Inter-Task Effects Induce Bias in Crowdsourcing

Edward Newell* and **Derek Ruths**
School of Computer Science, McGill University, Montreal, Canada
*edward.newell@mail.mcgill.ca

Abstract

Microtask platforms allow researchers to engage participants quickly and inexpensively. Workers on such platforms probably perform many tasks in succession, so we investigate interactions between earlier tasks and later ones, which we call *inter-task* effects. Existing research investigates many task design factors, such as *framing*, on the quality of responses, but to our knowledge, does not address inter-task effects. We used a canonical image-labeling task on Amazon Mechanical Turk to measure the impact of inter-task and framing effects on the focus and specificity of labels that workers provide. We found that inter-task effects had a much stronger impact than framing, and that workers provided more specific labels when labeling a series of images that were similar to one another.

Keywords: priming, framing, crowdsourcing.

Introduction

Microtask crowdsourcing platforms like Amazon Mechanical Turk (MTurk) make it possible to submit batches of small tasks to a large pool of human workers, who do the tasks for fun, a sense of purpose, and remuneration (Kazai, Kamps, and Milic-Frayling 2013; Antin and Shaw 2012). Originally used to distribute clerical work, these platforms increasingly serve as a fast and cheap means to engage experimental participants in a research setting (Snow et al. 2008).

The task requester can interact with the platform like a compute server, seamlessly integrating human and machine computation. Researchers have put forward the term HPU (Human co-Processing Unit), viewing the introduction of microtask platforms as a new computing architecture (Davis et al. 2010).

Here, we highlight an important way in which HPUs differ from CPUs, with serious implications for the design of tasks. It is well-known that people are subject to priming effects (Warren and Morton 1982; Noguera 2007; Beller 1971) and, in particular, task-repetition effects (Gass et al. 1999; Sohn and Anderson 2001). We investigate the effect that previously completed tasks have on workers' responses during subsequent ones. We call such effects *inter-task effects*. Inter-task effects would amount to a kind of *hysteresis*, meaning that HPU output is not only a function of the current input, but also of the history of inputs.

There has been considerable investigation into the factors that affect the quality and quantity of micro-task completion. These include the level of pay (Kazai, Kamps, and Milic-Frayling 2013), training (Le et al. 2010), screening of workers (Paolacci, Chandler, and Ipeirotis 2010), and user-interface design (Finnerty et al. 2013). Researchers have also investigated *framing*, by testing the effects of disclosing the workflow context (Kinnaird, Dabbish, and Kiesler 2012), and the purpose of tasks (Chandler and Kapelner 2013). To our knowledge, no study has investigated inter-task effects.

We investigated inter-task effects on the MTurk platform, using image-labeling tasks, one of the most common kinds of tasks on MTurk (Chandler and Kapelner 2013; Berinsky, Huber, and Lenz 2012; Finnerty et al. 2013; Paolacci, Chandler, and Ipeirotis 2010). Workers were required to label images featuring food and culture. We regard these tasks as consisting of an *initial* and a *test set*, but, crucially, no distinction was made between these sets from the view of the worker. We varied the images in the initial tasks, while keeping those in the test set the same, to analyze the effects that the initial tasks had on the content and specificity of labels attributed in the test set.

As a point of comparison, we subjected some groups of workers to a kind of *framing*, by disclosing a fictitious, semantically-loaded name for the requester funding the tasks. The names were chosen to suggest the requester's interest in a specific aspect of the image content. We expected this would cause workers to provide more labels, and greater specificity, relating to this "preferred" content.

Surprisingly, we found that inter-task effects were much stronger than framing. Our results show that initial tasks can significantly alter the focus of worker's labels. Interestingly, we find that inter-task effects can be used to induce greater specificity. Our results suggest that workers attribute more specific labels when labeling a series of images that that are more similar to one another. This suggests that careful consideration should be given to the bundling of tasks when designing a study using a microtask platform.

Experimental Setup

We solicited 900 MTurk workers to perform image-labeling tasks relating to food and culture. The workers were ran-

Treatment	Funder	Initial Image Set
AMBG	None	Ambiguous
CULT_{img}	None	Cultural
CULT_{fund}	Cultural	Cultural
$\text{CULT}_{fund,img}$	Cultural	Cultural
INGR_{img}	None	Ingredients
INGR_{fund}	Nutritional	Ingredients
$\text{INGR}_{fund,img}$	Nutritional	Ingredients

Table 1: Workers were assigned uniformly at random to one of the treatments listed above. The full funder names used were "The Global Foundation for Cultural Recognition" and "The National Foundation for Nutritional Awareness". The ambiguous, cultural, and ingredients initial image sets are shown in **Figs. S2**, **S3**, and **S4**.

domly assigned to one of the treatments shown in **Table 1**. The treatments $\text{INGR}_{fund,img}$ and $\text{CULT}_{fund,img}$ used (respectively) the image sets of INGR_{img} and CULT_{img}, but also incorporated framing. The addition of framing did not have a substantial impact on the results for these treatments, so we do not discuss $\text{INGR}_{fund,img}$ and $\text{CULT}_{fund,img}$ further.

Workers from all treatments were shown brief instructions. Depending on their treatment, workers were then shown the name of a research funder, or this step was skipped. Next, workers were given a series of ten image-labeling tasks. Each task required workers to privide five discriptive labels for an image. For the purpose of analysis, we divided the tasks into *initial* and *test* sets, comprising respectively the first five and last five tasks. From the perspective of the worker, there was no distinction or interruption between the initial and test sets. Depending on the treatment, one of three sets of images was used for the initial tasks, but the images in the test set were always the same.

The images from the test set contained prepared meals and featured a prominent, identifiable culture (see **Fig. S1**). To identify the initial image sets, we use the names "ambiguous", "cultural", and "ingredients". The ambiguous set was chosen to be similar to the test set, in the sense that it consisted of images of prepared meals (see **Fig. S2**), but its cultural features were less prominent. The cultural set featured iconic cultural scenes, but no food at all (see **Fig. S3**). Images from the ingredients set depicted separated ingredients, but, like the ambiguous set, avoided prominent cultural features (see **Fig. S4**).

Results

Earlier tasks oriented workers' focus during later tasks. Since the initial images were chosen to emphasize either food (ingredients set) or culture (cultural set), we looked for effects on the number of culture- and food-oriented labels that workers attributed to the test image set.

To this end, we constructed an ontology from the labels attributed to the test images. In the ontology, edges point from more general labels to more specific ones. For example, the ontology contains the path `food` $\rightarrow$ `ingredients` $\rightarrow$ `vegetables` $\rightarrow$ `tomato`.

Since food is a central feature of culture, our ontology contains many labels that have both `food` and `culture` in their ancestry. Nevertheless, there were many food-oriented labels, such as `bread`, which lacked specific cultural connections, as well as non-food, culture-oriented labels, such as `russian dolls`.

When we tallied labels attributed to the first image of the test set, we found that workers from CULT_{img} produced significantly more culture-oriented labels and less food-oriented ones than those from AMBG (see **Fig. 1A**). The inter-task effects were so strong that the proportion of food- and culture-oriented labels in CULT_{img} was essentially the reverse of that in AMBG, showing that inter-task effects *can* profoundly alter workers' focus.

Inter-task effects were stronger than framing effects. On the other hand, the labels attributed by INGR_{fund} were not significantly different in composition from those attributed by CULT_{fund}. Workers from the these treatments were told that the tasks were funded by, respectively, the "Foundation for Nutritional Awareness" and the "Foundation for Cultural Recognition". We find it remarkable that inter-task effects were stronger than those brought about by framing the tasks with reference to specific image content.

Earlier tasks influenced workers' level of specificity. The ontology described in the previous section allows us to define the relative specificity of two labels ℓ_1 and ℓ_2. We say that ℓ_2 is more specific than ℓ_1 if there is a *directed path* from ℓ_1 to ℓ_2. If there is no directed path between labels, we say they are *non-comparable*. For example, `tomato` was more specific than `food`, while `statue` and `food` were non-comparable.

We can then define the relative specificity of two workers with respect to test-image i as $s_i(u, v)$:

$$s_i(u,v) = \sum_{\ell \in u(i)} \sum_{m \in v(i)} \left(\mathbf{1}_{[\ell > m]} - \mathbf{1}_{[m > \ell]} \right), \qquad (1)$$

where $u(i)$ denotes the set of labels attributed by worker u to image i, and $\mathbf{1}_{[\ell > m]}$ evaluates to 1 if ℓ is more specific than m, and 0 otherwise. We can then define the relative specificity of two treatments, $\mathcal{U}$ and $\mathcal{V}$, with respect to the ith image, denoted $S_i(\mathcal{U}, \mathcal{V})$, to be the mean relative specificity of two uniformly drawn workers:

$$\hat{S}_i(\mathcal{U}, \mathcal{V}) = \frac{1}{|\mathcal{U}||\mathcal{V}|} \sum_{u \in \mathcal{U}} \sum_{v \in \mathcal{V}} s_i(u, v). \qquad (2)$$

The relative specificities of various treatments, based on **Eq. 2**, are shown in **Fig. 1B**. We found that workers from AMBG were more specific than workers from either CULT_{img} or INGR_{img}. Comparing CULT_{img} to INGR_{img}, we found that workers from INGR_{img} were more specific. This shows that inter-task effects do substantially influence the specificity of labels that workers provide. We will return to this point below, where we propose a mechanism to explain these differences in specificity.

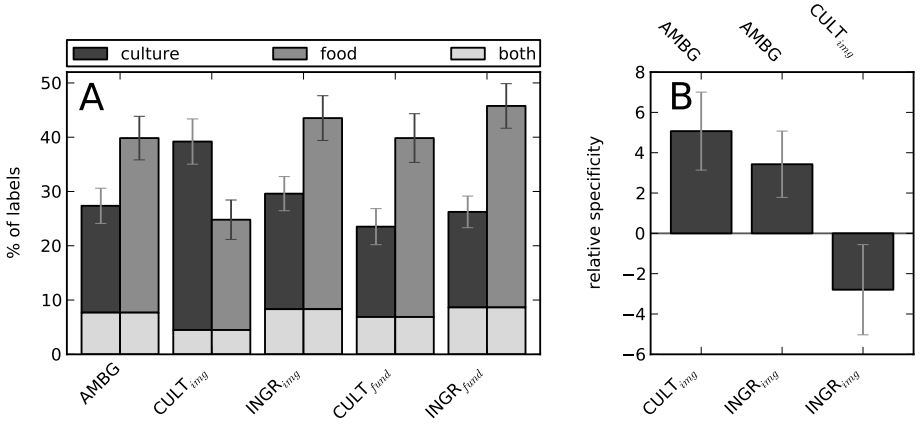

Figure 1: **A**) Percent label composition (culture- vs food-oriented labels) for various treatments. **B**) Relative specificities of treatments, indicated along the abscissa, compared to those indicated above the plot, according to **Eq. 2**. The size of the bar indicates how much more specific the labels from one treatment are compared to the other, and points toward the more specific treatment. Error bars indicate 95% confidence intervals.

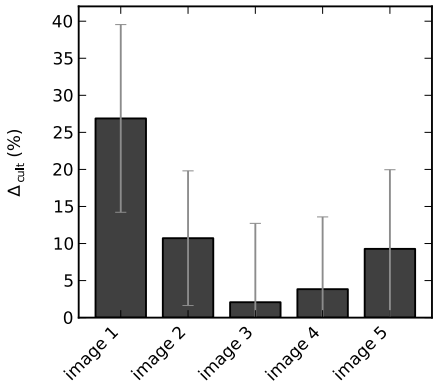

Figure 2: Excess cultural orientation ($\Delta_{\mathbf{cult}}$) of labels attributed by CULT$_{img}$ relative to those attributed by AMBG ($\Delta_{\mathbf{cult}}$ is defined in **Eq. 3**). Error bars indicate 95% confidence intervals.

Inter-task effects "washed out" quickly. It would stand to reason that, as workers proceed through the test images, priming from the initial images would be "washed out", diminishing the observed inter-task effects.

To investigate the evolution of inter-task effects, we define the *excess cultural orientation* to be the number of culture-oriented labels minus the number of food oriented ones. This measures how culture-oriented a given image is. Of course, we must account for the fact that some images inherently carry more cultural content than others. In keeping with our notion of priming *difference*, we calculate the excess cultural content for both CULT$_{img}$ and AMBG, and take their difference to be the *relative* excess cultural content, Δ_{cult}. Formally,

$$\Delta_{cult}(i) = \frac{1}{N}\left[\sum_{w\in \text{CULT}_{img}}\left(N^{(i)}_{w,cult} - N^{(i)}_{w,food}\right) - \sum_{w\in \text{AMBG}}\left(N^{(i)}_{w,cult} - N^{(i)}_{w,food}\right)\right], \tag{3}$$

where $N^{(i)}_{w,cult}$ stands for the number of culture-oriented labels attributed by worker w to image i, while $N^{(i)}_{w,food}$ similarly counts food-oriented labels, and N is the total number of labels in a treatment.

We found Δ_{cult} was largest for the first test image, but dropped off rapidly, remaining positive but not to a statistically significant extent (see **Fig. 2**).

Inter-task similarity encouraged more specific labels. To continue with the analysis, we sought a measure of image similarity. The characterization of image content is a deeply complex issue that has been approached by many disciplines (Panofsky 1939; Shatford 1986; Tversky 1977; Jaimes and Chang 2000). However, in the present study we are more interested how similar two sets of images are, *with respect to the labeling task*, which is simpler to operationalize than general perceptual similarity. For this purpose we measured the similarity between two sets of images by looking at the fraction of labels that they shared. Formally, to measure the similarity between two *sets of images*, X and Y, we computed the Jaccard index between the sets of labels attributed to them:

$$\text{Sim}(X,Y) = \frac{L(X)\cap L(Y)}{L(X)\cup L(Y)}, \tag{4}$$

where $L(X)$ denotes the set of labels attributed to X.

Image set	Ambig.	Cultural	Ingr.	Test
Ambiguous	1	0.0418	0.142	0.167
Cultural	0.0418	1	0.0347	0.0561
Ingredients	0.142	0.0347	1	0.110
Test	0.167	0.0561	0.110	1

Table 2: Pairwise similarities of each image set based on the labels attributed to them (see **Eq. 4**).

The pairwise similarities of the image sets are presented in **Table 2**. In particular, we draw the reader's attention to the similarity between the three initial sets and the test set. The ambiguous set was the most similar to the test set, followed by the ingredients set, while the cultural set was most different. Note the correspondence between these degrees of similarity and the ensuing relative specificity of labels: the more similar the initial images were to those in the the test set, the more specific were the labels attributed to the test set (c.f. **Fig. 1B**).

This suggests that presenting a series of very similar images elicits more specific labels. Such a phenomenon would be consistent with the psychological mechanism known as *negative priming*. Negative priming occurs when a person becomes desensitized to non-salient stimuli to which she is repeatedly exposed (Versace and Allain 2001; Mayr and Buchner 2007; De Zubicaray et al. 2008). Consider that workers who initially labeled the ambiguous image set had already seen five images showing prepared meals once they labeled the first test image. At that point, a worker might not regard the generic labels `food` or `meal` to be salient, and opt instead for `bread`, or `pasta`.

We are suggesting that, although workers are not instructed to compare images in any way, prior tasks nevertheless create a frame of reference relative to which later tasks are judged. This in turn influences the perception of salience. Thus, in a series of subjective characterization tasks that have very similar content, workers' focus will tend to be directed away from generic, shared attributes, toward those attributes that are specific and distinguishing.

Conclusions

Inter-task effects should be considered during task design. Our results show that inter-task effects can have a strong influence on how workers label images. In particular, we observed that prior tasks influence the specificity and content of labels. Surprisingly, inter-task effects were much stronger than framing the tasks by disclosing a semantically-loaded name for the task-funder.

We caution those designing studies using human computation: even if the requester has eliminated surrounding influences to every practical extent, *the greatest source of bias might lurk in the tasks themselves*. Due consideration should be given to how tasks are bundled together.

Batching tasks for better HPU performance. Our proposed connection between similarity and specificity during image-labeling might be used to tune the specificity of labels. For example, if one seeks very nuanced labeling, our results suggest that the images should first be sorted into batches based on their similarity. This could be accomplished by beginning with a first, coarse labeling of unsorted images, followed by bundling based on the similarity of coarse labels. Then, bundles of similar images could be served for a second round of finer labeling. The sorting and re-labeling could in principle be repeated.

Such a workflow involves serial processing, which points to an interesting potential difference between HPUs and CPUs. In general, whenever an aspect of a problem can be parallelized when employing CPUs, one gains efficiency. This is, of course, without any sacrifice to precision. But here, because of HPU hysteresis, one might gain precision by using HPUs with a more serialized algorithm. Further testing is needed to determine the gain in precision from this approach.

References

[Antin and Shaw 2012] Antin, J., and Shaw, A. 2012. Social desirability bias and self-reports of motivation: A study of amazon mechanical turk in the us and india. 2925–2934. cited By (since 1996)2.

[Beller 1971] Beller, H. K. 1971. Priming: Effects of advance information on matching. *Journal of Experimental Psychology* 87(2):176.

[Berinsky, Huber, and Lenz 2012] Berinsky, A.; Huber, G.; and Lenz, G. 2012. Evaluating online labor markets for experimental research: Amazon.com's mechanical turk. *Political Analysis* 20(3):351–368. cited By (since 1996)75.

[Chandler and Kapelner 2013] Chandler, D., and Kapelner, A. 2013. Breaking monotony with meaning: Motivation in crowdsourcing markets. *Journal of Economic Behavior & Organization* 90:123–133.

[Davis et al. 2010] Davis, J.; Arderiu, J.; Lin, H.; Nevins, Z.; Schuon, S.; Gallo, O.; and Yang, M.-H. 2010. The hpu. In *Computer Vision and Pattern Recognition Workshops (CVPRW), 2010 IEEE Computer Society Conference on*, 9–16.

[De Zubicaray et al. 2008] De Zubicaray, G. I.; McMahon, K. L.; Eastburn, M. M.; and Pringle, A. J. 2008. Negative priming in naming of categorically related objects: An fmri study. *cortex* 44(7):881–889.

[Finnerty et al. 2013] Finnerty, A.; Kucherbaev, P.; Tranquillini, S.; and Convertino, G. 2013. Keep it simple: Reward and task design in crowdsourcing. cited By (since 1996)0.

[Gass et al. 1999] Gass, S.; Mackey, A.; Alvarez-Torres, M.; and Fernández-García, M. 1999. The effects of task repetition on linguistic output. *Language Learning* 49(4):549–581. cited By (since 1996)26.

[Jaimes and Chang 2000] Jaimes, A., and Chang, S.-F. 2000. Conceptual framework for indexing visual information at multiple levels. volume 3964, 2–15. cited By (since 1996)30.

[Kazai, Kamps, and Milic-Frayling 2013] Kazai, G.; Kamps, J.; and Milic-Frayling, N. 2013. An analysis of human factors and label accuracy in crowdsourcing relevance judgments. *Information Retrieval* 1–41.

[Kinnaird, Dabbish, and Kiesler 2012] Kinnaird, P.; Dabbish, L.; and Kiesler, S. 2012. Workflow transparency in a microtask marketplace. 281–284. cited By (since 1996)0.

[Le et al. 2010] Le, J.; Edmonds, A.; Hester, V.; and Biewald, L. 2010. Ensuring quality in crowdsourced search relevance evaluation: The effects of training question distribution. In *SIGIR 2010 workshop on crowdsourcing for search evaluation*, 21–26.

[Mayr and Buchner 2007] Mayr, S., and Buchner, A. 2007. Negative priming as a memory phenomenon: A review of 20 years of negative priming research. *Zeitschrift für Psychologie/Journal of Psychology* 215(1):35.

[Noguera 2007] Noguera, Carmen; Ortells, J. J. A. M. J. 2007. Semantic priming effects from single words in a lexical decision task. *Acta Psychologica* 125:175–202.

[Panofsky 1939] Panofsky, E. 1939. *Studies in iconology*. New York.

[Paolacci, Chandler, and Ipeirotis 2010] Paolacci, G.; Chandler, J.; and Ipeirotis, P. G. 2010. Running experiments on amazon mechanical turk. *Judgment and Decision making* 5(5):411–419.

[Shatford 1986] Shatford, S. 1986. Analyzing the subject of a picture: a theoretical approach. *Cataloging & classification quarterly* 6(3):39–62.

[Snow et al. 2008] Snow, R.; O'Connor, B.; Jurafsky, D.; and Ng, A. Y. 2008. Cheap and fast—but is it good?: evaluating non-expert annotations for natural language tasks. In *Proceedings of the conference on empirical methods in natural language processing*, 254–263. Association for Computational Linguistics.

[Sohn and Anderson 2001] Sohn, M.-H., and Anderson, J. R. 2001. Task preparation and task repetition: two-component model of task switching. *Journal of Experimental Psychology: General* 130(4):764.

[Tversky 1977] Tversky, A. 1977. Features of similarity. *Psychological Review* 84(4):327–352. cited By (since 1996)2274.

[Versace and Allain 2001] Versace, R., and Allain, G. 2001. Negative priming in a gender decision task and in a semantic categorization task. *Acta psychologica* 108(1):73–90.

[Warren and Morton 1982] Warren, C., and Morton, J. 1982. The effects of priming on picture recognition. *British Journal of Psychology* 73(1):117–129.

Figure S1: Testing image set. These images were presented to all workers in the order shown after the initial set of images.

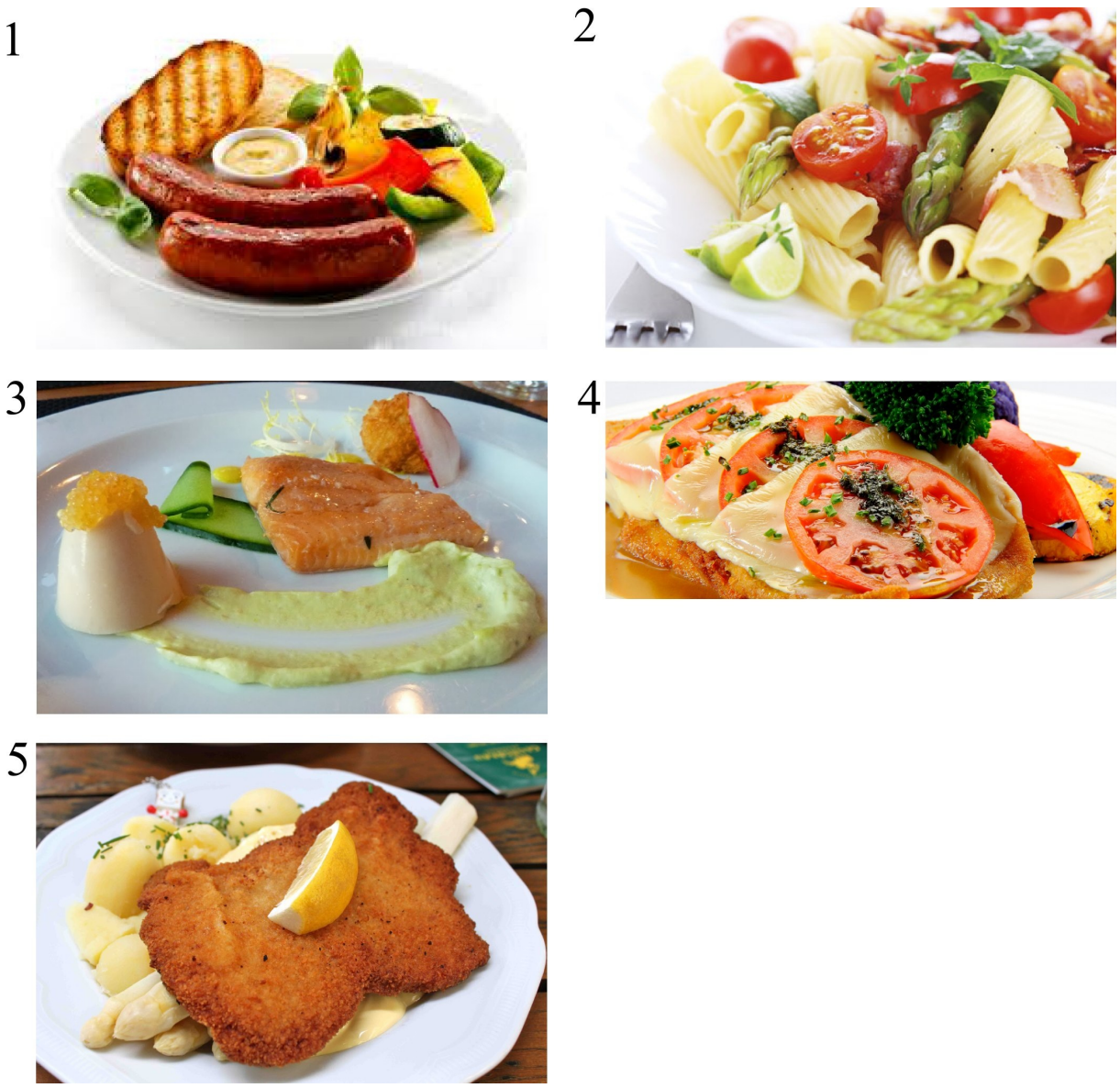

Figure S2: Ambiguous image set. These images were presented to workers from certain treatments (see **Table 1**) in the main text.

Figure S3: Cultural image set. These images were presented to workers from certain treatments (see **Table 1**) the main text.

Figure S4: Ingredients image set. These images were presented to workers from certain treatments (see **Table 1**) in the main text.