

Original Paper

Crowdsourcing for Food Purchase Receipt Annotation via Amazon Mechanical Turk: A Feasibility Study

Wenhua Lu¹, PhD; Alexandra Guttentag², BA; Brian Elbel^{3,4}, PhD, MPH; Kamila Kiszko³, MPH; Courtney Abrams³, MA; Thomas R Kirchner², PhD

¹Department of Childhood Studies, Rutgers, The State University of New Jersey, Camden, NJ, United States

²College of Global Public Health, New York University, New York, NY, United States

³School of Medicine, New York University, New York, NY, United States

⁴Robert F Wagner Graduate School of Public Service, New York University, New York, NY, United States

Corresponding Author:

Wenhua Lu, PhD

Department of Childhood Studies

Rutgers, The State University of New Jersey

329, Cooper Street

Camden, NJ, 08102

United States

Phone: 1 856 225 6083

Email: w.lu@rutgers.edu

Abstract

Background: The decisions that individuals make about the food and beverage products they purchase and consume directly influence their energy intake and dietary quality and may lead to excess weight gain and obesity. However, gathering and interpreting data on food and beverage purchase patterns can be difficult. Leveraging novel sources of data on food and beverage purchase behavior can provide us with a more objective understanding of food consumption behaviors.

Objective: Food and beverage purchase receipts often include time-stamped location information, which, when associated with product purchase details, can provide a useful behavioral measurement tool. The purpose of this study was to assess the feasibility, reliability, and validity of processing data from fast-food restaurant receipts using crowdsourcing via Amazon Mechanical Turk (MTurk).

Methods: Between 2013 and 2014, receipts (N=12,165) from consumer purchases were collected at 60 different locations of five fast-food restaurant chains in New Jersey and New York City, USA (ie, Burger King, KFC, McDonald's, Subway, and Wendy's). Data containing the restaurant name, location, receipt ID, food items purchased, price, and other information were manually entered into an MS Access database and checked for accuracy by a second reviewer; this was considered the *gold standard*. To assess the feasibility of coding receipt data via MTurk, a prototype set of receipts (N=196) was selected. For each receipt, 5 turkers were asked to (1) identify the receipt identifier and the name of the restaurant and (2) indicate whether a beverage was listed in the receipt; if yes, they were to categorize the beverage as cold (eg, soda or energy drink) or hot (eg, coffee or tea). Interturker agreement for specific questions (eg, restaurant name and beverage inclusion) and agreement between turker consensus responses and the gold standard values in the manually entered dataset were calculated.

Results: Among the 196 receipts completed by turkers, the interturker agreement was 100% (196/196) for restaurant names (eg, Burger King, McDonald's, and Subway), 98.5% (193/196) for beverage inclusion (ie, hot, cold, or none), 92.3% (181/196) for types of hot beverage (eg, hot coffee or hot tea), and 87.2% (171/196) for types of cold beverage (eg, Coke or bottled water). When compared with the gold standard data, the agreement level was 100% (196/196) for restaurant name, 99.5% (195/196) for beverage inclusion, and 99.5% (195/196) for beverage types.

Conclusions: Our findings indicated high interrater agreement for questions across difficulty levels (eg, single- vs binary- vs multiple-choice items). Compared with traditional methods for coding receipt data, MTurk can produce excellent-quality data in a lower-cost, more time-efficient manner.

(*J Med Internet Res* 2019;21(4):e12047) doi: [10.2196/12047](https://doi.org/10.2196/12047)

KEYWORDS

Amazon Mechanical Turk; food purchase receipt; crowdsourcing; feasibility; reliability; validity

Introduction

The decisions that individuals make about the food and beverage products they purchase and consume directly influence their energy intake and dietary quality and may lead to excess weight gain and obesity [1-3]. Research supports the notion that decision making related to food consumption may act as a potential mediator between the neighborhood food environment and individual dietary intake [4-7], but assessment of dietary behavior can be problematic [1]. Leveraging new sources of data on food and beverage purchase behavior, therefore, could provide novel insights into food and beverage decision making.

Food purchase receipts contain information about all foods and beverages purchased by individuals and households from different sources, such as fast-food restaurants, grocery stores, and convenience or corner stores [1]. Compared with retrospective self-reports, receipts can contribute more objective data, thereby avoiding social desirability influence and recall bias [8]. Unfortunately, accurately and reliably annotating large numbers of receipts and images has been a logistical bottleneck inhibiting their widespread use. Typically, academic researchers depend on undergraduate and graduate research assistants to extract data; research progress then depends on the ebb and flow of the semester. Further, each receipt must be carefully reviewed, which takes several minutes. As a result, it can take weeks, months, or even years to process receipt data, especially when large datasets are being handled and/or subjective reasoning is needed.

In the past decade, crowdsourcing has become increasingly popular due to its time-saving and cost-effective qualities [9]. In crowdsourcing, potentially large jobs are broken into many microtasks that are then outsourced directly to individual workers via public solicitation [10]. As the leading and most well-established online crowdsourcing service, Amazon Mechanical Turk (MTurk) enables researchers and businesses, identified as requesters, to recruit anonymous online workers (ie, turkers) worldwide to complete Human Intelligence Tasks (HITs) (ie, tasks that cannot be entirely automated and require human intelligence) at relatively low cost [11]. MTurk offers a basic user interface for simple tasks and a powerful application programming interface for developers to build a platform that uses their services [10,11].

Since its inception, MTurk has been used primarily by researchers in nonmedical fields (eg, psychology, marketing, management, business, political science, computer science, and neuroscience) to do data processing, including data extraction, transcription, translation, and sentiment analysis [12-18]. Emerging studies in recent years have also applied MTurk in various disciplines of health [12-14]. For example, a group of researchers pioneered the use of crowdsourcing technology in

public health research and utilized a custom MTurk interface for analyzing mobile phone photographs of retail point-of-sale tobacco marketing [19,20]. Over the course of one typical implementation, 299 turkers completed more than 23,000 tasks at a total cost of US \$2500 in less than 24 hours. Results of the crowdsourced photo-only assessments had an excellent level of correspondence to the traditional field survey data, which demonstrated the tremendous potential and reliability of MTurk as a medium for analyzing health-related data in a low-cost, time-efficient way [19,20].

Despite its growing popularity, MTurk has not yet been used to annotate data from food and beverage purchase receipts. This study, therefore, takes an initial foray into assessing the feasibility, reliability, and validity of processing fast-food restaurant receipt data using MTurk.

Methods

Overview

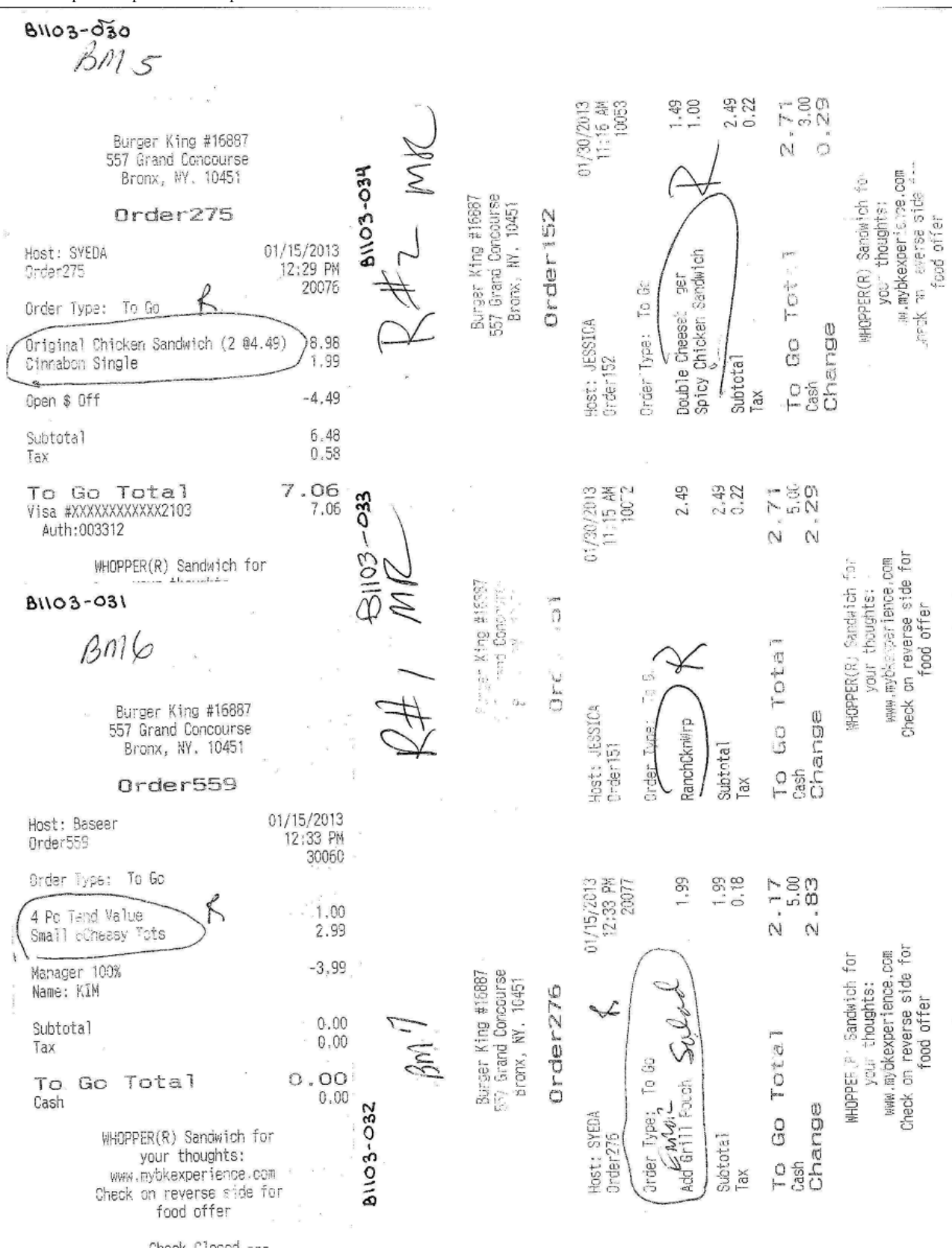
Our study consisted of three phases: First, data from a large number of food and beverage purchase receipts were obtained through a traditional in-laboratory, manual data extraction method and were confirmed for accuracy to serve as the *gold standard*; second, an MTurk project was set up and a group of turkers were recruited to extract some prespecified required data from a representative sample of the receipts; and third, the data processed through MTurk were compared with the gold standard and evaluated for reliability and validity. Details of each step are described in the following sections.

Step 1: Data Collection and Manual Data Extraction

Between August 2013 and May 2014, receipts were collected from consumer purchases at 60 different locations of five fast-food restaurant chains in New Jersey and New York City, USA: Burger King, KFC, McDonald's, Subway, and Wendy's. Data collectors stood outside of the restaurants and asked entering customers to save their receipts; upon leaving, they were asked to hand over their receipt and identify which items on the receipt they purchased for their own consumption. Altogether, three rounds of data collection were conducted within the project period, and a total of 12,165 receipts were collected. Detailed data collection procedure, the number of locations surveyed, and receipts collected by location and restaurant chain have been described previously [21].

After each round of receipt collection, research assistants pasted receipts on single printer paper sheets next to each other—about 4-6 per page—and scanned them into a database (see Figure 1). Unusual receipts were flagged and reported in the database; for example, those where details of exactly what was purchased by the customers were missing (ie, the receipt was not clearly marked, the ink rubbed off, or the receipt was not itemized).

Figure 1. Sample food purchase receipts.



Following that, the data containing the receipt identifier, restaurant name, food items purchased, purchase date, total price, etc, were extracted from individual receipts; manually entered into an MS Access database by a research assistant; and checked for accuracy by a second research assistant 1-10 months

following the initial data entry. The receipt data obtained through such a traditional in-laboratory manual data extraction method served as the gold standard in this study.

Step 2: Setting Up the Amazon Mechanical Turk Task and Crowdsourcing Workflow

To assess the feasibility of MTurk as an alternative tool for processing receipt data, some of the data extracted manually above was recollected using crowdsourcing via MTurk.

Specifically, two separate tasks were set up on MTurk. The first was an *expert* task, in which turkers were asked to crop one receipt per page at a time from the original pages with 4-6 receipts per page. An *expert* task in MTurk means that requesters trust one turker to do the assignment, rather than having multiple turkers do it and agree on an answer; such tasks are usually simple and do not involve extensive human reasoning and interpretation. First, scanned PDF documents (8.5 x 11 inches) with multiple receipts in different orientations on each page were uploaded onto MTurk. Following that, the MTurk Expert HIT was launched and turkers were recruited to crop each individual receipt using a crop tool and orient the receipts in a readable fashion. An instructional video was included to guide turkers in using the software correctly. This task was completed preceding this study for all receipts (N=12,165), with one receipt on one page.

The next MTurk HIT was a *consensus* task—the focus of this study—which required multiple turkers working on the same assignment and then checking the agreement among their responses. For this task, entitled “Receipt Information,” turkers were requested to identify required information from the food purchase receipts and respond to a series of questions based on the information they identified. As illustrated in Figure 2, a brief description was included under the title of the project: “Please gather the following information related to food purchase from a receipt.” For each receipt, turkers were asked to answer questions based on the following four tasks: (1) write down the receipt identifier, (2) choose the name of the restaurant from a drop-down list, (3) indicate whether a beverage was listed in the receipt, and (4) if a beverage was listed, categorize the beverage as cold beverage or hot beverage.

Specifically, question 1 required textual responses; for each individual receipt, turkers were requested to type in a unique identifier composed of letters and numbers (eg, B1103-036 and S2109A-022). Questions 2-4 included multiple-choice questions, which required subjective judgment at different difficulty levels (ie, single- vs binary-choice items). Considering that most information on food purchase receipts can be obtained through either textual responses or multiple-choice questions, it is reasonable to assume that if a turker can understand and respond accurately to these four exemplary questions, he or she could identify other data from food purchase receipts as well. For demonstration purposes, therefore, instead of using all of the 12,165 receipts, a prototypical sample of receipts (N=196) were used for this study, all of which were clearly marked with zero or only one beverage item on each receipt.

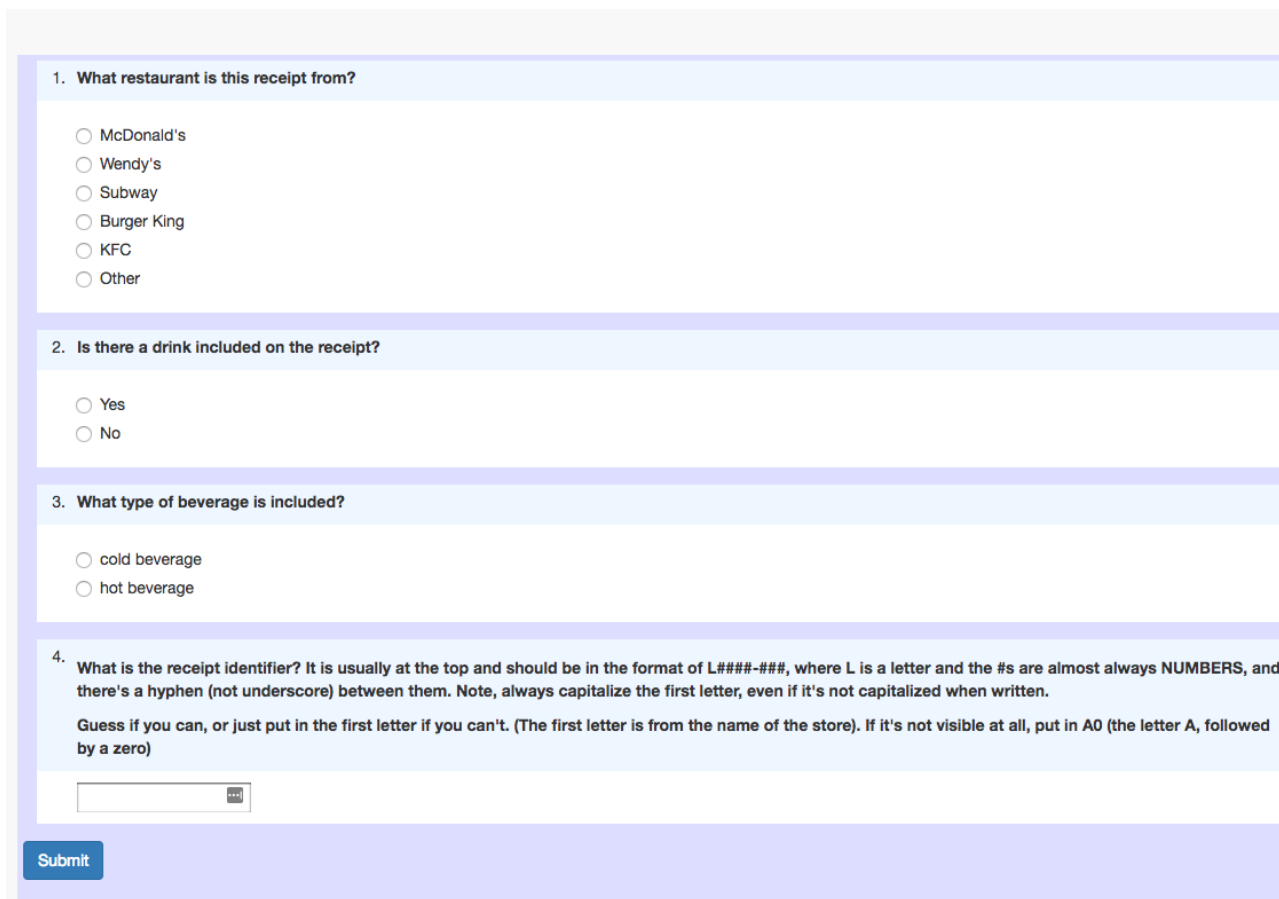
After the prototypical receipts were selected and the MTurk HIT was set up, we started to invite turkers to work on the

consensus task. To avoid spammers and control the quality of turkers, we screened the turkers by setting the minimum prior approval rating to 99%, meaning that at least 99% of a turker's answers to the MTurk HITs that they have completed to date were approved by the requester. Turkers' locations were further restricted to the United States, as previous studies have suggested that language, cultural background, and ethnicity can significantly influence people's comprehension of culture-related information such as food choices [16,22,23]. Turkers were paid US \$0.06 for interpreting each receipt, which was anticipated to take 60-90 seconds. Once a turker began processing a receipt, he or she had a maximum of 5 minutes to complete it. A turker could analyze as many receipts as he or she wanted.

One critical question in using crowdsourcing via MTurk is setting the minimum and maximum number of turkers who will complete each assignment (ie, the number of repetitions that each receipt will receive until consensus is achieved for each question in the study). Intuitively, 2 would be the absolute minimum in order to reach an agreement on responses to a question. However, setting 2 as the minimum number of repetitions can incur an incorrect agreed-upon answer if both turkers provide the same incorrect responses to a question. A minimum of 3 creates a majority, but the question agreement threshold (QAT) of 67% (2/3) is insufficiently low and an incorrect consensus can still be reached if 2 of the 3 turkers agree on an incorrect answer. A minimum of 4 is acceptable with an interturker agreement of 75% (3/4), but a consensus cannot be achieved if 2 turkers agree on one response while the other 2 turkers agree on another response.

Thus, to reach an agreement with high accuracy, a minimum of 5 turkers is required with a QAT of 60% (3/5). The maximum number of repetitions should also be set as 5, because after starting with a QAT of 60%, the probability of getting an interturker agreement of 80% will then decrease and it will unnecessarily delay the item consensus calculation process if continuing to add more turkers to complete the task. In this study, therefore, 5 was set as both the minimum and maximum number of repetitions that each receipt received with a QAT of 80% (4/5), meaning that a receipt would continue to be available for turkers to interpret until it was assessed 5 times by 5 turkers and at least 4/5 (80%) turkers agreed on a response.

Before the formal launch of the MTurk tasks, several trial runs were performed using small subsets of pictures (ie, 10-20 receipt images at a time) to confirm that the tasks would be manageable by turkers and to confirm that the questions were easily understood and completed. While doing assignments, turkers could reach out to the requester directly by email or through the *Report a Problem with this Task* tab on the MTurk survey interface. During the trial tasks, we received several email inquiries from turkers regarding uncertainties about numbers versus letters in receipt identifiers, receipt number cutoff from the image at the top, etc. No turker expressed problems with interpreting other receipt information, either during the trial runs or the formal task.

Figure 2. Screenshot of the Amazon Mechanical Turk consensus task.


1. What restaurant is this receipt from?

McDonald's

Wendy's

Subway

Burger King

KFC

Other

2. Is there a drink included on the receipt?

Yes

No

3. What type of beverage is included?

cold beverage

hot beverage

4. What is the receipt identifier? It is usually at the top and should be in the format of L####-###, where L is a letter and the #s are almost always NUMBERS, and there's a hyphen (not underscore) between them. Note, always capitalize the first letter, even if it's not capitalized when written.

Guess if you can, or just put in the first letter if you can't. (The first letter is from the name of the store). If it's not visible at all, put in A0 (the letter A, followed by a zero)

Submit

Step 3: Evaluating the Reliability and Validity of Amazon Mechanical Turk

After all assignments were completed, we started evaluating the reliability and validity of MTurk for processing food purchase receipt information. This was conducted in three steps. First, interturker agreement was examined on responses to the four questions asked (ie, receipt identifier, restaurant name, beverage included or not, and type of beverage), and a majority response for each individual question was identified. Following that, turkers' majority responses were compared with gold standard values in the manually entered dataset to evaluate the reliability and validity of MTurk. Finally, we conducted sensitivity testing to assess whether and how the number of turkers completing each assignment would influence the agreement between turkers' majority responses and the gold standard. All analyses were conducted using R version 3.2.4 (R Foundation) [24].

Results

In total, 209 turkers participated in the *consensus* task and initiated or attempted 1346 assignments, among which 983

(73.03%) were approved or completed. On average, each turker contributed 4.7 assignments (SD 1.5). It took an average of 93.12 seconds (SD 70.5, median 65.0) for a turker to analyze a receipt; the entire project was completed within 40 minutes after we launched it on MTurk, with a total cost of US \$80.80.

Table 1 lists the descriptive characteristics of the 196 prototypical receipts completed by 5 turkers. Among the 196 receipts that we sampled, one beverage item was listed in 140 receipts (71.4%), including 101 cold beverages (51.5%) and 39 hot beverages (19.9%). Among the 101 receipts with cold beverages, soda drinks were listed in 75 receipts (74.3%), including Coca-Cola, Sprite, Pepsi, Diet Coke, and generic drinks. The rest of the cold-beverage receipts included sweet tea (6/26, 23%), bottled water (5/26, 19%), milkshake or smoothie (5/26, 19%), iced coffee or coffee drinks (4/26, 15%), lemonade (3/26, 12%), and juice or juice beverages (3/26, 12%). Among the 39 receipts with hot beverages, hot coffee was listed in 36 receipts (92%), and the other 3 receipts included 1 with hot chocolate (3%) and 2 with hot tea (5%).

Table 1. Descriptive characteristics of the prototypical sample of receipts (N=196).

Restaurant	Number of receipts, n (%)	Number of receipts with beverages out of all receipts from the restaurant, n (%)	Number of hot beverages ^a out of all receipts from the restaurant, n (%)	Number of cold beverages ^b out of all receipts from the restaurant, n (%)
Burger King	15 (7.7)	12 (80)	0 (0)	12 (80)
KFC	40 (20.4)	29 (73)	8 (20)	21 (53)
McDonald's	102 (52.0)	78 (76.5)	31 (30.4)	47 (46.1)
Subway	21 (10.7)	5 (24)	0 (0)	5 (24)
Wendy's	18 (9.2)	16 (89)	0 (0)	16 (89)

^aHot beverages included hot coffee, hot chocolate, and hot tea.

^bCold beverages were mostly soda drinks, sweet tea, bottled water, and coffee drinks.

Turkers showed high agreement on their responses to the four questions that we asked. Specifically, among the 196 receipts that we sampled, the proportion of receipts with a QAT of at least 80% (ie, 4/5 interturker agreement) was 100% (196/196) for receipt identifier, 100% (196/196) for restaurant names (eg, Burger King, McDonald's, or Subway), 98.5% (193/196) for beverage inclusion (ie, yes or no), 92.3% (181/196) for hot beverage (eg, hot coffee or hot tea), and 87.2% (171/196) for cold beverage (eg, soda or bottled water). At a QAT of 100%, the proportions of receipts with unanimous (ie, 5/5) agreement among the turkers was 100% (196/196) for receipt identifier, 90.8% (178/196) for restaurant names, 75.5% (148/196) for beverage inclusion, 69.4% (136/196) for hot beverages, and 51.0% (100/196) for cold beverages.

We further checked the disagreement pattern among turkers for specific questions. For the two questions on receipt identifiers and restaurant names, no disagreement was observed among turkers. When asked to indicate whether a beverage was included or not, disagreements started to emerge. For some cases, turkers overlooked beverages, especially soda drinks that were included in a combo rather than being listed as separate items. For others, some turkers wrongly categorized receipts with smoothies as *beverage not included*. Consequently, when it came to coding the specific type of beverage (ie, cold or hot beverage), more discrepancies were noted.

When comparing turkers' majority responses with the gold standard data, the agreement rate was 100% (196/196) for receipt identifier, 100% (196/196) for restaurant name, 99.5% (195/196) for beverage inclusion, and 99.5% (195/196) for beverage types. We further tested whether and how the number of turkers influenced the agreement level between turkers' majority responses and the gold standard data. Based on the analysis, when 3 turkers completed the project, the agreement between their consensus response and the gold standard data was 100% (196/196) for receipt identifier, 100% (196/196) for restaurant name, 99.5% (195/196) for beverage inclusion, and 99.5% (195/196) for beverage type, which were the same as the proportions when 5 turkers completed the assignments.

Discussion

This study is the first effort to assess whether MTurk, a popular crowdsourcing platform, can be used for processing data from food purchase receipts. In general, findings from this study

supported the feasibility, reliability, and validity of MTurk as a cost-effective and time-efficient tool for processing food purchase receipt data.

Findings from this study demonstrated that, with minimal training, the MTurk workforce can categorize and analyze receipt data in a timely and cost-effective way. Despite the low compensation rate (ie, US \$0.06 for every assignment), turkers in this study completed the entire task in less than 40 minutes, and the data extracted were of excellent quality, which was consistent with evidence from previous evaluation studies [9,13,18,25,26]. In fact, turkers in previous studies have expressed other motivations that enticed them to complete tasks. For example, many turkers felt it was a productive way to spend available free time, was mentally engaging, was oftentimes interesting, and offered a source of entertainment [27-30]. Compared with manual data extraction, which is often time-consuming, expensive, and difficult to scale up, MTurk can greatly enhance the widespread use of receipts as an assessment of food purchase and dietary behaviors.

Our findings further supported the reliability and validity of using MTurk for annotating receipt data, with high interrater agreement for both textual and multiple-choice questions. Previous studies have noted that as data coding tasks became more subjectively difficult, it got harder to achieve interpretive convergence [26]. Consistently in this study, we found perfect agreement (ie, 100%) among 5 turkers for the two easier questions that did not require subjective judgement (ie, receipt identifier and restaurant name), but we found increased disagreements for the two questions regarding beverage inclusion and beverage type. Nevertheless, when turkers' majority responses were compared with the manually extracted gold standard, perfect or close-to-perfect agreements were observed, which confirmed the reliability of the number of 5 turkers that we requested for annotating individual receipts.

Our study has limitations. First, although we purposely selected receipts with clearly identifiable information, receipts used in health data analysis could sometimes be more difficult to read due to rips, pen markings, or small font, which would likely affect the agreement rates of turkers. Second, we only allowed turkers with prior approval ratings of 99% to participate in the task. Although this helped to ensure that turkers provide quality work, it also narrowed down the number of turkers available and likely increased the time for task completion. We did not

test whether or how lowering the approval rating would affect the reliability and validity of data processing. Third, for demonstration purposes, the tasks we selected were objectively easy; future studies are warranted to determine if the same success rates can be obtained with more complicated tasks.

Despite the limitations, findings from this study hold important practical and research implications. First, our work confirmed the feasibility and accuracy of using MTurk as an innovative approach for processing data from food purchase receipts. In the future, the traditional model of manually annotating food purchase receipts as the gold standard for comparison may be flipped. Instead, crowdsourcing platforms could be used with appropriate task qualification requirements (eg, requiring turkers with prior approval ratings of 95% or 99%) to identify majority or consensus responses, followed by manually annotating a proportion of the receipts to confirm the reliability and validity. This feasibility study demonstrates the scalable and sustainable nature of this approach. Second, the accuracy of crowdsourced receipt annotation in this study lends strong support to the

appropriateness of the number of turkers that we requested for each task. To get reliable consensus or majority responses among turkers when annotating image data on MTurk, we recommend future researchers set 5 as both the minimum and maximum number of repetitions for each image, with a question agreement threshold of 80%. Lastly, and most importantly, findings from this study point to the great potential of crowdsourcing for processing data in public health research, particularly tasks that cannot be entirely automated by computer programs and require human intelligence. A recent study has confirmed that objectively documented household food purchases from receipts can yield an unbiased and reasonably accurate estimate of overall diet quality as measured through 24-hour diet recalls [31]. With its time-saving and cost-effective qualities, crowdsourcing will vastly increase capacity for large-scale and high-quality receipt annotation, which in turn will advance our understanding of environmental influence on human health behaviors and ultimately lead to better health prevention and intervention efforts.

Acknowledgments

This study was funded by the National Institutes of Health, National Institute of Diabetes and Digestive and Kidney Diseases (grant number R01099241), the New York State Health Foundation (grant number 12-01682), and the Robert Wood Johnson Foundation (grant number 70823). The collaboration was supported by the American Academy of Health Behavior mentorship program.

Conflicts of Interest

None declared.

References

1. French SA, Shimotsu ST, Wall M, Gerlach AF. Capturing the spectrum of household food and beverage purchasing behavior: A review. *J Am Diet Assoc* 2008 Dec;108(12):2051-2058. [doi: [10.1016/j.jada.2008.09.001](https://doi.org/10.1016/j.jada.2008.09.001)] [Medline: [19027408](https://pubmed.ncbi.nlm.nih.gov/19027408/)]
2. Drewnowski A, Rehm CD. Energy intakes of US children and adults by food purchase location and by specific food source. *Nutr J* 2013 May 08;12:59 [FREE Full text] [doi: [10.1186/1475-2891-12-59](https://doi.org/10.1186/1475-2891-12-59)] [Medline: [23656639](https://pubmed.ncbi.nlm.nih.gov/23656639/)]
3. National Restaurant Association. 2016. National statistics: Restaurant industry facts at a glance URL: <https://www.restaurant.org/News-Research/Research/Facts-at-a-Glance> [accessed 2018-08-06] [WebCite Cache ID 71TL09Nny]
4. French SA, Wall M, Mitchell NR, Shimotsu ST, Welsh E. Annotated receipts capture household food purchases from a broad range of sources. *Int J Behav Nutr Phys Act* 2009 Jul 01;6:37 [FREE Full text] [doi: [10.1186/1479-5868-6-37](https://doi.org/10.1186/1479-5868-6-37)] [Medline: [19570234](https://pubmed.ncbi.nlm.nih.gov/19570234/)]
5. Gordon-Larsen P. Food availability/convenience and obesity. *Adv Nutr* 2014 Nov;5(6):809-817 [FREE Full text] [doi: [10.3945/an.114.007070](https://doi.org/10.3945/an.114.007070)] [Medline: [25398746](https://pubmed.ncbi.nlm.nih.gov/25398746/)]
6. Laska MN, Hearst MO, Forsyth A, Pasch KE, Lytle L. Neighbourhood food environments: Are they associated with adolescent dietary intake, food purchases and weight status? *Public Health Nutr* 2010 Nov;13(11):1757-1763 [FREE Full text] [doi: [10.1017/S1368980010001564](https://doi.org/10.1017/S1368980010001564)] [Medline: [20529405](https://pubmed.ncbi.nlm.nih.gov/20529405/)]
7. White M. Food access and obesity. *Obes Rev* 2007 Mar;8 Suppl 1:99-107. [doi: [10.1111/j.1467-789X.2007.00327.x](https://doi.org/10.1111/j.1467-789X.2007.00327.x)] [Medline: [17316311](https://pubmed.ncbi.nlm.nih.gov/17316311/)]
8. Elbel B, Kersh R, Brescoll VL, Dixon LB. Calorie labeling and food choices: A first look at the effects on low-income people in New York City. *Health Aff (Millwood)* 2009;28(6):w1110-w1121. [doi: [10.1377/hlthaff.28.6.w1110](https://doi.org/10.1377/hlthaff.28.6.w1110)] [Medline: [19808705](https://pubmed.ncbi.nlm.nih.gov/19808705/)]
9. Ranard BL, Ha YP, Meisel ZF, Asch DA, Hill SS, Becker LB, et al. Crowdsourcing: Harnessing the masses to advance health and medicine, a systematic review. *J Gen Intern Med* 2014 Jan;29(1):187-203 [FREE Full text] [doi: [10.1007/s11606-013-2536-8](https://doi.org/10.1007/s11606-013-2536-8)] [Medline: [23843021](https://pubmed.ncbi.nlm.nih.gov/23843021/)]
10. Mason W, Watts DJ. Financial incentives and the "performance of crowds". *SIGKDD Explor* 2009 Dec;11(2):100-108. [doi: [10.1145/1809400.1809422](https://doi.org/10.1145/1809400.1809422)]
11. Amazon Mechanical Turk. URL: <https://www.mturk.com/> [accessed 2018-06-18] [WebCite Cache ID 70GgopezZ]
12. Khare R, Good BM, Leaman R, Su AI, Lu Z. Crowdsourcing in biomedicine: Challenges and opportunities. *Brief Bioinform* 2016 Jan;17(1):23-32 [FREE Full text] [doi: [10.1093/bib/bbv021](https://doi.org/10.1093/bib/bbv021)] [Medline: [25888696](https://pubmed.ncbi.nlm.nih.gov/25888696/)]

13. Créquit P, Mansouri G, Benchoufi M, Vivot A, Ravaud P. Mapping of crowdsourcing in health: Systematic review. *J Med Internet Res* 2018 May 15;20(5):e187 [FREE Full text] [doi: [10.2196/jmir.9330](https://doi.org/10.2196/jmir.9330)] [Medline: [29764795](https://pubmed.ncbi.nlm.nih.gov/29764795/)]
14. Bohannon J. Psychology: Mechanical Turk upends social sciences. *Science* 2016 Jun 10;352(6291):1263-1264. [doi: [10.1126/science.352.6291.1263](https://doi.org/10.1126/science.352.6291.1263)] [Medline: [27284175](https://pubmed.ncbi.nlm.nih.gov/27284175/)]
15. Saunders DR, Bex PJ, Woods RL. Crowdsourcing a normative natural language dataset: A comparison of Amazon Mechanical Turk and in-lab data collection. *J Med Internet Res* 2013 May 20;15(5):e100 [FREE Full text] [doi: [10.2196/jmir.2620](https://doi.org/10.2196/jmir.2620)] [Medline: [23689038](https://pubmed.ncbi.nlm.nih.gov/23689038/)]
16. Yu B, Willis M, Sun P, Wang J. Crowdsourcing participatory evaluation of medical pictograms using Amazon Mechanical Turk. *J Med Internet Res* 2013 Jun 03;15(6):e108 [FREE Full text] [doi: [10.2196/jmir.2513](https://doi.org/10.2196/jmir.2513)] [Medline: [23732572](https://pubmed.ncbi.nlm.nih.gov/23732572/)]
17. Leroy G, Endicott JE, Kauchak D, Mouradi O, Just M. User evaluation of the effects of a text simplification algorithm using term familiarity on perception, understanding, learning, and information retention. *J Med Internet Res* 2013 Jul 31;15(7):e144 [FREE Full text] [doi: [10.2196/jmir.2569](https://doi.org/10.2196/jmir.2569)] [Medline: [23903235](https://pubmed.ncbi.nlm.nih.gov/23903235/)]
18. Kuang J, Argo L, Stoddard G, Bray BE, Zeng-Treitler Q. Assessing pictograph recognition: A comparison of crowdsourcing and traditional survey approaches. *J Med Internet Res* 2015 Dec 17;17(12):e281 [FREE Full text] [doi: [10.2196/jmir.4582](https://doi.org/10.2196/jmir.4582)] [Medline: [26678085](https://pubmed.ncbi.nlm.nih.gov/26678085/)]
19. Brabham DC, Ribisl KM, Kirchner TR, Bernhardt JM. Crowdsourcing applications for public health. *Am J Prev Med* 2014 Feb;46(2):179-187. [doi: [10.1016/j.amepre.2013.10.016](https://doi.org/10.1016/j.amepre.2013.10.016)] [Medline: [24439353](https://pubmed.ncbi.nlm.nih.gov/24439353/)]
20. Ilakkuvan V, Tancelosky M, Ivey KC, Pearson JL, Cantrell J, Vallone DM, et al. Cameras for public health surveillance: A methods protocol for crowdsourced annotation of point-of-sale photographs. *JMIR Res Protoc* 2014 Apr 09;3(2):e22 [FREE Full text] [doi: [10.2196/resprot.3277](https://doi.org/10.2196/resprot.3277)] [Medline: [24717168](https://pubmed.ncbi.nlm.nih.gov/24717168/)]
21. Cantor J, Torres A, Abrams C, Elbel B. Five years later: Awareness of New York City's calorie labels declined, with no changes in calories purchased. *Health Aff (Millwood)* 2015 Nov;34(11):1893-1900. [doi: [10.1377/hlthaff.2015.0623](https://doi.org/10.1377/hlthaff.2015.0623)] [Medline: [26526247](https://pubmed.ncbi.nlm.nih.gov/26526247/)]
22. Dowse R, Ehlers MS. The evaluation of pharmaceutical pictograms in a low-literate South African population. *Patient Educ Couns* 2001 Nov;45(2):87-99. [Medline: [11687321](https://pubmed.ncbi.nlm.nih.gov/11687321/)]
23. Kim H, Nakamura C, Zeng-Treitler Q. Assessment of pictographs developed through a participatory design process using an online survey tool. *J Med Internet Res* 2009 Feb 24;11(1):e5 [FREE Full text] [doi: [10.2196/jmir.1129](https://doi.org/10.2196/jmir.1129)] [Medline: [19275981](https://pubmed.ncbi.nlm.nih.gov/19275981/)]
24. R Project. The R Project for statistical computing URL: <https://www.r-project.org/> [WebCite Cache ID 71TPIKnhO]
25. Azzam T, Jacobson MR. Finding a comparison group. *Am J Eval* 2013 Jun 18;34(3):372-384. [doi: [10.1177/1098214013490223](https://doi.org/10.1177/1098214013490223)]
26. Mitra T, Hutto C, Gilbert E. Comparing person- and process-centric strategies for obtaining quality data on Amazon Mechanical Turk. In: Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems. 2015 Presented at: 33rd Annual ACM Conference on Human Factors in Computing Systems; April 18-23, 2015; Seoul, Republic of Korea p. 1345-1354. [doi: [10.1145/2702123.2702553](https://doi.org/10.1145/2702123.2702553)]
27. Snow R, O'Connor B, Jurafsky D, Ng A. Cheap and fast: But is it good? Evaluating non-expert annotations for natural language tasks. In: Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing. 2008 Presented at: 2008 Conference on Empirical Methods in Natural Language Processing; October 25-27, 2008; Honolulu, Hawaii p. 254-263 URL: <http://www.aclweb.org/anthology/D08-1027>
28. Chandler D, Kapelner A. Breaking monotony with meaning: Motivation in crowdsourcing markets. *J Econ Behav Organ* 2013 Jun;90:123-133. [doi: [10.1016/j.jebo.2013.03.003](https://doi.org/10.1016/j.jebo.2013.03.003)]
29. Horton JJ, Rand DG, Zeckhauser RJ. The online laboratory: Conducting experiments in a real labor market. *Exp Econ* 2011 Sep;14(3):399-425. [doi: [10.1007/s10683-011-9273-9](https://doi.org/10.1007/s10683-011-9273-9)]
30. Paolacci G, Chandler J, Ipeirotis P. Running experiments on Amazon Mechanical Turk. *Judgm Decis Mak* 2010 Aug;5(5):411-419 [FREE Full text]
31. Appelhans BM, French SA, Tangney CC, Powell LM, Wang Y. To what extent do food purchases reflect shoppers' diet quality and nutrient intake? *Int J Behav Nutr Phys Act* 2017 Dec 11;14(1):46 [FREE Full text] [doi: [10.1186/s12966-017-0502-2](https://doi.org/10.1186/s12966-017-0502-2)] [Medline: [28399887](https://pubmed.ncbi.nlm.nih.gov/28399887/)]

Abbreviations

- HIT:** Human Intelligence Task
MTurk: Amazon Mechanical Turk
QAT: question agreement threshold
-

Edited by G Eysenbach; submitted 27.08.18; peer-reviewed by P Liu, FA Shafie; comments to author 06.10.18; revised version received 15.12.18; accepted 16.12.18; published 05.04.19

Please cite as:

Lu W, Guttentag A, Elbel B, Kiszko K, Abrams C, Kirchner TR

Crowdsourcing for Food Purchase Receipt Annotation via Amazon Mechanical Turk: A Feasibility Study

J Med Internet Res 2019;21(4):e12047

URL: <http://www.jmir.org/2019/4/e12047/>

doi: [10.2196/12047](https://doi.org/10.2196/12047)

PMID: [30950801](https://pubmed.ncbi.nlm.nih.gov/30950801/)

©Wenhua Lu, Alexandra Guttentag, Brian Elbel, Kamila Kiszko, Courtney Abrams, Thomas R Kirchner. Originally published in the Journal of Medical Internet Research (<http://www.jmir.org>), 05.04.2019. This is an open-access article distributed under the terms of the Creative Commons Attribution License (<https://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research, is properly cited. The complete bibliographic information, a link to the original publication on <http://www.jmir.org/>, as well as this copyright and license information must be included.