


Using Attention Testing to Select Crowdsourced Workers and Research Participants

Social Science Computer Review
2021, Vol. 39(1) 84-104
© The Author(s) 2019
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/0894439319848726
journals.sagepub.com/home/ssc


Anja S. Göritz¹, Kathrin Borchert², and Matthias Hirth³

Abstract

Crowdsourcing offers fast and cost-effective access to human labor for business projects as well as to research participants for scientific projects. Due to the loose links between crowdsourcing employers and workers, quality control is even more important than in the off-line realm. We developed and validated the web-delivered attention test attentiveWeb in two versions (1) to come up with advance filters to identify workers who produce low-quality results and (2) to gauge the attention of workers who pass the advance filter. We apply attentiveWeb in three parallel user studies: one in the crowdsource Microworkers ($N = 539$), another one in Figure Eight ($N = 333$), and a third one in the online panel WiSoPanel ($N = 1,837$). The user studies confirm that it is useful to apply advance filtering to screen out poor workers. We propose an easily computed filter based on objective user behavior involving attentiveWeb. With regard to attention, despite the more severe advance filtering with Microworkers, their attention was lowest, followed by workers from Figure Eight, and it was highest in WiSoPanel. The platform differences in attention were not entirely explained by known differences—demographic and others—among the users of the three platforms. The attention test attentiveWeb has high Cronbach's α and split-half reliability. The first version of attentiveWeb predicted performance of the same crowdworkers in the second version of attentiveWeb 2 years later. We release attentiveWeb for assessing crowdworkers' attention into the research community and the wider public domain. The attention test attentiveWeb is open-source and can be used for free.

Keywords

crowdsourcing, online panel, quality, attention test, user study, attentiveWeb

The digitization of work has accelerated since the spreading of the Internet and, most recently, of smart mobile devices. These technological advances enable workers all over the world to collaborate with remote colleagues or to control systems in geographically distant places. Digitized work grants workers more flexibility with respect to where they live, their working hours, and the way they perform their job.

¹Freiburg University, Freiburg, Germany

²University of Würzburg, Würzburg, Germany

³Technical University of Ilmenau, Ilmenau, Germany

Corresponding Author:

Anja S. Göritz, Freiburg University, Engelbergerstr. 41, Freiburg D-79085, Germany.
Email: goeritz@psychologie.uni-freiburg.de

A far-reaching realization of digitized work is crowdsourcing. Crowdsourcing offers fast and cost-effective access to human labor for academic and business projects by distributing work via the Internet to a large and remote group of people (i.e., the crowd). Unlike traditional forms of work, with microtask crowdsourcing, there is no direct interaction between employer and worker; rather, both parties interact on a standardized platform. The granularity of work is finer than with traditional work, as many crowdsourced microtasks can be completed within minutes, and the payment per task usually ranges from a few cents to a few dollars.

Both the fine granularity of work and the anonymity render the relationship between employers and workers a loose one (Kuhn, 2016). Although most crowdsourcing microtasks are easy to fulfill, quality control is important to identify poor work by workers due to, for example, misunderstanding instructions, sloppiness, or faking. One approach to reducing poor work results is to design the microtask in a manner appropriate to crowdsourcing. Such task design might include a detailed description of the work steps or, if the task is tedious, restricting the number of items on which to work. An approach other than task design is to identify workers who are fit to accomplish a given microtask. For a translation task, for example, workers' language proficiency might be tested in advance. In addition to the task design and the worker's skill, a worker's diligence affects the quality of the work results.

There is concern that crowdsourcing employers approve work results by their workers more than they should, thereby inflating workers' reputation levels on the platform (Peer, Vosgerau, & Acquisti, 2014). Because of this inflation, reputation levels might become barely indicative of the quality of a worker's work results. Thus, an ad hoc assessment process on a per-task basis of "good" workers might be needed to avoid poor work results. Among other characteristics, a good worker is attentive, that is, she is able and willing to mentally focus on the microtask at hand.

Similar requirements have already been present in traditional work and research settings, which has given rise to several psychological tests to assess the attention of workers or research participants. However, most of the current tests to assess attention are not suited to crowdsourcing, as they are paper-and-pencil based, require controllable surroundings, or rely on complex instructions. To address this gap in existing attention tests, our work focuses on measuring a person's attention via the Internet. Inspired by the method of the established paper-and-pencil attention test d2-R, we created attentiveWeb, which is suited to be administered in a crowdsourcing setting. The attention test attentiveWeb is realized as a customizable web application that can be used for assessments of globally distributed users, also outside of crowdsourcing. It is open source, freely available and free to use (<https://github.com/linfo3/attentiveWeb/>).

To illustrate applications of this online attention test, as well as to shed light on what one can expect regarding the attention of crowdsourcing workers, we evaluate the attention of workers from two crowdsourcing platforms and from one online panel. By comparing the three platforms in terms of their workers' attention, we gain insights into the performance of people who stem from different platforms to do microtasks. Furthermore, we validate attentiveWeb by determining its predictive power for crowdworkers to perform in another version of attentiveWeb more than 2 years after taking the first version of attentiveWeb.

Crowdsourcing and Quality Control in Crowdsourcing

Crowdsourcing has become a valuable environment for work and research domains requiring a multitude of views, votes, judgments, or other human input. Hoßfeld, Hirth, and Tran-Gia (2011) describe crowdsourcing as an extension of the outsourcing approach that is characterized by yet finer granularity of the work and by decoupling the individuals who devise work from those who complete it. In traditional outsourcing, companies use subcontractors to outsource parts of the production process for efficiency. Microtask crowdsourcing carries this idea further and distributes tiny shares

of work to external workers. These microtasks are often repetitive but still cannot be completed using algorithms.

While crowdsourcing platforms offer several advantages for the employer or researcher, such as access to a large pool of diverse workers or participants (Casler, Bickel, & Hackett, 2013), due to the unsupervised and uncontrolled environment several challenges must be met (Buhrmester, Talaifar, & Gosling, 2018). One of these challenges is ensuring the quality of workers' results. Several approaches exist that try to tackle the issue of quality control (Daniel, Kucherbaev, Cappiello, Benatallah, & Allahbakhsh, 2018). A widely used technique is prior filtering of the workers based on gold standards (Oleson et al., 2011), other qualification tests or workers' reputation on the platform. Gold standards are questions or tests for which the optimal outcome is known. In a gold standard transcription task, for example, an expert has already transcribed a hand-written text. The worker's results are compared to the gold standard, and a worker is considered qualified if her results match the gold standard. While these questions or tests by their nature are not necessarily task-specific, in empirical reality, they often are. If gold standard qualification tests are to be task-specific, they have the drawback that the gold standard data must be generated for each task individually. Furthermore, a gold standard test is not suitable for all types of tasks, especially tasks with no prior known outcome, such as surveys (Gadiraju, Kawase, Dietze, & Demartini, 2015). Another downside is that workers can circumvent certain gold standard tests (Checco, Bates, & Demartini, 2018). Likewise, workers' reputation scores on the crowdsourcing platform are sometimes calculated based on having administered gold standard tests. Thus, the drawbacks of gold standard tests, at least in part, spill over onto the reputation score. One approach to avoiding the need for testing workers prior to each microtask type and to avoiding having to generate task-specific gold standards is assessing more general and context-independent skills or abilities such as intelligence or attention.

So far, little research has been done on assessing the attention of crowdsourced workers and/or research participants. One of the reasons for the scarcity of this research is that existing attention tests are not usable in a crowdsourcing environment. Initial studies on the attentiveness of crowdworkers were performed by Hauser and Schwarz (2016) and Goodman, Cryder, and Cheema (2013). By using instructional manipulation checks, they found that crowdsourced and noncrowdsourced participants did not differ in attention when following instructions. Nevertheless, even if workers read the instructions carefully and the microtasks are easy, after a while, these tasks tend to become tedious due to their repetitive nature, resulting in low-quality work (Gadiraju, Siehndel, Fetahu, & Kawase, 2015). Rothwell, Carter, Elshenawy, and Braga (2015) showed that workers who were filtered in advance, for example, by reputation or gold standard tests, and workers who were not filtered in advance, do not clearly differ in attentiveness. Due to inattentiveness when reading instructions, workers in both groups equally failed the attention checks. Additionally, Peer, Vosgerau, and Acquisti (2014) highlight that attention checks may affect the outcome of tasks negatively.

To sum up, while several approaches exist to screen out poor workers or research participants such as using their reputation on the platform, instructional manipulation checks, task-specific gold standards, and consistency checks by repeating slightly rephrased questions or by identifying incompatibilities (e.g., a worker who claims to be 18 years old and later to hold a supervisor position in her job), these approaches have individual drawbacks, for example, some are fakeable, some suffer from low reliability and validity, some are task-specific and thus need to be devised on a per-task basis, some cannot be applied prior to the actual crowdsourcing microtask, and some are limited to a survey-like task. In this work, we investigate the applicability of a psychometric attention test as a selection mechanism. The newly developed test is applied prior to the actual microtask, it is fairly context independent and hence works for many ensuing crowdworking tasks, it cannot be faked up, it is public domain, it is free of charge, and it is not limited to tasks that require participation in

academic surveys and experiments but may be used as a selection test for crowdworkers hired for commercial purposes to do repetitive, low-skill microtasks (labeling pictures, etc.).

Attention Testing

Attention is a basic cognitive function. It is defined as the sustained focus of cognitive resources on relevant stimuli while ignoring irrelevant stimuli. Although psychology distinguishes different forms of attention, this work is concerned with sustained selective attention, also called concentration. Sustained selective attention is the ability to maintain a consistent behavioral response during a continuous and repetitive activity. During this activity, a person tries to detect the appearance of a target stimulus while suppressing a response to nontarget stimuli. Crowdsourcing microtasks usually require concentration on the part of workers: These tasks are often simple but repetitive. That is why attention tests lend themselves to the context of microtask crowdsourcing.

There are many attention tests such as KLT-R (Düker & Lienert, 2001), ZRF_20 (Jacobs, 2013, 2014), CAPT (Starzacher, Nubel, & Grohmann, 2006), IMT/DMT (Dougherty, Marsh, & Mathias, 2002), GU (Jacobs, 2013, 2014, 2015), KONT-P (Satow, 2011), and d2 (Brickenkamp, 1981). These tests differ in the complexity of their setup as well as in the degree to which they are established and validated. Being attention tests, however, they share the characteristic that the basic task (e.g., identifying target stimuli) is easy. Their independence of IQ makes the tests applicable to a large segment of the population. A low attention test score indicates a low degree of attentiveness, either resulting from a low (state or trait) ability to mentally focus or low motivation to take the test (e.g., feeling that one has better things to do than taking the test) or low ability to take this test (e.g., not understanding instructions, motor impairments). Our idea is to make use of an attention test in a crowdsourcing setting to distinguish crowdsourcing workers who are attentive from those who are inattentive. For the practical purpose of identifying good versus bad workers on a crowdsourcing platform, it is not necessary to sound out why an individual worker scored low on the attention test (e.g., whether low attentiveness resulted from motivation and/or ability).

There are requirements of any attention test that is to be used in crowdsourcing or to be used in an online setting more generally: Workers are paid per time, so the test must be short, including its warm-up phase. Thus, the test should take a few minutes at most, as screening crowdsourcing workers with the help of the test is only a preparation for the ensuing crowdsourcing microtask. Moreover, due to the global distribution of crowdsourcing workers (Martin et al., 2017), the instructions must be provided online without personal interaction and must be understandable to users of various languages and diverse cultural and educational backgrounds. Consequently, the test should not require a superb command of the language the test is in, instructions should be easy to follow, and the test itself should tap pure attention without requiring further skills on the part of the testtaker (e.g., calculus). Furthermore, the test should not require the testtaker to install any software (e.g., browser plug-in) and should not require more than average hardware. The latter two requirements imply that text-based attention tests are particularly suitable. Lastly, the test should feature a large and broad norm sample: When testing only one sample from any given platform, the sample's attention score must be related to a suitable norm sample to get a sense of the level of attentiveness on this platform.

When perusing available offline tests (i.e., KLT-R, ZRF_20, CAPT, IMT/DMT, GU, KONT-P, and d2-R) to see whether they meet the requirements of an online attention test, most tests do not fit the purpose of assessing attention in an online environment. For example, KLT-R takes too long, ZRF_20 has a narrow norm sample, CAPT has a narrow norm sample and is not text based, IMT/DMT takes too long and captures impulsivity in particular, GU suffers from strong practice effects and takes too long, and KONT-P requires computational ability and is somewhat too long.

The d2-R paper-and-pencil test comes close to fulfilling the requirements. It is a validated and well-established test of sustained selective attention. It is a paper-and-pencil cancellation task in

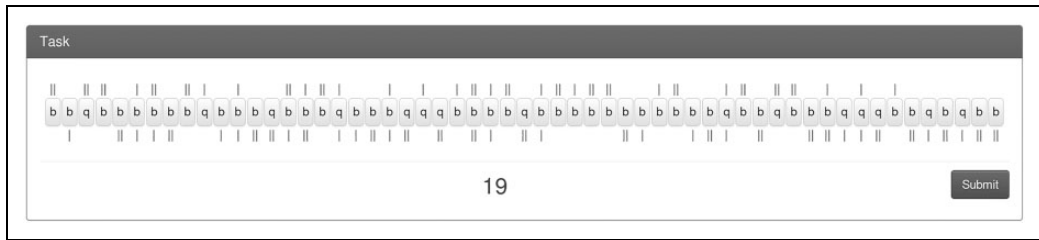


Figure 1. Sample page of the attentiveWeb attention test.

which participants cross out, in any order, any letter *d* surrounded by two marks. The distractors are similar to the target stimulus; for example, a *p* with two marks or a *d* with one or three marks. The original d2 test is by Brickenkamp (1981). While it is not language free, the test does not require a superior command of the test language, does not require elaborate cognitive skills, is text based, is short, and a norm sample is available. Our work was informed by the method of the revised version d2-R (Brickenkamp, Schmidt-Atzert, & Liepmann, 2010). A computerized version of the d2 test (called d2c) is described in Jacobs (2014). However, this version is not web-enabled and not publicly available (for details, see Jacobs, 2014, pp. 13–16).

Running an attention test online implies that it is administered in an unsupervised manner. Due to the absence of an instructor, the test should include detailed yet clear instructions. In our developed web-based attention test, we ensure that the testtaker does not need to install any software or needs more than average hardware. Furthermore, testtakers may vary in the device on which they take the test. Differences in screen size or input mode may influence testtakers' performance. Moreover, the particular surroundings and the situation in which testtakers complete the test may lead to distraction. In addition, hidden influences such as workers' disabilities might affect test results. In addition, crowdsourcing-specific challenges must be considered: The rewards earned through completing the task may encourage tricking the system or rushing.

Method

The benchmark d2-R paper-and-pencil test requires the testtaker to find and cross out any letter *d* with two marks surrounding it in a field of 14 rows, each row with 57 letters. The time for processing a row is 20 s. Informed by the method of d2-R, we developed attentiveWeb; however its realization differs from d2-R. In attentiveWeb, the letters are realized as buttons labeled with the targets and distractors. The buttons are arranged in rows separated by small spaces between the buttons. By clicking the buttons, the letters are marked as being crossed out. A button click cannot be reversed. The size of the display area is fixed, so the row of buttons is not wrapped on small displays. In addition, the size of each worker's display is assessed. If the display happens to be too small, the participant is prompted to use a larger display to accommodate an entire row of buttons.

We created two versions of attentiveWeb as follows: In Version 1, the target letter is *d* and the distractor letter is *p*, whereas in Version 2, the target letter is *b* and the distractor letter is *q*.

In contrast to d2-R, not all rows are visible at once, but one row per screen is shown at a time. This allows enforcing a maximum working time of 20 s per row. After 20 s, a popup notifies the testtaker, and input is disabled. If the testtaker needs less than 20 s to process a row, a submit button enables manual submission. After a countdown of 3 s, the next row is displayed automatically. In total, the testtaker completes 14 pages with one row each. As feedback on the progress, the number of processed and number of total rows are shown at the bottom of each page. Figure 1 depicts a sample page. Instructions are presented at the beginning of the test. The instructions are given in simple English.

The client side of attentiveWeb is an interactive web application based on common web technologies, including HTML ~ 5 and JavaScript. The data processing and storing of the results are realized by a server-side component using the PHP Flow Framework (<https://flow.neos.io/>). For our implementation, we used a MySQL server. Both the client- and server-side code is publicly available (<https://github.com/lsinfo3/attentiveWeb/>) under GPL-3.0 license.

The customary metrics for evaluating a testtaker's performance in the benchmark d2-R are the concentration performance (CP) and the error percentage $E\%$. These metrics rest on the number of clicked target items TN and the number of errors E . The errors are either omission errors $E1$ or confusion errors $E2$. The CP considers the number of processed target items TN as well as the errors E ; thus, it represents quantity of correct performance given the available time. As the total number of errors ($E1 + E2$) can be larger than the number of clicked target items TN, CP may be negative.

$$CP = TN - (E1 + E2).$$

The error percentage $E\%$ is the errors ($E1 + E2$) divided by the number of clicked target items; thus, it represents inaccuracy.

$$E\% = \frac{E1 + E2}{TN} \times 100$$

As the total number of errors ($E1 + E2$) can be larger than the number of clicked target items, $E\%$ ranges from 0% to more than 100%. For the calculation of both performance metrics in attentiveWeb, just as with the d2-R (Brickenkamp et al., 2010), the first and the last row are omitted. The metrics in d2-R require that the testtaker has worked on the test, which means that at least one target item should have been processed per row. Additionally, the computation of the d2-R metrics assumes that the testtaker processes the items from left to right. Due to the unsupervised administration of the test, the evaluation metrics in attentiveWeb were made more robust. As there is no guarantee that each participant works on each row beginning on the left, first, we determine the start point in each row. Further, we adapt CP and $E\%$ to tolerate incomplete page submissions as follows: In each row, we distinguish if the testtaker has processed no item at all or if the testtaker has clicked at least one button. If no button has been clicked, TN is calculated by the total number of target items shown in the row; hence, the number of omission errors equals TN. If no item has been processed, the number of confusion errors is zero. Then, CP is zero, and $E\%$ is 100%. If at least 1 item has been processed, the metrics must be adapted only if no target item has been clicked. If the testtaker failed to click any target item, TN is the number of target items from the first button to the last-clicked nontarget item, and the number of omission errors is the number of unprocessed target items until the last-clicked nontarget item. An adaptation of $E2$ is not necessary. When using the adapted values of TN and $E1$, CP equals $E2$, and $E\%$ is larger than 100%.

User Studies

We conducted three parallel user studies to (1) illustrate the applicability of the developed attentiveWeb, (2) determine its reliability, and (3) evaluate the attention of workers in two crowdsourcing platforms and in one online panel. We complemented attentiveWeb (Version 1 with target letter d) with two questionnaires—one given prior to and one given after the attention test—to capture potential influences on the workers' attention. In the first questionnaire, the workers indicated their gender, age, country of residence, and where they were completing the test. Moreover, we collected self-reports of participants' state of attentiveness and mouse skills, as both might correlate with the attention test results. In the second questionnaire, the workers indicated their state of attentiveness again and answered some slightly rephrased questions from the first questionnaire. Repeating questions captured the participants' degree of diligence, such as whether they had been randomly

clicking. For example, prior to attentiveWeb, we asked participants for their country of residence, and after attentiveWeb, we inquired about their continent of residence. The full set of questions is in the Appendix.

The user studies were fielded from May 30 to June 3, 2016. We recruited participants from three platforms: (1) Microworkers, a large and internationally available crowdsourcing platform; (2) Figure Eight, also a large and internationally available crowdsourcing platform that was formerly known as Crowdfunder; and (3) WiSoPanel, an online panel that holds Germans from all walks of life who have agreed to take part in noncommercial web-based studies (Görizt, 2014). On the two crowdsourcing platforms, workers' were solicited by paid tasks. Starting with 30 tasks, the number of tasks increased by 30 positions every 12 hr on Days 1 and 2 of the study. As workers had completed all tasks within hours after submitting them, we changed the interval to 20 tasks every 6 hr to obtain a more diverse group of testtakers. Thus, an almost constant stream of participants trickled in during the field period. The task was announced as taking 10–15 min with a reward of US\$0.20. This corresponds to a typical campaign on crowdsourcing platforms (Hirth, 2016). In total, 340 tasks were placed on each platform. Due to technical issues at the beginning of the study, more workers started the test than positions were available on Microworkers. Thus, 539 Microworkers visited the test, of which 420 submitted their answers to the final questionnaire. On Figure Eight, 333 workers started the test, of which 308 submitted their answers to the final questionnaire. To administer the reward in the two crowdsourcing platforms, on the final page workers were shown a personal payment code. Workers get paid by entering this code on the crowdsourcing platform; by not entering this code, they have the freedom not to collect the reward. In WiSoPanel, too, the task was announced to take 10–15 min, but with a reward of EUR 0.50. The participants from the WiSoPanel are used to a higher payment than is customary in crowdsourcing platforms (i.e., about EUR 3–4 per hour); hence, compared to the two crowdsourcing platforms, we favored a customary but unequal reward over an equal but unusually low reward. To mimic the wave-like manner in which the crowdsourcing workers were solicited, WiSoPanel users were invited via e-mail in four waves starting on the morning of May 30. Regardless of when particular WiSoPanel users were sent their invitation, the test was open until June 3. We invited 12,237 WiSoPanel users, of which 1,837 called up the first page of the test. Of those, 1,352 submitted their answers to the final questionnaire. To administer the reward in WiSoPanel, on the final page users were shown a checked box saying that their WiSoPanel account would be credited with EUR 0.50. To give up the reward, participants had the freedom to uncheck the box.

Validation Studies

More than 2 years after the user studies, we conducted two parallel validation studies to establish the predictive validity of attentiveWeb in one crowdsourcing platform and in the online panel. We used Version 2 of attentiveWeb, wherein all *d*'s are replaced by *b*'s and all *p*'s by *q*'s. Everything else was kept the same as in the user studies except for doing away with a few items in the two surrounding questionnaires whose answers were unlikely to have changed in the meantime such as gender and age (see Appendix). The validation studies were fielded September 4–16, 2018. We sought to solicit those participants on Microworkers and on WiSoPanel who had taken attentiveWeb (Version 1) as part of the user studies more than 2 years earlier. It was impossible to resolicit the workers from Figure Eight because we had no longer access to our account after the platform had changed its cost scheme.

Results

First, this section reports on the impact of advance filters used to detect workers with extremely low-quality work results. Second, we report on the demographics of the workers from the three platforms,

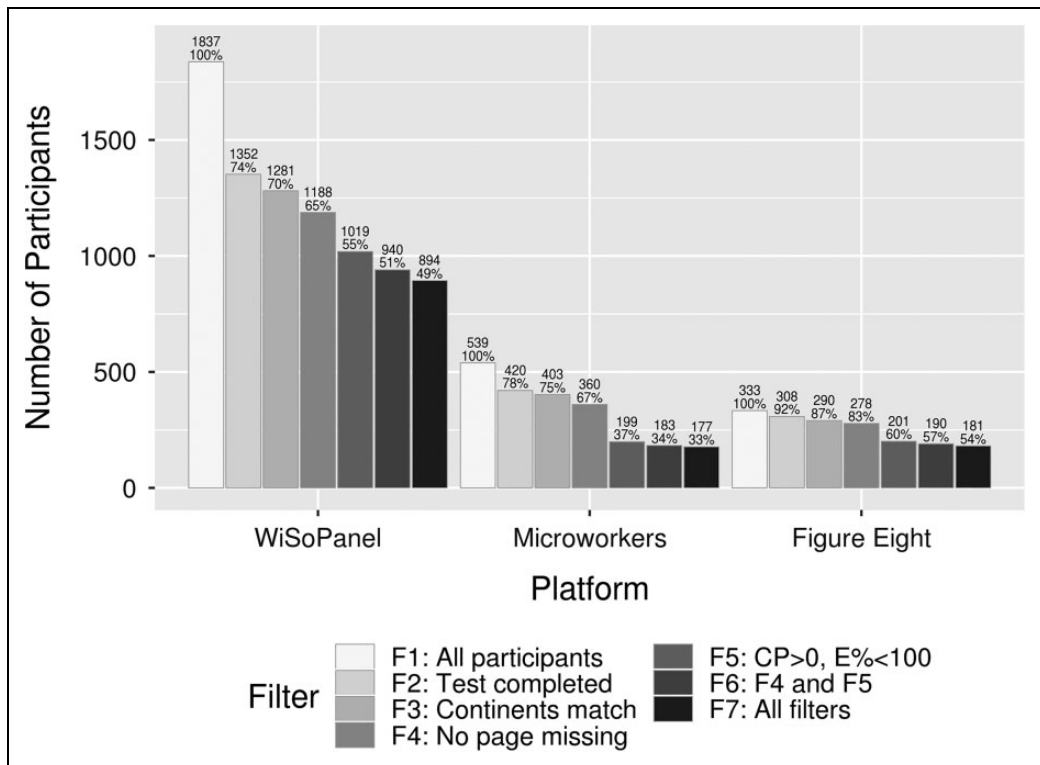


Figure 2. Advance filters by platform.

as their demographics might influence the work results. Third, we report the results of assessing workers' attention and test potential influences on the attention test results. Fourth, we report the internal consistency and the split-half reliability of attentiveWeb. Finally, we report results of checking the predictive validity of attentiveWeb.

Advance Filters

To identify low-quality testtaking of different kinds and degrees, we consider seven advance filters *F1*–*F7* (see Figure 2).

F1: All participants involves no filtering and thus correspond to the total number of unique participants, regardless of whether they completed or aborted the task. *F2: Test completed* leaves participants who submitted the final questionnaire. The following filters are applied on top of *F2*: The filter *F3: Continents match* leaves participants who submitted a correct tuple of country of residence in the initial questionnaire and continent of residence in the final questionnaire. *F4: No page missing* leaves participants who clicked at least one button per row. *F5: CP > 0 and E% < 100* leaves participants who clicked more target items than nontarget items in the attention test. *F6: F4 and F5* leave participants who clicked at least one button per row and clicked more target than nontarget items. *F7: All filters* leaves participants who passed all previously mentioned filters.

Comparing the filters shows that filtering by *F2: Test completed* is useful (i.e., across all three platforms, 77% of workers remain) but not thorough enough, as stricter filters identify many more poor workers. Filters that consider answers given by the participants such as *F3* or missing pages such as *F4* result in only slightly stricter filtering than *F2: Test completed*. Moreover, the filters that

Table 1. Country of Residence by Platform.

Country of Residence	WiSoPanel	Microworkers	Figure Eight
Austria	10 (1%)	—	—
Bangladesh	—	63 (34%)	3 (2%)
Brazil	—	2 (1%)	8 (4%)
Croatia	—	5 (3%)	2 (1%)
Germany	883 (94%)	2 (1%)	3 (2%)
India	—	23 (13%)	10 (5%)
Indonesia	—	6 (3%)	7 (4%)
Italy	1 (<1%)	4 (2%)	7 (4%)
Macedonia	—	4 (2%)	4 (2%)
Nepal	—	7 (4%)	—
Portugal	—	2 (1%)	8 (4%)
Romania	—	2 (1%)	7 (4%)
Russian Federation	—	2 (1%)	7 (4%)
Serbia	—	16 (9%)	24 (13%)
Spain	—	—	8 (4%)
Sri Lanka	—	5 (3%)	—
Switzerland	12 (1%)	—	—
United States	1 (<1%)	4 (2%)	6 (3%)
Venezuela	—	1 (<1%)	18 (9%)
Other	11 (1%)	35 (19%)	68 (36%)
Not specified	22 (2%)	—	—

consider answers given by the participant such as $F3$ are error prone, as any pre–post-match of answers might result from chance. The filter $F6$, which combines filter $F5$: $CP > 0$ and $E\% < 100$ and the behavioral filter $F4$: *No page missing*, requires little computational effort, thereby being only slightly more lenient than $F7$: *All filters*. Moreover, the results of applying $F6$ are error free, as they rely on objective user input and are not subject to chance. Thus, we use advance filter $F6$ in any further analyses of the results. When comparing the three platforms in terms of filtering with $F6$, the smallest share of participants is filtered out in Figure Eight (57% of workers are left), the second smallest share is filtered out in WiSoPanel (51% are left), and the highest share is filtered out in Microworkers (34% are left). A χ^2 test establishes that the extent of filtering when applying $F6$ varies by platform ($p < .001$).

Demographics

Table 1 shows users' most frequent countries of residence broken down by platform. WiSoPanel holds primarily Germans (94%). The two crowdsourcing platforms are more heterogeneous with regard to workers' country of residence. The three most frequent countries of Microworkers are Bangladesh, India, and Serbia. In Figure Eight, Serbia is most represented, followed by Venezuela and India. Comparing the two crowdsourcing platforms, Microworkers originate mainly in Asia, whereas the workers from Figure Eight are more international.

With regard to age, as far as specified by each worker (Figure 3), in WiSoPanel, age varies more widely, and the median age-group (i.e., 41–50) is higher than in the two crowdsourcing platforms (Microworkers: 20–30 and Figure Eight: 31–40). A Kruskal–Wallis test confirmed that the three samples do not originate from the same age distribution ($p < .001$).

With regard to gender, a slight majority of users recruited from WiSoPanel are women (55%). In contrast, most workers on the two crowdsourcing platforms are men (Microworkers: 77% and Figure

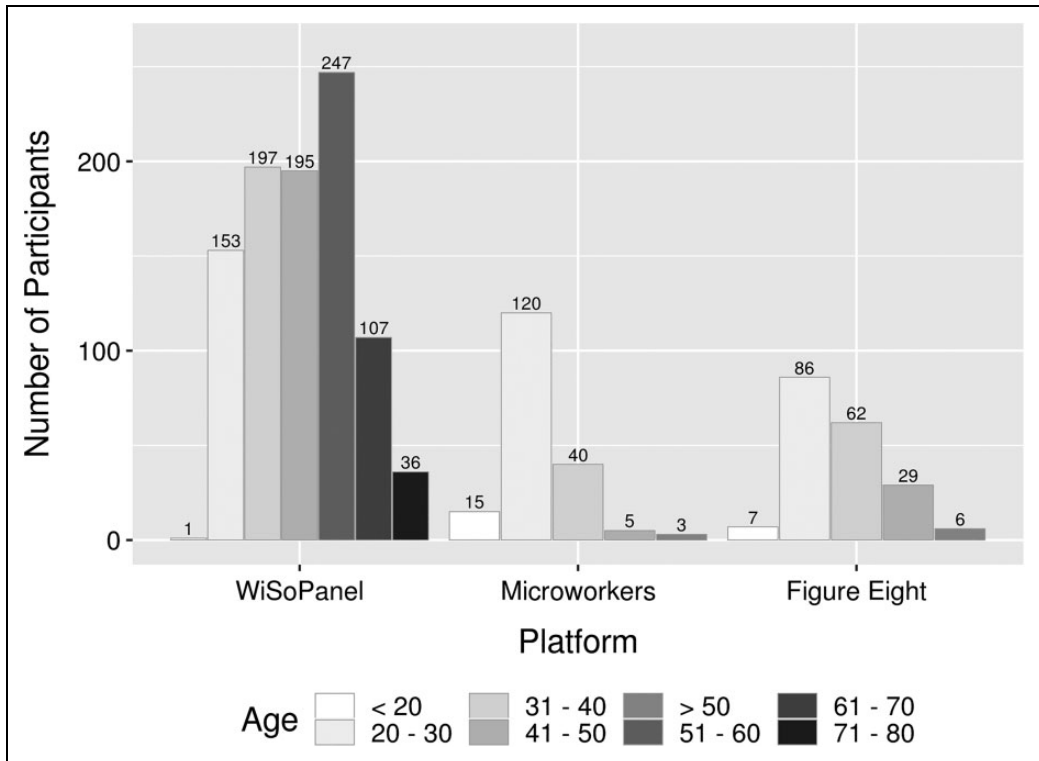


Figure 3. Age-group by platform.

Table 2. Language Skills by Platform.

Language Skills	WiSoPanel	Microworkers	Figure Eight
Beginner	1 (<1%)	42 (23%)	36 (19%)
Advanced	30 (3%)	117 (64%)	135 (71%)
Native speaker	900 (96%)	24 (13%)	18 (10%)
Not specified	9 (<1%)	–	1 (<1%)

Eight: 69%). The inequality of the gender distribution across the three platforms is significant, $\chi^2 = 91.35, df = 2, p < .001$.

As it was known that most of the WiSoPanel users live in German-speaking countries (Göritz, 2014), the attention test was presented in German. Due to the international audience, the working language of most crowdsourcing platforms is English. That is why in the two crowdsourcing platforms, the attention test was in English. Given the reported countries of residence of the crowdsourcing workers, this implies that the attention test was not in most workers’ mother tongue. The language barrier may lead to misunderstandings, and hence, the results obtained from the two crowdsourcing platforms may be affected by workers having worked in a foreign language (Goodman, Cryder, & Cheema, 2013). Table 2 shows the participants’ self-rated skills in the language of the test. As expected, most of the WiSoPanel users are German native speakers; thus, they took the test in their mother tongue. On the crowdsourcing platforms, around one fifth of the workers self-identified as having poor English skills, and about two thirds reported having good English skills.

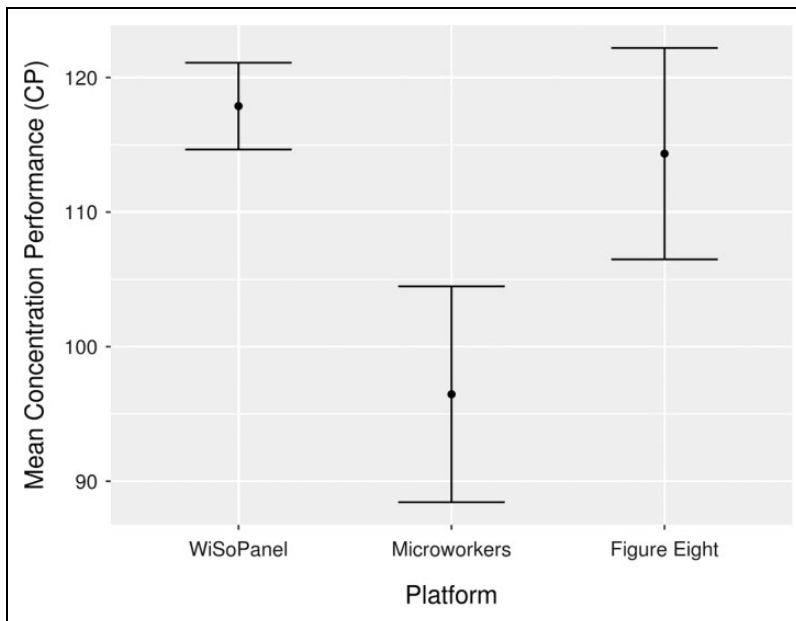


Figure 4. Concentration performance by platform with 95% confidence interval.

The language skills did not differ between the two crowdsourcing platforms, Spearman's $\rho = .007$, $n = 372$, $p = .90$.

Attention Test

Figure 4 gives the CP by platform. Users of the WiSoPanel obtained a mean CP of 117.9, those from Figure Eight 114.3 and those from Microworkers 96.5. CP differs significantly among the three platforms, $Welch(2,320.33) = 11.92$, $p < .001$. Post hoc tests revealed that Microworkers had a significantly lower average CP than WiSoPanel workers (Tamhane = 21.42, $p < .001$) and Figure Eight workers (Tamhane = 17.88, $p = .003$), while WiSoPanel and Figure Eight did not differ ($p = .80$). With regard to the variance in CP, workers from WiSoPanel and Figure Eight perform more uniformly. There is inhomogeneity of the variance of CP across platforms, $Levene(2,1310) = 4.32$, $p = .013$, with the source of the difference being between WiSoPanel and Microworkers, $Levene(1,1121) = 7.41$, $p = .007$.

Figure 5 shows the mean error percentage $E\%$ broken down by platform. The users of WiSoPanel worked most accurately with an average $E\%$ of 24.6%. The $E\%$ with Microworkers is 39.4% and with Figure Eight 33.9%. Among the three platforms, $E\%$ differs significantly, $Welch(2,311.88) = 24.93$, $p < .001$. Post hoc tests revealed that workers in WiSoPanel had a significantly lower $E\%$ than those in Microworkers (Tamhane = 14.79, $p < .001$) and in Figure Eight (Tamhane = 9.83, $p < .001$), while the crowdsourcing platforms do not significantly differ ($p = .21$).

Thus, on the one hand, there are differences in the demographics of the workers of the three platforms, while on the other hand, there are differences in the attention test results among the platforms. To bring the differences in the demographics and in the attention test results together, we conducted stepwise regression analyses to test, in a first step, whether any known differences among the workers of the three platforms are relevant to their attention test performance using backward elimination of nonsignificant predictors and, in a second step, whether accounting for attention-relevant differences among the workers of the three platforms sufficiently explains their attention

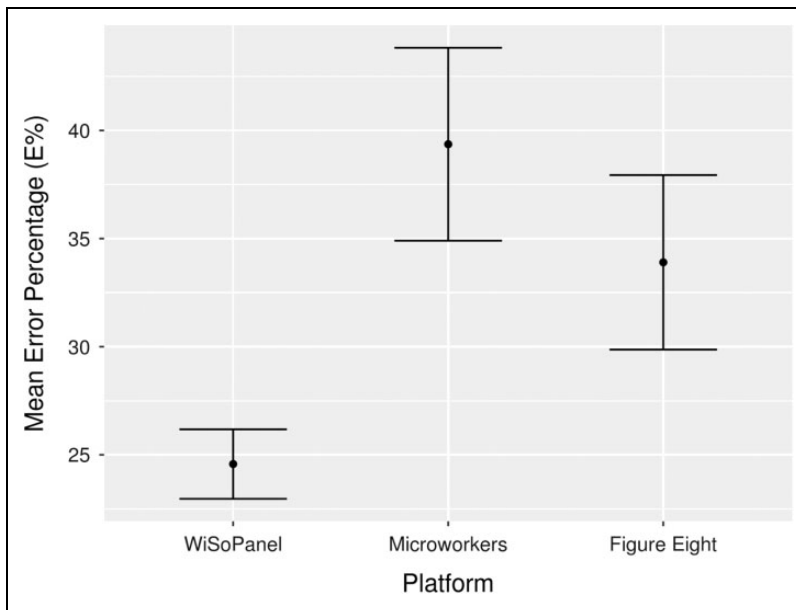


Figure 5. Error percentage by platform with 95% confidence interval.

test results or whether despite taking those differences into account there remain inherent platform-related differences with regard to attention. In Step 1 of each of the two regression analyses, all known predictors were evaluated, namely, gender (women vs. men), age (up to 40 years vs. over 40 years), country (Western European culture [Australia, Austria, Belgium, Canada, France, Germany, Greece, Ireland, Italy, Malta, the Netherlands, Portugal, Spain, Sweden, Switzerland, the United Kingdom, and the United States] vs. non-Western-European culture [all others]), language (native language vs. foreign language), current place (at home vs. elsewhere), people around (alone vs. not alone), self-rated mental focus before attention test (five levels), self-rated mental focus after attention test (five levels), self-rated mouse skills (five levels), and d2 done before (no vs. yes).

The CP was higher among younger testtakers ($\beta = -.24$), among people who live in a Western European culture ($\beta = .26$), among those who participated not from home ($\beta = .06$) and with high self-rated mouse skills ($\beta = .08$). Leaving these four significant predictors in the model and testing whether dummy-coded platforms explain additional variance in attention reveals that Figure Eight achieves a higher CP than the other two platforms together ($\beta = .12$), while WiSoPanel by tendency is higher than the other two platforms together, although this tendency fails a conventional level of significance ($\beta = .09$). The final model, which includes the four predictors from Step 1 and the two dummy-coded platform variables, fits well, $F(6, 1,287) = 20.49, p < .001$, adjusted $r^2 = .083$.

The error percentage $E\%$ was higher among older testtakers ($\beta = .15$) and among those who live in a non-Western-European culture ($\beta = -.29$). Leaving these two significant predictors in the model and testing whether dummy-coded platforms explain additional variance reveals that WiSoPanel has a lower $E\%$ than the other two platforms together ($\beta = -.14$), while Figure Eight is not lower than the other two platforms together ($\beta = .06$). The final model, which includes the two predictors from Step 1 and the two dummy-coded platform variables, fits well, $F(4, 1,287) = 23.94, p < .001$, adjusted $r^2 = .067$.

To determine attentiveWeb's internal consistency, for a faithful comparison with the benchmark d2-R, we calculated Cronbach's α such that each quarter of attentiveWeb was treated as an item. In WiSoPanel ($n = 940$), attentiveWeb's internal consistency was .91 with CP and .87 with $E\%$. In

Microworkers ($n = 183$), Cronbach's α was .89 with CP and .85 with $E\%$. In Figure Eight ($n = 190$), Cronbach's α was .90 with CP and .87 with $E\%$. The internal consistency with CP did not differ among the three platforms, $\chi^2 = 1.75$, $df = 2$, $p = .42$, nor did the internal consistency differ among the platforms with $E\%$, $\chi^2 = 1.78$, $df = 2$, $p = .41$ (Diedenhofen & Musch, 2016). Thus, across all platforms together ($N = 1,313$), Cronbach's α was .91 with CP and .87 with $E\%$.

As regard split-half reliability, in WiSoPanel ($n = 940$), it was .90 with CP and .85 with $E\%$. In Microworkers ($n = 183$), split-half reliability was .85 with CP and .85 with $E\%$. In Figure Eight ($n = 190$), split-half reliability was .87 with CP and .85 with $E\%$. The split-half reliability with CP did not differ among the three platforms, $\chi^2 = 2.68$, $df = 2$, $p = .26$, nor did the split-half reliability differ among the platforms with $E\%$, $\chi^2 = 0.01$, $df = 2$, $p = .99$. Thus, across all platforms together ($N = 1,313$), split-half reliability was .89 with CP and .86 with $E\%$.

Validation of Attention Test

We sought to solicit those participants from Microworkers ($n = 183$) and from WiSoPanel ($n = 940$) who had passed advance filter $F6$ (see Figure 2) more than 2 years earlier. Of the 183 eligible Microworkers, 33 (18.0%) visited the validation study, of which 25 (75.8%) submitted the final questionnaire, and of which 21 (84.0%) passed advance filter $F6$ in the validation study. Of the 940 eligible people from WiSoPanel, 909 (96.7%) had a valid e-mail address and were thus invited to the validation study, of which 635 (69.9%) called up the first page of the validation study, of which 540 (85.0%) submitted the final questionnaire, and of which 453 (83.9%) passed advance filter $F6$ in the validation study.

In WiSoPanel ($n = 453$), attentiveWeb's predictive validity was $r = .58$ with regard to CP and $r = .40$ with regard to $E\%$. In Microworkers ($n = 21$), attentiveWeb's validity was $r = .48$ with CP and $r = .38$ with $E\%$. The validity with CP did not significantly differ between the two platforms, $z = .58$, nor did the validity with $E\%$, $z = .09$. In the two platforms together, attentiveWeb (Version 1) predicted crowdworkers' performance ($N = 474$) more than 2 years later in attentiveWeb (Version 2) at $r = .57$ with CP and at $r = .40$ with $E\%$.

Discussion

Crowdsourcing has gained tremendous importance for soliciting human work. Given the anonymous and short-lived work contract between crowdsourcing employers and crowdsourcing workers, the quality of the work results may vary greatly. Hence, ensuring quality is a major concern with crowdsourcing, and—if successful—confers notable economic benefit. A cost-efficient approach to ensuring quality is to assess general properties of the crowdsourcing workers such as their attention and to select workers for crowdsourced microtasks depending on the outcome of this assessment. This approach amounts to a personnel selection, which has a long tradition in classic work arrangements. To assess workers' attention, we developed and applied the web-based attention test attentiveWeb as inspired by the method of the well-known attention test d2-R. The feasibility of attentiveWeb was demonstrated on two crowdsourcing platforms, Microworkers and Figure Eight, and on an online panel geared to academic data collection, the WiSoPanel. The web-based attention test attentiveWeb can be downloaded from the web and used at no cost.

Three user studies applying attentiveWeb were run in parallel on these three platforms. The studies confirm that it is useful to apply advance filtering to screen out poor workers. In the studies at hand, a filter worked well that combined skipping of at least one page with values of the attention test metrics that indicate extremely poor performance. This filter was easy to compute and did not rely on workers' self-reports but on their objective behavior. Comparing the three platforms in applying this filter shows that Microworkers was more in need of advance filtering than were Figure Eight and WiSoPanel.

When comparing the performance in attentiveWeb among workers who had passed this advance filter, there were differences in the quantity and quality of the work results. Despite the more severe advance filtering, Microworkers' attention was lowest, followed by workers from Figure Eight, and it was highest in WiSoPanel. This pattern varied slightly depending on which of the two metrics of attentiveWeb was inferentially tested: While CP was lower in Microworkers than in both other platforms, the somewhat higher CP in WiSoPanel compared to Figure Eight failed a conventional level of statistical significance. Furthermore, while the error percentage was lower in WiSoPanel than in Microworkers and Figure Eight, the latter two platforms did not significantly differ. The same pattern emerged with regard to the variance in the two metrics.

These platform differences in attention were partly but not entirely explained by known differences—demographic and others—among the users of the three platforms: The CP was higher among younger testtakers and those who live in a Western European culture. While younger age is somewhat more represented in Figure Eight and markedly more represented in Microworkers, living in a Western European culture is somewhat less represented in Figure Eight and markedly less represented in Microworkers; thus, age and culture to some degree offsetting each other. The finding of cultural differences ties in with Litman, Robinson, and Rosenzweig (2015) who observed that India-based workers submitted data of a lesser quality than U.S. workers. Furthermore, the CP was higher if a testtaker participated from elsewhere than from home and for those with higher self-evaluated mouse skills. Above and beyond these influences on CP, Figure Eight workers showed a higher CP than the other two platforms together, whereas WiSoPanel by tendency was higher than the other two platforms together, but not significantly so. The error percentage was higher among older testtakers and among those who live in a non-Western-European culture. Again, while older age is somewhat less represented in Figure Eight and markedly less represented in Microworkers, living in a non-Western-European culture is somewhat more represented in Figure Eight and markedly more represented in Microworkers, thus counterbalancing each other. Beyond these identified influences on error percentage, workers from WiSoPanel worked more accurately than workers from both crowdsourcing platforms together.

Using either attentiveWeb metric shows that workers' age impacted attention test performance and that performance profited from younger age. This likely reflects the often made finding that performance in timed tests lessens with age because of declining perceptual motor skills (Eckert, 2011; Kennedy, Partridge, & Raz, 2008). We assume that the CP profited somewhat more from younger age than the error percentage because age shifts the speed-accuracy tradeoff toward accuracy (Forstmann et al., 2011). Moreover, we assume that users who participated from elsewhere but their home had a better CP because of self-selection pertaining to motivation: Those who decide to comply with a work request despite having to deal with the discomfort of not being at home have a stronger motivation to participate and consequently attain better results. The better performance with higher self-evaluated mouse skills is self-explanatory. There were no influences of gender on either of the two metrics; hence, the different gender proportions across the three platforms had no bearing on observed differences in attention. Moreover, the varying levels of language skills had no effect on the attention test performance, since any possible influence of language skill likely had been absorbed by the variable coding for culture. Finally, it did not matter if a worker had previously done a d2.

To sum up the comparison of the platforms, despite accounting for attention-relevant differences among workers on the three platforms, there remained inherent differences across platforms with regard to attention. Our two main findings, first: marked demographic differences of workers in an online panel and crowdsourcing workers, and second: marked differences in the quality of their work results, tie in with Smith, Roster, Golden & Albaum (2016) who compared workers in an online panel with Amazon Mechanical Turk (MTurkers). An explanation for the observed work quality differences might be the types of microtasks that are typically offered on these platforms. Perhaps users of Figure Eight and WiSoPanel are more used to tasks that resemble our attention test than are

Microworkers. Furthermore, online panels (e.g., WiSoPanel) and crowdsourcing platforms (e.g., Figure Eight and Microworkers) differ in characteristics that might be relevant for their users' motivation or ability when carrying out a microtask (Göritz & Neumann, 2016). In online panels, people have expressed their interest in participating in web-delivered research studies, whereas in crowdsourcing platforms, people have expressed their interest in carrying out different kinds of web-delivered work more generally. The motivation to carry out special work for which one has expressed interest is likely to be higher than the motivation to carry out work that falls within a broad spectrum of work for which one has signed up. Moreover, participating in research studies, unlike in various kinds of web-delivered work, promises to fulfill motives such as being entertained, learning something about oneself, or contributing to discover scientific facts. Seeking to have such motives fulfilled by one's participation makes submitting sloppy work or cheating quite pointless. By contrast, workers of a big crowdsourcing platform described the platform as a labor market (Brawley & Pury, 2016), and Litman and colleagues (2015) showed that the motivation of crowdsourcing workers has shifted from being primarily intrinsic to being mainly extrinsic. Furthermore, as most online panels are smaller than crowdsourcing platforms, they represent close-knit networks, and their participants are likely to be treated more personally. Thus, participants in online panels may have a stronger identification with and sense of membership in the platform than participants in crowdsourcing platforms. Lastly, as more educated people are less likely to rely on earning a secondary income, they have a higher motivation to sign up with an online panel than with a crowdsourcing platform compared to people who rely on earning a secondary income, which is more likely to be true for the less educated (cf. Kuhn & Maleki, 2017).

The test attentiveWeb captured crowdworkers' attention both reliably—which was established on all three platforms—and validly—which was established on two of the platforms. With regard to internal consistency, attentiveWeb's Cronbach's α was .91 with CP and .87 with $E\%$. By way of comparison, according to the d2-R manual (Brickenkamp et al., 2010), d2-R has a Cronbach's α of .96 with CP and .87 with $E\%$. Thus, attentiveWeb was a little less internally consistent than d2-R as regard CP but measures up as regard $E\%$. Split-half reliability of attentiveWeb was .89 with CP and .86 with $E\%$. According to the manual, d2-R has a split-half reliability of .94 as regard CP and .85 as regard to $E\%$. Thus, attentiveWeb was somewhat less split-half reliable than d2-R as regard CP but measures up as regard $E\%$. With regard to validity, attentiveWeb (Version 1) predicted crowdworkers' performance 2 years later in attentiveWeb (Version 2) at $r = .57$ as regard CP and at $r = .40$ as regard $E\%$. According to the d2-R manual, after a retest interval of 10 days, d2-R showed a retest reliability of .85 with CP and .47 with $E\%$. The attentiveWeb validity test at hand used a much longer re-test interval of more than 2 years, and the validity test at hand was not based on readministering the same task but on administering a similar attention task; thus, the validation studies at hand tested for predictive validity rather than retest validity. As predictive validities are lower than retest reliabilities because predictive validity refers to a different test taken at a later time and retest reliability refers to the same test taken at a later time and as reliabilities tend to be higher the shorter the retest interval (except for very short retest intervals where fatigue may play a role), attentiveWeb's predictive validity of .57 with CP is to be considered very good and the .40 with $E\%$ outstanding. Thus, applying attentiveWeb on crowdworkers allows for predicting their attention more than 2 years later.

The fact that attentiveWeb's internal consistency and split-half reliability have been established on three platforms (WiSoPanel, Microworkers, and Figure Eight) and its predictive validity on two platforms (WiSoPanel and Microworkers) lends confidence in its quality when used on other crowdsourcing platforms as a personnel selection test for tasks that require sustained attention. Furthermore, attentiveWeb has proven its robustness: It showed similar test quality when used on several platforms, wherein it was administered in different language versions (i.e., English and German) and to audiences that differed in demographics and in other characteristics. With attentiveWeb being in the public domain, its costs are zero. However, when using attentiveWeb as a test for

preselecting crowdworkers, it takes a few minutes to take part in, for which the employer needs to pay the tested crowdworkers. Moreover, the employer needs to set up attentiveWeb on her server in the first place, which takes a while when done for the first time. On balance, given the established usefulness of attentiveWeb, its advantages likely outweigh the additional cost and effort involved, which is even more true the more often an employer administers this test. In case, that attentiveWeb becomes a globally adopted selection test that any crowdworker can expect to occasionally be confronted with, unlike tests that suffer from considerable training effects or that can be faked easily, the benchmark test d2-R merely suffers from mild training effects (Brickenkamp et al., 2010), and it cannot be “faked up.” Although “faking down” is possible, it is not in the interest of any crowdworker to appear less attentive than they are. When attentiveWeb gains in popularity, to win the arms race with crowdsourcers who try to have a program take the test, developing a version in which the target and distractor stimuli are shuffled before being presented becomes worthwhile. Any such or other modifications of attentiveWeb are encouraged thanks to its open-source license. Moreover, using the two versions of attentiveWeb (i.e., Version 1 with d as the target and Version 2 with b as the target) is expected to reduce training effects.

There are several modes in which attentiveWeb might be used. One mode is to administer it before completion of every new task. This would primarily be the case if the employer has rarely solicited crowdworkers for any microtask; hence, he or she has not yet build a personal pool of reliable workers. The enlarged quality of the work results when using workers for the entailing crowdsourced microtask who have been selected on the basis of attentiveWeb is likely to make paying for the few minutes of its administration by the employer worthwhile. Another mode of attentiveWeb’s usage is if an employer frequently contracts out microtasks that require attention on the part of the crowdsourcing worker. In that case, the employer may build their personal pool of reliable workers by administering attentiveWeb to each applying crowdworker once, and for later tasks solicit workers from that pool. Finally, if attentiveWeb is used on the level of the crowdsourcing platform in that the platform endorses its use as part of a screening for workers’ suitability, a worker’s attentiveWeb results feed into his or her reputation score on the platform and can thus be used by all employers on that platform.

Based on the results of three user studies, when soliciting Microworkers as compared to workers from Figure Eight or WiSoPanel, requesters of crowdsourced work should be prepared that more Microworkers need to be filtered out and that the not-filtered-out workers carry out work with less attention. However, given that the platforms are dynamic environments and that the operators have the possibility to change policies and ways of operation, in the present work, the differences among platforms could only be captured as a snapshot and hence should be reevaluated every once in a while.

This study’s conclusions are restricted by limitations, which call for more research. There are online panels other than WiSoPanel and there are crowdsourcing platforms other than Microworkers and Figure Eight—for example, Amazon MTurk—that we did not examine. Although we do not see any reason of why the outcomes of this study should not apply to other platforms, this must be tested empirically. We settled on an advance filter that combines that a page was skipped or that particularly unlikely values of the two test metrics were observed. This filter was objective to implement, and it was strict, but not maximally so. In the end, the choice of advance filter was researcher-made, and the results would have been somewhat different had one settled on a more lenient filter or no advance filtering of workers at all. Furthermore, some of the data that were used in this study the crowdworkers had reported themselves. The self-reports are probably less accurate than the behavioral data that were collected or if a standardized language test had been administered. In particular, whenever the likelihood to be hired seemed to depend on the workers’ self-reports (e.g., language skill), the self-reports were probably enhanced in the putatively desired direction. However, we cannot think of reasons of why the enhancement (e.g., overreporting language skills) should have been vastly different in one platform compared to the others. Furthermore, although many

crowdsourced microtasks profit if crowdworkers invest their attention in the task—and consequently attentiveWeb is expected to be a fairly task-independent crowdworker selection test—there are crowdsourced tasks that can be completed without investing sustained attention, or some crowdworkers might compensate for their wavering attention by spending more time on the task. Hence, depending on the nature of the task that an employer wishes to source to the crowd, attentiveWeb varies in its utility as a basis to select crowdsourced workers. Context independence does not imply relevance for any context. While being beyond the scope of this work, it is desirable that further studies sound out attentiveWeb’s relevance with different types of entailing crowdsourcing tasks. More research into attentiveWeb’s criterion validity is warranted, whereby different types of ensuing work tasks are potential criteria. Furthermore, applying an attention test is by no means the only way of sifting “good workers from bad ones.” To maximize the effectiveness and efficiency of selecting crowdworkers, it would be interesting to combine attention testing and existing approaches of quality control (e.g., using workers’ reputation on the platform or instructional manipulation checks or task-specific gold standards) to balance out the individual limitations of each of the different approaches. Finally, given that this work has established attentiveWeb’s applicability and quality as a necessary first step, next it would be interesting to compare this crowdworker selection tool to other tests and approaches of quality control in crowdsourcing.

To conclude, this work supports the potential of crowdsourcing by providing the community with a free, open source, and versatile personnel selection test to ensure the quality of crowdsourced work. We exemplified the use of the newly developed and validated attention test attentiveWeb to “Reject the Bad” (Kuhn, 2015) by deriving advance filters based on this test as well as to “Select the Good” (Kuhn, 2015) by assessing crowdworkers’ sustained attention in a crowdsourced task. Although this work has shed some light on the complexity of quality control in crowdsourcing, it is unlikely to be the last word on this dynamically developing type of work arrangement.

Appendix

Questions Asked Before attentiveWeb	Possible Answers
Please select your gender ^a	— Male Female
Please select your age ^a	— <20 20–30 31–40 41–50 >50
Please select your country ^a	— List of countries taken from https://maxmind.com/
Is English [WiSoPanel: German] your mother tongue? ^a	— Yes No
Please rate your English [WiSoPanel: German] skills ^a	— Beginner Advanced Native speaker
Where do you use English [WiSoPanel: German]? ^a	— At school/at work Daily life Vacations

(continued)

Appendix. (continued)

Questions Asked Before attentiveWeb	Possible Answers
Where are you at the moment?	— At home At work Internet café Somewhere else
How many people are around you?	— 0 1–3 4–10 >10
How mentally focused are you at the moment?	— 1: Not at all 2 3 4 5: Very highly
How skilled are you with the mouse? ^a	— 1: Unskilled 2 3 4 5: Highly skilled
Please select your continent ^a	— List of continents
Where are you at the moment? ^a	— At home At work Internet café Somewhere else
Have you ever done this test before? ^a	— Yes No
If you did the test before, when was it? ^a	— <1 week 1 week–1 month 1 month–1 year >1 year
Which strategy did you use for the test?	— As fast as possible, accepting mistakes Slower but correct answers
How mentally focused are you at the moment?	— 1: Not at all 2 3 4 5: Very highly
Did you do the test seriously?	— Yes No
Feedback	Free text

^aQuestions skipped in the validation studies, which were conducted more than 2 years after the user studies.

Authors' Note

The data are available in the folder “data” (<https://github.com/lsinfo3/attentiveWeb-studies-data>). The developed test attentiveWeb is available open source and for free (<https://github.com/lsinfo3/attentiveWeb/>). The figures were created with the open source and free statistical software R (Version 3.2.3; <https://www.r-project.org/>) using the library ggplot2 (Version 3.1.0; <https://ggplot2.tidyverse.org/>).

Declaration of Conflicting Interests

The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The authors disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This research received funding from the Deutsche Forschungsgemeinschaft (DFG) under Grants HO4770/2-2 and TR257/38-2.

References

- Brawley, A. M., & Pury, C. L. (2016). Work experiences on MTurk: Job satisfaction, turnover, and information sharing. *Computers in Human Behavior, 54*, 531–546.
- Brickenkamp, R. (1981). *Test d2. Aufmerksamkeits-Belastungs-Test*. Göttingen, Germany: Hogrefe.
- Brickenkamp, R., Schmidt-Atzert, L., & Liepmann, D. (2010). *Test d2-Revision: Aufmerksamkeits-und Konzentrationstest*. Göttingen, Germany: Hogrefe.
- Buhrmester, M. D., Talaifar, S., & Gosling, S. D. (2018). An evaluation of Amazon’s Mechanical Turk, its rapid rise, and its effective use. *Perspectives on Psychological Science, 13*, 149–154.
- 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*, 2156–2160.
- Checco, A., Bates, J., & Demartini, G. (2018). All that glitters is gold—An attack scheme on gold questions in crowdsourcing. In Y. Chen & G. Kazai (Eds.), *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing* (pp. 2–11). Palo Alto: AAAI Press.
- Daniel, F., Kucherbaev, P., Cappiello, C., Benatallah, B., & Allahbakhsh, M. (2018). Quality control in crowdsourcing: A survey of quality attributes, assessment techniques, and assurance actions. *ACM Computing Surveys, 51*, 7.
- Diedenhofen, B., & Musch, J. (2016). cocron: A web interface and R package for the statistical comparison of Cronbach’s alpha coefficients. *International Journal of Internet Science, 11*, 51–60.
- Dougherty, D., Marsh, D., & Mathias, C. W. (2002). Immediate and delayed memory tasks: A computerized behavioral measure of memory, attention, and impulsivity. *Behavioral Research Methods: Instruments and Computers, 34*, 391–398.
- Düker, H., & Lienert, G. A. (2001). *Konzentrations-Leistungs-Test: KLT-R*. Göttingen, Germany: Hogrefe.
- Eckert, M. A. (2011). Slowing down: Age-related neurobiological predictors of processing speed. *Frontiers in Neuroscience, 5*, 25.
- Forstmann, B. U., Tittgemeyer, M., Wagenmakers, E.-J., Derrfuss, J., Imperati, D., & Brown, S. (2011). The speed-accuracy tradeoff in the elderly brain: A structural model-based approach. *Journal of Neuroscience, 31*, 17242–17249.
- Gadiraju, U., Kawase, R., Dietze, S., & Demartini, G. (2015). Understanding malicious behavior in crowdsourcing platforms: The case of online surveys. In B. Begole, J. Kim, K. Inkpen, & W. Woo (Eds.), *Proceedings of the 33rd ACM Conference on Human Factors in Computing Systems* (pp. 1631–1640). NY: ACM.
- Gadiraju, U., Siehdnel, P., Fetahu, B., & Kawase, R. (2015). Breaking bad: Understanding behavior of crowd workers in categorization microtasks. In Y. Yesilada, R. Farzan, & G. J. Houben (Eds.), *Proceedings of the 26th ACM Conference on Hypertext and Social Media* (pp. 33–38). NY: ACM.

- Goodman, J. K., Cryder, C. E., & Cheema, A. (2013). Data collection in a flat world: The strengths and weaknesses of Mechanical Turk samples. *Journal of Behavioral Decision Making*, *26*, 213–224.
- Göritz, A. S. (2014). Determinants of the starting rate and the completion rate in online panel studies. In M. Callegaro, R. Baker, J. Bethlehem, A. S. Göritz, J. A. Krosnick, & P. J. Lavrakas (Eds.), *Online panel research: A data quality perspective* (pp. 154–170). Chichester, England: Wiley.
- Göritz, A. S., & Neumann, B. P. (2016). The longitudinal effects of incentives on response quantity in online panels. *Translational Issues in Psychological Science*, *2*, 163–173.
- Hauser, D. J., & Schwarz, N. (2016). Attentive turkers: MTurk participants perform better on online attention checks than do subject pool participants. *Behavior Research Methods*, *48*, 400–407.
- Hirth, M. (2016). *Modeling crowdsourcing platforms—A use-case driven approach*. University of Würzburg, Würzburg, Germany. Retrieved from <https://opus.bibliothek.uni-wuerzburg.de/frontdoor/index/index/start/0/rows/10/sortfield/score/sortorder/desc/searchtype/simple/query/Matthias+Hirth+/docId/14072>
- Hirth, M., Hoßfeld, T., & Tran-Gia, P. (2011). Anatomy of a crowdsourcing platform—Using the example of microworkers.com. In K. Chae, A. Nakao, & H. D. J. Jeong (Eds.), *Workshop on Future Internet and Next Generation Networks (FINGNet)*. Seoul, Korea: Conference Publishing Service.
- Hoßfeld, T., Hirth, M., & Tran-Gia, P. (2011). Modeling of crowdsourcing platforms and granularity of work organization in future internet. In C. Kalmanek, D. Medhi, Å. Arvidsson, G. de Veciana, S. Low, G. Agrawal, & A. Clemm (Eds.) *Proceedings of ITC 2011 International Telegrafic Congress* (pp. 142–149). San Francisco, CA: ITC Publications.
- Jacobs, B. (2013). *Erprobung zweier Online-Konzentrationstests mit Zahlen an Studierenden des Lehramts*. Retrieved April 16, 2019, from <http://bildungswissenschaften.uni-saarland.de/personal/jacobs/diagnostik/tests/konzentration/konzentrationstests.html>
- Jacobs, B. (2014). *Analyse von Testgütekriterien und Übungseffekten zweier Online-Konzentrationstests*. Retrieved April 16, 2019, from http://bildungswissenschaften.uni-saarland.de/personal/jacobs/diagnostik/tests/konzentration/uebung/uebungseffekte_gu_und_zrf_20.pdf
- Jacobs, B. (2015). *Der Zahlenreihenfolgetest*. Retrieved April 16, 2019, from http://bildungswissenschaften.uni-saarland.de/personal/jacobs/diagnostik/tests/konzentration/gu_neu/gu.pdf
- Kennedy, K. M., Partridge, T., & Raz, N. (2008). Age-related differences in acquisition of perceptual-motor skills: Working memory as a mediator. *Aging, Neuropsychology, and Cognition*, *15*, 165–183.
- Kuhn, K. M. (2015). Selecting the good vs. rejecting the bad: Regulatory focus effects on staffing decision making. *Human Resource Management*, *54*, 131–150.
- Kuhn, K. M. (2016). The rise of the “gig economy” and implications for understanding work and workers. *Industrial and Organizational Psychology*, *9*, 157–162.
- Kuhn, K. M., & Maleki, A. (2017). Micro-entrepreneurs, dependent contractors, and Instaservers: Understanding online labor platform workforces. *The Academy of Management Perspectives*, *31*, 183–200.
- Le, J., Edmonds, A., Hester, V., & Biewald, L. (2010). Ensuring quality in crowdsourced search relevance evaluation: The effects of training question distribution. In V. R. Carvalho, M. Lease, & E. Yilmaz (Eds.), *Proceedings of the ACM SIGIR 2010 Workshop on Crowdsourcing for Search Evaluation (CSE 2010)* (pp. 17–20). NY: ACM.
- Litman, L., Robinson, J., & Rosenzweig, C. (2015). The relationship between motivation, monetary compensation, and data quality among US- and India-based workers on mechanical turk. *Behavior Research Methods*, *47*, 519–528.
- Martin, D., Carpendale, S., Gupta, N., Hoßfeld, T., Naderi, B., Redi, J., . . . Wechsung, I. (2017). Understanding the crowd: Ethical and practical matters in the academic use of crowdsourcing. In D. Archambault, H. Purchase, & T. Hoßfeld (Eds.), *Evaluation in the crowd: Crowdsourcing and human-centered experiments* (pp. 27–69). Heidelberg: Springer.
- Oleson, D., Sorokin, A., Laughlin, G. P., Hester, V., Le, J., & Biewald, L. (2011). Programmatic gold: Targeted and scalable quality assurance in crowdsourcing. *Human Computation*, *11*. Retrieved April 16, 2019, from https://publicassets.s3.amazonaws.com/papers/HCOMP2011_philosopher_stone.pdf

- Peer, E., Vosgerau, J., & Acquisti, A. (2014). Reputation as a sufficient condition for data quality on Amazon Mechanical Turk. *Behavior Research Methods*, *46*, 1023–1031.
- Rothwell, S., Carter, S., Elshenawy, A., & Braga, D. (2015). *Job complexity and user attention in crowdsourcing microtasks* (AAAI Technical Report WS-15-24). Retrieved April 16, 2019, from <https://www.aaai.org/ocs/index.php/HCOMP/HCOMP15/paper/viewFile/11735/12334>
- Satow, L. (2011). *Psychomeda Konzentrationstest (KONT-P)*. Retrieved June 24, 2018, from <https://www.psychomeda.de/online-tests/Psychomeda-Konzentrationstest.pdf>
- Smith, S. M., Roster, C. A., Golden, L. L., & Albaum, G. S. (2016). A multi-group analysis of online survey respondent data quality: Comparing a regular USA consumer panel to MTurk samples. *Journal of Business Research*, *69*, 3139–3148.
- Starzacher, E., Nubel, K., & Grohmann, G. (2006). *Continuous attention performance test*. Göttingen, Germany: Hogrefe.

Author Biographies

Anja S. Göritz is a full professor of Occupational and Consumer Psychology at the University of Freiburg in Germany. She has published in journals including *Behavior Research Methods*, *Leadership Quarterly* and *Journal of Business Ethics*. Her research focusses on online data collection and consumer behavior. For more information please visit www.goeritz.net.

Kathrin Borchert is a PhD student at the Chair of Communication Networks at the University of Würzburg, Germany. Her research focuses on the task design and workflow optimization of microtasking jobs. Furthermore, she investigates the realization of enterprise crowdsourcing, for example, subjective user studies for evaluating the users' quality of experience in thin-client architectures as deployed in business infrastructures.

Matthias Hirth is an assistant professor at TU Ilmenau, Germany and heading the research group on user-centric analysis of multimedia data. His research focuses on the large-scale assessment of the quality of experience (QoE) of the technical system, the modeling of QoE, and the optimization of technical systems from a user's perspective. He is also working on the application of crowdsourcing for subjective assessments and the optimization of crowdsourcing workflows.