots, x_0).$$

Induction step for induction on γ: Now suppose our hypothesis of $f(x) = f^*(x)$ is true for all $\{x | \gamma(x) \leq \gamma_0 - 1\}$ for some $\gamma_0 \in \{2, \ldots, B\}$. We will now prove that the hypothesis $f(x) = f^*(x)$ is also true for all $\{x | \gamma(x) = \gamma_0\}$. Towards this goal, we now induct on σ, that is, we prove the hypothesis separately for every value of σ.

Base case for induction on σ: Observe the following set-relation $\{x|\gamma(x) = \gamma_0 - 1\} = \{x|\gamma(x) = \gamma_0, \sigma = 0\}$. Due to the assumed induction hypothesis $f(x) = f^*(x)$ for every $\{x|\gamma(x) = \gamma_0 - 1\}$, we have $f(x) = f^*(x)$ for every $\{x|\gamma(x) = \gamma_0, \sigma = 0\}$, thereby proving the base case of $\sigma = 0$.

Induction step for induction on σ: Now suppose that the hypothesis is true for all $\{x|\gamma(x) = \gamma_0, \sigma(x) \leq \sigma_0 - 1\}$ for some $\sigma_0 \in [B-1]$. We will prove that the hypothesis remains true for all $\{x|\gamma(x) = \gamma_0, \sigma(x) = \sigma_0\}$. We prove this statement is true for all values of β via an induction on β.

Base case for induction on β: Recall that we have restricted our attention to those x that have their elements in a descending order. Observe that the element with the minimum value of β in the set $\{x|\gamma(x) = \gamma_0, \sigma(x) = \sigma_0\}$ is $(\gamma_0 + \sigma_0 - 1, \ldots, \sigma_0 + 1, 1, \ldots, 1)$. We will prove the hypothesis for this element as the base case for our induction on β. Applying Lemma B.1 with $y = (\gamma_0 + \sigma_0 - 1, \ldots, \sigma_0 + 2, \sigma_0 + 1, 1, \ldots, 1)$ and $y' = (\gamma_0 + \sigma_0 - 1, \ldots, \sigma_0 + 2, \sigma_0, 1, \ldots, 1)$ gives the inequality

$$\begin{aligned} & c_1 f(\gamma_0 + \sigma_0 - 1, \ldots, \sigma_0 + 2, \sigma_0 + 1, 1, \ldots, 1) + c_1' f(\gamma_0 + \sigma_0 - 1, \ldots, \sigma_0 + 2, 1, 1, \ldots, 1) \\ & \quad + \sum_{s \subsetneq \{\gamma_0 + \sigma_0 - 1, \ldots, \sigma_0 + 2\}} (c_s f(s, 1, 1, \ldots, 1) + c_s' f(s, \sigma_0 + 1, 1, \ldots, 1)) \\ & \geq c_1 (1-\rho) f(\gamma_0 + \sigma_0 - 1, \ldots, \sigma_0 + 2, \sigma_0, 1, \ldots, 1) + c_1' f(\gamma_0 + \sigma_0 - 1, \ldots, \sigma_0 + 2, 1, 1, \ldots, 1) \\ & \quad + \sum_{s \subsetneq \{\gamma_0 + \sigma_0 - 1, \ldots, \sigma_0 + 2\}} (c_s f(s, 1, 1, \ldots, 1) + c_s' (1-\rho) f(s, \sigma_0, 1, \ldots, 1)), \end{aligned} \tag{15}$$

for some positive constants c_1, c_1', c_s, c_s' (which represent the probabilities of the respective set of G questions being chosen as the G gold standard questions). Now, for any $s \subsetneq \{\gamma_0 + \sigma_0 - 1, \ldots, \sigma_0 + 2\}$, observe that $\gamma(s, \sigma_0 + 1, 1, \ldots, 1) \leq \gamma_0 - 1$ and $\gamma(s, \sigma_0, 1, \ldots, 1) \leq \sigma_0 - 1$. Thus, we have

$$f(s, \sigma_0 + 1, 1, \ldots, 1) \overset{(i)}{=} f^*(s, \sigma_0 + 1, 1, \ldots, 1) = (1-\rho) f^*(s, \sigma_0, 1, \ldots, 1) \overset{(ii)}{=} (1-\rho) f(s, \sigma_0, 1, \ldots, 1), \tag{16}$$

where equations (i) and (iii) follow from our induction hypothesis. Also, $\gamma(\gamma_0 + \sigma_0 - 1, \ldots, \sigma_0 + 2, \sigma_0, 1, \ldots, 1) = \gamma_0$ and $\sigma(\gamma_0 + \sigma_0 - 1, \ldots, \sigma_0 + 2, \sigma_0, 1, \ldots, 1) = \sigma_0 - 1$. Consequently, our induction hypothesis yields

$$\begin{aligned} f(\gamma_0 + \sigma_0 - 1, \ldots, \sigma_0 + 2, \sigma_0, 1, \ldots, 1) &= f^*(\gamma_0 + \sigma_0 - 1, \ldots, \sigma_0 + 2, \sigma_0, 1, \ldots, 1) \\ &= (1-\rho)^{\gamma_0 + \sigma_0 - 2 + \cdots + \sigma_0 + 1 + \sigma_0 - 1} \alpha. \end{aligned} \tag{17}$$

Substituting Equations (16) and (17) in Equation (15) and canceling out common terms yields the inequality

$$\begin{aligned} f(\gamma_0 + \sigma_0 - 1, \ldots, \sigma_0 + 2, \sigma_0 + 1, 1, \ldots, 1) &\geq (1-\rho)^{\gamma_0 + \sigma_0 - 2 + \cdots + \sigma_0} \alpha \\ &= f^*(\gamma_0 + \sigma_0 - 1, \ldots, \sigma_0 + 2, \sigma_0 + 1, 1, \ldots, 1). \end{aligned} \tag{18}$$

We now derive a matching upper bound on $f(\gamma_0 + \sigma_0 - 1, \ldots, \sigma_0 + 2, \sigma_0 + 1, 1, \ldots, 1)$. Applying Lemma B.1 with $y = (\gamma_0 + \sigma_0 - 1, \ldots, \sigma_0 + 1, 2, \ldots, 2)$ and $y' = (\gamma_0 + \sigma_0 - 1, \ldots, \sigma_0 + 1, 1, \ldots, 1)$ gives

$$\begin{aligned} & c_1 f(\gamma_0 + \sigma_0 - 1, \ldots, \sigma_0 + 1, 2, \ldots, 2) + \sum_{s \subsetneq \{\gamma_0 + \sigma_0 - 1, \ldots, \sigma_0 + 1\}} c_s f(s, 2, \ldots, 2) \\ & \geq c_1 (1-\rho)^{G - \gamma + 1} f(\gamma_0 + \sigma_0 - 1, \ldots, \sigma_0 + 1, 1, \ldots, 1) + \sum_{s \subsetneq \{\gamma_0 + \sigma_0 - 1, \ldots, \sigma_0 + 1\}} c_s (1-\rho)^{G - |s|} f(s, 1, \ldots, 1), \end{aligned} \tag{19}$$

for some positive constants c_1, c_s. Now, for any $s \subsetneq \{\gamma_0 + \sigma_0 - 1, \ldots, \sigma_0 + 2\}$, observe that $\gamma(s, 2, \ldots, 2) \leq \gamma_0 - 1$ and $\gamma(s, 1, \ldots, 1) \leq \sigma_0 - 1$. Thus, we have

$$f(s, 2, \ldots, 2) \overset{(i)}{=} f^*(s, 2, \ldots, 2) = (1-\rho)^{G-|s|} f(s, 1, \ldots, 1) \overset{(ii)}{=} (1-\rho)^{G-|s|} f(s, 1, \ldots, 1), \tag{20}$$

where equations (i) and (ii) follow from our induction hypothesis. Also note that $\gamma(\gamma_0 + \sigma_0 - 1, \ldots, \sigma_0 + 1, 2, \ldots, 2) \leq \gamma_0$ and $\sigma(\gamma_0 + \sigma_0 - 1, \ldots, \sigma_0 + 1, 2, \ldots, 2) = \sigma_0 - 1$, which allows us to apply our induction hypothesis to get

$$\begin{aligned} f(\gamma_0 + \sigma_0 - 1, \ldots, \sigma_0 + 1, 2, \ldots, 2) &= f^*(\gamma_0 + \sigma_0 - 1, \ldots, \sigma_0 + 1, 2, \ldots, 2) \\ &= (1-\rho)^{\gamma_0+\sigma_0-2+\cdots+\sigma_0+G-\gamma+1} \alpha. \end{aligned} \tag{21}$$

Substituting Equations (24) and (21) in Equation (19) and canceling out common terms yields the upper bound

$$\begin{aligned} f(\gamma_0 + \sigma_0 - 1, \ldots, \sigma_0 + 2, \sigma_0 + 1, 1, \ldots, 1) &\leq (1-\rho)^{\gamma_0+\sigma_0-2+\cdots+\sigma_0} \alpha \\ &= f^*(\gamma_0 + \sigma_0 - 1, \ldots, \sigma_0 + 2, \sigma_0 + 1, 1, \ldots, 1). \end{aligned} \tag{22}$$

The bounds (18) and (22) in conjunction imply that the hypothesis is true for $x = (\gamma_0 + \sigma_0 - 1, \ldots, \sigma_0 + 2, \sigma_0 + 1, 1, \ldots, 1)$, which is the base case for our induction on β.

Induction step for induction on β: Now consider some x^* such that $\gamma(x^*) = \gamma_0$, $\sigma(x^*) = \sigma_0$ and $\beta(x^*) = \beta_0$, for some β_0. One can verify that any such x^* must necessarily take the form

$$x^* = (x_1^*, \ldots, x_m^*, \underbrace{\sigma_0 + x_G^*, \ldots, \sigma_0 + x_G^*}_{m_1}, x_G^*, \ldots, x_G^*), \tag{23}$$

with $x_1^* \geq x_2^* \geq \cdots \geq x_m^* > \sigma_0 + x_G^*$ for some $m \geq 0$, $m_1 \geq 1$, $m + m_1 < G$.

Suppose the hypothesis $f(x) = f^*(x)$ is true for every $\{x | \gamma(x) = \gamma_0, \sigma(x) = \sigma_0, \beta(x) \leq \beta_0 - 1\}$. In what follows, we show that we must have $f(x) = f^*(x)$ for every $\{x | \gamma(x) = \gamma_0, \sigma(x) = \sigma_0, \beta(x) = \beta_0\}$.

Applying Lemma B.1 to f with the choices $y = (x_1^*, \ldots, x_m^*, \underbrace{\sigma_0 + x_G^*, \ldots, \sigma_0 + x_G^*}_{m_1}, x_G^*, \ldots, x_G^*)$ and $y' = (x_1^*, \ldots, x_m^*, \underbrace{\sigma_0 + x_G^* - 1, \ldots, \sigma_0 + x_G^* - 1}_{m_1}, x_G^*, \ldots, x_G^*)$ gives the inequality

$$\begin{aligned} &c_1 f(x_1^*, \ldots, x_m^*, \underbrace{\sigma_0 + x_G^*, \ldots, \sigma_0 + x_G^*}_{m_1}, x_G^*, \ldots, x_G^*) + \sum_{s \subsetneq \{x_1^*, \ldots, x_m^*, \underbrace{\sigma_0 + x_G^*, \ldots, \sigma_0 + x_G^*}_{m_1}\}} c_s f(s, x_G^*, \ldots, x_G^*) \\ &\quad \geq c_1 (1-\rho)^{m_1} f(x_1^*, \ldots, x_m^*, \underbrace{\sigma_0 + x_G^* - 1, \ldots, \sigma_0 + x_G^* - 1}_{m_1}, x_G^*, \ldots, x_G^*) \\ &\qquad + \sum_{s \subsetneq \{x_1^*, \ldots, x_m^*, \underbrace{\sigma_0 + x_G^* - 1, \ldots, \sigma_0 + x_G^* - 1}_{m_1}\}} c_s (1-\rho)^{\sum_i \mathbf{1}\{s_i = \sigma_0 + x_G^* - 1\}} f(s, x_G^*, \ldots, x_G^*), \end{aligned} \tag{24}$$

for some positive constants c_1, c_s. Recall that x^* takes the form (23) and has $\gamma(x^*) = \gamma_0$, $\sigma(x^*) = \sigma_0$ and $\beta(x^*) = \beta_0$. Thus, we have

$$\gamma(x_1^*, \ldots, x_m^*, \underbrace{\sigma_0 + x_G^* - 1, \ldots, \sigma_0 + x_G^* - 1}_{m_1}, x_G^*, \ldots, x_G^*) = \begin{cases} \gamma_0 - 1 & \text{if } \sigma_0 = 1 \\ \gamma_0 & \text{otherwise,} \end{cases}$$

and hence the induction hypothesis is satisfied in the first case of $\sigma_0 = 1$. In the second case of $\sigma_0 \neq 1$, we have

$$\sigma(x_1^*, \ldots, x_m^*, \underbrace{\sigma_0 + x_G^* - 1, \ldots, \sigma_0 + x_G^* - 1}_{m_1}, x_G^*, \ldots, x_G^*) = \sigma_0 - 1,$$

and hence the induction hypothesis is satisfied in the second case as well. Thus, we have

$$\begin{aligned} f(x_1^*, \ldots, x_m^*, &\underbrace{\sigma_0 + x_G^* - 1, \ldots, \sigma_0 + x_G^* - 1}_{m_1}, x_G^*, \ldots, x_G^*) \\ &= f^*(x_1^*, \ldots, x_m^*, \underbrace{\sigma_0 + x_G^* - 1, \ldots, \sigma_0 + x_G^* - 1}_{m_1}, x_G^*, \ldots, x_G^*) \\ &= (1-\rho)^{\sum_{i=1}^{m}(x_i^*-1) + m_1(\sigma_0 + x_G^* - 2) + (G - m_1 - m)(x_G^* - 1)} \alpha. \end{aligned} \tag{25}$$

Consider any $s \subsetneq \{x_1^*, \ldots, x_m^*, \underbrace{\sigma_0 + x_G^* - 1, \ldots, \sigma_0 + x_G^* - 1}_{m_1}\}$. We claim that the induction hypothesis is satisfied for $(s, x_G^*, \ldots, x_G^*)$. To this end, define the quantity $\mathfrak{m}_1(s)$ as

$$\mathfrak{m}_1(s) := \sum_i \mathbf{1}\{s_i = \sigma_0 + x_G^* - 1\}.$$

Observe that if $\mathfrak{m}_1(s) > 0$, then either $\gamma(s, x_G^*, \ldots, x_G^*) \leq \gamma_0 - 1$ or $\sigma(s, x_G^*, \ldots, x_G^*) \leq \sigma_0 - 1$; if $\mathfrak{m}_1(s) = 0$, then $\gamma(s, x_G^*, \ldots, x_G^*) \leq \gamma_0 - 1$. For any $s \subsetneq \{x_1^*, \ldots, x_m^*, \underbrace{\sigma_0 + x_G^*, \ldots, \sigma_0 + x_G^*}_{m_1}\}$, also define the quantity $\widetilde{\mathfrak{m}}_1(s)$ as

$$\widetilde{\mathfrak{m}}_1(s) := \sum_i \mathbf{1}\{s_i = \sigma_0 + x_G^*\}.$$

Observe that if $\widetilde{\mathfrak{m}}_1(s) > 0$, then either $\gamma(s, x_G^*, \ldots, x_G^*) \leq \gamma_0 - 1$ or $\beta(s, x_G^*, \ldots, x_G^*) \leq \beta_0 - 1$; if $\widetilde{\mathfrak{m}}_1(s) = 0$, then $\gamma(s, x_G^*, \ldots, x_G^*) \leq \gamma_0 - 1$. Consequently, from our induction hypothesis, we have the series of equations

$$\begin{aligned} \sum_{s \subsetneq \{x_1^*, \ldots, x_m^*, \underbrace{\sigma_0 + x_G^*, \ldots, \sigma_0 + x_G^*}_{m_1}\}} c_s f(s, x_G^*, \ldots, x_G^*) &= \sum_{s \subsetneq \{x_1^*, \ldots, x_m^*, \underbrace{\sigma_0 + x_G^*, \ldots, \sigma_0 + x_G^*}_{m_1}\}} c_s f^*(s, x_G^*, \ldots, x_G^*) \\ &= \sum_{s \subsetneq \{x_1^*, \ldots, x_m^*, \underbrace{\sigma_0 + x_G^* - 1, \ldots, \sigma_0 + x_G^* - 1}_{m_1}\}} c_s (1-\rho)^{\sum_i \mathbf{1}\{s_i = \sigma_0 + x_G^* - 1\}} f^*(s, x_G^*, \ldots, x_G^*) \\ &= \sum_{s \subsetneq \{x_1^*, \ldots, x_m^*, \underbrace{\sigma_0 + x_G^* - 1, \ldots, \sigma_0 + x_G^* - 1}_{m_1}\}} c_s (1-\rho)^{\sum_i \mathbf{1}\{s_i = \sigma_0 + x_G^* - 1\}} f(s, x_G^*, \ldots, x_G^*). \end{aligned} \tag{26}$$

Substituting Equations (25) and (26) in Equation (24) and canceling out common terms gives

$$\begin{aligned} f(x_1^*, \ldots, x_m^*, &\underbrace{\sigma_0 + x_G^*, \ldots, \sigma_0 + x_G^*}_{m_1}, x_G^*, \ldots, x_G^*) \\ &\geq (1-\rho)^{m_1} f(x_1^*, \ldots, x_m^*, \underbrace{\sigma_0 + x_G^* - 1, \ldots, \sigma_0 + x_G^* - 1}_{m_1}, x_G^*, \ldots, x_G^*) \\ &= (1-\rho)^{\sum_{i=1}^{m}(x_i^*-1) + m_1(\sigma_0 + x_G^* - 1) + (G - m_1 - m)(x_G^* - 1)} \alpha. \end{aligned} \tag{27}$$

We now employ Lemma B.1 to derive a matching upper bound to Equation (27). Setting the value $y = (x_1^*, \dots, x_m^*, \underbrace{\sigma_0 + x_G^*, \dots, \sigma_0 + x_G^*}_{m_1}, x_G^* + 1, \dots, x_G^* + 1)$ and $y' = (x_1^*, \dots, x_m^*, \underbrace{\sigma_0 + x_G^*, \dots, \sigma_0 + x_G^*}_{m_1}, x_G^*, \dots, x_G^*)$ in Lemma B.1 yields the inequality

$$\begin{aligned} & c_1 f(x_1^*, \dots, x_m^*, \underbrace{\sigma_0 + x_G^*, \dots, \sigma_0 + x_G^*}_{m_1}, x_G^* + 1, \dots, x_G^* + 1) \\ & \qquad + \sum_{s \subsetneq \{x_1^*, \dots, x_m^*, \underbrace{\sigma_0 + x_G^*, \dots, \sigma_0 + x_G^*}_{m_1}\}} c_s f(s, x_G^* + 1, \dots, x_G^* + 1) \\ & \geq c_1 (1-\rho)^{m_1} f(x_1^*, \dots, x_m^*, \underbrace{\sigma_0 + x_G^*, \dots, \sigma_0 + x_G^*}_{m_1}, x_G^*, \dots, x_G^*) \\ & \qquad + \sum_{s \subsetneq \{x_1^*, \dots, x_m^*, \underbrace{\sigma_0 + x_G^*, \dots, \sigma_0 + x_G^*}_{m_1}\}} c_s (1-\rho)^{G-|s|} f(s, x_G^*, \dots, x_G^*), \end{aligned} \tag{28}$$

for some positive constants c_1, c_s. Observe that

$$\gamma(x_1^*, \dots, x_m^*, \underbrace{\sigma_0 + x_G^*, \dots, \sigma_0 + x_G^*}_{m_1}, x_G^* + 1, \dots, x_G^* + 1) = \begin{cases} \gamma_0 - 1 & \text{if } \sigma_0 = 1 \\ \gamma_0 & \text{otherwise,} \end{cases}$$

and that the induction hypothesis is satisfied in the first case of $\sigma = 1$. In the second case of $\sigma \neq 1$, we have

$$\sigma(x_1^*, \dots, x_m^*, \underbrace{\sigma_0 + x_G^*, \dots, \sigma_0 + x_G^*}_{m_1}, x_G^* + 1, \dots, x_G^* + 1) = \sigma_0 - 1,$$

and hence the induction hypothesis is satisfied in this case as well. Thus, from our induction hypothesis, we have

$$\begin{aligned} & f(x_1^*, \dots, x_m^*, \underbrace{\sigma_0 + x_G^*, \dots, \sigma_0 + x_G^*}_{m_1}, x_G^* + 1, \dots, x_G^* + 1) \\ & \qquad = f^*(x_1^*, \dots, x_m^*, \underbrace{\sigma_0 + x_G^*, \dots, \sigma_0 + x_G^*}_{m_1}, x_G^* + 1, \dots, x_G^* + 1) \\ & \qquad = (1-\rho)^{\sum_{i=1}^{m}(x_i^*-1) + m_1(\sigma_0 + x_G^* - 1) + (G - m_1 - m)(x_G^* - 2)} \alpha. \end{aligned} \tag{29}$$

Now consider any $s \subsetneq \{x_1^*, \dots, x_m^*, \underbrace{\sigma_0 + x_G^*, \dots, \sigma_0 + x_G^*}_{m_1}\}$. We now show that our induction hypothesis is satisfied for $(s, x_G^*, \dots, x_G^*)$ as well as $(s, x_G^* + 1, \dots, x_G^* + 1)$. To this end, recall our notation of $\widetilde{m}_1(s) := \sum_i \mathbf{1}\{s_i = \sigma_0 + x_G^*\}$. If $\sigma_0 = 1$ or if $\widetilde{m}_1(s) = 0$, then $\gamma(s, x_G^* + 1, \dots, x_G^* + 1) \leq \gamma_0 - 1$; if $\sigma > 1$ and $\widetilde{m}_1(s) > 0$, then $\gamma(s, x_G^* + 1, \dots, x_G^* + 1) \leq \gamma_0$ and $\sigma(s, x_G^* + 1, \dots, x_G^* + 1) \leq \sigma_0 - 1$. If $\widetilde{m}_1(s) = 0$, then $\gamma(s, x_G^*, \dots, x_G^*) \leq \gamma_0 - 1$, otherwise $\gamma(s, x_G^*, \dots, x_G^*) \leq \gamma_0$, $\sigma(s, x_G^*, \dots, x_G^*) = \sigma_0$ and $\beta(s, x_G^*, \dots, x_G^*) \leq \beta_0 - 1$. These terms thus satisfy our induction hypothesis and hence, we have

$$\begin{aligned} f(s, x_G^* + 1, \ldots, x_G^* + 1) &= f^*(s, x_G^* + 1, \ldots, x_G^* + 1) \\ &= (1-\rho)^{G-|s|} f^*(s, x_G^*, \ldots, x_G^*) \\ &= (1-\rho)^{G-|s|} f(s, x_G^*, \ldots, x_G^*). \end{aligned} \tag{30}$$

Substituting Equations (29) and (30) in Equation (28) gives us an upper bound

$$\begin{aligned} f(x_1^*, \ldots, x_m^*, \underbrace{\sigma_0 + x_G^*, \ldots, \sigma_0 + x_G^*}_{m_1}, x_G^*, \ldots, x_G^*) &\leq (1-\rho)^{\sum_{i=1}^m (x_i^*-1) + m_1(\sigma_0 + x_G^* - 1) + (G - m_1 - m)(x_G^* - 1)} \alpha \\ &= f^*(x_1^*, \ldots, x_m^*, \underbrace{\sigma_0 + x_G^*, \ldots, \sigma_0 + x_G^*}_{m_1}, x_G^*, \ldots, x_G^*). \end{aligned} \tag{31}$$

The matching bounds (27) and (31) together complete the proof of induction on β. Moving back up in our nesting of inductions, this also completes the proof for $\{x | x_i \geq 0 \forall\, i \in [G]\}$.

Case of $x_G < 0$: We now address the remaining case of $\{x \mid \min_{i \in [G]} x_i < 0\}$ and show that $f(x) = f^*(x) = 0$ for all such x. The arguments above for the case $\{x \mid \min_{i \in [G]} x_i > 0\}$ imply that for any strictly proper scoring rule f, the inequality (14a) in the statement of Lemma B.1 must be satisfied with an equality. This allows us to employ the second part of Lemma B.1 and we do so in the following manner: For every $i \in [G]$, let $y_i = y_i' = x_i$ if $x_i > 0$, and $y_i - 1 = y_i' = |x_i|$ otherwise; set $y_i = y_i' = B$ for all $i \in \{G+1, \ldots, N\}$. Then Equation (14b) in the statement of Lemma B.1 yields $f(x_1, \ldots, x_G) = 0$, thus completing the proof.

B.5.2 Part (b). We assume without loss of generality that $\alpha_{\min} = 0$.

First consider the case of any x such that $x_i < 0$ for some $i \in [G]$. For this case, we have

$$f(x) \geq 0 = f^*(x),$$

thereby proving our claim.

Now consider the remaining case where $x_i > 0$ for every $i \in [G]$. For the setting $N = G$ under consideration, the inequality (14a) simplifies to

$$f(y_1, \ldots, y_G) \geq (1-\rho)^{|\mathcal{I}|} f(y_1', \ldots, y_G'),$$

for any set $\mathcal{I} \in [G]$, and any $y, y' \in [B]^G$ such that $y_i = y_i' + 1$ for every $i \in \mathcal{I}$ and $y_i = y_i'$ otherwise. A repeated application of this inequality yields the bound

$$f(y_1, \ldots, y_G) \geq (1-\rho)^{\sum_{i=1}^G (y_i - 1)} f(1, \ldots, 1) = (1-\rho)^{\sum_{i=1}^G (y_i - 1)} \alpha_{\max} = f^*(y_1, \ldots, y_G),$$

for any $y \in [B]^G$, thereby proving the claimed result.

B.5.3 Part (c). The proof immediately follows from the observations that

$$f^*(x_1, \ldots, x_G) = \alpha_{\min},$$

whenever any of the evaluations $x_1, \ldots, x_G$ take a negative value, and that

$$f(x_1, \ldots, x_G) \geq \alpha_{\min},$$

for every evaluation $(x_1, \ldots, x_G)$.

B.5.4 Part (d). We consider the regime $N > G$, since the case of $N = G$ has already been studied in part (b) of this proof.

First consider the case of any x such that $x_i < 0$ for some $i \in [G]$. For this case, we have

$$f(x) \geq 0 = f^*(x),$$

thereby proving our claim.

Now consider the remaining case where $x_i > 0$ for every $i \in [G]$. Define a function $\nu : [B]^G \rightarrow \{0, \ldots, G\}$ that measures the number of entries equaling 1 in the input, that is,

$$\nu(x) = \sum_{i=1}^{G} \mathbf{1}\{x_i = 1\},$$

for every $x \in [B]^G$.

First suppose that $\nu(x) = G$. In this case, we must have $x = (1, \ldots, 1)$. Consequently, we have $f(x) = \alpha_{\max} = f^*(x)$.

Now for some value $\nu_0 \in [G]$, suppose that $f(x) = f^*(x)$ for every $\{x \mid \nu(x) \geq \nu_0\}$, as in the statement of the theorem. In what follows, we show that any x satisfying $\nu(x) = \nu_0 - 1$ must also satisfy $f(x') > f^*(x')$.

Consider any $x \in [B]^G$ such that $\nu(x) = \nu_0 - 1$. We assume, without loss of generality, that $x_1 \geq \cdots \geq x_G$ (≥ 0). Then, we have $x_{G-\nu_0+2} = \cdots = x_G = 1$ and $x_{G-\nu_0+1} > 1$. Define a vector $y = (x, 1, \ldots, 1) = (x_1, \ldots, x_{G-\nu_0+1}, 1, \ldots, 1)$. We now apply Equation (14a) from Lemma B.1 with $y' = (x_1 - 1, x_2, \ldots, x_{G-\nu_0+1}, 1, \ldots, 1)$ and $\mathcal{I} = \{1\}$ to obtain the bound

$$\begin{aligned} c_1 & \sum_{(j_2, \ldots, j_G) \subseteq \{2, \ldots, N\}} f(x_1, y_{j_2}, \ldots, y_{j_G}) + c_2 \sum_{(j_1, \ldots, j_G) \subseteq \{2, \ldots, N\}} f(y_{j_1}, \ldots, y_{j_G}) \\ & \geq c_1 \sum_{(j_2, \ldots, j_G) \subseteq \{2, \ldots, N\}} (1-\rho) f(x_1 - 1, y_{j_2}, \ldots, y_{j_G}) + c_2 \sum_{(j_1, \ldots, j_G) \subseteq \{2, \ldots, N\}} f(y_{j_1}, \ldots, y_{j_G}), \end{aligned}$$

for some constants $c_1 > 0$ and $c_2 > 0$ whose values depend only on N and G. Canceling out common terms, we are left with

$$\sum_{(j_2, \ldots, j_G) \subseteq \{2, \ldots, N\}} f(x_1, y_{j_2}, \ldots, y_{j_G}) \geq \sum_{(j_2, \ldots, j_G) \subseteq \{2, \ldots, N\}} (1-\rho) f(x_1 - 1, y_{j_2}, \ldots, y_{j_G}). \tag{32}$$

Both the left- and right-hand sides of this inequality involve linear combinations of the function f evaluated at various points. For our chosen values of y and y', observe that whenever $\{2, \ldots, G - \nu_0 + 1\} \nsubseteq \{j_2, \ldots, j_G\}$, we must have $\nu(x_1, y_{j_2}, \ldots, y_{j_G}) \geq \nu_0$ and $\nu(x_1 - 1, y_{j_2}, \ldots, y_{j_G}) \geq \nu_0$. Consequently, for any such value of $(j_2, \ldots, j_G)$, we have

$$f(x_1, y_{j_2}, \ldots, y_{j_G}) = f^*(x_1, y_{j_2}, \ldots, y_{j_G}) = (1-\rho) f^*(x_1 - 1, y_{j_2}, \ldots, y_{j_G}) = (1-\rho) f(x_1 - 1, y_{j_2}, \ldots, y_{j_G}).$$

For any remaining value of $(j_2, \ldots, j_G)$ (such that $\{2, \ldots, G - \nu_0 + 1\} \subseteq \{j_2, \ldots, j_G\}$), we have $(y_{j_2}, \ldots, y_{j_G}) = (x_2, \ldots, x_{G-\nu_0+1}, 1, \ldots, 1)$. Substituting these relations in Equation (32) and canceling out common terms leaves us with the bound

$$f(x_1, x_2, \ldots, x_{G-\nu_0+1}, 1, \ldots, 1) \geq (1-\rho) f(x_1 - 1, x_2, \ldots, x_{G-\nu_0+1}, 1, \ldots, 1).$$

A repeated application of this inequality for different values of x_1 then yields

$$\begin{aligned} f(x_1, x_2, \ldots, x_{G-\nu_0+1}, 1, \ldots, 1) & \geq (1-\rho)^{x_1 - 1} f(1, x_2, \ldots, x_{G-\nu_0+1}, 1, \ldots, 1) \\ & \overset{(i)}{=} (1-\rho)^{x_1 - 1} f^*(1, x_2, \ldots, x_{G-\nu_0+1}, 1, \ldots, 1) \\ & = f^*(x_1, x_2, \ldots, x_{G-\nu_0+1}, 1, \ldots, 1), \end{aligned}$$

as claimed, where equation (i) is because $\nu(1, x_2, \ldots, x_{G-\nu_0+1}, 1, \ldots, 1) \geq \nu_0$.

B.5.5 Proof of Lemma B.1. From Corollary 4.4, we know that the set of all strictly proper scoring rules for relative thresholding is a (not necessarily strict) subset of the set of all strictly proper scoring rules for support elicitation under the coarse beliefs assumption. Consequently, we consider f as any strictly proper scoring rule for eliciting support under the coarse beliefs assumption, and the result for the more general setting of relative thresholding follows directly.

Consider a real number $\rho_0 \in (\rho, \frac{1}{B})$ whose precise value will be specified later. Consider a worker such that for every question $i \in \mathcal{I}$, her belief is ρ_0 for the first option, is $\frac{1-\rho_0}{y_i-1}$ for each of the last $(y_i - 1)$ options, and is 0 for the remaining options. For every question $i \notin \mathcal{I}$, her belief is uniformly distributed among the first y_i options. Note that this satisfies the coarse beliefs assumption, since

$$\frac{1-\rho_0}{y_i - 1} \overset{(i)}{\geq} \frac{1-\rho_0}{B-1} \overset{(ii)}{>} \frac{1-\frac{1}{B}}{B-1} \geq \frac{1}{B} \overset{(iii)}{>} \rho,$$

where the inequality (i) follows from the fact that $y_i \leq B$, inequality (ii) follows from the assumption $\rho_0 < \frac{1}{B}$, and inequality (iii) is an assumption on ρ (see Definition 4).

If this worker selects precisely the support of her beliefs for every question, then her expected payment Ψ_1 is

$$\Psi_1 = \frac{1}{\binom{N}{G}} \sum_{(j_1,\ldots,j_G)\subseteq[N]} f(y_{j_1}, \ldots, y_{j_G}). \tag{33}$$

We will compare the aforementioned action to another action, where for each question $i \in \mathcal{I}$, the worker selects only the last $y'_i = (y_i - 1)$ options but not the first option; for each question $i \notin \mathcal{I}$, the worker selects the support of her belief. Under this action, the expected payment Ψ_2 equals

$$\begin{aligned}\Psi_2 = \frac{1}{\binom{N}{G}} & \sum_{\substack{(j_1, \ldots, j_G)\\ \subseteq[N]}} \\ & \times \sum_{\substack{(\epsilon_1, \ldots, \epsilon_G)\\ \in \{-1, 1\}^G}} \mathbf{1}\{\{j_i \mid \epsilon_i = -1\} \subseteq \mathcal{I}\}(1-\rho_0)^{|\mathcal{I}\cap\{j_i|\epsilon_i=1\}|}\rho_0^{|\mathcal{I}\cap\{j_i|\epsilon_i=-1\}|} f(\epsilon_1 y'_{j_1}, \ldots, \epsilon_G y'_{j_G}).\end{aligned} \tag{34}$$

(34)

In the expression (34), the outer summation represents the expectation over the random choice of the G gold standard questions among the N questions. The inner summation represents the expectation with respect to the correctness or incorrectness of the answers to the G gold standard questions: for any question i, $\epsilon_i = 1$ captures the event where the ith question in the gold standard is answered correctly and $\epsilon_i = -1$ represents the event of this question being answered incorrectly. The term $\mathbf{1}\{\{j_i \mid \epsilon_i = -1\} \subseteq \mathcal{I}\}$ ensures that only the questions in $\mathcal{I}$ can be wrong, since it is only these questions for which the worker has selected a subset of her belief's support.

Since $f(x) \geq 0$ for every valid argument x, we can lower bound Ψ_2 as

$$\Psi_2 \geq \frac{1}{\binom{N}{G}} \sum_{(j_1,\ldots,j_G)\subseteq[N]} (1-\rho_0)^{|\mathcal{I}\cap\{j_1,\ldots,j_G\}|} f(y'_{j_1}, \ldots, y'_{j_G}). \tag{35}$$

A strictly proper scoring rule must incentivize the worker to perform the first action (over the second), i.e, must have $\Psi_1 > \Psi_2$. Thus, from Equations (33) and (35), we get

$$\sum_{(j_1,\ldots,j_G)\subseteq[N]} f(y_{j_1}, \ldots, y_{j_G}) > \sum_{(j_1,\ldots,j_G)\subseteq[N]} (1-\rho_0)^{|\mathcal{I}\cap\{j_1,\ldots,j_G\}|} f(y'_{j_1}, \ldots, y'_{j_G}). \tag{36}$$

Note that Equation (36) must hold for all $\rho_0 > \rho$. The left-hand side of Equation (36) does not involve ρ_0, whereas the right-hand side is continuous in ρ_0. It follows that

$$\sum_{(j_1,\ldots,j_G)\subseteq[N]} f(y_{j_1},\ldots,y_{j_G}) \geq \sum_{(j_1,\ldots,j_G)\subseteq[N]} (1-\rho)^{|\mathcal{I}\cap\{j_1,\ldots,j_G\}|} f(y'_{j_1},\ldots,y'_{j_G}). \tag{37}$$

This proves the first part (14a) of the lemma.

We now move on to prove the second part (14b) of the lemma. We prove the claimed result by means of a contradiction argument. Suppose that $f(\epsilon_1 y'_{j_1},\ldots,\epsilon_G y'_{j_G})$ is strictly positive for some $(j_1,\ldots,j_G)\subseteq[N], \{(\epsilon_1,\ldots,\epsilon_G)\in\{-1,1\}^G\backslash\{1\}^G \mid \epsilon_i = 1 \text{ whenever } j_i \notin \mathcal{I}\}$. Then, using the fact that $f(x) \geq 0$ for all x, from Equation (34), we obtain the inequality

$$\Psi_2 \geq \sum_{(j_1,\ldots,j_G)\subseteq[N]} (1-\rho_0)^{|\mathcal{I}\cap\{j_1,\ldots,j_G\}|} f(y'_{j_1},\ldots,y'_{j_G}) \\ + (1-\rho_0)^{|\mathcal{I}\cap\{j_i|\epsilon_i=1\}|}\rho_0^{|\mathcal{I}\cap\{j_i|\epsilon_i=-1\}|} f(\epsilon_1 y'_{j_1},\ldots,\epsilon_G y'_{j_G}).$$

Note that this inequality is the equivalent of Equation (35) in the first part of the lemma, but also accounts for the additional strictly positive term. Following arguments identical to those in the first part, we have the following tighter version of Equation (37):

$$\sum_{(j_1,\ldots,j_G)\subseteq[N]} f(y_{j_1},\ldots,y_{j_G}) \geq \sum_{(j_1,\ldots,j_G)\subseteq[N]} (1-\rho)^{|\mathcal{I}\cap\{j_1,\ldots,j_G\}|} f(y'_{j_1},\ldots,y'_{j_G}) \\ + (1-\rho)^{|\mathcal{I}\cap\{j_i|\epsilon_i=1\}|}\rho^{|\mathcal{I}\cap\{j_i|\epsilon_i=-1\}|} f(\epsilon_1 y'_{j_1},\ldots,\epsilon_G y'_{j_G}).$$

Since $f(\epsilon_1 y'_{j_1},\ldots,\epsilon_G y'_{j_G}) > 0$, we then have

$$\sum_{(j_1,\ldots,j_G)\subseteq[N]} f(y_{j_1},\ldots,y_{j_G}) > \sum_{(j_1,\ldots,j_G)\subseteq[N]} (1-\rho)^{|\mathcal{I}\cap\{j_1,\ldots,j_G\}|} f(y'_{j_1},\ldots,y'_{j_G}),$$

thereby contradicting the hypothesis of the equality in Equation (14a) assumed in the second part of the lemma, and hence proving the claimed result.

B.6 Proof of Proposition A.1: Impossibility of Eliciting Support in Absence of Coarse Beliefs Assumption

We assume that there indeed exists some strictly proper scoring rule f and prove a contradiction.

Let us first consider the special case of $N = G = 1$ and $B = 2$. Since $N = G = 1$, there is only one question. Let $p_1 > 0.5$ be the probability, according to the belief of the worker, that option 1 is correct; the worker then believes that option 2 is correct with probability $(1 - p_1)$.

When $p_1 = 1$, we need the worker to select option 1 alone. Thus, we need

$$f(1) > f(2).$$

When $p_1 \in (0.5, 1)$, we require the worker to select options 1 and 2, as opposed to selecting option 1 alone. For this, we need

$$p_1 f(1) + (1-p_1) f(-1) < f(2).$$

It follows that we need

$$(1-p_1)(f(1)-f(-1)) > f(1)-f(2). \tag{38}$$

However, the inequality (38) is satisfied only when $f(1) > f(-1)$ and $(1-p_1) > \frac{f(1)-f(2)}{f(1)-f(-1)}$. Thus, for any given payment function f, a worker with belief $(1-p_1) \in (0, \frac{f(1)-f(2)}{f(1)-f(-1)})$ will not be incentivized to select the support of her belief. This yields a contradiction.

We now move on to the general case of $N \geq G \geq 1$ and $B \geq 2$. Consider a worker who is clueless about questions 2 through N (i.e., her belief is uniform across all options for these questions). Suppose this worker selects all B options for these questions as desired. For the first question, suppose that the worker is sure that options $3, \ldots, B$ are incorrect. We are now left with the first question and the first two options for this question. Letting X denote a random variable representing the evaluation of the worker's response to the first question, the expected payment then is

$$\frac{G}{N}\mathbb{E}[f(X, B, \ldots, B)] + \left(1 - \frac{G}{N}\right) f(B, \ldots, B).$$

The expectation in the first term is taken with respect to the randomness in X. Defining

$$\tilde{f}(X) := \frac{G}{N} f(X, B, \ldots, B) + \left(1 - \frac{G}{N}\right) f(B, \ldots, B),$$

and applying the same arguments to $\tilde{f}$ as those for f for the case of $N = G = 1$, $B = 2$ above gives the desired contradiction. This thus completes the proof of impossibility.

B.7 Proof of Proposition A.2: Impossibility of Incentivizing Exact σ-belief

Let us first prove the result for the case of $N = G = 1$. The result of Theorem 3.2 implies that if there does exist a strictly proper scoring rule for this setting, then it must be Scoring rule 1 up to a constant shift and positive scale. Consider a worker with the belief $p_1 = 1 - \sigma$, $p_2 = \sigma$, and $p_3 = \cdots p_B = 0$. Since $\sigma < \frac{1}{2}$, under a strictly proper scoring rule, the expected payment must be strictly larger if the worker selects only option 1 as compared to the expected payment when the worker selects options 1 and 2. However, one can compute that under Scoring rule 1, the expected payment in the two cases is identical. It follows that under any possible strictly proper scoring rule, the expected payment must be identical in the two following two actions of the worker (a) selecting only option 1, and (b) selecting options 1 and 2. It follows that there is no strictly proper scoring rule.

We now move on to the general case of $N \geq G \geq 1$. Let f denote any strictly proper scoring rule for the setting at hand. Consider a worker who knows the answers to questions 2 through N with a belief of 1 in each case. Suppose that for each of these $(N - 1)$ questions, this worker selects the respective options that she thinks are correct. We are now left with the first question. Letting X denote a random variable representing the evaluation of the worker's response to the first question, the expected payment from the worker's point of view is

$$\frac{G}{N}\mathbb{E}[f(X, 1, \ldots, 1)] + \left(1 - \frac{G}{N}\right) f(1, \ldots, 1).$$

The expectation in the first term is taken with respect to the randomness in X. Defining

$$\tilde{f}(X) := \frac{G}{N} f(X, 1, \ldots, 1) + \left(1 - \frac{G}{N}\right) f(1, \ldots, 1),$$

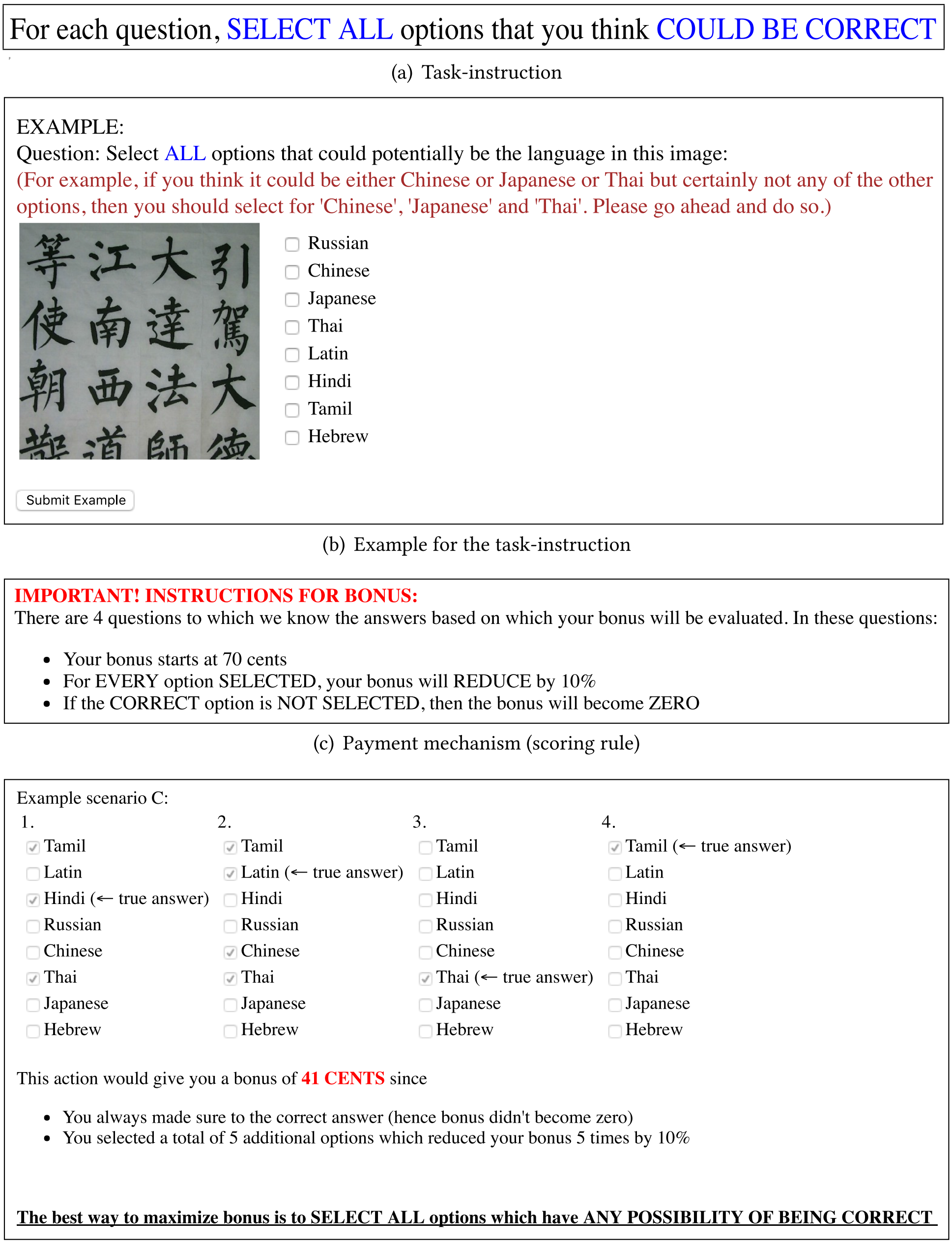
and applying the same arguments to $\tilde{f}$ as those for f for the case of $N = G = 1$ above gives the desired contradiction. This completes the proof.

C ADDITIONAL DETAILS OF EXPERIMENTS

In this section, we provide additional details of the experiments discussed in Section 5 in the main text. We also remind the reader that the complete set of interfaces presented to the workers as well as the data obtained from the workers is available on the website of the first author at:

https://cs.cmu.edu/~nihars/data/data_approval.zip.

We recruited 215 each for the three experiments. Only those workers with more than 500 tasks previously accepted and an acceptance rate of at least 95% were allowed (these are standard quality



For each question, SELECT ALL options that you think COULD BE CORRECT

(a) Task-instruction

EXAMPLE:
Question: Select ALL options that could potentially be the language in this image:
(For example, if you think it could be either Chinese or Japanese or Thai but certainly not any of the other options, then you should select for 'Chinese', 'Japanese' and 'Thai'. Please go ahead and do so.)

- □ Russian
- □ Chinese
- □ Japanese
- □ Thai
- □ Latin
- □ Hindi
- □ Tamil
- □ Hebrew

(b) Example for the task-instruction

IMPORTANT! INSTRUCTIONS FOR BONUS:
There are 4 questions to which we know the answers based on which your bonus will be evaluated. In these questions:

- Your bonus starts at 70 cents
- For EVERY option SELECTED, your bonus will REDUCE by 10%
- If the CORRECT option is NOT SELECTED, then the bonus will become ZERO

(c) Payment mechanism (scoring rule)

Example scenario C:

1.	2.	3.	4.
☑ Tamil	☑ Tamil	□ Tamil	☑ Tamil (← true answer)
□ Latin	☑ Latin (← true answer)	□ Latin	□ Latin
☑ Hindi (← true answer)	□ Hindi	□ Hindi	□ Hindi
□ Russian	□ Russian	□ Russian	□ Russian
□ Chinese	☑ Chinese	□ Chinese	□ Chinese
☑ Thai	☑ Thai	☑ Thai (← true answer)	□ Thai
□ Japanese	□ Japanese	□ Japanese	□ Japanese
□ Hebrew	□ Hebrew	□ Hebrew	□ Hebrew

This action would give you a bonus of **41 CENTS** since

- You always made sure to the correct answer (hence bonus didn't become zero)
- You selected a total of 5 additional options which reduced your bonus 5 times by 10%

The best way to maximize bonus is to SELECT ALL options which have ANY POSSIBILITY OF BEING CORRECT

(d) Examples illustrating the scoring rule

Fig. 4. The interface for the instructions presented to the workers in the experiments.

control settings on Amazon Mechanical Turk). In the experiments on identifying animals, languages, and textures, respectively, 12, 2, and 12 workers exited the system before submitting their answers. Thus, we had a total of 203, 213, and 203 workers, respectively, in the three experiments. Each worker in any experiment was randomly assigned to one of the four interfaces/scoring-rules. The breakdown of the number of workers in the four interfaces/scoring-rules is as follows:

	Single+additive	Skip+multiplicative	Approval+fixed	Approval+Scoring rule 2
Animals	54	49	53	57
Languages	54	40	64	45
Textures	67	40	45	51

The four components of the interface, referred to in Section 5, are illustrated in Figure 4. This illustration pertains to the approval voting interface with the multiplicative Scoring rule 2, and we note that the general pattern of providing three examples to illustrate the scoring rule, etc., remains consistent throughout the three experiments and the four settings tested.

REFERENCES

[1] Nima Anari, Gagan Goel, and Afshin Nikzad. 2014. Mechanism design for crowdsourcing: An optimal 1-1/e competitive budget-feasible mechanism for large markets. In *Proceedings of the Symposium on Foundations of Computer Science (FOCS'14)*. 266–275.

[2] Steven J. Brams and Peter C. Fishburn. 1978. Approval voting. *Amer. Polit. Sci. Rev.* 72, 03 (1978), 831–847.

[3] Steven J. Brams and D. Marc Kilgour. 2014. Satisfaction approval voting. In *Voting Power and Procedures.* Springer, 323–346.

[4] Glenn W. Brier. 1950. Verification of forecasts expressed in terms of probability. *Month. Weath. Rev.* 78, 1 (1950), 1–3.

[5] Yang Cai, Constantinos Daskalakis, and Christos H. Papadimitriou. 2015. Optimum statistical estimation with strategic data sources. In *Proceedings of the Conference on Learning Theory.*

[6] Ioannis Caragiannis, Dimitris Kalaitzis, and Evangelos Markakis. 2010. Approximation algorithms and mechanism design for minimax approval voting. In *Proceedings of the AAAI Conference on Artificial Intelligence.*

[7] Xi Chen, Paul N. Bennett, Kevyn Collins-Thompson, and Eric Horvitz. 2013. Pairwise ranking aggregation in a crowdsourced setting. In *Proceedings of the ACM International Conference on Web Search and Data Mining.* 193–202.

[8] Leverne S. Collet. 1971. Elimination scoring: An empirical evaluation. *J. Educ. Meas.* 8, 3 (1971), 209–214.

[9] Clyde H. Coombs. 1953. On the use of objective examinations. *Educ. Psychol. Meas.* 13, 2 (1953), 308–310.

[10] Clyde H. Coombs, John Edgar Milholland, and Frank Burton Womer. 1956. The assessment of partial knowledge. *Educ. Psychol. Meas.* 16, 1 (1956), 13–37.

[11] Anirban Dasgupta and Arpita Ghosh. 2013. Crowdsourced judgement elicitation with endogenous proficiency. In *Proceedings of the 22nd International Conference on World Wide Web.* International World Wide Web Conferences Steering Committee, 319–330.

[12] Alexander Philip Dawid and Allan M. Skene. 1979. Maximum likelihood estimation of observer error-rates using the EM algorithm. *Appl. Statist.* 28, 1 (1979), 20–28.

[13] Ofer Dekel, Felix Fischer, and Ariel D. Procaccia. 2008. Incentive compatible regression learning. In *Proceedings of the ACM-SIAM Symposium on Discrete Algorithms.* 884–893.

[14] Jean-Paul Doignon, Aleksandar Pekeč, and Michel Regenwetter. 2004. The repeated insertion model for rankings: Missing link between two subset choice models. *Psychometrika* 69, 1 (2004), 33–54.

[15] J.-C. Falmagne and Michael Regenwetter. 1996. A random utility model for approval voting. *J. Math. Psychol.* 40, 2 (1996), 152–159.

[16] Boi Faltings, Radu Jurca, Pearl Pu, and Bao Duy Tran. 2014. Incentives to counter bias in human computation. In *Proceedings of the 2nd AAAI Conference on Human Computation and Crowdsourcing.*

[17] Fang Fang, Maxwell Stinchcombe, and Andrew Whinston. 2007. "Putting your money where your mouth is"—A betting platform for better prediction. *Rev. Netw. Econ.* 6, 2 (2007).

[18] Rafael M. Frongillo, Yiling Chen, and Ian A. Kash. 2014. Elicitation for aggregation. *arXiv preprint arXiv:1410.0375* (2014).

[19] Ujwal Gadiraju, Ricardo Kawase, Stefan Dietze, and Gianluca Demartini. 2015. Understanding malicious behavior in crowdsourcing platforms: The case of online surveys. In *Proceedings of the 33rd ACM Conference on Human Factors in Computing Systems.* ACM, 1631–1640.

[20] Jean D. Gibbons, Ingram Olkin, and Milton Sobel. 1979. A subset selection technique for scoring items on a multiple choice test. *Psychometrika* 44, 3 (1979), 259–270.

[21] Tilmann Gneiting and Adrian E. Raftery. 2007. Strictly proper scoring rules, prediction, and estimation. *J. Amer. Statist. Assoc.* 102, 477 (2007), 359–378.

[22] Chien-Ju Ho, Shahin Jabbari, and Jennifer W. Vaughan. 2013. Adaptive task assignment for crowdsourced classification. In *Proceedings of the International Conference on Machine Learning (ICML'13).* 534–542.

[23] Chien-Ju Ho, Yu Zhang, Jennifer Vaughan, and Mihaela Van Der Schaar. 2012. Towards social norm design for crowdsourcing markets. In *Proceedings of the AAAI Workshops.*

[24] Paul Horst. 1932. The chance element in the multiple choice test item. *J. Gen. Psychol.* 6, 1 (1932), 209–211.

[25] Harold Hotelling. 1951. A generalized T test and measure of multivariate dispersion. In *Proceedings of the Second Berkeley Symposium on Mathematical Statistics and Probability.* https://digitalassets.lib.berkeley.edu/math/ucb/text/math_s2_article-03.pdf.

[26] Panagiotis G. Ipeirotis, Foster Provost, Victor S. Sheng, and Jing Wang. 2014. Repeated labeling using multiple noisy labelers. *Data Mining Knowl. Disc.* 28, 2 (2014), 402–441.

[27] Shaili Jain, Yiling Chen, and David C. Parkes. 2009. Designing incentives for online question and answer forums. In *Proceedings of the 10th ACM Conference on Electronic Commerce.* 129–138.

[28] W. Paul Jones and Scott A. Loe. 2013. Optimal number of questionnaire response categories more may not be better. *SAGE Open* 3, 2 (2013), 2158244013489691.

[29] Radu Jurca and Boi Faltings. 2009. Mechanisms for making crowds truthful. *J. Artif. Intell. Res.* 34 (2009), 209–253.

[30] David R. Karger, Sewoong Oh, and Devavrat Shah. 2011. Iterative learning for reliable crowdsourcing systems. In *Proceedings of the Conference on Advances in Neural Information Processing Systems.* 1953–1961.

[31] Gabriella Kazai, Jaap Kamps, Marijn Koolen, and Natasa Milic-Frayling. 2011. Crowdsourcing for book search evaluation: Impact of HIT design on comparative system ranking. In *Proceedings of the ACM SIGIR Conference on Research and Development in Information Retrieval.* 205–214.

[32] John Kellett and Kenneth Mott. 1977. Presidential primaries: Measuring popular choice. *Polity* 9, 4 (1977), 528–537.

[33] Ashish Khetan and Sewoong Oh. 2016. Reliable crowdsourcing under the generalized Dawid-Skene model. *arXiv preprint arXiv:1602.03481* (2016).

[34] Nicolas Lambert and Yoav Shoham. 2008. Truthful surveys. In *Proceedings of the International Workshop on Internet and Network Economics.* Springer, 154–165.

[35] Nicolas Lambert and Yoav Shoham. 2009. Eliciting truthful answers to multiple-choice questions. In *Proceedings of the 10th ACM Conference on Electronic Commerce.* 109–118.

[36] Jean-François Laslier and Karine Van der Straeten. 2008. A live experiment on approval voting. *Experim. Econ.* 11, 1 (2008), 97–105.

[37] Qiang Liu, Jian Peng, and Alexander T. Ihler. 2012. Variational inference for crowdsourcing. In *Proceedings of the Conference on Advances in Neural Information Processing Systems (NIPS'12).* 701–709.

[38] A. A. J. Marley. 1993. Aggregation theorems and the combination of probabilistic rank orders. In *Probability Models and Statistical Analyses for Ranking Data.* Springer, 216–240.

[39] Jordi Massó and Marc Vorsatz. 2008. Weighted approval voting. *Econ. Theor.* 36, 1 (2008), 129–146.

[40] George A Miller. 1956. The magical number seven, plus or minus two: Some limits on our capacity for processing information.*Psychol. Rev.* 63, 2 (1956), 81.

[41] Nolan Miller, Paul Resnick, and Richard Zeckhauser. 2005. Eliciting informative feedback: The peer-prediction method. *Manag. Sci.* 51, 9 (2005), 1359–1373.

[42] Sendhil Mullainathan, Joshua Schwartzstein, and Andrei Shleifer. 2008. Coarse thinking and persuasion. *Quart. J. Econ.* 123, 2 (2008), 577–619.

[43] Ritesh Noothigattu, Nihar Shah, and Ariel Procaccia. 2018. Loss functions, axioms, and peer review. *arXiv preprint arxiv:1808.09057* (2018).

[44] Guy Ottewell. 1977. The arithmetic of voting. *In Defence of Variety* (1977). http://www.universalworkshop.com/ARVOfull.htm.

[45] Dražen Prelec. 2004. A Bayesian truth serum for subjective data. *Science* 306, 5695 (2004), 462–466.

[46] Ariel D. Procaccia and Nisarg Shah. 2015. Is approval voting optimal given approval votes? In *Proceedings of the Conference on Advances in Neural Information Processing Systems (NIPS'15).*

[47] Vikas C. Raykar, Shipeng Yu, Linda H. Zhao, Gerardo Hermosillo Valadez, Charles Florin, Luca Bogoni, and Linda Moy. 2010. Learning from crowds. *J. Mach. Learn. Res.* 11 (2010), 1297–1322.

[48] Michel Regenwetter and Ilia Tsetlin. 2004. Approval voting and positional voting methods: Inference, relationship, examples. *Soc. Choice Welf.* 22, 3 (2004), 539–566.

[49] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. 2015. Imagenet large scale visual recognition challenge. *Int. J. Comput. Vis.* 115, 3 (2015), 211–252.

[50] Thomas L. Saaty and Mujgan S. Ozdemir. 2003. Why the magic number seven plus or minus two. *Math. Comput. Modell.* 38, 3 (2003), 233–244.

[51] Leonard J. Savage. 1971. Elicitation of personal probabilities and expectations. *J. Amer. Statist. Assoc.* 66, 336 (1971), 783–801.

[52] Nihar B. Shah, Sivaraman Balakrishnan, Joseph K. Bradley, Abhay Parekh, Kannan Ramchandran, and Martin Wainwright. 2015. Estimation from pairwise comparisons: Sharp minimax bounds with topology dependence. In *Proceedings of the International Conference on Artificial Intelligence and Statistics (AIStats'15).*

[53] Nihar B. Shah, Sivaraman Balakrishnan, and Martin J. Wainwright. 2016. A permutation-based model for crowd labeling: Optimal estimation and robustness. *arXiv preprint arXiv:1606.09632* (2016).

[54] Nihar B. Shah, Behzad Tabibian, Krikamol Muandet, Isabelle Guyon, and Ulrike Von Luxburg. 2018. Design and analysis of the Advances in Neural Information Processing Systems NIPS 2016 review process. *J. Mach. Learn. Res.* 19, 1 (2018), 1913–1946.

[55] Nihar B. Shah and Dengyong Zhou. 2015. Double or nothing: Multiplicative incentive mechanisms for crowdsourcing. In *Proceedings of the Conference on Advances in Neural Information Processing Systems (NIPS'15).*

[56] R. M. Shiffrin and R. M. Nosofsky. 1994. Seven plus or minus two: A commentary on capacity limitations. *Psychol. Rev.* 101, 2 (1994), 357.

[57] Hammad Siddiqi. 2011. Does coarse thinking matter for option pricing? Evidence from an experiment. *IUP J. Behav. Fin.* 8, 2 (2011).

[58] Ivan Stelmakh, Nihar B. Shah, and Aarti Singh. 2019. PeerReview4All: Fair and accurate reviewer assignment in peer review. In *Proceedings of the 30th International Conference on Algorithmic Learning Theory.* 828–856.

[59] Jeroen Vuurens, Arjen P. de Vries, and Carsten Eickhoff. 2011. How much spam can you take? An analysis of crowdsourcing results to increase accuracy. In *Proceedings of the ACM SIGIR Workshop on Crowdsourcing for Information Retrieval.* 21–26.

[60] Paul Wais, Shivaram Lingamneni, Duncan Cook, Jason Fennell, Benjamin Goldenberg, Daniel Lubarov, David Marin, and Hari Simons. 2010. Towards building a high-quality workforce with Mechanical Turk. In *Proceedings of the NIPS Workshop on Computational Social Science and the Wisdom of Crowds.*

[61] Robert J. Weber. 1977. *Comparison of Voting Systems. Cowles Foundation Discussion Paper A* 498. New Haven, CT.

[62] Jacob Whitehill, Paul Ruvolo, Ting-fan Wu, Jacob Bergsma, and Javier Movellan. 2009. Whose vote should count more: Optimal integration of labels from labelers of unknown expertise. In *Proceedings of the Conference on Advances in Neural Information Processing Systems.* 2035–2043.

[63] Yichong Xu, Han Zhao, Xiaofei Shi, and Nihar Shah. 2019. On strategyproof conference review. In *Proceedings of the International Joint Conference on Artificial Intelligence.*

[64] Man-Ching Yuen, Irwin King, and Kwong-Sak Leung. 2011. Task matching in crowdsourcing. In *Proceedings of the IEEE International Conference on Cyber, Physical and Social Computing.* 409–412.

[65] Yu Zhang and Mihaela van der Schaar. 2012. Reputation-based incentive protocols in crowdsourcing applications. In *Proceedings of the IEEE Conference on Computer Communications (INFOCOM'12).* IEEE, 2140–2148.

[66] Dengyong Zhou, Qiang Liu, John C. Platt, Christopher Meek, and Nihar B. Shah. 2015. Regularized minimax conditional entropy for crowdsourcing. *arXiv preprint arXiv:1503.07240* (2015).

[67] Dengyong Zhou, John Platt, Sumit Basu, and Yi Mao. 2012. Learning from the wisdom of crowds by minimax entropy. In *Proceedings of the Conference on Advances in Neural Information Processing Systems.* 2204–2212.

[68] James Zou, Reshef Meir, and David Parkes. 2014. Approval voting behavior in Doodle polls. In *Proceedings of the 5th Workshop on Computational Social Choice.*

Received September 2015; revised March 2020; accepted April 2020