Behavior Research Methods
https://doi.org/10.3758/s13428-023-02152-y

# Comparing data quality from an online and in-person lab sample on dynamic theory of mind tasks

**Anne C. Krendl**[1] · **Kurt Hugenberg**[1] · **Daniel P. Kennedy**[1]

Accepted: 24 May 2023


## Abstract

Nearly half the published research in psychology is conducted with online samples, but the preponderance of these studies rely primarily on self-report measures. The current study validated data quality from an online sample on a novel, dynamic task by comparing performance between an in-lab and online sample on two dynamic measures of theory of mind—the ability to infer others' mental states. Theory of mind is a cognitively complex construct that has been widely studied across multiple domains of psychology. One task was based on the show *The Office*®, and has been previously validated by the authors with in-lab samples. The second was a novel task based on the show *Nathan for You*®, which was selected to account for familiarity effects associated with *The Office*. Both tasks measured various dimensions of theory of mind (inferring beliefs, understanding motivations, detecting deception, identifying faux pas, and understanding emotions). The in-person lab samples ($N = 144$ and 177, respectively) completed the tasks between-subject, whereas the online sample ($N = 347$ from Prolific Academic) completed them within-subject, with order counterbalanced. The online sample's performance across both tasks was reliable (Cronbach's $\alpha = .66$). For *The Office*, the in-person sample outperformed the online sample on some types of theory of mind, but this was driven by their greater familiarity with the show. Indeed, for the relatively unfamiliar show *Nathan for You*, performance did not differ between the two samples. Together, these results suggest that crowdsourcing platforms elicit reliable performance on novel, dynamic, complex tasks.

**Keywords** Theory of mind · Crowdsourcing · Social cognition

Crowdsourcing (collecting research data via online platforms) has become increasingly popular in academic research. Recent reviews of consumer research, cognitive science, and social psychology suggest that nearly half of the published research in these fields is conducted with online samples (Anderson et al., 2019; Goodman & Paolacci, 2017; Stewart et al., 2017), with a noticeable uptick in the prevalence of crowdsourcing in recent years (Sassenberg & Ditrich, 2019). Two key benefits of such crowdsourcing is that it affords large samples relatively quickly (Keith et al., 2022; Klein et al., 2014), and it increases sample diversity (Behrend et al., 2011; Casler et al., 2013), which both address important limitations in traditional, in-person samples (Henrich et al., 2010). However, these benefits have been offset by concerns about data quality (Chandler & Shapiro, 2016; Hossain & Kauranen, 2015; Pickering & Blaszczynski, 2021). For example, online samples engage in more problematic behaviors (e.g., multitasking) than in-lab samples (Necka et al., 2016), and their data are more variable (Keith et al., 2022).

Given the increased reliance on crowdsourcing, it is important to demonstrate whether online samples yield valid and reliable data. This is particularly important for novel, complex tasks, which are largely underrepresented in crowdsourcing research. For example, crowdsourced studies in social psychology rely primarily on self-report measures (Sassenberg & Ditrich, 2019). In the field of cognitive science, crowdsourced studies are still limited primarily to measures that have been frequently studied (e.g., reaction times) (Stewart et al., 2017). As psychological research calls for methods that better capture the cognitive complexity of everyday life (Hamilton et al., 2022; Osborne-Crowley, 2020), validating the data quality from crowdsourcing platforms on novel, dynamic tasks that may accomplish this goal is an important gap in the literature. This is the focus of the current study.

✉ Anne C. Krendl
akrendl@indiana.edu

[1] Department of Psychological & Brain Sciences, Indiana University, Bloomington, 1101 E. 10th St., Bloomington, IN 47405, USA

Published online: 15 June 2023

Two approaches have generally been taken to address concerns about data quality in crowdsourcing research. One standard approach has been to administer well-validated tasks or surveys to online samples, and determine whether they yield similar results to established, previously published findings in the literature (Behrend et al., 2011; Briones & Benham, 2017; Keith et al., 2022; Klein et al., 2014; Miller et al., 2017). For example, a meta-analysis examined a variety of scales from organizational science, and found that although online and traditional samples had similar means, there was higher variability in the former than the latter (Keith et al., 2022). In a large-scale replication study, an open science project replicated several classic findings in the field of social psychology (e.g., imagined intergroup contact reduces bias) in online samples (Klein et al., 2014), suggesting that this participant sample yielded valid and replicable results. Though effective, there are two important limitations to this approach. First, it cannot be applied to novel tasks that do not have pre-established norms. Second, because this approach compares performance to norms that, in some cases, were established several decades ago (Klein et al., 2014), its accuracy may be limited due to potential differences between the original, in-person samples and the contemporary online samples.

These limitations are partially resolved in a second validation approach that directly compares performance between online and traditional, in-lab samples (e.g., undergraduate populations) in the same study (Armitage & Eerola, 2020; Behrend et al., 2011; Casler et al., 2013; Lutz, 2015; Sasaki & Yamada, 2019). For example, one study conducted a priming task in-lab and online to compare participants' reaction times, and found the two samples to have comparable response times on the task (Armitage & Eerola, 2020). Another study administered a frustration task in lab and across five online platforms, and found that effects (frustration tasks evoke anger and aggression) replicated across all sources, though effect sizes were smaller in the online samples (Lutz, 2015). In a similar approach, Casler et al. (2013) replicated well-established preferences for novel versus familiar objects in both in-lab and online samples. Though direct comparison has several strengths, a limitation of this approach is that has been primarily restricted to static, relatively unidimensional tasks. One reason for this might be to address concerns about poorer attention in online samples (Necka et al., 2016). However, these tasks may have less ecological validity than dynamic, multidimensional tasks because they do not fully capture the conceptual complexities that people face in everyday life. Demonstrating that novel, dynamic tasks yield high quality data from online samples would thus expand the scope and potential impact of online research.

The current study applied the strengths of these prior approaches to validate the data quality from an online sample on a novel, dynamic, complex task. Specifically, we administered a recently validated dynamic task as well as a novel dynamic task to an online and in-lab sample. By directly comparing the performance of the two samples on these tasks, our approach addressed potential confounds associated with relying on validations that were conducted several years prior. Moreover, by using a recently validated *and* a novel dynamic task, this approach grounds the performance on the novel task in an established literature. Finally, task complexity was achieved by focusing on theory of mind—the ability to infer others' mental states (Frith & Frith, 2005). Theory of mind is a conceptually complex construct (Apperly, 2012) that has been a widely studied topic across multiple domains of psychology (e.g., social cognition, clinical psychology, cognitive science, developmental science) (Brüne et al., 2007; Demichelis et al., 2020; Henry et al., 2013; Peterson et al., 2009), thereby making it ideally suited to the present study.

## Theory of mind: Constructs and measures

Theory of mind includes a wide range of domains, including, but not limited to, the ability to infer beliefs or intentions, understand others' emotions, detect deception, and identify social faux pas (Baron-Cohen, 2001; Quesque & Rossetti, 2020). The preponderance of theory of mind research tends to focus on a single type of theory of mind, or collapse different subcomponents of theory of mind (e.g., inferring beliefs and intentions) into a single measure of theory of mind (Fischer et al., 2017; Henry et al., 2013; Wang & Su, 2013). This has prompted critiques that traditional measures of theory of mind lack specificity (Quesque & Rossetti, 2020; Schaafsma et al., 2015).

An additional limitation of standard measures of theory of mind is that that they do not often capture known theory of mind deficits (e.g., Scheeren et al., 2013). Thus, recent work has shifted toward using dynamic stimuli that better reflect the complexity of real-world social interactions, e.g., (Byom & Mutlu, 2013; Dziobek et al., 2006; Grainger et al., 2019; Johansson Nolaker et al., 2018). Such tasks provide an opportunity to measure theory of mind in a more ecologically valid manner that does not rely on a single modality (e.g., reading a story, looking at a cartoon) (Kliemann & Adolphs, 2018). One recent study used this approach by asking older adults to answer questions that required theory of mind based on a mockumentary-style popular television show (Krendl et al., 2022; Krendl et al., in press). The task assessed multiple subcomponents of theory of mind (understanding others' affective states, beliefs, thoughts, or intentions, and detecting deception), allowing for comparisons

within sample across multiple domains. In addition to finding the predicted age deficits in theory of mind performance, the authors also found that older adults' performance predicted real-world social outcomes, notably the size and structure of older adults' social networks (Krendl et al., 2022).

In the current study, we examined whether online and in-lab samples performed comparably on two dynamic measures of theory of mind—one that has been employed multiple times in previous work on dynamic theory of mind in a traditional (college undergraduate) and community (older adult) samples (Krendl et al., 2022; Krendl et al., in press) and a second novel dynamic task. Both tasks were based on mockumentary style popular television shows (*The Office*® and *Nathan for You*®, respectively). A benefit of using a mockumentary style show is that, by portraying fictional worlds in a realistic manner, the underlying premise of the show is based on deception, which is a subcomponent of theory of mind. Moreover, the fictional aspect of the show also engages theory of mind. Indeed, prior work has shown that participants' theory of mind improved after watching fictional television series versus documentaries (Black & Barnes, 2015). We employed both a previously validated (i.e., based on *The Office*) and a novel (i.e., based on *Nathan for You*) dynamic theory of mind task for two reasons. First, the novel dynamic task allowed us to minimize familiarity effects on performance, which had been observed in previous work employing *The Office* (Krendl et al., 2022; Krendl et al., in press). Second, because the dynamic task employing *Nathan for You* was novel, it provided the opportunity to determine whether performance was similar between an online and traditional in-lab sample across multiple conceptually related but distinct measures.

The design for both dynamic tasks was similar: participants viewed clips of each television show, and completed about 60 multiple choice questions about the people and situations viewed in the clip. The questions included control questions and unique types of theory of mind (e.g., inferring beliefs, inferring intentions, understanding emotions, detecting deception, and identifying social faux pas). An in-lab sample (undergraduate) and online sample (through Prolific Academic) were recruited to complete the tasks. The in-lab sample completed additional measures, including a traditional theory of mind task. Due to time constraints, the two dynamic theory of mind tasks were completed between-subject. However, the online sample completed both tasks within-subject. The online sample also completed a self-report measure to determine whether both dynamic tasks elicited theory of mind. We predicted that the online samples' performance would be reliable across both tasks. We also predicted that the online and in-lab samples would perform similarly across both tasks.

## Methods

### Participants

The lab sample completed the two dynamic tasks between-subject, as well as several additional measures including a standard theory of mind task. For data collected online, *The Office* and *Nathan for You* tasks were completed within-subject, with order counterbalanced across participants. To address these design differences, comparisons between samples only included the online participants who completed the respective task first (e.g., lab participants who completed *The Office* were compared to online participants who completed *The Office* first). However, we used the full online sample to examine their reliability in performance across both tasks. Power estimates were calculated based on both approaches, with the former requiring the highest power (and thus is reported here). Between-group comparisons were examined using a mixed model ANOVA, with question type (control, emotion, belief, motivation, faux pas, deception) as the within-subject variable and data source (in-lab, online) as a between-subject variable. Because prior work by the authors has shown that performance is affected by familiarity with the shows, show familiarity (yes or no) was added as an additional within-subject measure. Power analyses were conducted in G*Power (Faul et al., 2007) using a small effect size ($f = .15$), $a = .05$, and assuming a moderate correlation (.4) between the measures targeted an $N$ of 248 for 80% power.

Two different undergraduate samples ($N_{Office} = 141$; $N_{Nathan\ for\ you} = 177$) were recruited for each of the dynamic tasks. The demographics of these two groups was similar, with both being about 70% female, about 75% White, with an average age of 18.9 years. See Table 1 for demographics by study. Participants received partial course credit for participating. A total of 347 participants were recruited from the online platform Prolific Academic (www.prolific.ac) (Palan & Schitter, 2018; Peer et al., 2017) for a one-hour study. Participants from Prolific Academic were selected to reflect a representative sample of the U.S. population. Consistent with Prolific Academic's practice of ethical pricing (Newman et al., 2021), participants each received $12. Prolific Academic is commonly used for research in a wide range of disciplines (Newman et al., 2021). The same group of participants completed both dynamic tasks, with 160 seeing *Nathan for You* first, and 187 seeing *The Office* first. Participants ranged in age from 18 to 84 years, and the average age was 45.8 years (SD = 15.7) About 47.6% of the sample ($N = 165$) identified as male, and 50.1% ($N = 174$) as female. More than half the sample ($N = 271$; 78.1%) was White, and the majority were well-educated, with 86.7% ($N = 301$) reporting having some college education or higher. See Table 1 for sample demographics.

**Table 1** Demographics for participants from Prolific Academic and undergraduate samples completing *The Office* and *Nathan for You*

| | | Prolific Academic ($N=347$) | In-lab *The Office* ($N=141$) | In-lab *Nathan for you* ($N=177$) |
|---|---|---|---|---|
| Mean age | | 45.89 years (15.7) | 18.95 (.94) | 18.88 (2.52) |
| Gender | Male | 165 (47.6%) | 40 (28.4%) | 45 (25.4%) |
| | Female | 174 (50.1%) | 100 (70.9%) | 128 (72.3%) |
| | Other/NB | 8 (2.3%) | 1 (.71%) | 4 (2.26%) |
| Race | White | 271 (78.1%) | 104 (74.3%) | 135 (76.3%) |
| | Non-White | 76 (21.9%) | 37 (26.2%) | 42 (23.7%) |
| Education | HS or less | 46 (13.3%) | - | - |
| | Some college or more | 301 (86.7%) | - | - |

For age, SD (). For all other demographics, data reflect the total *N* of the sample in each category with the percent of the sample in (). Both in-lab participant samples consisted of college undergraduates.

Data collection was approved by the Indiana University Institutional Review Board.

For the online sample, data from four participants were removed because they reported having difficulty watching the video clips and indicated that it negatively impacted their ability to complete the task. We also set an a priori criterion to exclude participants whose performance on any task was more than three standard deviations from the mean; 11 participants (10 online; 1 in-lab) were excluded for this reason.

## Materials

Two dynamic theory of mind tasks were employed. These were based on two U.S. mockumentary-style television shows: *The Office*® (which aired from March 2005 to May 2013) and *Nathan for You*® (which aired from February 2013 to November 2017). *The Office* task has been used in prior work (Krendl et al., 2022; Krendl et al., in press), and older adults' performance on this task has been previously shown to predict real-world social outcomes, notably the size and structure of their social network (Krendl et al., 2022). *Nathan For You* was selected because familiarity has been shown to affect performance on *The Office* (Krendl et al., 2022; Krendl et al., in press), and viewership of *Nathan for You* was much lower across multiple populations. Lower familiarity with *Nathan for You* was confirmed through two pilot analyses: one on Prolific Academic, conducted in May 2022 with 125 individuals, and one from 61 undergraduates at Indiana University (I.U.) who were surveyed in spring 2022. Pilot participants were excluded from the current task. For Prolific Academic, we found that 17.2% ($N=21$) had seen the show before, whereas only 8.2% ($N=5$) of I.U. undergraduates had seen the show before.

The design was the same for both dynamic tasks. In each task, participants viewed multiple short clips (20–30 seconds each) of a single episode, presented in sequential order. Following each clip, participants responded to a series of multiple-choice questions about what they had just seen. Across all participants, questions were presented in a fixed order, but the order of the answer options was randomized. Questions advanced once the participant had provided a response. For each question, a picture of the characters referenced in the question and/or response options was presented on the screen along with the question to ensure participants correctly identified who the question referenced. At the end of the task, participants were asked if they had ever seen *The Office/Nathan for You* before (response options: yes or no) and, if so, how familiar they were with the series. The sequential nature of the clips allowed participants to understand the basic structure of the narrative in the episode, while allowing us to measure theory-of-mind-related inferences, such as why a particular character was performing a behavior.

A separate set of questions was developed for each episode to capture five different aspects of theory of mind: inferring beliefs, detecting deception, understanding emotions, inferring motivations, and detecting faux pas. Control questions were also included that did not rely on theory of mind; rather, they were factually related to what a character had said or done. However, to correctly answer the theory of mind questions, respondents needed to use contextual or nonverbal cues to make inferences about characters' internal states. Questions were categorized to the relative theory of mind domain by full consensus of the three authors, who are experts in social cognition, including theory of mind (see Krendl et al., 2022, for similar approach). If consensus was not reached, the question was removed or modified to achieve full consensus. Full consensus was also required in evaluating the suitability of the response options.

*The Office* task was adapted from prior work with the same episode (Krendl et al., 2022; Krendl et al., in press). Modifications included adding new questions (e.g., related

to faux pas detection), and modifying clips to align with the new questions. In the current task, response options were standardized to three for all questions (1 correct, 2 foils). Approximately 12 minutes of Season 1, Episode 4 ("The Alliance") was divided into 25 clips ranging from 9 to 55 seconds in length ($M_{Length}$ = 29 seconds, SD = 9 seconds). The show was edited to follow two key plotlines (an office "alliance" and a "birthday party"); unrelated plotlines were removed. See Appendix A Table 6 for approximate time codes of each clip. As with prior iterations of this task, participants watched clips in sequential order. Following each clip, they responded to 1 to 5 questions about what they had just seen. There were a total of 64 questions.

In *The Office* task, there were nine questions pertaining to inferring beliefs (e.g., "What does Pam think about having a birthday party for Meredith?"), 10 questions related to detecting deception (e.g., "Is Jim telling Dwight the truth about why he was talking to Pam?"), 10 questions related to understanding the character's emotions (e.g., "After talking to Michael, how does Dwight feel about his job?"), 10 questions pertaining to inferring the motivations of others (e.g., "Why does Dwight want to keep the alliance secret?"), and 10 questions related to detecting if a faux pas had occurred (e.g., "Was it inappropriate for Michael to suggest an ice cream cake?"). There were also 15 control questions in which participants were asked a question about something they had just seen or heard (e.g., "When is Meredith's birthday?"). Critically, control questions did not require additional contextual cues, and simply measured comprehension. See Appendix B for full task.

In *Nathan for You* task, approximately 6.5 minutes of Season 3, Episode 3 ("The Antique Shop") was divided into 18 clips ranging from 15 to 45 seconds ($M_{Length}$ = 23 seconds, SD = 4 seconds). The show was edited to follow the key plotline. Following each clip, participants responded to 2 to 6 questions about what they had just seen. There were a total of 63 questions. To reduce potential ceiling effects, each multiple-choice question had four possible answers (1 correct, 3 foils).

In the *Nathan for You* task, there were 11 questions that measured belief inference (e.g., "What does Nathan think about some of the items in Emily's store?"), 11 that measured deception detection (e.g., "Why did Nathan want his glass to be refilled with apple juice?"), 10 that measured understanding emotions (e.g. "How does Emily feel about having bars and nightclubs in the areas?"), 10 for inferring motivations (e.g., "Why does Nathan want Emily to extend her hours?"), and 10 for detecting faux pas (e.g., "Did someone say or do something inappropriate in this clip?"). An additional 11 control questions were also included (e.g., "What is the name of Emily's business?"). See Appendix B for full task.

For both tasks, several steps were taken to minimize task demands that could emerge across domains. First, both tasks included control questions. These questions asked about specific events that had occurred, but did not require theory of mind to respond (e.g., "What does the birthday card say?"). Second, to reduce potential memory demands, each question included photographs of any character who was referenced in either the question or the response options (range for *The Office*: 1–4 characters, M = 2.01, SD = 79; range for *Nathan for You:* 0–3 characters, M = 1.73, SD = 70). The names of the characters were included below each respective photograph to remind participants who was being referenced in the question. Third, though efforts were made to standardize the length of the questions across domains, there was a slight deviation in the overall question length across domains in each task (e.g., the deception questions for the *Nathan for You* task were the longest, whereas the faux pas questions for *The Office* were the longest). See Table 2 for the average number of words per question and response options by domain for each task. To address these differences, the questions were self-paced rather than timed.

**Table 2** Mean number of words for questions and response options by domain for *The Office* and *Nathan for You.* SD ()

| | Domain | Question | Correct answer | Foils |
|---|---|---|---|---|
| *The Office* | Control | 8.13 (3.07) | 5.73 (2.99) | 5.47 (2.28) |
| | Belief | 9.6 (3.47) | 5.50 (1.84) | 6.35 (2.84) |
| | Deception | 8.3 (2.31) | 5.40 (1.51) | 5.8 (1.92) |
| | Emotion | 9.0 (1.33) | 7.5 (3.21) | 6.60 (1.82) |
| | Faux Pas | 9.78 (1.56) | 5.89 (2.47) | 7.06 (2.72) |
| | Motivation | 9.33 (2.60) | 7.78 (2.44) | 7.83 (2.01) |
| *Nathan for You* | Control | 7.82 (2.48) | 3.55 (2.30) | 3.15 (1.87) |
| | Belief | 8.36 (2.25) | 7.91 (1.92) | 6.70 (1.87) |
| | Deception | 10.40 (2.27) | 3.90 (3.358) | 4.50 (3.73) |
| | Emotion | 6.30 (2.00) | 9.80 (2.94) | 7.57 (1.88) |
| | Faux Pas | 9.64 (2.062) | 6.36 (2.84) | 6.27 (2.94) |
| | Motivation | 8.40 (1.84) | 8.60 (2.67) | 7.13 (1.72) |

For the online sample only, participants completed 12 additional questions at the end of each respective dynamic task. First, participants were asked if they had had any problems viewing the clips (response options: yes or no). If they responded yes, they were then asked whether the technical issues they experienced hindered their ability to understand the clips (response options: yes or no). As noted above, the four participants who responded yes to this question were removed from the analyses. Following these questions, they completed a five-item self-report measure that evaluated their beliefs about the extent to which the task activated theory of mind: "In order to understand this show, how important do you think it was to know what the characters were thinking?/what the characters were feeling?/what motivated what the characters were doing?/whether the characters were

**Table 3** Mean relatability, theory of mind for *The Office* and *Nathan for You*, as rated by online sample ($N = 333$). SD ()

| | | *The Office* | *Nathan for You* | *t statistic* | 95% CI | |
|---|---|---|---|---|---|---|
| | | | | | Upper | Lower |
| Relatability | Relate to content | 4.40 (1.73) | 2.88 (1.71) | −15.83** | −1.71 | −1.34 |
| | Relate to characters | 4.53 (1.74) | 3.37 (1.69) | −11.32** | −1.37 | −0.96 |
| | Socially complex | 4.81 (1.47) | 3.59 (1.71) | −12.35** | −1.42 | −1.03 |
| | Enjoy | 5.46 (1.71) | 4.24 (2.04) | −10.50** | −1.46 | −1.00 |
| | Find funny | 5.38 (1.76) | 3.94 (2.17) | −11.70** | −1.68 | −1.20 |
| Theory of mind | Thinking | 5.73 (1.35) | 5.28 (1.54) | −5.51** | −0.61 | −0.29 |
| | Feeling | 5.64 (1.31) | 5.5 (1.37) | −1.81 | −0.29 | 0.02 |
| | Motivation | 6.03 (1.11) | 5.56 (1.39) | −7.01** | −0.61 | −0.34 |
| | Deception | 5.76 (1.41) | 5.57 (1.47) | −2.49* | −0.33 | −0.04 |
| | Faux pas | 5.32 (1.63) | 5.13 (1.68) | −2.30* | −0.35 | −0.03 |

*$p < .05$, **$p < .001$

being inappropriate/whether the characters were being dishonest or misleading?" Response options ranged from 1 (not at all important) to 7 (very important). Here, responses greater than 4 would indicate that the shows engaged theory of mind. For both shows, theory of mind was required across all domains of theory of mind (range of $M_{Ratings}$: 5.13–6.02). See Table 3 for means.[1]

Finally, we asked five questions about the relatability of the shows: "Based on your own personal experiences, how much could you relate to this show's content?", "Based on your own personal experiences, how much could you relate to this show's characters?", "How much did you enjoy this show?", "How funny did you find this show to be?", "How complex were the social relationships in this show?". Response options were made on a scale of 1 (not at all) to 7 (very much). Overall, participants found *The Office* to be more relatable, socially complex, funny, and enjoyable than *Nathan for You* (all *t*s > 10.49, *p*s < .001). See Table 3 for means.

## Procedure

For all participants, the video tasks were administered through Qualtrics. After completing the informed consent, participants were told they would be watching video clips and answering questions about them. For each video, participants were also told they would first view a practice video clip, during which they should adjust their audio to a comfortable volume. They then watched a brief clip, after which they were asked "Was the audio good?" (response options: yes or no). If they responded no to the question, there were immediately taken back to the instructions for the practice clip and watched it again to adjust their audio as needed. This loop continued until they indicated that their audio was suitably adjusted. After answering yes, they were prompted to press a button to begin the task. After watching all videos and responding to all questions, participants were thanked and debriefed.

For data collected online, *The Office* and *Nathan for You* tasks were completed within subject, with order counterbalanced across participants. A post-hoc *t*-test for the online sample found that order did not affect overall performance on *The Office,* $t(335) = 1.06$, $p = .292$, 95% CI, −.008, .025. However, it did affect overall performance on *Nathan for You*, $t(335) = 2.69$, $p = .008$, 95% CI, .005, .033 such that the online sample performed better on the *Nathan for You* task if they saw it second ($M_{Accuracy} = .89$, $SD = .07$) than if they saw it first ($M_{Accuracy} = .87$, $SD = .06$). This pattern suggests that fatigue did not

[1] We also conducted a 5 (theory of mind type: infer belief, infer intention, understand emotion, detect deception, detect faux pas) × 2 (show: *The Office* v. *Nathan for You*) repeated-measures ANOVA on the subjective theory of mind questions completed by the online sample. There was a main effect of theory of mind type, $F(4,1328) = 25.376$, $p < .001$, $\eta^2_{partial} = .07$, and show, $F(1,1328) = 30.95$, $p < .001$, $\eta^2_{partial} = .09$. The main effect of show emerged because *The Office* was perceived as requiring more theory of mind than *Nathan for You*, whereas the main effect of theory of mind type emerged because, overall, detecting faux pas was rated as being least important for understanding the show, (all *t*s > 3.74, *p*s < .001), whereas understanding their motivations was most important, (all *t*s > 2.29, *p*s < .023). Understanding what others were thinking, feeling, or detecting deception generally fell in the middle (see Table 3).

The main effects were qualified by a show × theory of mind type interaction, $F(1,1328) = 6.36$, $p < .001$, $\eta^2_{partial} = .02$. Here, theory of mind types necessary for understanding *Nathan for You* separated into two groups: understanding what others were thinking or when they committed a faux pas were considered to be least important, and understanding others' motivations, deceptions, and feelings being most important. However, for *The Office*, the types of theory of mind needed for performance emerged more as a hierarchy, with detecting faux pas being rated the least important, and understanding motivation the most important. Knowing what others were thinking, feeling, or when they were being deceptive feel between the two. See Table 4.

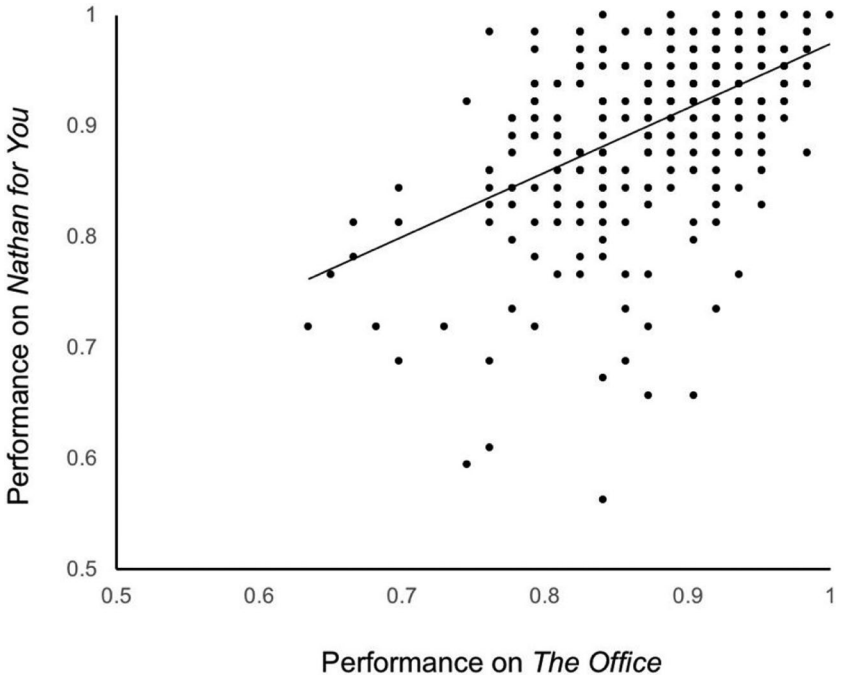


**Fig. 1** A scatterplot showing the relationship between performance on *The Office* and performance on *Nathan for You* for each online respondent. Performance is scored as proportion accuracy (#correct divided by total number of questions), ranging from 0 to 1. For online participants, performance on the two dynamic tasks was correlated, $r(333) = .50$, $p < .001$, consistent with a large effect size by Cohen's conventions

necessarily undermine performance. However, given these differences, between-group comparisons (online versus in-lab sample) for *The Office* and *Nathan for You* focus on the subset of online participants who completed the respective task first. This approach aligns with the between-subject design experienced by the laboratory participants.

## Results

Consistent with prior work, the online sample was older ($M_{Age} = 45.89$ years) than the undergraduate sample ($M_{Age} = 19$ years both tasks), included more males (online = 50.1% male, traditional = 25.4–28.4% males), and had greater variability in education levels (e.g., 13.3% had a high school education or lower). The two samples were similar in their racial diversity (~75% White across samples). See Table 1 for demographic information.

### Reliability of online sample across both video tasks

Performance on both dynamic tasks was calculated by dividing the total number of correct responses by the overall number of questions for each task. For group comparisons, we also calculated performance by each type of theory of mind. However, since we did not have predictions about reliability differing across types of theory of mind, we examined the online sample's performance reliability in their overall performance on *The Office* ($M_{Accuracy} = .91$, SD = .08) with their overall performance on *Nathan for You* ($M_{Accuracy} = .88$, SD = .07). Using Cronbach's α, we found that reliability between the two tasks was acceptable, α = .66, suggesting that the online sample's performance was consistent across both tasks. See Fig. 1.

Given the relative diversity in the online sample, we also used linear regressions to determine whether age, gender, race, education, task order, or show familiarity predicted task performance. We dichotomized all variables, (0 = male, 1 = female), race (0 = non-White, 1 = White), education (0 = lower than college, 1 = college or higher), and order (0 = *Nathan for You* first, 1 = *Nathan for You* second). Both models were significant (*The Office*: $F(6,331) = 3.40$, $p = .003$, $R^2 = .06$; *Nathan for You*: $F(6,332) = 5.51$, $p < .001$, $R^2 = .09$). For *The Office*, age, race, and familiarity contributed to this effect, all *ß*s > .11, whereas gender, race, and order contributed to the performance effects on *Nathan for You*. See Table 4 for all regression statistics.

### Examining performance on the office for online versus in-lab participants

We next examined whether performance for *The Office* differed across participant samples. We included question type

**Table 4** Summary of linear regression analysis for variables predicting performance on *The Office* and *Nathan for You* from the online sample

| *The Office* | | | | | *Nathan for You* | | | |
|---|---|---|---|---|---|---|---|---|
| Variable | $\beta$ | $t$ | $R$ | $R^2$ | $\beta$ | $t$ | $R$ | $R^2$ |
| | | | .24 | .06 | | | .29 | .09 |
| Age | −0.12 | −2.00* | | | −0.07 | −1.22 | | |
| Gender | 0.06 | 1.03 | | | 0.12 | 2.25* | | |
| Race | 0.13 | 2.23* | | | 0.21 | 3.83** | | |
| Education | 0.05 | 0.79 | | | 0.04 | 0.74 | | |
| Order | −0.03 | −0.52 | | | 0.16 | 2.90* | | |
| Familiarity | 0.14 | 2.38* | | | 0.07 | 1.25 | | |

Analysis includes five bivariate variables – race (0 = non-White, 1= White), order (0 = *Nathan for You* first, 1 = *The Office* first), gender (0 = male, 1 = female), education (0 = less than college, 1 = college or higher) familiarity with the respective show (0 = no, 1 = yes). Age is continuous. $N = 333$

*$p < .05$; ** $p < .001$

and prior familiarity with the show in these analyses as these have previously been shown to affect performance (Krendl et al., 2022; Krendl et al., in press). We thus conducted a 6 (question type: control, emotion, belief, motivation, faux pas, deception) × 2 (data source: Prolific versus in-lab) × 2 (show familiarity: yes or no) mixed-ANOVA with question type as a repeated measure. A main effect of question type emerged, $F(5,1580) = 72.59$, $p < .001$, $\eta^2_{partial} = .19$, as well as a main effect of show familiarity, $F(1,316) = 31.91$, $p < .001$, $\eta^2_{partial} = .09$. However, there was no main effect of source, $F < 1$. The main effects were qualified by two interactions: a question type X source interaction, $F(5,1580) = 3.06$, $p = .009$, $\eta^2_{partial} = .01$, and a question type X show familiarity interaction, $F(5,1580) = 5.34$, $p < .001$, $\eta^2_{partial} = .02$. There was no source X familiarity interaction, $F(1,316)$ 2.79, $p = .096$, $\eta^2_{partial} = .009$, but there was a three-way interaction, $F(5,1580) = 3.91$, $p = .002$, $\eta^2_{partial} = .012$.

The main effect of question type emerged because participants performed worse on the emotion questions compared to all other question types, and best on the control questions (see Table 5 for means by question type and source). The main effect of familiarity emerged because people performed better if they had seen the show before versus if they had not (see Table 5). This effect was particularly pronounced in the lab sample, who performed worse on all channels except for understanding motivation (all $t$s > 2.00, $p$s < .05) if they had not seen *The Office* as compared to if they had. However, for the online sample, prior familiarity with the show only affected performance on understanding deception, $t(180) = 4.43$, $p < .001$, 95% CI .04, .14, whereas familiarity did not affect performance on any other question type. For the in-lab sample, familiarity had a particularly pronounced effect for understanding emotions, with performance being much lower for the unfamiliar ($M_{Emotion} = .70$, SD = .19) than familiar ($M_{Emotion} = .86$, SD = .12) groups, $t(137) = 5.39$, $p < .001$, 95% CI .10, .22. It is important to note, however, the in-lab participants were more likely to have seen *The Office* (82.7%, $N = 115$) than online participants (66.7%, $N = 222$), $\chi^2(1,320) = 18.98$, $p < .001$.[2] Thus, given the small number of in-lab participants who had not seen the show before, the familiarity effects observed in this sample should be interpreted with caution.

## Examining performance on Nathan for you for online versus in-lab participants

We next examined performance on *Nathan for You* task using a 6 (question type: control, emotion, belief, motivation, faux pas, deception) × 2 (source: Prolific versus in-lab) × 2 (show familiarity: yes or no) mixed-ANOVA with question type as a repeated measure. As noted above, pilot testing revealed that both the in-lab and online samples were relatively unfamiliar with *Nathan for You.* As with *The Office*, a main effect of question type emerged, $F(5,1620) = 96.38$, $p < .001$, $\eta^2_{partial} = .23$, as well as a main effect of show familiarity, $F(1,324) = 4.34$, $p = .038$, $\eta^2_{partial} = .013$. However, there was no main effect of source, $F(1,324) = 1.65$, $p = .199$, $\eta^2_{partial} = .005$. No two-way or three-way interactions emerged, all $F$s < 1, $p$s > .36, $\eta^2_{partials} < .005$.

The main effect of question type emerged because, similar to *The Office* task, participants performed worse on the emotion questions compared to all other question types (all

[2] Online participants who had seen *The Office* were significantly younger ($M_{age} = 42.86$ , SD = 15.32) than those who had not ($M_{Age} = 49.92$, SD = 15.14), $t(180) = 3.05$, $p = .003$, 95% CI, 2.50, 11.61. Thus, the younger age of the in-lab sample may have accounted for their higher familiarity with the show. To confirm this, we examined familiarity with the show among online participants who were under the age of 30 ($N = 69$), since this age range was comparable to our in-lab participants. Among this group, 84.1% reported having seen *The Office* before, which was comparable to the in-lab sample.

**Table 5** Mean performance and skewness by channel, separated by source (Academic Prolific, in lab) for both tasks (*The Office*, left, and *Nathan for You*, right). SD ()

| | *The Office* | | | | *Nathan For You* | | | |
|---|---|---|---|---|---|---|---|---|
| | Prolific ($N = 181$) | | In lab ($N = 140$) | | Prolific ($N = 151$) | | In lab ($N = 177$) | |
| | Mean | Skewness | Mean | Skewness | Mean | Skewness | Mean | Skewness |
| Control | 0.93 (0.1) | −2.03 | 0.96 (0.07) | −1.78 | 0.91 (0.1) | −1.02 | 0.92 (0.1) | 1.23 |
| Emotion | 0.81 (0.17) | −.64 | 0.84 (0.15) | −1.01 | 0.72 (0.16) | −.33 | 0.74 (0.16) | −.44 |
| Belief | 0.92 (0.13) | −1.88 | 0.95 (0.1) | −1.61 | 0.93 (0.1) | −1.72 | 0.92 (0.1) | −1.35 |
| Motivation | 0.92 (0.11) | −1.53 | 0.93 (0.11) | −1.51 | 0.92 (0.1) | −1.29 | 0.95 (0.1) | −2.52 |
| Faux Pas | 0.94 (0.09) | −1.24 | 0.92 (0.11) | −1.21 | 0.79 (0.17) | −.73 | 0.76 (0.17) | −.79 |
| Deception | 0.92 (0.15) | −2.07 | 0.94 (0.11) | −1.13 | 0.93 (0.07) | −.87 | 0.95 (0.07) | −1.27 |
| Overall | 0.91 (0.09) | −1.52 | 0.92 (0.08) | −1.29 | 0.87 (0.07) | −.84 | 0.88 (0.08) | −1.30 |

$ts > 4.38$, $ps < .001$). With only one exception, participants also performed better on the control questions relative to all other question types (all $ts > 3.12$, $ps \leq .002$). Participants' performance on control questions did not differ from their performance on inferring beliefs, $t < 1$, $p = .335$, and they performed better on questions related to inferring beliefs than faux pas or deception-related questions, both $ts > 3.48$, $ps < .001$ (see Table 5 for all means). The main effect of familiarity emerged because people performed better if they had seen the show before versus if they had not. However, an important caveat to this finding is that less than 10% ($N = 33$) of the 328 participants who responded to the familiarity item indicated that they had previously seen the show. Although about 13.2% of online participants indicated prior familiarity with the show versus 7.3% of the in-lab participants, this difference was not significant, $\chi^2(1,328) = 3.14$, $p = .077$, and thus was not examined further.

## Discussion

Across two separate dynamic theory of mind tasks, the current study demonstrated that online samples yield consistent and reliable results as compared to traditional in-lab samples. Specifically, we found that the participants recruited from an online platform had similar performance to in-lab participants across two dynamic theory of mind tasks. Though in-lab participants had higher performance than the online sample on one task, this effect was nuanced and due primarily to the in-lab sample's greater familiarity with the show. Indeed, when using a novel task based on a show with which participants were relatively unfamiliar, no differences emerged between the two samples. Though a secondary goal of the current study, data from the in-lab and online samples also verified that both tasks engaged theory of mind.

Consistent with prior work in other fields that has replicated key findings across an online and in-lab sample (Armitage & Eerola, 2020; Behrend et al., 2011; Casler et al., 2013; Lutz, 2015; Sasaki & Yamada, 2019), the current study found that online participants performed similarly to in-lab samples on a previously validated (Krendl et al., 2022; Krendl et al., in press) and novel measure of dynamic theory of mind. Moreover, we found that results were robust across both samples, unlike prior work that has found that effect sizes were smaller in the online samples (Lutz, 2015). This work thus makes two important contributions to the extant work examining data quality in crowdsourcing studies. First, we demonstrated that performance on a novel but procedurally similar task was similar in in-person and online samples, thus providing important evidence about the suitability of testing crowdsourcing platforms for novel methods. Second, the current work provides a potential template for developing and validating novel, dynamic tasks for online research. Given the growth of the prevalence of crowdsourcing in psychological research (Anderson et al., 2019; Goodman & Paolacci, 2017; Stewart et al., 2017), this is an important domain of investigation, particularly with the increased use of crowdsourcing platforms since the onset of the COVID-19 pandemic (Obschonka et al., 2022).

An important caveat from these findings is that, consistent with prior work demonstrating that data from online samples is more variable (Keith et al., 2022), the proportion of data that was excluded in online sample (about 4%) was higher than our in-lab sample (< 1%). This finding may reflect the fact that the online sample performed both video tasks, whereas the in-lab sample only did one task. However, the in-lab sample completed additional tasks (e.g., Reading the Mind in the Eyes) that are beyond the scope of this investigation, resulting in the in-lab and online samples both completing about one hour of work. Thus, it is unlikely that fatigue contributed to the different exclusion rates. Indeed, we did not find order effects for the online sample that would suggest fatigue (e.g., worse performance on the second task). Another possibility for the differences in exclusion rates

is that our online sample may have been more distracted. Indeed, prior work has found that online samples engage in more problematic behaviors (e.g., multitasking) that might affect data quality (Necka et al., 2016). Thus, researchers using crowdsourcing platforms should continue to adhere to best practices to minimize and exclude low quality data (Aguinis et al., 2021).

Related to the above point, it is important to note that our results were collected through Prolific Academic and online participants recruited through Prolific Academic have higher motivation and are more attentive than participants recruited from other online platforms (d'Eon et al., 2019). Prolific Academic also yields higher quality data than other online platforms, including Amazon's Mechanical Turk (MTurk) (Adams et al., 2020; Peer et al., 2022). Finally, participant samples recruited through Prolific are more diverse than samples recruited through MTurk (Peer et al., 2017). In addition, this platform has the benefit of having a relatively more diverse participant pool than MTurk (Peer et al., 2017). However, given that MTurk is one of the primary online platforms used in psychological crowdsourcing research (Anderson et al., 2019), future work may extend these findings to other online platforms, though it will be important to bear in mind in this work that different motivations have been noted across platforms (Bakici, 2020).

The fact that performance was strongly correlated across both video tasks, but differed in the domains where performance was highest was likely driven by differences in the nature of the two tasks. Indeed, performance deficits differed across the types of theory of mind, with emotion being disrupted on both tasks, but faux pas being particularly disrupted on *Nathan for You.* Moreover, self-reports from the on-line participants suggested that different types of theory of mind were required for the different tasks. For understanding *Nathan for You*, respondents felt it was most important to understand others' motivations, deceptions, and feelings. However, for understanding *The Office*, they felt understanding motivation was the most important. This finding may speak to the broader complexities of theory of mind, e.g., Apperly (2012). Specifically, the manner in which individuals may infer beliefs or emotional states of others may differ based on the context in which the judgments are being made, the relatability of the context, and their familiarity with the situations being depicted. Future work should disentangle these differences.

Our results also suggest that both tasks have unique benefits. Because the familiarity affected performance on *The Office* task, this task might be less effective in measuring theory of mind among younger individuals, as familiarity was high in this group. However, the fact that participants found *The Office* to be more relatable, socially complex, funny, and enjoyable than *Nathan for You* may make it more engaging for other age groups. It is important to note that our results do not indicate whether either task is a more effective measure of theory of mind. Indeed, participants indicated that both shows elicit theory of mind, albeit in different ways. Specifically, understanding others' motivations, deceptions, and feelings was perceived as the most important for *Nathan for You,* whereas understanding motivations was perceived as being the most important for *The Office*. Further reinforcing the notion that the two tasks capture slightly different constructs, reliability in overall performance between *The Office* and *Nathan for You* was only acceptable. This finding is consistent with the reported reliability on other commonly used measures of theory of mind such as Baron-Cohen et al. (2001), likely reflecting the conceptual complexity of theory of mind (Apperly, 2012). Though speculative, one possibility is that individuals who are able to perform well on both tasks may have more cognitive flexibility, and are thus better able to engage theory of mind in real-world contexts. Indeed, prior work has shown that cognitive flexibility is positively related to better theory of mind performance (Champagne-Lavau et al., 2012; Sami et al., 2023). Conversely, individuals who performed relatively well on one but less well on the other may be less adept at implementing the optimal theory of mind strategies in different social situations. Because prior work has established that performance on *The Office* predicts real-world outcomes (e.g., the structure of older adults' personal social networks; Krendl et al., 2022), future work should extend this finding to performance on *Nathan for You* to disentangle these possibilities.

Directly related to the above point, the fact that familiarity affected performance on *The Office*, but not *Nathan for You* raises important considerations about how differences in the characteristics of on-line versus in-lab samples might affect performance. Indeed, we found that the older participants from our on-line sample were less familiar with *The Office* (a finding that replicates our work with older adults from community samples (Krendl et al., 2022)), and familiarity drove performance differences between the two samples. Thus, it may be that some differences that emerge between "classic" findings from in-lab samples and crowdsourcing platforms may be related, as least in part, to differences in the sociodemographic characteristics of the samples (see below for further discussion). These findings suggest that it may be particularly important to assess prior familiarity when using mainstream stimuli. Moreover, since familiarity may differ as a function of sociodemographic factors (e.g., age), there may be advantages to using unfamiliar or less familiar stimuli. Indeed, our concerns related to familiarity with *The Office* promoted us to develop an additional task with the less familiar show *Nathan for You.*

There are several limitations to the current study. First, there is an inherent tradeoff between improving ecologically valid and losing specificity in theory of mind measures.

Though numerous efforts were made to ensure the overall rigor of the tasks (e.g., by using a within-subject design, including control questions, making the tasks self-paced), the stimuli in these tasks were less controlled than they are in standard theory of mind tasks. Because theory of mind engages multiple cognitive resources, including memory (Fernandes et al., 2021; Laillier et al., 2019) and executive function (Bailey & Henry, 2008; Charlton et al., 2009; Wang & Su, 2013), individual differences in these domains could affect performance. However, a benefit of the within-subject design is that deficits in either domain would be distributed across all sub-types of theory of mind that were assessed within task.

A second limitation of this study is that, though the consensus-based approach for developing the tasks was consistent with the manner in which other standard theory of mind tasks were developed (Baron-Cohen et al., 2001; Saxe & Kanwisher, 2003), this approach could limit the potential generalizability of the tasks. Specifically, the consensus approach could have been biased by potential differences in how the authors interpret dynamic social stimuli as compared to how individuals with different sociodemographic characteristics or life experiences might interpret the same stimuli. Though the benefit of having experts on social cognition develop the task is that it better ensures that the task accurately reflects the core theoretical constructs associated with the unique theory of mind domains being examined, differences associated with sociodemographic variables did emerge on this task. Specifically, even when controlling for familiarity, age and race were associated with better performance on *The Office* task, whereas race and gender were associated with better performance on the *Nathan for You* task. Though the authors differed in gender, they did not differ in race or age, raising the possibility that the task construction may reflect inherent biases. Conversely, it is possible that the tasks themselves, which featured young-to-middle aged White men and women, may have driven the sociodemographic differences that emerged. Simply put, some work suggests that theory of mind performance is more accurate for ingroup versus outgroup members (Gönültaş et al., 2020). While this possibility underscores the importance of considering and controlling for sociodemographic factors in research on social cognition, future work should consider these limitations and strive toward developing more inclusive tasks.

A final caveat to our results is that performance across both tasks was quite high, suggesting potential ceiling effects. Ceiling effects have been frequently observed in theory of mind tasks with adults. Indeed, two meta-analyses found that healthy adults scored >90% accuracy on most standard theory of mind tasks (Bora et al., 2009; Chung et al., 2014). Given that theory of mind is a fundamental social cognitive skill (Frith & Frith, 2005), it is therefore unsurprising that performance is quite high on these tasks. However, the fact that poorer performance has been shown to predict real-world outcomes (Krendl et al., 2022) speaks to the importance of using tasks that may be relatively easy for some, but challenging for others. However, future work should explore whether more challenging theory of mind tasks uncover important nuances in social behavior.

Together, the results of the current study suggest that crowdsourcing elicits reliable performance on novel social cognitive tasks. Critically, by leveraging a task that has been used in other work by the research team (Krendl et al., 2022; Krendl et al., in press) and a novel task, we were able to confirm that online samples elicit similar results on traditional in-lab participants on complex social cognitive tasks.

## Appendix A

**Table 6** Twenty-five clips were extracted from Season 1, Episode 4, "The Alliance", from *The Office®,* and 18 were extracted from Season 3, Episode 4 "The Antique Shop" from *Nathan for You®*

| | *The Office* | | *Nathan For You* | |
|---|---|---|---|---|
| | Duration | Time code | Duration | Time code |
| Clip 1 | 0:28 | 0:01–0:29 | 0:22 | 0:00–0:22 |
| Clip 2 | 0:48 | 1:33–2:21 | 0:29 | 0:22–0:48 |
| Clip 3 | 0:39 | 2:21–3:00 | 0:22 | 0:47–1:09 |
| Clip 4 | 0:26 | 3:41–4:07 | 0:24 | 1:09–1:33 |
| Clip 5 | 0:55 | 4:15–5:10 | 0:15 | 1:33–1:49 |
| Clip 6 | 0:26 | 5:12–5:38 | 0:21 | 1:47–2:08 |
| Clip 7 | 0:27 | 5:39–6:06 | 0:19 | 2:06–2:25 |
| Clip 8 | 0:09 | 6:09–6:18 | 0:22 | 2:26–2:48 |
| Clip 9 | 0:31 | 10:15–10:46 | 0:21 | 2:45–3:07 |
| Clip 10 | 0:22 | 7:05–7:27 | 0:27 | 3:08–3:35 |
| Clip 11 | 0:17 | 8:20–8:37 | 0:27 | 3:35–4:02 |
| Clip 12 | 0:31 | 8:33–9:04 | 0:19 | 4:08–4:29 |
| Clip 13 | 0:45 | 9:21–10:06 | 0:21 | 4:30–4:51 |
| Clip 14 | 0:16 | 11:08–11:24 | 0:24 | 4:51–5:12/5:12–5:22 |
| Clip 15 | 0:10 | 12:30–12:40 | 0:26 | 5:22–5:49 |
| Clip 16 | 0:16 | 13:18–13:34 | 0:21 | 5:49–6:11 |
| Clip 17 | 0:41 | 13:35–14:16 | 0:19 | 8:24–8:42 |
| Clip 18 | 0:42 | 14:18–15:00 | 0:45 | 8:42–9:30 |
| Clip 19 | 0:46 | 15:01–15:47 | - | - |
| Clip 20 | 0:21 | 16:10–16:31 | - | - |
| Clip 21 | 0:24 | 16:43–17:07 | - | - |
| Clip 22 | 0:50 | 17:07–17:57 | - | - |
| Clip 23 | 0:23 | 19:35–19:58 | - | - |
| Clip 24 | 0:30 | 19:59–20:29 | - | - |
| Clip 25 | 0:14 | 20:36–20:50 | - | - |

Approximate time codes from which each clip was extracted are listed below. Start time codes are based on a 0:00 start time for the episode, and no commercials. Duration (clip length) of each clip is also provided.

## Appendix B

Question numbers (corresponding to order in which question was presented in the task), response options, and corresponding clip number for *The Office*® and *Nathan for You*® theory of mind tasks. Each subcomponent of theory of mind (control, detecting deception, understanding emotions, inferring beliefs, inferring intentions) is denoted for the question. For clarity, correct answers are always listed first, but answers on the task were presented in random order across participants.

**"The Office"**

**Clip 1**

1) Why did Michael seem annoyed when he saw Dwight waiting? (faux pas)

   a) He thinks it wasn't appropriate for Dwight to wait by the bathroom.
   b) He hadn't seen Dwight in a long time
   c) He had been expecting to see someone else

2) After talking to Michael, how does Dwight feel about his job? (emotion)

   a) Dwight is worried about losing his job
   b) Dwight is happy there will not be downsizing
   c) Dwight is no longer worried about losing his job

3) Why does Michael say, "No, no, no… Maybe." (inference)

   a) He thinks there could be downsizing
   b) He does not think there will be downsizing
   c) He thinks Dwight is nosy

**Clip 2**

4) Why does Michael want to have a birthday party for someone at the office? (control)

   a) He wants to improve morale
   b) He does not want to work
   c) He likes parties

5) When is Meredith's birthday? (control)

   a) The next month
   b) The next week
   c) The next day

**Clip 3**

6) What does Pam think about having a birthday party for Meredith? (inference)

   a) Pam thinks it's a bad idea
   b) Pam thinks it will boost morale
   c) Pam is disappointed it won't be for her

7) Did someone say something awkward in this clip? (faux pas)

   a) Yes, Michael called Pam a wet blanket
   b) Yes, Pam asked if there would be a party
   c) No, nothing was awkward

**Clip 4**

8) What will Meredith think of having an ice cream cake? (inference)

   a) She will not want the cake
   b) She will be excited about the cake
   c) She will hope the cake is mint chocolate chip

9) Why does Michael suggest having an ice cream cake? (motivation)

   a) He wants an ice cream cake
   b) He thinks Meredith would like it
   c) He has never had an ice cream cake

10) Was it inappropriate for Michael to suggest a mint chocolate chip ice cream cake? (faux pas)

   a) Yes, because Meredith is allergic to dairy
   b) Yes, because Meredith doesn't like mint chocolate chip
   c) No, the suggestion was appropriate

**Clip 5**

11) What does Jim say about forming an alliance? (control)

a) Jim says it's a good opportunity to get back at Dwight
b) Jim says the alliance will protect him from downsizing
c) Jim says the alliance will keep him from getting arrested

12) Did someone say or do something awkward in this clip? (faux pas)

a) Yes, Dwight made a joke about his muscles
b) Yes, Jim talked about getting arrested
c) No, nothing was awkward

**Clip 6**

13) Why does Dwight want to keep the alliance secret? (motivation)

a) He does not want others to know about it
b) He does not want others to gossip about them
c) He does not want to keep it a secret

14) Why do you think Dwight is watching Jim? (motivation)

a) He wants to know what Jim is telling Pam
b) He wants to talk to Jim about paper products
c) He needs Jim's help spying on someone

**Clip 7**

15) Is Jim telling Dwight the truth about why he was talking to Pam? (deception)

a) No, Jim is lying to Dwight
b) Yes, Jim is telling Dwight the truth
c) No, Jim does not understand his conversation

16) Why do you think Jim tells Dwight to ignore it if he talks with Pam? (motivation)

a) Jim does not want Dwight to interfere if he talks to Pam
b) Jim is worried Dwight will tell Pam about the alliance
c) Jim does not want Dwight to be worried

**Clip 8**

17) Why is Michael laughing? (emotion)

a) Michael thinks the card is funny
b) Michael is excited to celebrate Meredith's birthday
c) Michael thinks Meredith will not understand the card

18) What does the birthday card say? (control)

a) The card says "Happy Bird Day"
b) The card says "Happy birthday"
c) The card was not shown

**Clip 9**

19) Why does Michael want to talk to Dwight? (control)

a) He wants to know something personal about Meredith
b) He wants to talk about the alliance
c) He wants to know about her hysterectomy

20) How many times has Meredith been divorced? (control)

a) She has been divorced twice
b) She has been divorced once
c) She has never been divorced

21) Was it appropriate for Dwight to bring up Meredith's surgery? (faux pas)

a) No, it was inappropriate
b) Yes, it was appropriate
c) No, Michael didn't know what it was

**Clip 10**

22) What does Jim believe his colleagues are up to in the kitchen? (inference)

a) Jim thinks they are having lunch
b) Jim agrees with Dwight's suspicions that they are up to something
c) Jim is not sure and wants to find out

23) What does Jim tell Dwight about his conversation in the kitchen? (deception)

a) Jim lied to Dwight
b) Jim told Dwight the truth
c) Jim did not remember his conversation

24) What does Jim ask his co-worker about in the kitchen? (control)

a) A sandwich
b) Downsizing
c) Their different departments

**Clip 11**

25) Why did Oscar go into Michael's office? (motivation)

a) He wants Michael to donate to his nephew's charity
b) His nephew is sick and he is raising money
c) He wants to talk to Michael about his nephew

**Clip 12**

26) Why does Michael agree to donate to the charity? (motivation)

a) Michael wants to look good
b) Michael cares about cerebral palsy
c) Michael cares about Oscar's nephew

27) What does Michael say about other people's donations? (control)

a) Michael suggests they did not donate enough
b) Michael says everyone was very generous
c) Michael says people care about the cause

28) How does Oscar feel about Michael's donation? (emotion)

a) Oscar is happy
b) Oscar is annoyed
c) Oscar is annoyed

**Clip 13**

29) Why does Pam want to talk to Jim? (deception)

a) Pam and Jim are trying to fool Dwight
b) Pam is upset at Michael
c) Pam wants Jim's advice

30) Has Pam been working with Michael and corporate about downsizing? (deception)

a) No, Pam is lying about Michael
b) Yes, Pam is being truthful about Michael
c) Probably, she would need to help with the downsizing

31) What does Jim mean when he says Pam is so great? (emotion)

a) Jim has a crush on Pam
b) Jim thinks Pam is a great actress
c) Jim is being sarcastic

**Clip 14**

32) Where are Jim and Dwight talking? (control)

a) In an office
b) Sitting at their desks
c) In the warehouse

33) Why does Jim tell Dwight that one of the alliances is going to meet in the warehouse? (deception)

a) Jim is lying to Dwight
b) Jim wants Dwight's help
c) Jim wants to warn Dwight

**Clip 15**

34) Why does Jim tell Dwight "that's good"? (deception)

a) Jim is playing along with Dwight
b) Jim thinks Dwight is being very clever
c) Jim wants to go to Meredith's birthday party

**Clip 16**

35) Why does Meredith not want any cake? (control)

a) She is allergic to dairy
b) She does not like cake
c) She already had some

36) Did someone say something inappropriate in this clip? (faux pas)

a) Yes, Michael said he'd hate to be allergic to dairy
b) Yes, Meredith talked about her dairy allergy
c) No, the comments were appropriate

37) How does Michael feel about the cake? (emotion)

a) He is happy to have the cake
b) He is worried about Meredith
c) He is relieved that there's more cake for him

**Clip 17**

38) Why does Pam go downstairs? (deception)

a) Pam is trying to fool Dwight
b) Pam needs to make a phone call
c) Pam wants to get Dwight to come out of the box

39) Who is Pam talking to on the phone? (deception)

a) No one
b) Jim
c) An unknown friend

40) Why does Pam run out of the warehouse? (motivation)

a) She can't stop laughing
b) She needs to get back to the party
c) She is scared of the box

**Clip 18**

41) How does Michael feel about his donation? (emotion)

a) Michael is unhappy about his donation
b) Michael is proud of his donation
c) Michael is embarrassed of his donation

42) What had Michael believed when he made his donation? (inference)

a) He thought the donations were a set amount of money
b) He thought the donations were per mile
c) He thought he donated the same amount as everyone

43) Why does Michael ask if Oscar's around? (motivation)

a) He wants to change his donation
b) He wants to give Oscar his donation
c) He wants his help finding Dwight

**Clip 19**

44) How does Michael feel about Oscar's nephew total miles last year? (emotion)

a) Michael is unhappy
b) Michael is impressed
c) Michael is jealous

45) Did Oscar deceive Michael in order to get the donation? (deception)

a) No, Michael made a mistake in his donation
b) Yes, Oscar intentionally lied to him to get him to donate more
c) No, Michael changed his mind about how much to donate

46) Did something awkward happen in this clip? (faux pas)

a) Yes, Michael and Oscar disagree about the donation
b) Yes, Michael and Oscar left the birthday party
c) No, nothing was awkward in the clip

**Clip 20**

47) Who does Michael think wrote the best comment on Meredith's card? (inference)

a) Michael
b) Dwight
c) Jim

48) 48 What was the message written on Meredith's card? (control)

a) "Because you work here, where time stands still."
b) "Happy birthday, you're the best."
c) "Good news, it's your birthday!"

49) How does Michael feel about the comment Meredith read? (emotion)

a) Michael feels annoyed
b) Michael thinks it's funny
c) Michael feels embarrassed

**Clip 21**

50) What does Meredith think about Michael's comment on her card? (inference)

a) Meredith does not think it's funny
b) Meredith does not understand the joke
c) Meredith thinks the joke is funny

51) How does Michael feel about Meredith reaction? (emotion)

a) Michael is unhappy
b) Michael is jealous
c) Michael is proud that he outsmarted her

52) Why did no one laugh at Michael's joke? (faux pas)

a) People found it offensive
b) Someone else made a funnier joke
c) No one understood the joke

**Clip 22**

53) Did people think Michael's joke about Meredith's divorces was funny? (faux pas)

a) No, it was offensive
b) No, because it wasn't true
c) No, because no one understood the joke

54) Why does Michael say that he got the joke off the internet? (motivation)

a) He doesn't want to take responsibility for the joke
b) He doesn't want to take responsibility for the joke
c) He didn't want people to know that he couldn't think of a joke

55) Why does Oscar say "Nice party, Michael"? (inference)

a) Oscar is being sarcastic
b) Oscar is having a good time at the party
c) Oscar wants to hear more of Michael's jokes

**Clip 23**

56) Where were Jim and Pam talking? (control)

a) Behind her desk
b) In Michael's office
c) In the parking lot

57) What reason does Jim give Dwight for going to Stamford? (control)

a) To spy on their other branch
b) To find out if the other branch is spying on them
c) To play a prank on Michael

58) Which best explains what Jim told Dwight? (deception)

a) Jim lied to Dwight
b) Jim told Dwight the truth
c) Jim is confused

**Clip 24**

59) How does Jim feel about seeing Pam's fiancé, Roy? (emotion)

a) Jim is unhappy to see Roy
b) Jim is excited to see Roy
c) Jim is annoyed at Roy

60) Why is Pam's fiancé, Roy, unhappy with Jim? (faux pas)

a) He doesn't think it's appropriate for Jim to be so friendly with Pam
b) He wants to be a part of the joke
c) He doesn't think Jim will let him join the alliance

61) What did Jim want Dwight to tell Pam's fiancé? (inference)

a) Jim wanted Dwight to talk about the alliance
b) Jim wanted Dwight to keep the alliance a secret
c) Jim wanted Dwight to play more office pranks

62) What does Dwight say he has no idea about the alliance? (motivation)

a) He wants to get back at Jim
b) He does not want Roy to be mad at him
c) He does not know what Jim is talking about

63) What reason does Pam give her fiancé, Roy, about why she and Jim are laughing? (control)

a) Office pranks
b) Dwight is going to dye his hair blonde
c) Jim convinced Dwight to go to Stamford

**Clip 25**

64) Why is Dwight's hair blonde? (control)

a) Because Jim told him to dye it to go undercover
b) He wanted to get back at Jim
c) He is making fun of Jim

**"Nathan For You"**

**Clip 1**

1. What type of businesses are around the antique shop? (control)

a. Bars
b. International restaurants
c. Antique shops
d. Grocery stores

2. What is Emily's business problem? (control)

a. She does not have a lot of customers
b. She cannot hire enough help
c. She sells poor quality goods
d. She does not like to drink

3. What is the name of Emily's business? (control)

a. Magnolia & Willow
b. Sport Bar
c. til 2 Club
d. P.B.S. Pub & Company

**Clip 2**

4. How does Emily feel about having bars and nightclubs in the area? (emotion)

a. It bothers her
b. It does not concern her
c. It makes her happy
d. It makes her angry

5. Why does Emily think overserving (serving too much alcohol) is a problem? (infer belief)

a. People might get rowdy or disruptive
b. People won't have money left to shop
c. The bars will run out of alcohol
d. The bars become overcrowded

6. Why does Emily think having bars nearby doesn't affect her business? (infer belief)

a. Her store closes before most people are drunk
b. She is closed in time to go to the bars after work
c. Bars and antique stores attract different customers
d. Their different hours make parking easier for her customers

7. What does Nathan want Emily to change about her business? (control)

a. Her hours
b. What she sells
c. Her location
d. The store's size

8. Why does Nathan want Emily to extend her hours? (motivation)

a. So drunk customers will come to her store
b. She needs to work harder
c. So she can hire people from the bar
d. To make sure no one breaks anything

9. What is the policy in Emily's store? (control)

a. You break it, you buy it
b. Don't touch the merchandise
c. No shoes, no service
d. Cash only

**Clip 3**

10. How does Emily feel about Nathan's plan? (emotion)

a. Reluctant
b. Enthusiastic
c. Intrigued
d. Agitated

11. Why does Nathan want drunk customers in Emily's store? (motivation)

a. Because they are likely to break her merchandise
b. He wants Emily to have company when she is open late
c. So Emily can have more people in the store
d. He thinks drunk customers will want to spend more money

12. Did someone say or do something inappropriate in this clip? (social norm violation)

    a. Yes, Nathan's plan for people to break things so they have to buy them is inappropriate
    b. Yes, it is inappropriate to make Emily work longer days
    c. No, there was nothing inappropriate in this clip
    d. Yes, Emily did not appreciate Nathan's help

**Clip 4**

13. Why does Nathan want "the right drunk" to come in Emily's shop? (motivation)

    a. They are more likely to break, then have to buy, a lot of items
    b. To get them interested in antiques
    c. To convince Emily of his plan
    d. They would want to spend a lot of money

14. Did something awkward happen in this clip? (social norm violation)

    a. Yes, Emily is uneasy by the plan, and Nathan does not notice
    b. Yes, Emily does not want to work late at night
    c. No, Emily and Nathan are both excited about the plan
    d. No, Nathan thinks Emily's business is doing well

15. How does Emily feel about selling broken items (emotion)

    a. She is not enthusiastic about the idea
    b. She loves the idea
    c. She is sorry she didn't think of it first
    d. She is frustrated by the idea

16. Did someone say something inappropriate? (social norm violation)

    a. Yes, Nathan said some of Emily's items aren't worth buying
    b. Yes, Emily says she would prefer not to sell broken items
    c. No, a sale is a sale
    d. No, Nathan and Emily agree

17. What does Nathan think about some of the items in Emily's store? (infer belief)

    a. He thinks some of them aren't worth buying
    b. He thinks they are in high demand
    c. He thinks they are attractive
    d. He thinks they are underpriced

18. How does Emily feel when Nathan says that some items wouldn't sell unless broken? (emotion)

    a. She is offended
    b. She is amused
    c. She is confused
    d. She is irritated

**Clip 5**

19. Why did Nathan narrow the aisles? (motivation)

    a. To increase the likelihood that people would break the items
    b. To allow space to showcase more items
    c. To confuse people when they enter the store
    d. To make space for a bar

20. Where did he put her poorer selling items? (control)

    a. On the edge of shelves
    b. In the front of the store
    c. On the floor
    d. In different aisles

21. Why did Nathan move Emily's poorer selling items? (motivation)?

    a. To make them easier to break
    b. To get people to notice them
    c. To make them easier to reach
    d. To hide items that sell easily

**Clip 6**

22. Why did Nathan say he wanted to guarantee Emily saw results that evening? (control)

    a. So she would use his plan
    b. So she would not have to sell broken items
    c. So she could sell undesirable items
    d. So he could have a drink

23. Why did Nathan head to a nearby bar? (motivation)

    a. To find someone drunk to take to the store
    b. To have a drink after the long day
    c. To ask the owners about their business success
    d. To find a nice place to show Emily

**Clip 7**

24. What type of patron did Nathan want to befriend (motivation)?

    a. Someone he thought he would be able to fool
    b. Someone who would have a drink with him
    c. Someone who wanted to chat with him about his plan
    d. Someone who likes antiques

25. Did something awkward happen in this clip? (social norm violation)

    a. Yes, the locals didn't want to talk to Nathan
    b. Yes, Nathan did not have a drink
    c. No, there was nothing awkward
    d. No, Nathan and the locals had fun

**Clip 8**

26. What does the JJ say is his favorite movie? (control)

    a. Forrest Gump
    b. Inception
    c. No favorite
    d. Antique Store

27. Why does JJ think Nathan is talking to him? (infer belief)

    a. Because JJ thinks Nathan is friendly
    b. Because Nathan sells antiques
    c. Because Nathan wants to borrow JJ's sunglasses
    d. Because Nathan hopes JJ will buy him a drink

28. Why is Nathan actually talking to JJ? (deception)

    a. Nathan thinks JJ would be easy to fool
    b. Nathan is friendly
    c. Nathan thinks JJ is an interesting person
    d. Nathan thinks JJ would like antiques

29. Why does Nathan want to stay sober? (motivation)

    a. So he can keep his focus
    b. So he does not give the bar more business
    c. So he can enjoy listening to JJ
    d. So he can drive JJ home

**Clip 9**

30. What is Nathan drinking (control)

    a. Apple juice
    b. Beer
    c. Liquor
    d. Water

31. Why did Nathan want his glass to be refilled with apple juice? (deception)

    a. Nathan wants to stay sober, but encourage JJ to drink
    b. Nathan prefers the taste of apple juice
    c. Nathan did not want to pay for alcohol
    d. Nathan wants to use the device he had made

32. What does JJ think Nathan is drinking? (infer belief)

    a. Liquor
    b. Apple juice
    c. Beer
    d. Water

**Clip 10**

33. What did JJ's roommates put in his pocket? (control)

    a. His address
    b. Money
    c. A recipe
    d. A shopping list

34. Did something awkward happen in this clip? (social norm violation)

    a. Yes, JJ tells Nathan he gets drunk every night
    b. Yes, Nathan should have shown JJ his address too
    c. No, Nathan and JJ enjoyed each other's company
    d. No, Nathan decided it was time to leave when JJ got tipsy

35. Why did Nathan think they were ready to head out? (infer belief)

    a. He thought JJ was drunk enough to break things in the shop
    b. He thought JJ needed to go home
    c. He wanted to see if Emily had gotten any business
    d. He was tired of spending time with JJ

**Clip 11**

36. Is there a costume party? (deception)

    a. No, but JJ thinks there is

b. Yes, and Nathan is taking JJ
c. No, but JJ knows that
d. Yes, but Nathan is going alone

37. Why are the cameras actually there (deception)

a. They are recording Nathan's show
b. To record a documentary on nightlife in Long Beach
c. They are security cameras for the bar
d. They belong to other patrons in the bar

38. Why did Nathan tell JJ there was a costume party? (deception)

a. To get JJ to wear a bulky outfit
b. To keep JJ out longer
c. Nathan likes costume parties
d. Nathan thought JJ would enjoy it

39. Why did Nathan want JJ to wear that specific costume? (motivation)

a. To make JJ more likely to break things in the store
b. To disguise him from Emily
c. To keep him warm while they walked
d. To make him look foolish

**Clip 12**

40. Why is Nathan wearing a costume? (deception)

a. To convince JJ that there was a costume party
b. Because he does not want JJ to look foolish
c. Because he was tired of wearing his jacket
d. Because he got his clothes dirty

41. Why does the antique shop advertise free pizza? (deception)

a. To draw people into the store
b. To compete with the bar
c. To expand their inventory
d. To make the store smell better

42. Who does JJ think put the pizza in the antique shop? (infer belief)

a. The store owner
b. Nathan
c. The bar owner
d. The cameraman

43. Why is the pizza in the back of the antique shop? (deception)

a. To trick people to walk down the aisle and break things
b. It's a thank you for paying customers
c. To help drunk people sober up
d. As a snack for Emily when she's working late

**Clip 13**

44. Why does Nathan tell JJ to be careful? (deception)

a. Nathan wants JJ to think he's being supportive
b. Nathan does not want JJ to eat all the pizza
c. Nathan doesn't want JJ to break anything
d. JJ is not looking where he is going

45. What does Emily think JJ will do? (infer belief)

a. She thinks he will break something
b. She thinks he will eat all the pizza
c. She thinks he will want to buy antiques
d. She thinks he will be careful

46. Did something awkward happen? (social norm violation)

a. Yes, JJ does not fit in the aisle
b. Yes, Nathan told JJ to be careful
c. No, Nathan walked into the store with JJ
d. No, Nathan warned JJ to be careful

**Clip 14**

47. Why would Nathan ask JJ if he wants the pizza? (deception)

a. Nathan wants JJ to break more things
b. Nathan thinks JJ is hungry
c. Nathan is worried about JJ
d. Nathan wants the pizza

48. Did something awkward happen? (social norm violation)

a. Yes, JJ broke items in the store
b. Yes, Nathan did not help JJ
c. No, Nathan helped JJ
d. No, Emily was happy to make the sale

49. How does JJ feel about the "break it, you buy it" policy? (emotion)

a. Unhappy
b. Excited
c. Surprised
d. Angry

50. How does Emily feel as antiques are being broken? (emotion)

a. She is uncomfortable with the situation
b. She is happy that Nathan's plan is working
c. She is angry that JJ didn't break more
d. She is worried that JJ will not be able to pay her

**Clip 15**

51. Why did Nathan tell JJ he was pretty clumsy? (deception)

a. So JJ would not realize Nathan tricked him
b. Nathan is mad at JJ for breaking so many things
c. So JJ would enjoy the pizza
d. To distract JJ from wanting to take off his costume

52. How much damage did JJ cause? (control)

a. About $300
b. Less than $100
c. About $1000
d. At least $500

53. How does JJ feel about the cost of the broken items? (emotion)

a. Defeated
b. Surprised
c. Relieved
d. Carefree

**Clip 16**

54. Did something inappropriate happen? (social norm violation)

a. No, it was appropriate for JJ to pay for the items he broke
b. Yes, JJ's credit card was denied
c. Yes, Emily should have charged JJ more
d. No, Nathan made JJ feel better

55. How does JJ feel about getting to take home the vase? (emotion)

a. He's unhappy because the vase isn't worth the money he paid
b. He's excited because it looks brand new
c. He's relieved that something good came out of it
d. He's angry that he did not get to take more items

56. How does Emily feel about having JJ buy the items he broke? (emotion)

a. She is uncomfortable with it
b. She is excited that it helped her business
c. She is surprised that she made such a large sale
d. She is upset JJ took her vase

**Clip 17**

57. What does Nathan think about how his plan went? (infer belief)

a. He thinks Emily will be excited about it
b. He thinks it was a bad idea
c. He wishes it had not cost JJ so much
d. He thinks it was uncomfortable

58. What does Emily think of Nathan's plan? (infer belief)

a. She is uneasy with the plan
b. She thought the plan was a great idea
c. She will definitely try the plan on her own
d. She thinks Nathan will continue to carry it out for her

59. Did something awkward happen in this clip? (social norm violation)

a. Yes, Nathan does not understand that Emily dislikes his plan
b. Yes, Emily does not appreciate Nathan's hard work
c. No, Nathan helped Emily's business
d. No, Nathan and Emily are on the same page

**Clip 18**

60. Did someone do something awkward in this clip? (social norm violation)

a. Yes, Nathan broke the plate
b. No, nothing was awkward in the clip
c. Yes, Nathan told Emily to have a good day
d. No, Emily thanked Nathan for his help

61. Why did Nathan break the plate? (motivation)

    a. To give Emily another sale
    b. He did not like the plate
    c. He wanted to give it to JJ
    d. It was an accident

62. Do you think Emily will continue to use Nathan's plan? (infer belief)

    a. No, she did not like it
    b. Yes, she will try it
    c. Yes, she thinks it was really great
    d. No, she wanted to sell more

63. How did Nathan feel about breaking the plate? (emotion)

    a. Proud
    b. Embarrassed
    c. Angry
    d. Fortunate

**Acknowledgments** This project was supported by R01s AG070931 and AG075044 from the NIA (PI: Krendl). The authors thank Amy Gourley and Dr. Lucas Hamilton for assistance with data collection and data cleaning. The authors have no conflicts to report. This study was approved by the Institutional Review Board at Indiana University (IRBs # 2008106329 and # 15297).

## References

Adams, T. L., Li, Y., & Liu, H. (2020). A replication of beyond the turk: Alternative platforms for crowdsourcing behavioral research–sometimes preferable to student groups. *AIS Transactions on Replication Research, 6*(1), 15.

Aguinis, H., Villamor, I., & Ramani, R. S. (2021). MTurk research: Review and recommendations. *Journal of Management, 47*(4), 823–837.

Anderson, C. A., Allen, J. J., Plante, C., Quigley-McBride, A., Lovett, A., & Rokkum, J. N. (2019). The MTurkification of social and personality psychology. *Personality and Social Psychology Bulletin, 45*(6), 842–850.

Apperly, I. A. (2012). What is "theory of mind"? Concepts, cognitive processes and individual differences. *Quarterly Journal of Experimental Psychology, 65*(5), 825–839.

Armitage, J., & Eerola, T. (2020). Reaction time data in music cognition: Comparison of pilot data from lab, crowdsourced, and convenience web samples. *Frontiers in Psychology, 10*, 2883.

Bailey, P. E., & Henry, J. D. (2008). Growing less empathic with age: Disinhibition of the self-perspective. *The Journals of Gerontology Series B: Psychological Sciences and Social Sciences, 63*(4), 219–226.

Bakici, T. (2020). Comparison of crowdsourcing platforms from social-psychological and motivational perspectives. *International Journal of Information Management, 54*, 102121.

Baron-Cohen, S. (2001). Theory of mind in normal development and autism. *Prisme, 34*(1), 74–183.

Baron-Cohen, S., Wheelwright, S., Hill, J., Raste, Y., & Plumb, I. (2001). The "Reading the mind in the eyes" test revised version: A study with normal adults, and adults with Asperger syndrome or high-functioning autism. *Journal of Child Psychology and Psychiatry, and Allied Disciplines, 42*, 241–251.

Behrend, T. S., Sharek, D. J., Meade, A. W., & Wiebe, E. N. (2011). The viability of crowdsourcing for survey research. *Behavior Research Methods, 43*(3), 800–813.

Black, J., & Barnes, J. L. (2015). Fiction and social cognition: The effect of viewing award-winning television dramas on theory of mind. *Psychology of Aesthetics, Creativity, and the Arts, 9*(4), 423.

Bora, E., Yucel, M., & Pantelis, C. (2009). Theory of mind impairment in schizophrenia: Meta-analysis. *Schizophrenia Research, 109*(1-3), 1–9.

Briones, E. M., & Benham, G. (2017). An examination of the equivalency of self-report measures obtained from crowdsourced versus undergraduate student sample. *Behavior Research Methods, 49*(1), 320–334.

Brüne, M., Abdel-Hamid, M., Lehmkämper, C., & Sonntag, C. (2007). Mental state attribution, neurocognitive functioning, and psychopathology: What predicts poor social competence in schizophrenia best? *Schizophrenia Research, 92*(1-3), 151–159.

Byom, L. J., & Mutlu, B. (2013). Theory of mind: Mechanisms, methods, and new directions. *Frontiers in Human Neuroscience, 7*, 413.

Casler, K., Bickel, L., & Hackett, E. (2013). Separate but equal? A comparison of participants and data gathered via Amazon's MTurk, social media, and face-to-face behavioral testing. *Computers in Human Behavior, 29*(6), 2156–2160.

Champagne-Lavau, M., Charest, A., Anselmo, K., Rodriguez, J. P., & Blouin, G. (2012). Theory of mind and context processing in schizophrenia: The role of cognitive flexibility. *Psychiatry Research, 200*(2-3), 184–192.

Chandler, J., & Shapiro, D. (2016). Conducting clinical research using crowdsourced convenience samples. *Annual Review of Clinical Psychology, 12*, 53–81.

Charlton, R. A., Barrick, T. R., Markus, H. S., & Morris, R. G. (2009). Theory of mind associations with other cognitive functions and brain imaging in normal aging. *Psychology and Aging, 24*(2), 338.

Chung, Y. S., Barch, D., & Strube, M. (2014). A meta-analysis of mentalizing impairments in adults with schizophrenia and autism spectrum disorder. *Schizophrenia Bulletin, 40*(3), 602–616.

d'Eon, G., Goh, J., Larson, K., & Law, E. (2019). Paying crowd workers for collaborative work. *Proceedings of the ACM on Human-Computer Interaction, 3*, 1–24.

Demichelis, O. P., Coundouris, S. P., Grainger, S. A., & Henry, J. D. (2020). Empathy and theory of mind in Alzheimer's disease: A meta-analysis. *Journal of the International Neuropsychological Society, 26*(10), 963–977.

Dziobek, I., Fleck, S., Kalbe, E., Rogers, K., Hassenstab, J., Brand, M., & Convit, A. (2006). Introducing MASC: A movie for the assessment of social cognition. *Journal of Autism and Developmental Disorders, 36*(5), 623–636.

Faul, F., Erdfelder, E., Lang, A., & Buchner, A. (2007). G*power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods, 39*(2), 175–191.

Fernandes, C., Barbosa, F., Martins, I. P., & Marques-Teixeira, J. (2021). Aging and social cognition: A comprehensive review of the literature. *Psychology & Neuroscience, 14*(1), 1.

Fischer, A. L., O'Rourke, N., & Loken Thornton, W. (2017). Age differences in cognitive and affective theory of mind: Concurrent contributions of neurocognitive performance, sex, and pulse pressure. *The Journals of Gerontology: Series B, 72*(1), 71–81.

Frith, C., & Frith, U. (2005). Theory of mind. *Current Biology, 15*(17), R644–R645.

Gönültaş, S., Selçuk, B., Slaughter, V., Hunter, J. A., & Ruffman, T. (2020). The capricious nature of theory of mind: Does mental state understanding depend on the characteristics of the target? *Child Development, 91*(2), e280–e298.

Goodman, J. K., & Paolacci, G. (2017). Crowdsourcing consumer research. *Journal of Consumer Research, 44*(1), 196–210.

Grainger, S. A., Steinvik, H. R., Henry, J. D., & Phillips, L. H. (2019). The role of social attention in older adults' ability to interpret naturalistic social scenes. *Quarterly Journal of Experimental Psychology, 72*(6), 1328–1343.

Hamilton, L. J., Gourley, A. N., & Krendl, A. C. (2022). They cannot, they will not, or we are asking the wrong questions: Re-examining age-related decline in social cognition. *Frontiers in Psychology, 13*.

Henrich, J., Heine, S. J., & Norenzayan, A. (2010). The weirdest people in the world? *Behavioral and Brain Sciences, 33*(2-3), 61–83.

Henry, J. D., Phillips, L. H., Ruffman, T., & Bailey, P. E. (2013). A meta-analytic review of age differences in theory of mind. *Psychology and Aging, 28*(3), 826.

Hossain, M., & Kauranen, I. (2015). Crowdsourcing: A comprehensive literature review. *Strategic Outsourcing: An International Journal*.

Johansson Nolaker, E., Murray, K., Happé, F., & Charlton, R. A. (2018). Cognitive and affective associations with an ecologically valid test of theory of mind across the lifespan. *Neuropsychology, 32*(6), 754.

Keith, M. G., Stevenor, B. A., & McAbee, S. T. (2022). Scale mean and variance differences in MTurk and non-MTurk samples: A meta-analysis. *Journal of Personnel Psychology*.

Klein, R. A., Ratliff, K., Vianello, M., Adams Jr, R. B., Bahník, S., & Bernstein, M. J. (2014). Investigating variation in replicability: A "many labs" replication project. *Open Science Framework*.

Kliemann, D., & Adolphs, R. (2018). The social neuroscience of mentalizing: Challenges and recommendations. *Current Opinion in Psychology, 24*, 1–6.

Krendl, A. C., Kennedy, D. P., Hugenberg, K., & Perry, B. L. (2022). Social cognitive abilities predict unique aspects of older adults' personal social networks. *The Journals of Gerontology: Series B, 77*(1), 18–28.

Krendl, A. C., Mannering, W., Jones, M. N., Hugenberg, K., & Kennedy, D. P. (in press). Determining whether older adults use similar strategies to young adults in theory of mind tasks. *Journal of Gerontology: Series B*.

Laillier, R., Viard, A., Caillaud, M., Duclos, H., Bejanin, A., de La Sayette, V., et al. (2019). Neurocognitive determinants of theory of mind across the adult lifespan, 103588. *Brain and Cognition, 136*.

Lutz, J. (2015). The validity of crowdsourcing data in studying anger and aggressive behavior. *Social Psychology*.

Miller, J. D., Crowe, M., Weiss, B., Maples-Keller, J. L., & Lynam, D. R. (2017). Using online, crowdsourcing platforms for data collection in personality disorder research: The example of Amazon's mechanical Turk. *Personality Disorders: Theory, Research, and Treatment, 8*(1), 26.

Necka, E. A., Cacioppo, S., Norman, G. J., & Cacioppo, J. T. (2016). Measuring the prevalence of problematic respondent behaviors among MTurk, campus, and community participants. *PLoS One, 11*(6), e0157732.

Newman, A., Bavik, Y. L., Mount, M., & Shao, B. (2021). Data collection via online platforms: Challenges and recommendations for future research. *Applied Psychology, 70*(3), 1380–1402.

Obschonka, M., Cai, Q., Chan, A. C., Marsalis, S., Basha, S. A., Lee, S. K., & Gewirtz, A. H. (2022). International psychological research addressing the early phase of the COVID-19 pandemic: A rapid scoping review and implications for global psychology. *International Journal of Psychology, 57*(1), 1–19.

Osborne-Crowley, K. (2020). Social cognition in the real world: Reconnecting the study of social cognition with social reality. *Review of General Psychology, 24*(2), 144–158.

Palan, S., & Schitter, C. (2018). Prolific. Ac—A subject pool for online experiments. *Journal of Behavioral and Experimental Finance, 17*, 22–27.

Peer, E., Brandimarte, L., Samat, S., & Acquisti, A. (2017). Beyond the Turk: Alternative platforms for crowdsourcing behavioral research. *Journal of Experimental Social Psychology, 70*, 153–163.

Peer, E., Rothschild, D., Gordon, A., Evernden, Z., & Damer, E. (2022). Data quality of platforms and panels for online behavioral research. *Behavior Research Methods, 54*(4), 1643–1662.

Peterson, C. C., Garnett, M., Kelly, A., & Attwood, T. (2009). Everyday social and conversation applications of theory-of-mind understanding by children with autism-spectrum disorders or typical development. *European Child & Adolescent Psychiatry, 18*(2), 105–115.

Pickering, D., & Blaszczynski, A. (2021). Paid online convenience samples in gambling studies: Questionable data quality. *nternational Gambling Studies, 21*(3), 516–536.

Quesque, F., & Rossetti, Y. (2020). What do theory-of-mind tasks actually measure? Theory and practice. *Perspectives on Psychological Science, 15*(2), 384–396.

Sami, H., Tei, S., Takahashi, H., & Fujino, J. (2023). Association of cognitive flexibility with neural activation during the theory of mind processing. *Behavioural Brain Research, 443*, 114332.

Sasaki, K., & Yamada, Y. (2019). Crowdsourcing visual perception experiments: A case of contrast threshold. *PeerJ, 7*, e8339.

Sassenberg, K., & Ditrich, L. (2019). Research in social psychology changed between 2011 and 2016: Larger sample sizes, more self-report measures, and more online studies. *Advances in Methods and Practices in Psychological Science, 2*(2), 107–114.

Saxe, R., & Kanwisher, N. (2003). People thinking about thinking people: The role of the temporo-parietal junction in "theory of mind". *Neuroimage, 19*(4), 1835–1842.

Schaafsma, S. M., Pfaff, D. W., Spunt, R. P., & Adolphs, R. (2015). Deconstructing and reconstructing theory of mind. *Trends in Cognitive Sciences, 19*(2), 65–72.

Scheeren, A. M., de Rosnay, M., Koot, H. M., & Begeer, S. (2013). Rethinking theory of mind in high-functioning autism spectrum disorder. *Journal of Child Psychology and Psychiatry, 54*(6), 628–635.

Stewart, N., Chandler, J., & Paolacci, G. (2017). Crowdsourcing samples in cognitive science. *Trends in Cognitive Sciences, 21*(10), 736–748.

Wang, Z., & Su, Y. (2013). Age-related differences in the performance of theory of mind in older adults: A dissociation of cognitive and affective components. *Psychology and Aging, 28*(1), 284.

**Open Practices Statement** The materials for the current study are fully available in Appendix A Table 6 & B of this manuscript. This study was not preregistered, but data are available at https://osf.io/t2wq8/?view_only=3c6bdaf810274a52874a1a138a387d8a.