Personality Assessment To-Go

Formal Aspects of Unproctored Web-Based Personality Assessment in Relation to Its Content and Quality

Tobias Altmann and Niket Kapoor

Department of Psychology, University of Duisburg-Essen, Essen, Germany

Abstract: A particular feature of unproctored Internet Testing (UIT) is the participants' freedom to decide on the formal aspects of their participation, such as time of day, device, and whether, how often, and for how long they might intermit their participation. A main point of discussion has been how these aspects alter the quality and content of an assessment. The issue remains understudied while simultaneously maintaining great importance for many fields. We examined this question in a UIT assessment of the Big Five personality factors in the present study. A sample of 441 participants who completed the assessment and 527 participants who aborted their participation was used to analyze quality (internal consistency, response styles) and content (mean score) differences. Results revealed several dependencies among small effect sizes. The discussion focuses on the potential practical implications of the present findings.

Keywords: personality assessment, web-based, unproctored, Big Five

Unproctored Internet Testing (UIT) is used in psychological assessment and testing to describe instruments brought to participants via the Internet, not monitored by a human or non-human test administrator (Tippins, 2006). This form of remote and unsupervised web-based data collection has already enjoyed massive popularity among researchers and practitioners alike (Barak, 2011; Epstein & Klinkenberg, 2001). The rise in popularity can be demonstrated by looking at the 50% annual growth rate (averaged across the past two decades) of the search term “online assessment” in the literature database PsycINFO. However, the term UIT is fairly vast because Internet-based studies, proctored or not, can vary greatly, ranging from web surveys to Internet-based experiments (Reips, 2013), each with its own set and extent of potential benefits and concerns with respect to its translation into a web-based format.

Many self-explanatory benefits come from UIT, especially when contrasting these benefits with the conditions involved in proctored in-person testing. Apart from obvious benefits (e.g., increased economic use, scalability, adaptability, and an easier way to gather international data; Barak, 2011), UIT may also increase participants' depth and breadth of self-disclosure due to increased anonymity and improved comfort (Joinson et al., 2008). In the clinical setting, it can be the necessary bridge to the proper remote assessment of patients who, due to anxiety disorders or health limitations, might not be able to complete an in-person assessment (Jones & Stokes, 2009).

Regardless, UIT formats are also associated with a plethora of problems, among them privacy issues, especially

when considering the often quiet personal data collected in the realm of psychology (Burgoon et al., 1989). Online anxiety, as well as the fear of a possible “digital divide” – a form of selection bias in which the people willing to engage in a UIT might have common characteristics (Langenfeld, 2020) – have also been discussed.

One of the most apparent issues in UIT is the lack of control over the testing environment (Epstein & Klinkenberg, 2001). In proctored testing, whether in-person or web-based, the context is considerably less variable, and hence, the derived data are better shielded from possible distortions. By contrast, with UIT, participants are freer to decide about the formal aspects of their participation, such as when, where, and how they choose to complete a given test or participate in a given survey. For instance, participants may decide to complete an assessment in the morning on their stationary PC in one go or at night lying in their bed, using their smartphone to participate with several intermissions. Whereas there is a reasonable discussion about whether this possibly should be considered a benefit due to an increase in ecological validity (Reips, 2002), others consider it one of the main disadvantages of online testing (Barak, 2011). Hence, researchers have been interested in investigating the equivalency of a given test when completed at home versus in a laboratory.

Most studies have concluded that the assessment of noncognitive psychological constructs (e.g., the Big Five) appear to be valid regardless of the administration environment (Barak & Hen, 2008). Many studies in equivalency research have utilized a between-subject design and

differentiated between a traditional paper-and-pencil design versus an online at-home condition. These studies have found equivalency, especially in the realm of noncognitive constructs (Beatty et al., 2011; Buchanan, 2009; Chuah et al., 2006; Cronk & West, 2002; Templer & Lange, 2008; Vecchione et al., 2011). Le Corff and colleagues (2017) applied a within-subject design to increase the comparability of results but still found only negligible differences.

However, this evidence is not unchallenged. Uncertainty remains from research showing differences in data distributions and psychometric quality. Unlike the aforementioned articles, this evidence does not support the general validity of unproctored assessment (Buchanan, 2003; Carstairs & Myers, 2009; Do, 2009; O'Neil & Penrod, 2001). Furthermore, studies concerning UIT equivalence usually do not isolate context variables (e.g., type of device, time of day of participation, continuation, and total participation time). Such a differentiated view becomes necessary when striving for a thorough analysis of potential distortions when comparing proctored and unproctored assessments and when conducting investigations across different variations of unproctored settings.

It can be assumed that variations in these formal aspects can result in varying data quality within UIT assessments. Tantamount to analyzing external variables of proctored testing, unproctored testing with its distinct class of formal aspects is severely understudied. This is especially true for personality assessments. Personality assessments in UIT make up an increasingly higher percentage of online studies, but their quality has not yet been studied with the same thoroughness with which cognitive constructs have been studied. Possible main and interaction effects of these formal assessments could also imply practical applications of UIT and aid decisions regarding partial variable fixation without overly limiting participants' freedom, which could be counterproductive (Reips, 2008).

Formal Aspects of Participation in Unproctored Internet Testing

To obtain a broad view of the current state of research on the topic of formal aspects of UIT, we decided to focus on the aforementioned variables, namely, type of device, time of day, continuation, and total participation time. Concerns regarding possible device effects were already voiced decades ago (Bartram & Bayliss, 1984) with very little research since. Discussions about device effects must consider that a mere dichotomous separation into digital and non-digital does not represent the individual parameters within the large spectrum of digital devices, as it could reference stationary personal computers and mobile

smartphones or tablets. These digital devices differ in many respects, such as user interface (e.g., touchscreen or keyboard) and mobility (e.g., stationary PCs or mobile smartphones).

Arthur and colleagues (2018) and Morelli and colleagues (2014) focused on interface and screen size and found no mode-based differences in results. However, Arthur, Keiser, Hagen, et al. investigated mobile versus nonmobile devices only in a proctored lab setting. Therefore, the study mainly focused on cognitive load stemming only from the device due to its structural characteristics (e.g., larger screen sizes, degree of permissibility). Similarly, Morelli et al. assessed only the work-related abilities and skills relevant to customer service advisors, so their results are hardly generalizable or applicable to personality assessment.

Apart from screen size and keyboard, another fundamental way that stationary and mobile devices differ is in the aspect of mobility. When using a predominantly stationary device, such as a PC or a laptop, one typically sits down and does not move the device (or oneself) very often. By contrast, purely mobile devices, such as smartphones or tablets, accompany many people in practically all everyday situations with almost unlimited mobility and can thus be used (e.g., when participating in UIT assessments), for instance, while riding on public transportation, waiting in line, or taking a walk. These differences in mobility may be associated with the attention that people allocate to the device or in their willingness to focus on the contents on the screen. As a result, data quality may be expected to differ between these device types. Interestingly, this fundamental difference between stationary and mobile devices has received practically no attention in (personality) research so far (one of the few examples is the study by Stachl et al., 2020). This is a severe oversight, considering the steep increase in online studies and the fact that participants are free to choose the device they want to use to participate in such studies.

For the remaining three formal aspects, research has also been sparse. Participation continuity was a topic of interest in Miller and colleagues (2002), where participants took breaks as a proxy for general disturbances due to uncontrolled conditions (e.g., fatigue or time pressure). These results were compared with a traditional paper-and-pencil condition and a web-based condition without interruptions. No differences in results were found, but the study assessed alcohol use, and thus, the findings might not apply to personality assessment (e.g., for a study on the weak associations between the Big Five and alcohol consumption, see Lackner et al., 2013).

Effects of time of day and total participation time have also been understudied in the literature. Lawrence et al. (2009) found no difference when looking at time of participation in a web-based personality assessment, but they

used unspecified measures of constructs such as “Achievement Orientation,” “Positive Affectivity,” and “Time Management.” With that said, effects of time of day have long been studied in the realm of traditional proctored personality research, especially for the Big Five (Deyoung et al., 2007). The Big Five is a well-established and widely accepted model of personality (McCrae & Allik, 2002). Not only is it used frequently, but it also has solid empirical backing, especially concerning its longitudinal stability and associated life outcomes (Soldz & Vaillant, 1999; Soto, 2019). A relevant concern is that limiting the time of day to certain time slots could bias the data on personality as there are well-studied preferential times for certain personality configurations. Exemplarily, Randler (2008) found significant correlations between time of day and the Big Five (e.g., a positive relationship between agreeableness and generally being a “morning person”). Therefore, limiting the time slot could result in a biased sample due to diurnal preferences or a lower quality data set when certain participants complete the study during an undesired time of day. Unproctored UIT studies that do not limit the time of day could hence be seen as advantageous in this regard.

In light of the increasing number of studies conducted online, it seems advisable to study the consequences of UIT in personality assessment. Several studies have already begun analyzing the effects of the formal aspects of participation, which are commonly unstandardized in UIT. However, both comprehensive evidence in general and evidence in the context of personality assessment are lacking. For this reason, there is a need for research on how formal aspects of UIT influence both assessment quality and results.

The Present Study

In the present study, we investigated the four formal aspects (1) type of device, (2) time of day, (3) continuation, and (4) total participation time in relation to content- and quality-related aspects of the Big Five. We used the type of device to operationalize mobility and dichotomized it into stationary and purely mobile devices. Stationary devices, such as PCs and laptops, allow the user only a limited degree of physical mobility and usually require the user to sit down in front of the device. This requirement may be accompanied by a higher level of attention allocated toward the device. By contrast, purely mobile devices, such as smartphones, allow practically unlimited flexibility and are often used accordingly, for instance, on public transportation or when waiting in line. Such mobility may be associated with lowered attention and greater responsiveness to potentially distracting stimuli.

Time of day represents the participants’ choice of when to begin participating in the study, which can be any time

of the day. Continuation represents the different variations of study completion, including completion in one sitting, completion but with an intermission, and participating without completion. Total participation time represents the time used to complete the study and applies only to the group of participants who completed the study in one sitting.

We operationalized content effects by testing for Big Five mean differences between the levels of the formal aspects (i.e., testing whether the levels of the aspects explained variance in the Big Five). Quality was operationalized with the use of internal consistency and response styles. Internal consistency (Cronbach’s α and McDonald’s ω) is a frequently used indicator or proxy of reliability or the interrelatedness of the items used to measure a latent construct (Tavakol & Dennick, 2011). In this study, the analysis of internal consistency scores allowed insights into differences in the amount of measurement error present in the Big Five factors at the different levels of the formal aspects. Response styles provide extensions of these quality evaluations. To our knowledge, response styles have been studied in various contexts (e.g., in cross-cultural comparisons; Johnson et al., 2005) but not in the context of UIT. In this study, we utilized participants’ total numbers of extreme responses and the standard deviations across the Big Five items as potential sources of measurement error (Cronbach, 1946). Taken together, these content and quality aspects may offer additional insights into participants’ response behavior.

Generally, we expected to find differences in all four formal aspects (i.e., type of device, time of day, continuation, and total participation time) about the quality- and content-related aspects of the Big Five. Specifically, we expected that the use of stationary devices would increase quality (Arthur, Keiser, & Doverspike, 2018; Arthur, Keiser, Hagen, et al., 2018). Furthermore, we expected that the internal consistencies would be highest in the morning due to increased attentiveness in comparison with nighttime or even afternoon completion, although this point again had to be analyzed exploratorily to some extent due to the lack of previous studies. Regarding total participation time, we expected total participation times farther from the median of the distribution to be associated with lower internal consistency. Few or no intermissions and hence continuous study participation should also be linked to higher data quality. In addition, we expected certain personality dimensions to have preferential participation times, such as agreeable people preferring to participate in the morning versus at noon (based on the time-preference studies, see Randler, 2008).

In general, with the lack of research on differences in the formal aspects of participation in UIT, by and large, and particularly concerning personality assessment, these

hypotheses should be characterized as rather speculative and the analyses as exploratory.

Method

Sample and Procedure

The sample was recruited from different universities in Germany using social media network sites and apps (e.g., Facebook, Instagram, WhatsApp), email forwarding, and newsletters. The advertisement informed potential participants that the study involved personality assessment. It included a weblink to the online assessment tool “Questback” also known as “Unipark,” used to collect the data. Before accessing the study, participants gave informed consent for their data to be used for scientific purposes. They could also have their data entirely deleted after their (partial) participation, but no participant opted for this.

Participants were free to choose the time of day to participate and how long they could take to complete each item. They were also free to stop participating at any time without any consequences. All assessments were in accordance with the local and national ethical guidelines. Participation was voluntary and not monetarily rewarded.

The primary sample of participants who completed the assessment consisted of 441 German participants (343 women). Their mean age was 25.0 years ($SD = 6.3$), ranging from 18 to 64. As their highest level of education, 7.7% reported that they had attended school for 8–10 years, 66.4% reported 12–13 years, and 25.9% reported holding a university degree; 62.1% were students, 35.1% were employed, and the rest were unemployed or retired.

A secondary sample of 527 participants began participating in the study but stopped before they reached the end of the assessment. Most of them (396) did not provide any data and aborted their participation just after accessing the initial “welcome” page, which described the study. Such participants are common in web-based studies and represent a relevant subsample on their own, but they are rarely included in the analyses.

Measures

After assessing personal details (gender, age, education, and employment), the Big Five personality factors were assessed using the German version of the NEO Five-Factor Inventory (NEO-FFI; Borkenau & Ostendorf, 2008). The NEO-FFI contains 60 items for assessing Agreeableness (e.g., “I try to be courteous to everyone I meet”), Conscientiousness (e.g., “I keep my belongings clean and neat”), Extraversion (e.g., “I like to have a lot of people around

me”), Neuroticism (e.g., “I often feel tense and jittery”), and Openness (e.g., “I often try new and foreign foods”). Participants rated each item on a 5-point Likert scale measuring agreement. Cronbach’s α s for the five factors ranged from .75 to .88. Further details are provided in the Results section.

In addition to these data, we also recorded meta-information about each time the participants accessed the study. These included the date and time when the participants first accessed the study, total participation time, and the type of device used to participate. Such information can be assessed by all online assessment tools and are commonly provided to the researchers by most of them.

Results

Calculation of Study Variables

Using the measures and meta-information about participation described above, we created variables about formal aspects of participation (i.e., time of day, device, continuation, and total participation time), content (Big Five mean scores), and quality (internal consistency [α and ω] measures of the Big Five as well as response styles). With respect to the formal aspects of participation, we used the meta-information about participation described above to create the four categorical variables time of day, device, continuation, and total participation time. Time of day represents the time of day of participation. It was created using the time when the study was first accessed and was clustered into the four categories night (23:00–04:59 o’clock), a.m. (05:00–12:59 o’clock), noon (13:00–15:59 o’clock), and p.m. (16:00–22:59 o’clock). The device represented the type of device used to participate and was clustered into two categories: primarily stationary devices (e.g., laptop and desktop computer) and primarily mobile devices (e.g., smartphone; cf. Arthur, Keiser, Hagen, et al., 2018). Continuation represents whether participants completed the assessment without an intermission, with an intermission, or aborted it entirely. Total participation time represents the time the participants took to complete the assessment, measured in seconds. For this study, the originally continuous distribution of total participation time was clustered into three categories: participants with the lowest 25%, the middle 50%, and the highest 25%. Means of total participation time of these three groups were 742 s ($SD = 109$) for the lowest 25%, 1,102 s ($SD = 140$) for the middle 50%, and 1,886 s ($SD = 527$) for the highest 25%. Table 1 presents the total numbers of participants at each level of each of the four formal aspects.

With respect to the content- and quality-related aspects of participation, we calculated the mean Big Five scores,

Table 1. Number of participants per formal aspect

Aspects and levels	<i>n</i>	%
Time of day		
Night	34	3.5
a.m.	223	23.0
Noon	344	35.5
p.m.	367	37.9
Device		
Stationary	182	18.8
Mobile	786	81.2
Continuation		
Aborted	527	54.4
Intermitted	36	3.7
Completed	405	41.8
Total participation time		
Lower 25%	102	23.1
Middle 50%	203	46.0
Upper 25%	100	22.7

their internal consistencies, and two response styles. Internal consistency was calculated as Cronbach's α and McDonald's ω using the R package "coefficient alpha" (Zhang & Yuan, 2016) with robust estimates and 1,000 samples for bootstrapping 95% confidence intervals. Additionally, we used the NEO-FFI items to calculate two response styles: frequency of extreme responses (i.e., responses at either end of the continuum: "1" or "5" on the scales ranging from "1" to "5") and the standard deviation for each participant across all NEO-FFI items (inverse items not reversed). The former indicates a willingness to agree strongly or disagree with statements (but may be susceptible to acquiescence), whereas the latter indicates the general variability of the responses. All subsequent calculations using the Big Five mean scores or the two response styles were controlled for participants' age and gender (except for the calculation of internal consistencies).

Contingency Between the Formal Aspects

To test whether the four formal aspects were independent of each other, we used χ^2 to estimate the bivariate contingencies. Table 2 summarizes these results.

We found contingencies between time of day and device, time of day and continuation, device and total participation time, and device and continuation (see Table 2). The dominant effects are summarized in the following. For noon (as one level of the time-of-day variable), more participants than expected used stationary devices (84 vs. 64.7 expected) and, accordingly, fewer than expected used mobile devices (260 vs. 279.3 expected). The opposite was true for p.m. (49 vs. 69.0 expected, and 318 vs. 298 expected, respectively). For noon, fewer participants than expected aborted their participation (153 vs. 187.3 expected) and, accordingly, more than expected completed the study without an intermission (177 vs. 143.9 expected). With respect to total participation time, more stationary-device users than expected were among those with the 25% lowest total participation times (34 vs. 24.4 expected) and, accordingly, fewer mobile-device users than expected (68 vs. 77.6 expected). This indicates that using a stationary device tended to be associated with shorter participation times. With respect to continuation, more mobile-device users than expected were among those who aborted their participation (448 vs. 427.9 expected) and, accordingly, more stationary-device users than expected were among those who completed the assessment without an intermission (97 vs. 76.1 expected). Overall, the effect sizes for all contingencies were comparable and small (see Table 2).

Formal Aspects in Relation to the Quality-Related Aspects of Participation

To analyze whether the formal aspects were associated with higher or lower quality of the assessed data, we used two aspects of data quality: internal consistency and response styles (as detailed above). For internal consistency, we calculated both Cronbach's α s and McDonald's ω s with confidence intervals (as described above) for each of the levels of the four formal aspects. Figure 1 provides a graphical overview of the results. Contrary to our expectations, we found only small differences.

To test these differences for statistical significance, we used the R package "cocron" (Diedenhofen & Musch, 2016), which implements the tests to compare Cronbach α scores as suggested by Feldt et al. (1987). Multiple

Table 2. Contingencies between the four formal aspects based on their marginal distributions

	<i>n</i>	χ^2	<i>df</i>	<i>p</i>	ϕ
Time of day – device	968	14.75	3	.002	.12
Time of day – continuation	968	28.08	6	< .001	.17
Time of day – total participation time	405	6.25	6	.396	.12
Device – continuation	968	12.16	2	.002	.11
Device – total participation time	405	6.71	2	.035	.13

Figure 1. Cronbach's α s and McDonald's ω s with 95% confidence intervals for the Big Five at each aspect level. Dots represent α/ω coefficient estimates. The left and right ends of the lines represent the lower and upper bounds of the respective 95% confidence intervals. Black vertical lines represent the respective coefficient estimates in the total sample. Brackets represent significant differences between levels, calculated only for Cronbach's α scores. Omega estimates for level *Time of day: Night* were not estimable due to sample size.

comparisons for the formal variables time of day, device, and total participation time yielded significant differences only for time of day with respect to Agreeableness ($\chi^2 = 9.09$, $df = 3$, $p = .028$), Conscientiousness ($\chi^2 = 36.79$, $df = 3$, $p < .001$), Extraversion ($\chi^2 = 26.33$, $df = 3$, $p < .001$), and Neuroticism ($\chi^2 = 24.50$, $df = 3$, $p < .001$). Pairwise comparisons showed several significant differences, which are indicated by brackets in Figure 1. It should be noted that most of these significant comparisons were with the level night (of the time-of-day variable), which only included a small number of participants. The level night also had larger confidence intervals for some of the Big Five (see Figure 1). Post hoc inspection of these Big Five dimensions revealed a few items that had surprisingly low or even negative correlations with the respective total score (e.g., Item 9 from Agreeableness, which is “I often find myself arguing with my family and colleagues,” had a surprisingly corrected item-total correlation of $-.53$).

With respect to the two response styles (frequency of extreme responses and individual standard deviation), we calculated five analysis of variance (ANOVA) models for

each dependent variable: univariate ANOVAs (Models 1–3), a multivariate additive model with main effects (Model 4), and a complete multivariate model including all interaction terms (Model 5), as presented in Table 3. The frequency of extreme responses differed only within the device levels (see Model 5), $F(1, 383) = 8.95$, $p < .05$, and only with a very small effect of $\eta^2 = .02$ (a higher frequency of extreme responses for participants who used a mobile device). Contrary to our expectations, the individual standard deviation did not depend on the formal aspects (see Table 3).

Formal Aspects in Relation to the Content-Related Aspects (Big Five)

We used the same structure of ANOVA models again to analyze the dependencies between the formal aspects and the Big Five (see Table 4). Contrary to our expectations, we did not find significant differences in the formal aspects with respect to Conscientiousness or Neuroticism. In line

Table 3. ANOVA models of the two response styles calculated across all NEO-FFI items

	<i>df</i>	Frequency of extreme responses		Individual standard deviation	
		<i>F</i>	η^2	<i>F</i>	η^2
Model 1					
Time of Day	3, 437	1.41	.01	0.99	.01
Model 2					
Device	1, 439	7.43**	.02	0.44	.00
Model 3					
Total Participation Time	2, 402	0.92	.01	0.54	.00
Model 4					
Time of day	3, 398	0.86	.01	1.17	.01
Device	1, 398	5.73*	.01	0.32	.00
Total Participation Time	2, 398	1.36	.01	0.77	.00
Model 5					
Time of Day	3, 383	0.21	.00	0.97	.01
Device	1, 383	8.95**	.02	2.61	.01
Total Participation Time	2, 383	0.39	.00	0.20	.00
Time of Day × Device	3, 383	1.98	.02	1.27	.01
Time of Day × Total Participation Time	6, 383	1.35	.02	1.25	.02
Device × Total Participation Time	2, 383	1.59	.01	1.21	.01
Time of Day × Device × Total Participation Time	4, 383	0.54	.01	0.13	.00

Note. We controlled for age and gender in the dependent variables. * $p < .05$; ** $p < .01$.

with our expectations, we found several differences with respect to Agreeableness, Extraversion, and Openness as follows.

Bonferroni-adjusted post hoc analysis of the interaction effect between time of day and total participation time for Agreeableness (see Model 5), $F(6, 383) = 2.51$, $p = .022$, $\eta^2 = .04$, revealed a significant difference between a.m. and noon among participants who belonged to the group with the 25% lowest total participation time rates, $t(68) = 3.63$, $p = .001$. Here, Agreeableness was higher for participants who completed the assessment in the morning before noon (5–12 o'clock) compared with at noon (13–15 o'clock). Despite the Bonferroni adjustment, this effect seemed extremely specific and may thereby have resulted by chance. This interpretation may also be adequate for the effect of the device on Agreeableness, which was significant in the additive Model 4, $F(1, 398) = 6.06$, $p = .014$, $\eta^2 = .02$, with higher scores for stationary devices. However, this effect was not found in the interaction effect Model 5, $F(1, 383) = 1.72$, $p = .191$.

Bonferroni-adjusted post hoc analysis of the main effect of time of day on Extraversion, $F(3, 383) = 4.28$, $p = .005$, $\eta^2 = .03$, revealed significantly lower scores, $t(279) = 3.02$, $p = .008$, for participants who completed the assessment at noon (13–15 o'clock) compared with a.m. (5–12 o'clock).

Finally, for Openness, there was a main effect of device in Model 4, $F(1, 398) = 7.18$, $p = .008$, $\eta^2 = .02$. Participants who used a mobile device scored significantly lower than

participants who used a stationary device. Again, this effect was prevalent in Model 4 but not in Model 5, $F(1, 383) = 1.53$, $p = .216$. In general, all effects were small.

Discussion

In this study, we were interested in the extent to which formal aspects of UIT influence the quality and content of personality assessments. We hypothesized that certain personality dimensions would have preferential participation times and that the data quality would vary based on the specific manifestation of the formal variables (e.g., greater quality for stationary PCs vs. mobile devices).

The internal consistencies and response styles as measures of data quality differed with small effect sizes across the various formal aspects, hence showing that contrary to our hypotheses, the effects for quality seem limited. This evidence corroborates that of previously mentioned studies that found equivalency for unproctored studies compared to proctored settings. If one desires specific recommendations for optimum data quality, our research shows that stationary devices tend to be associated with higher completion rates and shorter participation times. Furthermore, more participants completed the study at noon. Interestingly, stationary devices were preferably utilized at noon, further supporting a preference for this time slot.

Table 4. ANOVA models of the Big Five mean scores

	df	Agreeableness		Conscientiousness		Extraversion		Neuroticism		Openness	
		F	η^2	F	η^2	F	η^2	F	η^2	F	η^2
Model 1											
Time of Day	3, 437	3.68*	.03	0.22	.00	3.94**	.03	1.02	.01	2.72*	.02
Model 2											
Device	1, 439	3.21	.01	0.60	.00	6.08*	.01	2.97	.01	6.53*	.02
Model 3											
Total Participation Time	2, 402	1.20	.01	1.60	.01	2.69	.01	0.63	.00	1.44	.01
Model 4											
Time of Day	3, 398	3.50*	.03	0.23	.00	3.57*	.03	0.88	.01	2.07	.02
Device	1, 398	6.06*	.02	0.14	.00	2.39	.01	1.60	.00	7.18**	.02
Total Participation Time	2, 398	1.16	.01	1.69	.01	1.50	.01	0.33	.00	2.42	.01
Model 5											
Time of day	3, 383	2.95*	.02	0.13	.00	4.28**	.03	0.97	.01	0.91	.01
Device	1, 383	1.72	.00	0.10	.00	0.20	.00	0.21	.00	1.53	.00
Total Participation Time	2, 383	0.55	.00	0.17	.00	0.47	.00	0.17	.00	2.18	.01
Time of Day $\times$ Device	3, 383	0.59	.01	1.30	.01	0.10	.00	0.16	.00	0.28	.00
Time of Day $\times$ Total Participation Time	6, 383	2.51*	.04	1.14	.02	1.03	.02	1.20	.02	0.60	.01
Device $\times$ Total Participation Time	2, 383	0.92	.01	1.39	.01	0.42	.00	0.23	.00	2.52	.01
Time of Day $\times$ Device $\times$ Total Participation Time	4, 383	1.05	.01	1.63	.02	0.61	.01	1.79	.02	0.38	.00

Note. We controlled for age and gender in the dependent variables. * $p < .05$; ** $p < .01$.

But the fixation of these formal aspects comes with a trade-off, namely, that of reducing participants' freedom and reducing external validity (Reips, 2002). Especially when considering that the effect sizes of the formal aspects in contingency with one another and with respect to quality were small, it is not advisable to limit participation to such a dramatic extent for a comparably small increase in data quality. However, an exception may be studied using MTurk or Prolific samples, which can achieve a high rate of data collection in a concise period of time. In such cases, the time of day when the study is launched can result in unstable data quality and a limited range of participants when considering the links between personality and participation time.

In sum, quality and content did not vary substantially when comparing study completion between a.m. and noon, but more substantial differences in internal consistencies were found for p.m. and night study completion, indicating that study launches should aim for a time slot between 5 a.m. and 4 p.m. Such issues may also vary in relevance

according to the type of study being conducted. It may be more relevant when the maximum number of participants is fixed and can be achieved in a short period of time (e.g., using MTurk participants, see more discussion above and below) or when the assessment is bound to a specific time of day, for example, in ambulatory assessments when measures are assessed repeatedly over the course of a day with narrow and fixed time frames for the responses (e.g., to assess self-esteem stability; e.g., Kernis, 2005).

In addition, we found certain time slots in which certain personality dimensions seemed to have a diurnal preference, reminiscent of Randler's (2008) study. For instance, we found that agreeable individuals preferred to complete the assessment in the morning compared to noon. These findings also illustrate the methodological risk of limiting participation in unproctored settings, as this could result in a bias for certain personality configurations (e.g., a reluctance of people low in conscientiousness to participate in a study when participation is only possible in the morning).

Individuals high on openness preferred stationary devices, which is surprising, as one would expect individuals with high openness to be attracted to the flexibility and innovation of mobile devices and not so much to the rigidity of stationary ones. Stachl and colleagues (2020) found that app usage, a primarily mobile technology, predicted openness. We, therefore, assume that our finding was a matter of chance.

Limitations

Many online studies use professional participants recruited from websites such as Mechanical Turk (MTurk) or Prolific. An inherent problem with using such samples in empirical studies is that it confounds the particular online assessment methodology with assessing a particular type of participant. By contrast, in the present study, we recruited participants the same way we would have recruited them for a laboratory study. Therefore, our findings more closely resemble the case in which a known sample is assessed, but we used an online assessment methodology. However, this procedure entails the potential limitation that the present findings cannot be applied to other types of samples (e.g., MTurk or Prolific) with certainty. For instance, whereas it may be easier for students to participate in online studies in the morning or at noon, MTurk participants who participate at these hours may be more likely to be professional participants with lower data quality (Goodman et al., 2013; Peer et al., 2021).

Some further limitations must be mentioned. First, as Table 1 shows, the sample sizes of the subgroups varied extensively (e.g., 34 at night as opposed to 344 in the a.m.). These sample size differences could have affected the detectability of possible effects. Second, the sample potentially had an educational bias because the majority consisted of university students. Third, we could not control whether some participants accessed the study from a different time zone. This could have influenced the time-of-day variable because time was recorded as server time rather than local time. Considering that the study was conducted in German, spread across German students and in German-speaking social media groups, we believe this issue is likely minor. Nevertheless, we cannot exclude the possibility that some participants accessed the study from a different time zone, thus resulting in a certain degree of measurement error in the time variable.

The meta-information used in this study was rich in its informative value and allowed us to investigate formal aspects in detail. Yet, some concerns remain, for example, whether participants who were registered as participating “without an intermission” may have taken a break while leaving the website running. These concerns could be resolved by adding more items to the meta-information.

We decided not to use such items here, but they could be utilized in future studies.

The study was not conducted experimentally. Randomized-controlled experiments are considered the gold standard in psychological research for a good reason, as they allow for the assumption of the equal distribution of errors across the different sample groups and because of the potential to assess causality. Yet, experimental studies also decrease ecological validity. Especially in the given study where the aim was to test how different natural environments are associated with results, it would have been counterproductive to limit these environments to experimental set-ups. It could be argued that not randomizing the formal aspects (e.g., the devices people used for the study) could result in confounding effects stemming from selective distortions. But we argue that assigning device types (or any of the other formal aspects in the present study) to participants could feel unnatural to them, thus reducing the informative value of the study and creating artificial effects (e.g., possible effects might not be attributable to the devices but instead to whether the assigned device matched the participants’ preferences or not). Even with the expected trade-off of possible confounds, we prioritized ecological validity (i.e., so we could study actual behavior and its dependencies).

Finally, the present study was limited to effects on the Big Five, which are highly stable and consistent personality traits (McCrae & Allik, 2002). The influence of the formal aspects of participation might be more relevant in psychological constructs with higher expected day-to-day variation than the Big Five. An example of such a construct is self-esteem, which shows variability over the course of a day (Hank & Baltes-Götz, 2019) and weeks (Kernis, 2005). The study also focused on a questionnaire. Other measures are also used in personality psychology, most notably narrative measures. These could behave differently in a UIT setting and hence could be of interest in future studies. In general, further studies with a similar differentiation of formal aspects are welcomed as this is a necessary basis to justify using UIT-derived data.

Conclusion

In sum, the results are quite affirming as they do not call for an overly cautionary use of UIT. They show that even when isolating individual formal aspects with meta-information, differences based on these variations were not concerning. Therefore, our results provide more confidence in the use of data derived from unproctored web-based platforms in the realm of personality psychology. Debated concerns regarding device effects or diurnal preferences, for example, do not seem to hold equal bearing in the realm of noncognitive constructs.

References

- Arthur, W., Keiser, N. L., & Doverspike, D. (2018). An information-processing-based conceptual framework of the effects of unproctored Internet-based testing devices on scores on employment-related assessments and tests. *Human Performance*, 31(1), 1–32. <https://doi.org/10.1080/08959285.2017.1403441>
- Arthur, W., Keiser, N. L., Hagen, E., & Traylor, Z. (2018). Unproctored Internet-based device-type effects on test scores: The role of working memory. *Intelligence*, 67, 67–75. <https://doi.org/10.1016/j.intell.2018.02.001>
- Barak, A. (2011). Internet-based psychological testing and assessment. In R. Kraus, G. Stricker, & C. Speyer (Eds.), *Online counseling* (pp. 225–255). Academic Press. <https://doi.org/10.1016/B978-0-12-378596-1.00012-5>
- Barak, A., & Hen, L. (2008). Exposure in cyberspace as means of enhancing psychological assessment. In A. Barak (Ed.), *Psychological aspects of cyberspace: Theory, research, applications* (pp. 129–162). Cambridge University Press. <https://doi.org/10.1017/CBO9780511813740.007>
- Bartram, D., & Bayliss, R. (1984). Automated testing: Past, present and future. *Journal of Occupational Psychology*, 57(3), 221–237. <https://doi.org/10.1111/j.2044-8325.1984.tb00164.x>
- Beaty, J. C., Nye, C. D., Borneman, M. J., Kantrowitz, T. M., Drasgow, F., & Grauer, E. (2011). Proctored versus unproctored Internet tests: Are unproctored noncognitive tests as predictive of job performance? *International Journal of Selection and Assessment*, 19(1), 1–10. <https://doi.org/10.1111/j.1468-2389.2011.00529.x>
- Borkenau, P., & Ostendorf, F. (2008). *NEO-Fünf-Faktoren-Inventar (NEO-FFI) nach Costa und McCrae* [NEO-Five-Factor Inventory (NEO-FFI) by Costa and McCrae]. Hogrefe.
- Buchanan, T. (2003). Internet-based questionnaire assessment: Appropriate use in clinical contexts. *Cognitive Behaviour Therapy*, 32(3), 100–109. <https://doi.org/10.1080/16506070310000957>
- Buchanan, T. (2009). Personality testing on the Internet: What we know, and what we do not. In A. N. Joinson, K. McKenna, T. Postmes, & U.-D. Reips (Eds.), *Oxford handbook of Internet psychology* (pp. 447–459). Oxford University Press. <https://doi.org/10.1093/oxfordhb/9780199561803.013.0028>
- Burgoon, J., Parrott, R., Poire, B., Kelley, D., Walther, J., & Perry, D. (1989). Maintaining and restoring privacy through communication in different types of relationships. *Journal of Social and Personal Relationships*, 6, 131–158. <https://doi.org/10.1177/026540758900600201>
- Carstairs, J., & Myers, B. (2009). Internet testing: A natural experiment reveals test score inflation on a high-stakes, unproctored cognitive test. *Computers in Human Behavior*, 25(3), 738–742. <https://doi.org/10.1016/j.chb.2009.01.011>
- Chuah, S. C., Drasgow, F., & Roberts, B. W. (2006). Personality assessment: Does the medium matter? No. *Journal of Research in Personality*, 40(4), 359–376. <https://doi.org/10.1016/j.jrp.2005.01.006>
- Cronbach, L. J. (1946). Response sets and test validity. *Educational and Psychological Measurement*, 6(4), 475–494. <https://doi.org/10.1177/001316444600600405>
- Cronk, B. C., & West, J. L. (2002). Personality research on the Internet: A comparison of web-based and traditional instruments in take-home and in-class settings. *Behavior Research Methods, Instruments, & Computers*, 34(2), 177–180. <https://doi.org/10.3758/BF03195440>
- Deyoung, C., Hasher, L., Djikic, M., Criger, B., & Peterson, J. (2007). Morning people are stable people: Circadian rhythm and the higher-order factors of the Big Five. *Personality and Individual Differences*, 43, 267–276. <https://doi.org/10.1016/j.paid.2006.11.030>
- 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.
- Do, B.-R. (2009). Research on unproctored Internet testing. *Industrial and Organizational Psychology*, 2(1), 49–51. <https://doi.org/10.1111/j.1754-9434.2008.01107.x>
- Epstein, J., & Klinkenberg, W. D. (2001). From Eliza to Internet: A brief history of computerized assessment. *Computers in Human Behavior*, 17(3), 295–314. [https://doi.org/10.1016/S0747-5632\(01\)00004-8](https://doi.org/10.1016/S0747-5632(01)00004-8)
- Feldt, L. S., Woodruff, D. J., & Salih, F. A. (1987). Statistical inference for coefficient alpha. *Applied Psychological Measurement*, 11, 93–103.
- 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(3), 213–224. <https://doi.org/10.1002/bdm.1753>
- Hank, P., & Baltes-Götz, B. (2019). The stability of self-esteem variability: A real-time assessment. *Journal of Research in Personality*, 79, 143–150. <https://doi.org/10.1016/j.jrp.2019.03.004>
- Johnson, T., Kulesa, P., Cho, Y. I., & Shavitt, S. (2005). The relation between culture and response styles: Evidence from 19 countries. *Journal of Cross-Cultural Psychology*, 36(2), 264–277. <https://doi.org/10.1177/0022022104272905>
- Joinson, A. N., Paine, C., Buchanan, T., & Reips, U.-D. (2008). Measuring self-disclosure online: Blurring and non-response to sensitive items in web-based surveys. *Computers in Human Behavior*, 24(5), 2158–2171. <https://doi.org/10.1016/j.chb.2007.10.005>
- Jones, G., & Stokes, A. (2009). *Online counselling: A handbook for practitioners*. Palgrave Macmillan.
- Kernis, M. H. (2005). Measuring self-esteem in context: The importance of stability of self-esteem in psychological functioning. *Journal of Personality*, 73(6), 1569–1605. <https://doi.org/10.1111/j.1467-6494.2005.00359.x>
- Lackner, N., Unterrainer, H.-F., & Neubauer, A. C. (2013). Differences in Big Five personality traits between alcohol and polydrug abusers: Implications for treatment in the therapeutic community. *International Journal of Mental Health and Addiction*, 11(6), 682–692. <https://doi.org/10.1007/s11469-013-9445-2>
- Langenfeld, T. (2020). Internet-based proctored assessment: Security and fairness issues. *Educational Measurement: Issues and Practice*, 39(3), 24–27. <https://doi.org/10.1111/emip.12359>
- Lawrence, A., Quist, J., & O'Connell, M. (2009, April). *Unproctored Internet Testing: Examining the impact of test environment*. Paper presented at the 24th Annual Conference of the Society for Industrial and Organizational Psychology, New Orleans, LA.
- Le Corff, Y., Gingras, V., & Busque-Carrier, M. (2017). Equivalence of unproctored Internet testing and proctored paper-and-pencil testing of the Big Five. *International Journal of Selection and Assessment*, 25(2), 154–160. <https://doi.org/10.1111/ijsa.12168>
- McCrae, R. R., & Allik, J. (2002). *The five-factor model of personality across cultures*. Kluwer Academic.
- Miller, E., Neal, D., Roberts, L., Baer, J., Cressler, S., Metrik, J., & Marlatt, G. (2002). Test-retest reliability of alcohol measures: Is there a difference between Internet-based assessment and traditional methods? *Psychology of Addictive Behaviors*, 16, 56–63. <https://doi.org/10.1037/0893-164X.16.1.56>
- Morelli, N. A., Mahan, R. P., & Illingworth, A. J. (2014). Establishing the measurement equivalence of online selection assessments delivered on mobile versus nonmobile devices. *International Journal of Selection and Assessment*, 22(2), 124–138. <https://doi.org/10.1111/ijsa.12063>
- O'Neil, K. M., & Penrod, S. D. (2001). Methodological variables in web-based research that may affect results: Sample type, monetary incentives, and personal information. *Behavior*

- Research Methods, Instruments, & Computers*, 33(2), 226–233. <https://doi.org/10.3758/bf03195369>
- Peer, E., Rothschild, D., Gordon, A., Evernden, Z., & Damer, E. (2021). Data quality of platforms and panels for online behavioral research. *Behavior Research Methods*. <https://doi.org/10.3758/s13428-021-01694-3>
- Randler, C. (2008). Morningness-eveningness, sleep-wake variables and Big Five personality factors. *Personality and Individual Differences*, 45, 191–196. <https://doi.org/10.1016/j.paid.2008.03.007>
- Reips, U.-D. (2002). Standards for Internet-based experimenting. *Experimental Psychology*, 49, 243–256. <https://doi.org/10.1026/1618-3169.49.4.243>
- Reips, U.-D. (2008). How Internet-mediated research changes science. In A. Barak (Ed.), *Psychological aspects of cyberspace* (pp. 268–294). Cambridge University Press. <https://doi.org/10.5167/uzh-4569>
- Reips, U.-D. (2013). Internet-based studies. In M. D. Gellman & J. R. Turner (Eds.), *Encyclopedia of behavioral medicine* (pp. 1097–1102). Springer. https://doi.org/10.1007/978-1-4419-1005-9_28
- Soldz, S., & Vaillant, G. E. (1999). The Big Five personality traits and the life course: A 45-year longitudinal study. *Journal of Research in Personality*, 33(2), 208–232. <https://doi.org/10.1006/jrpe.1999.2243>
- Soto, C. J. (2019). How replicable are links between personality traits and consequential life outcomes? The life outcomes of personality replication project. *Psychological Science*, 30(5), 711–727. <https://doi.org/10.1177/0956797619831612>
- Stachl, C., Au, Q., Schoedel, R., Buschek, D., Völkel, S., Schuwerk, T., Oldemeier, M., Ullmann, T., Hussmann, H., Bischl, B., & Bühner, M. (2020). *Behavioral patterns in smartphone usage predict Big Five personality traits*. PsyArXiv. <https://doi.org/10.31234/osf.io/ks4vd>
- Tavakol, M., & Dennick, R. (2011). Making sense of Cronbach's alpha. *International Journal of Medical Education*, 2, 53–55. <https://doi.org/10.5116/ijme.4dfb.8dfd>
- Templer, K. J., & Lange, S. R. (2008). Internet testing: Equivalence between proctored lab and unproctored field conditions. *Computers in Human Behavior*, 24(3), 1216–1228. <https://doi.org/10.1016/j.chb.2007.04.006>
- Tippins, N. T. (2006). Unproctored Internet testing in employment settings. *Personnel Psychology*, 59(1), 189–225. <https://doi.org/10.1111/j.1744-6570.2006.00909.x>
- Vecchione, M., Alessandri, G., & Barbaranelli, C. (2011). Paper-and-pencil and web-based testing: The measurement invariance of the Big Five personality tests in applied settings. *Assessment*, 19(2), 243–246. <https://doi.org/10.1177/1073191111419091>
- Zhang, Z., & Yuan, K. H. (2016). Robust coefficients alpha and omega and confidence intervals with outlying observations and missing data: Methods and software. *Educational and Psychological Measurement*, 76(3), 387–411. <https://doi.org/10.1177/0013164415594658>

History

Received January 29, 2021

Revision received May 23, 2021

Accepted July 28, 2021

Published online December 17, 2021

Conflict of Interest

We have no known conflict of interest to declare.

Funding

Open access publication enabled by the University of Duisburg-Essen.

Tobias Altmann

Department of Psychology
University of Duisburg-Essen
Universitätsstr. 2
45141 Essen
Germany
tobias.altmann@uni-due.de