Erschienen in: Behavior research methods ; 50 (2018), 6. - S. 2267-2275
https://dx.doi.org/10.3758/s13428-017-0991-6

Can smartphones be used to bring computer-based tasks from the lab to the field? A mobile experience-sampling method study about the pace of life

Stefan Stieger[1] · **David Lewetz**[2] · **Ulf-Dietrich Reips**[1]

Abstract Researchers are increasingly using smartphones to collect scientific data. To date, most smartphone studies have collected questionnaire data or data from the built-in sensors. So far, few studies have analyzed whether smartphones can also be used to conduct computer-based tasks (CBTs). Using a mobile experience-sampling method study and a computer-based tapping task as examples ($N = 246$; twice a day for three weeks, 6,000+ measurements), we analyzed how well smartphones can be used to conduct a CBT. We assessed methodological aspects such as potential technologically induced problems, dropout, task noncompliance, and the accuracy of millisecond measurements. Overall, we found few problems: Dropout rate was low, and the time measurements were very accurate. Nevertheless, particularly at the beginning of the study, some participants did not comply with the task instructions, probably because they did not read the instructions before beginning the task. To summarize, the results suggest that smartphones can be used to transfer CBTs from the lab to the field, and that real-world variations across device manufacturers, OS types, and CPU load conditions did not substantially distort the results.

Keywords Pace of life · Experience sampling · Smartphone · Well-being · Psychological pressure

✉ Stefan Stieger
stefan.stieger@uni-konstanz.de

[1] Department of Psychology, University of Konstanz, Konstanz, Germany

[2] Department of Psychology, University of Vienna, Vienna, Austria

In recent years, researchers have been increasingly using smartphones to collect scientific data (Dufau et al., 2011; Miller, 2012; Raento, Oulasvirta, & Eagle, 2009). Smartphones offer several advantages over conventional data collection devices (e.g., printed diaries; e.g., Harari et al., 2016) and have the potential to broaden knowledge about psychological concepts by allowing researchers to do research in the field instead of in the lab (Wrzus & Mehl, 2015; for an early similar attempt using a microphone sensor, see Mehl, Pennebaker, Crow, Dabbs, & Price, 2001). To date, smartphones have mostly been used to collect questionnaire data and/or data from the built-in sensors. In contrast, the potential for using smartphones to conduct *computer-based tasks* (CBTs)—that is, tasks that rely on computer programming to collect some form of nonquestionnaire data (e.g., reaction times or the results of a sorting task)—has been largely neglected. In fact, to date, only a handful of studies have used smartphones to conduct CBTs (Dufau et al., 2011; Kassavetis et al., 2016; Lee et al., 2016). Thus, there is little methodological information about how successfully smartphones can be used to collect such data. In the present study, we therefore use the example of a smartphone-based "tapping task" to explore how well smartphones can be used to conduct CBTs according to a range of criteria, including technologically induced dropout, task compliance, and the accuracy of time measurements.

CBTs are used widely in psychological research. Well-known examples of CBTs include the implicit association test (IAT; Greenwald, McGhee, & Schwartz, 1998) and the Stroop task (Stroop, 1935).[1] Meanwhile, many CBTs are designed so that the task can be accessed from the Internet and completed

[1] www.millisecond.com/download/library/

with a desktop computer. These CBTs are either created using specific Web browser plugins and technologies (e.g., Flash, Shockwave, Java Applet), or require participants to install specific players in order to perform the task on their own computers (e.g., Inquisit from Millisecond). Although these approaches to CBT design have their advantages, they also suffer from several drawbacks such as discontinued support of certain web technologies (e.g., Java Applets, Flash), the refusal of participants to install unknown software, or participants' inability to install software due to missing computer administrator rights. It is well established that CBT designs can lead to technologically induced dropout, which can potentially bias the results (e.g., Schwarz & Reips, 2001; Stieger, Göritz, & Voracek, 2011).

Unlike CBTs designed for desktop computers, CBTs designed for smartphones usually do not rely on an Internet browser. Instead, smartphone tasks typically use apps (i.e., applications), which are installed on the smartphone itself. Apps can be used to present questionnaires, retrieve information from the built-in sensors, store a participant's data and send them via the Internet or phone line to the researcher's server, and send bings/signals (i.e., reminders) in longitudinal studies (e.g., experience-sampling method [ESM] designs; Bolger & Laurenceau, 2013), along with many other functions. As such, smartphone apps can be used for a variety of functions, which can greatly increase the richness of the data that can be collected. For example, with digital devices such as personal digital devices (PDAs) or smartphones, it is possible to assess the time when the participant completes a questionnaire, offering the possibility to get more information about the participants' compliance (Stone, Shiffman, Schwartz, Broderick, & Hufford, 2002).

Interestingly, despite the potential advantages of smartphone data collection procedures, to date few studies have used smartphone apps to conduct CBTs (for exceptions, see Dufau et al., 2011; Kassavetis et al., 2016; Lee et al., 2016). Thus, currently little information is available about how successfully smartphones can be used to conduct CBTs from a methodological point of view. Dufau and colleagues conducted a lexical decision task using Internet-connected iPhone/iPad devices. They found that response time distributions in their online tasks were very similar to those found in lab studies, but it is unclear if this also applies to smartphones using an Android operating system (OS), which are heterogeneous with regard to OS type and manufacturer (Götz, Stieger, & Reips, 2017). Furthermore, it is largely unknown whether using smartphones to conduct CBTs might lead to technologically induced or design-specific dropout or measurement inaccuracy e.g., due to incompatibilities or tasks not being displayed as intended by the researcher. For example, Stisen and colleagues (2015) found substantial differences in the accuracy of the accelerometer sensor across different smartphone devices (i.e., there might also be differences in the accuracy and precision of different sensors).

In the present study, we addressed the gap in methodological knowledge about how well smartphones can be used to conduct CBTs. We embedded our methodological questions into a project about the pace of life and its correlates (e.g., Garhammer, 2002; Levine & Bartlett, 1984; Levine & Norenzayan, 1999; Rosa, 2003). Specifically, we conducted a smartphone app study using an ESM design. We assessed *pace of life* twice a day for three weeks. We measured pace of life in two ways. First, we used the classical direct approach by asking participants to use a visual analogue scale (VAS; Reips & Funke, 2008) to answer the question, "How do you perceive your pace of life at the moment?" Second, we developed a computer-based tapping task in which participants had to tap on the smartphone's touchscreen display for 10 s according to their current pace of life. Past research has shown that walking speed can be used as a measure of pace of life (e.g., Levine & Bartlett, 1984). Hence, a faster pace of life seems to be reflected in the speed of body movements. If this is the case, then tapping with one's finger might also be an indicator of one's pace of life. Tapping tasks in general have frequently been used in functional neuroimaging studies (e.g., Witt, Laird, & Meyerand, 2008) and within the medical sciences for assessing motor system abnormalities such as with Parkinson's disease (e.g., Lee et al., 2016).

We evaluated how well the CBT task could be implemented using smartphones, on the basis of a number of criteria. First, we explored whether there were indications of technological problems (e.g., technologically induced dropout). Given that CBTs are generally not as self-explanatory as simple questionnaires (e.g., sets of questions with a predefined answering format such as Likert-type scales), we assessed participants' task compliance. In the laboratory, experimenters can easily provide instructions and address any arising problems, but this is difficult when tasks are conducted on the Internet or in the field. With Internet-based tasks, participants are usually given written instructions, but it is questionable whether participants always comply with such instructions (see Stieger & Reips, 2010). Furthermore, we assessed the accuracy of millisecond measurements. Millisecond measurements formed the basis of our dependent variable (the number of finger taps, described below) and are a critical part of many other CBTs (e.g., reaction time tasks). The running of many parallel processes on a computer or smartphone can result in distorted time measurements. In lab experiments using desktop computers, it is possible to turn off any potentially interfering processes. The number of parallel processes can hardly be controlled in smartphone studies. Thus, inaccurate millisecond measurement represents a potential problem associated with using smartphones to conduct CBTs. Finally, we compared the predictive validity of the CBT to the VAS.

Method

Participants

The sample was recruited through word of mouth in southern Germany. A total of 295 people installed the study app and indicated informed consent. After excluding people who did not participate or only participated once during the longitudinal part of the study, the data from 246 participants remained for analyses (53% women, 44% men, 3% did not disclose their sex). Reported participant age ranged from 16 to 70 years ($M = 25.2$, $SD = 10.9$). About half of the participants indicated that they were students (54.1%).

The participants who dropped out of the study ($n = 49$) did not differ from the participants who filled in the longitudinal questionnaire at least twice ($n = 246$) with regard to reported sex ($\chi^2 < 0.1$, $df = 1$, $p = .88$; odds ratio $OR = 1.06$, 95% CI: 0.48, 2.33) or age (Mann–Whitney U: $z = -1.28$, $p = .20$). However, drop-outs were more frequently iOS users ($\chi^2 = 5.5$, $df = 1$, $p = .02$; odds ratio $OR = 2.23$, 95% CI: 1.13, 4.40).

Procedure

The smartphone study on pace of life used a mobile experience sampling methodology (mESM; real-time and multiple time point measurements using mobile devices). The app prompted participants to provide ratings twice a day. The longitudinal part of the study lasted three weeks for a total of 42 measurement occasions. Throughout the study, participants were in their natural surroundings and evaluated their present situation. Real-time data is usually more accurate than retrospective self-report data (e.g., Conner, Tennen, Fleeson, & Barrett, 2009). The design allowed us to analyze the methodological characteristics of the task longitudinally (e.g., whether possible reaction time measurement distortions appeared only once or systematically across time).

After completing the longitudinal part of the study, participants completed an Internet-based post-test questionnaire, which is not part of this study. Participation was compensated with either the chance to win one of two Amazon gift cards worth €20 each (entry was optional) or with course credit. The study was conducted in German.

Measures

Pace of life We used two different measures to assess pace of life. First, participants were asked "How do you perceive your pace of life at the moment?," and they used a VAS to indicate their answer (0 = *very slow*, 100 = *very fast*). Second, we developed a computer-based tapping task in which participants had to tap on the smartphone touchscreen display for 10 s according to their current pace of life (for screenshots, see Fig. 1). Participants first read a short paragraph describing how to perform the task (see Fig. 1, lower part). On their screen, participants saw a large green circle with the instruction "Tap! with the speed of your current pace of life." The tap counter started after the first tap on the green circle. After the first tap, the color of the circle changed from green to red, and the word "start" appeared to indicate that that the counter had started. After 10 s, the app automatically moved on to the next question and saved the data on the backend server. The data consisted of a stream of millisecond values for each tap. In a strict sense, each tap had two values, one for the duration the finger rested on the touchscreen, and one value for the duration the finger was away from the touchscreen. On the basis of the number of millisecond values, we could determine the number of taps, which was the measure of pace of life for all subsequent analyses. Additionally, the millisecond measurement allowed us to control whether the app really moved on to the next questionnaire page after 10 s.

Paradata: Time to complete the questionnaire[2] As an indirect measure of pace of life, we assessed the time that participants needed to complete the whole app questionnaire. Pace of life seems to be associated with a faster working speed (see Levine & Bartlett, 1984). Participants with a faster pace of life should therefore also complete the questionnaire more quickly than participants with a comparatively slower pace of life.

Correlates of pace of life from previous studies: Well-being and psychological pressure To explore the predictive validity of the tap measure, we assessed well-being and psychological pressure as two variables that had been related to pace of life in previous research (Garhammer, 2002). Well-being was assessed with the item, "How is your current well-being?" and a VAS (0 = *very bad*; 100 = *very good*). Psychological pressure was assessed with the item, "How strongly are you currently bothered by things you actually have to do, but haven't yet?" and a VAS (0 = *not bothered*, 100 = *very bothered*).

Smartphone app

We programmed a hybrid smartphone app called *pace of life*, which was designed for this project and made freely available through the Google App Store (for screenshots, see Fig. 1) as well as the Apple store. Hybrid apps differ slightly from native apps. Native apps are programmed in the preferred

[2] We also assessed the time to react to reminder as another potential measure of pace of life. Unfortunately, because of a JavaScript failure the timestamps could not be used. Furthermore, in line with the accuracy of public clocks from the study by Levine and Norenzayan (1999), we also assessed the accuracy of the smartphones' clocks. Basically, all smartphones set their clock either through the telephone network or the Internet, which resulted in a median deviation of only 1 s. Therefore, we did not use this additional measure for further analyses.

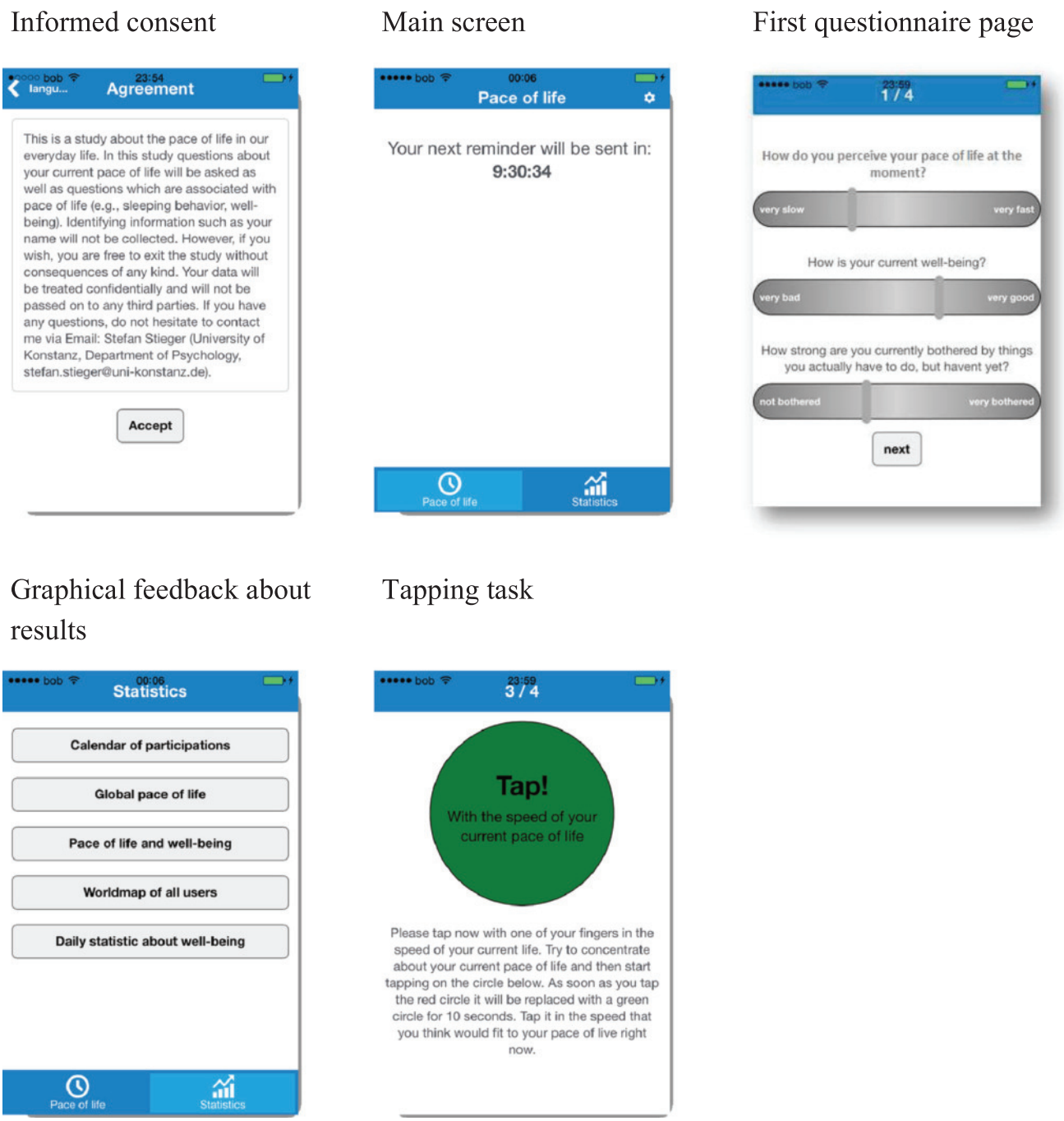

Fig. 1 Screenshots of the smartphone app

programming language of the smartphone OS (e.g., Java for Android). Hybrid apps are web applications (programmed in HTML and JavaScript), which run in a native app only to provide a frame for the web application. Participants could anonymously download the app directly onto their smartphones (82% used Google Android, 18% Apples iOS). Back-end server software was used for the communication with the app, specifically, to store the data and provide participants with personal statistical charts (e.g., their overall pace of life; daily well-being). To be able to merge datasets, the app generated a random and anonymous participant number after first installation.

When the app was first opened, participants had to provide informed consent and were asked about their basic demographics (age, sex, and country of origin). The first two screens (informed consent, demographics) were only presented once, during the first administration. Afterward, the main screen appeared. The main screen showed a counter indicating when the next measurement would take place. The app randomly produced a trigger (i.e., in-app reminder) to conduct the

measurement within two predefined time windows (morning 5:00 to 12:00; afternoon/evening 15:00 to 24:00). Participants could select their own time frame within these predefined times (i.e., personal time windows could only be smaller, not larger).[3] Furthermore, on the main screen, participants had the possibility to request personal statistics in a graphical format (for an example, see Fig. 1).[4]

When the app generated a trigger, the participant's smartphone alerted him or her that it was time to answer the study questions. After the participant tapped on the alert, the app automatically started, and the participant was presented with the study questions on three successive screens (see Fig. 1).

Statistical analyses

First, we examined how many participants dropped out of the study. We also analyzed whether technological problems with the tapping task might have led to dropout. The tapping task was on the third page and was followed by another page with questionnaire items. We thus compared the rate of missing data on the different pages. If missing data were higher on the fourth than on the other pages, it would suggest that the automatic forwarding after 10 s from the tapping task page to the last page did not work.

To assess task compliance, we examined the distributions and the descriptive statistics of the VAS and tap measures of pace of life. We examined the frequency of responses in the middle (~50) of the VAS as being potentially indicative of noncompliance (for a discussion of the unspecific middle-category with Likert-type scales, see Kulas & Stachowski, 2009). We used the frequency of very low numbers of taps (<4) as an indicator of noncompliance on the tapping task (for more details, see the gray rectangle in Fig. 2C). This procedure is supported by comparing the correlations between the VAS and the tapping task, both assessing pace of life (0–3 taps: $r = .017$, $p = .748$; >3 taps: $r = .331$, $p < .001$). We also examined whether the frequency of such low tap values changed over time. If the frequency of low tap values was due to participants not reading the instructions when first performing the task, the frequency of low tap values should decrease over time, as participants became aware of how the task worked. To explore how the frequency of low tap values changed over time, we first analyzed bar charts of the tap values (low, high) at each measurement occasion. We used a χ^2 test to evaluate whether the frequency of low tap values differed across measurement occasions. We calculated the Kendall's tau–b correlation for ordered variables between measurement occasion and tap value (low, high). We used the standardized residuals from the cross-tabulation to assess whether the residuals were higher or lower than would be expected for the measurement occasions at the beginning versus the end of the study. Finally, we examined the number of participants with high rates of low tap values across the duration of the study as an indication of either technological problems or noncompliance.

To assess millisecond measurement accuracy, we calculated how long a particular participant at a particular measurement occasion had his or her finger on and away from the touchscreen (i.e., the sum of all durations that the finger rested on the touchscreen, plus the sum of all durations the finger was away from the touchscreen). If the millisecond measurement was perfectly accurate, the total duration should sum up to 10,000 ms, the time at which the app automatically loaded the next questionnaire page. We examined the distribution and central tendencies of the total recorded time on and away from the touchscreen during the tapping task to determine the extent to which the millisecond measurements were accurate.

Finally, to examine the predictive validity of the tap measure and the VAS measure of pace of life, we first calculated the correlation between the two measures. Next, we assessed the extent to which fill-in time, psychological well-being, and psychological pressure predicted each of the two measures of pace of life. The distribution of fill-in time was skewed. We therefore log-transformed this variable prior to analysis (log + 1). We also included the time-dependent number of the assessment as a predictor to assess a possible effect of time on pace of life, in which case the separation of the between- from within-subjects variance would be biased but solvable through detrending (Curran & Bauer, 2011). Because the current dataset represents a mixture of independent (data across participants) as well as dependent (re-tests within participants) data, we used R (package "nlme") to calculate two multilevel models with random intercepts and coefficients and the two pace of life measures as the dependent variables. Daily observations (Level 1) were nested within participants (Level 2). To maximize information from the available data and to separate within- from between-participants variance, we followed the CWC(M) approach (i.e., centered within context with reintroduction of the mean at Level 2; Curran & Bauer, 2011). Specifically, we centered the Level 1 variables around the person-means (i.e., the personal average of each individual) and included

[3] During the installation process, 70 participants accepted the early time frame 5:00 to 12:00, and 101 participants accepted the later time frame 15:00 to 24:00 as is (61 participants accepted both time frames). All the other participants used the option to adjust the time frames. During the study, 57 participants changed the time frames; 45 changed them once, six changed them twice, and another six changed more than twice. The mean time frames chosen by all participants were from 7:36 (SD = 1:56) to 11:17 (SD = 1:24) and from 16:18 (SD = 1:48) to 22:10 (SD = 1:57).

[4] Participants requested the following graphics at least once: 86% the calendar, 83% the mean pace-of-life score across all participants, 72% a line chart of one's own well-being over the course of a day, 75% a scatterplot with a regression line displaying the association between pace of life and well-being, and 71% a world map with data about the participants' countries.

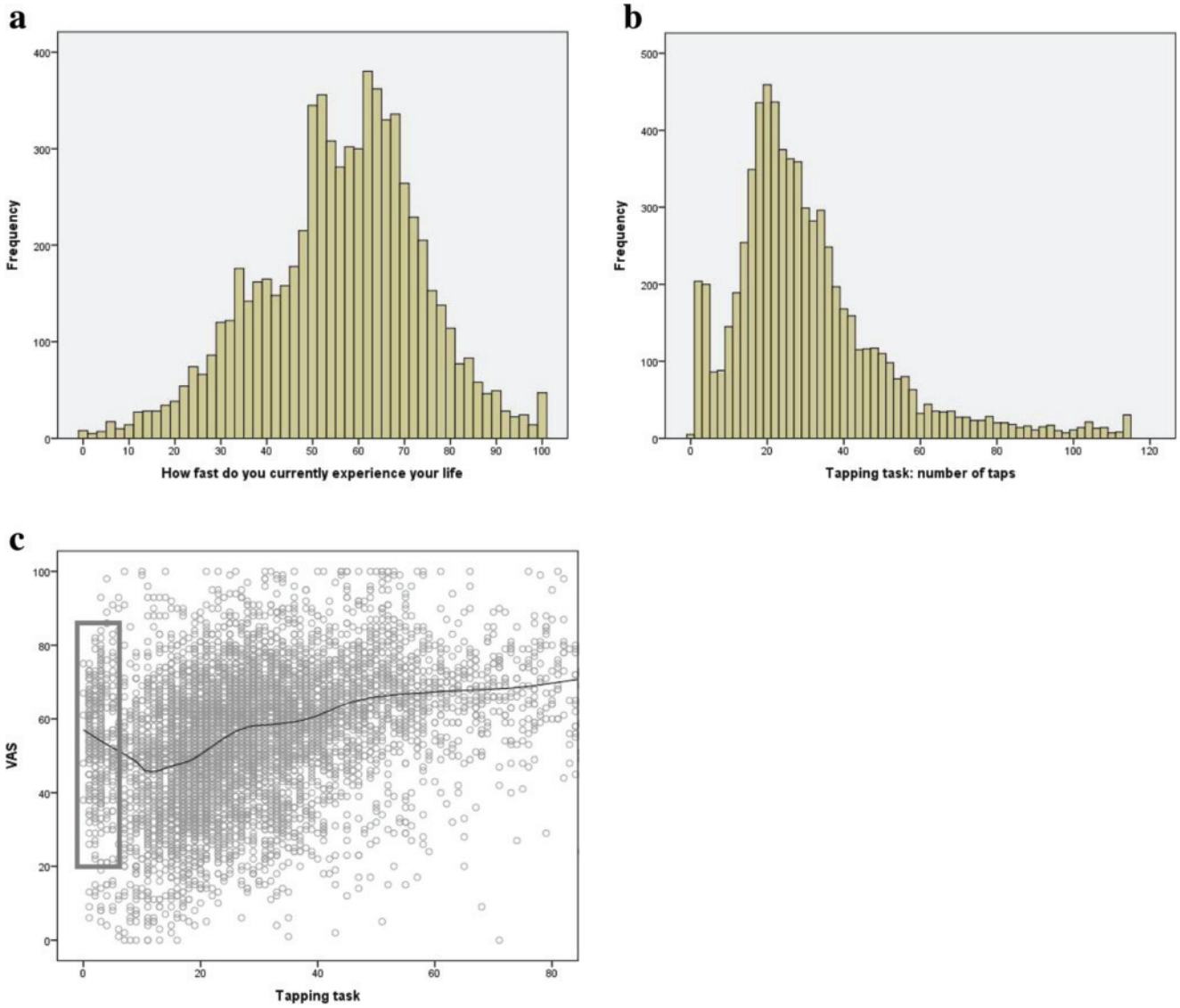

Fig. 2 Descriptive statistics of the visual analog scale (VAS) and tapping task measures of pace of life. (A) Frequency of pace-of-life values measured with the VAS, where higher values represent a faster pace of life. (B) Frequency of pace-of-life values measured with the tapping task (where higher values again represent a faster pace of life). (C) Association between the VAS and tapping task measures of pace of life (included is a nonlinear regression line with an Epanechnikov kernel density function of 30%). The gray rectangle represents a cluster of low tapping rates, possibly indicating noncompliance with the task

the person means as predictors at Level 2. Thus, the Level 1 variables capture the *within*-person variance (i.e., the extent to which a particular measurement deviated from that person's personal average) whereas the variable means on Level 2 capture the *between*-person variance (i.e., differences between participants). Because Level 1 variables represent data from multiple retests, which are often correlated, we controlled for autocorrelations. We compared the results of the multilevel model to the results of earlier research on pace of life (Garhammer, 2002). We also examined the intraclass correlations (ICC) for the VAS and tap measures to assess the extent to which the variance could be attributed to between- and within-subjects differences. Because we questioned the validity of low tap values (see above), we excluded cases with fewer than four taps from these analyses.

Results and discussion

Technologically induced problems and dropout

The dropout rate was very low for a longitudinal design. Only 14 participants dropped out before the end of the first week, and another 11 participants by the end of the second week. Of the 6,000+ completed questionnaires, there were only 18

instances in which a participant did not complete the fourth (last) questionnaire page. This suggests that the automatic forwarding after 10 s from the tapping task page to the last page functioned correctly.

Task compliance

As can be seen from Fig. 2 (panel A), the VAS data were just about normally distributed, but there were slightly more responses in the middle of the scale (= 50) than in a normal distribution. Participants may have more frequently used the middle category because they really did have a "medium" pace of life more frequently, or because they were not willing or able to provide information about their pace of life (i.e., a sign of noncompliance; Kulas & Stachowski, 2009).

Figure 2B displays the distribution of the number of taps from the tapping task. This distribution is rather skewed, with substantially more values between 1 and 3 than would be expected in a normal distribution. This probably occurred because some participants did not read the instructions (i.e., they first tapped the circle a couple of times, which started the counter, and *then* read the instructions). Furthermore, the high frequency of low tap values could also represent technical problems during the task (for more details, see Fig. 2C, gray rectangle).

Inspection of the bar charts indicated a high frequency of low tap values (<4) still at the last measurement occasion (graphs have been omitted for brevity). Thus, the bar graphs did not suggest that low tap values were due to a misunderstanding of the task instructions fading over time. The frequency of low tap values did not differ significantly across the different time points, $\chi^2 = 19.6$, $df = 43$, $p = .99$. However, we did observe a significant, positive Kendall's tau–b correlation between low-versus-high tap values, on the one hand, and measurement occasion, on the other hand ($r = .023$, $p = .028$). Specifically, there were more low tap values at the beginning of the study, although the magnitude of the relationship was of very small effect size. The standardized residuals from the cross-tabulation were higher than expected for the low taps at the beginning of the study (standardized residuals #1, #3, #5, and #6 were larger than 1), and lower than expected at the end of the study (e.g., 16 of the last 20 low tap counts had negative standardized residuals—i.e., counts were lower than expected). This again suggested that low tap values were more frequent at the beginning than at the end of the study. Nevertheless, low tap values were recorded throughout the entire course of the study (e.g., on the last day of the study, six participants [2.4%] still had fewer than four taps). Furthermore, 11 participants (4.5%) had very high rates of low tap values over the entire course of the study, potentially representing either technological problems or noncompliance.

Millisecond measurement accuracy

The median time recorded on and away from the touchscreen during the tapping task was 9,779 ms (M = 9,289, SD = 1,700). Closer inspection of the data revealed that the deviation from 10,000 occurred due to the last millisecond value sometimes not being stored because the app had already loaded the next question. In 306 cases (4.1%), the sum of all milliseconds was more than 10,000, with a median of 10,019 ms (M = 10,436, SD = 988; range 10,001 to 19,034). It is highly probable that values higher than 10,000 occurred because other processes were running on the smartphone's CPU, which led to a slowdown of the processor. In general, there were few cases in which the sum was greater than 10,000 ms, and the bias was very small (the mode was 10,001). This is indicative of the high accuracy of the millisecond measurements.

Predictive validity of the VAS and tapping task measures of pace of life

As expected, the tap values and VAS scores for pace of life were positively correlated ($r = .33$, $p < .001$). As can be seen in Table 1, there were significant relationships between psychological pressure, fill-in time, and time of assessment with within-person variation for both the VAS and tap measures of pace of life (Level 1). However, the magnitudes of the effects

Table 1 Results of the multilevel model analysis with visual analog scale (VAS) score and tap measures of pace of life as the dependent measures

MLM	Predictors	VAS β (standardized)	Tapping Task
Within subjects	Well-being	.003	– .011
	Psychological pressure	.165***	.079***
	Fill-in time	.027**	.014+
	Time of assessment	.086***	.057***
Between subjects	Well-being	.127***	– .011
	Psychological pressure	.393***	.241***
	Fill-in time	.013	– .066

*** $p < .001$, ** $p < .01$, * $p < .05$, + $p < .10$.

were substantially lower for the tap value than for the VAS measure. For example, within-person variation in psychological pressure explained 2.66% of the within-person variation of pace of life as measured with the VAS, but only 0.69% of the within-persons variation in pace of life as measured with the tapping task (~ 4 times less variance). Psychological well-being and psychological pressure were significant predictors of between-person variance in pace of life as measured with the VAS, but only psychological pressure was a significant predictor of pace of life as measured with the tapping task. Similar to the Level 1 results, psychological pressure on Level 2 explained 15.52% of the variance in pace of life *between* participants as measured with the VAS, but just 5.95% of the variance in pace of life between participants as measured with the tapping task (2.6 times less variance). Furthermore, the ICC for the VAS measure was 30.3%, whereas the ICC for the tapping task was 62.0%. In other words, most of the variance (~ 70%) in the VAS measure can be attributed to within-person differences (i.e., fluctuations in how a person rated his or her pace of life at different measurement occasions), whereas most of the variance in the tap measure stems from differences between people; that is, relative to the tap measure, the VAS measure was more sensitive to changes over time (fluctuating state aspects, as opposed to stable trait aspects of pace of life).

Conclusion

In the present study, we used a tapping task to measure pace of life to analyze whether smartphones can be used to successfully transfer CBTs from the lab to the field (e.g., Dufau et al., 2011). We found that the smartphone CBT functioned quite well. First, although some participants did not appear to read the instructions and/or did not comply with the task's requirements (for a similar result when using online questionnaires, see Stieger & Reips, 2010), task noncompliance was small and decreased over time. Nevertheless, researchers need to be aware that more complex CBTs might produce higher rates of noncompliance.

Second, we found low rates of dropout, and no evidence of more dropout on the tapping task page relative to the other pages. This is a very encouraging result because technology-induced dropout has been frequently found in online questionnaire studies using technologies other than HTML (Schwarz & Reips, 2001; Stieger et al., 2011). Nevertheless, we recommend that researchers using smartphones to conduct CBTs check dropout rates in great detail.

Third, in the present study we found evidence that smartphones could measure milliseconds quite accurately. Thus, other processes running in parallel to the study app did not appear to substantially influence the accuracy of smartphone millisecond measurements. This result is in line with the results of Dufau et al. (2011) who used iPads and iPhones to measure milliseconds. In the present study, participants had smartphones from many different manufacturers with many different Android operating system versions. The results with regards to measurement accuracy therefore appear to generalize to Android smartphones as well (see also Götz, Stieger, & Reips, 2017). Importantly, it seems that using smartphones as opposed to computers will not substantially affect the results of CBTs involving millisecond measurements (e.g., Implicit Association Test; Greenwald et al., 1998) or exact timing (e.g., speeded computer tasks; MouseTracker; Freeman & Ambady, 2010). Hence, smartphones appear to have considerable potential for allowing researchers to transfer CBTs from the lab to the field. Nevertheless, more systematic research will be needed before smartphones can be used for tasks in which single-millisecond measurements are of importance (for an example regarding Web experiments, see Keller, Gunasekharan, Mayo, & Corley, 2009).

Taken together, our results suggest that smartphones can be used effectively to conduct CBTs. The results should, however, be interpreted in light of some limitations. First, we programmed hybrid apps by using HTML5 and JavaScript. Further tests will be necessary to analyze the possible negative effects of using native apps that are programmed in the operating system's official programming language (e.g., Java for Android apps). Second, the observed methodological characteristics apply only to the tapping task used in the present study. Although the present results demonstrate the efficiency of tapping tasks, it may be more problematic to transfer other CBTs to smartphones (but see Dufau et al., 2011). Therefore, we cannot speculate in the generalizability of our results. Third, we could not compare the results from the smartphone study with the results of a similar study conducted in the laboratory. Lab studies are usually standardized not only regarding the procedure but also regarding the technology used (e.g., only one computer and operating system type). Therefore, some differences are to be expected.

Despite these limitations, we believe that the present results point to the potential of using smartphones to conduct CBT studies. The results suggest that variations in smartphone manufacturers, OS types, and CPU load conditions probably do not substantially distort the results of CBTs when transferred from the lab to the field.

References

Bolger, N., & Laurenceau, J.-P. (2013). Intensive longitudinal methods: An introduction to diary and experience sampling research. New York, NY: Guilford.

Conner, T. S., Tennen, H., Fleeson, W., & Barrett, L. F. (2009). Experience sampling methods: A modern idiographic approach to personality. *Social and Personality Psychology Compass*, *3*, 292–313. https://doi.org/10.1111/j.1751-9004.2009.00170.x

Curran, P. J., & Bauer, D. J. (2011). The disaggregation of within-person and between-person effects in longitudinal models of change. *Annual Review of Psychology*, *62*, 583–619. https://doi.org/10.1146/annurev.psych.093008.100356

Dufau, S., Duñabeitia, J. A., Moret-Tatay, C., McGonigal, A., Peeters, D., Alario, F.-X., . . . Grainger, J. (2011). Smart phone, smart science: How the use of smartphones can revolutionize research in cognitive science. *PLoS ONE*, *6*, e24974. https://doi.org/10.1371/journal.pone.0024974

Freeman, J. B., & Ambady, N. (2010). MouseTracker: Software for studying real-time mental processing using a computer mouse-tracking method. *Behavior Research Methods*, *42*, 226–241. https://doi.org/10.3758/BRM.42.1.226

Garhammer, M. (2002). Pace of life and enjoyment of life. *Journal of Happiness Studies*, *3*, 217–256. https://doi.org/10.1023/A:1020676100938

Götz, F. M., Stieger, S., & Reips, U.-D. (2017). Users of the main smartphone operating systems (iOS, Android) differ only little in personality. *PLoS ONE*, *12*, e0176921. https://doi.org/10.1371/journal.pone.0176921

Greenwald, A. G., McGhee, D. E., & Schwartz, J. L. K. (1998). Measuring individual differences in implicit cognition: The Implicit Association Test. *Journal of Personality and Social Psychology*, *74*, 1464–1480.

Harari, G. M., Lane, N. D., Wang, R., Crosier, B. S., Campbell, A. T., & Gosling, S. D. (2016). Using Smartphones to collect behavioral data in psychological science: Opportunities, practical considerations, and challenges. *Perspectives on Psychological Science*, *11*, 838–854. https://doi.org/10.1177/1745691616650285

Kassavetis, P., Saifee, T. A., Roussos, G., Drougkas, L., Kojovic, M., Rothwell, J. C., . . . Bhatia, K. P. (2016), Developing a tool for remote digital assessment of Parkinson's disease. *Movement Disorders Clinical Practice*, *3*, 59–64. https://doi.org/10.1002/mdc3.12239

Keller, F., & Gunasekharan, S., Mayo, N., & Corley, M. (2009). Timing accuracy of Web experiments: A case study using the WebExp software package. *Behavior Research Methods*, *41*, 1–12. https://doi.org/10.3758/BRM.41.1.12

Kulas, J. T., & Stachowski, A. A. (2009). Middle category endorsement in Likert-type response scales: Associated item characteristics, response latency, and intended meaning. *Journal of Research in Personality*, *43*, 489–493. https://doi.org/10.1016/j.jrp.2008.12.005

Lee, C. Y., Kang, S. J., Hong, S.-K., Ma, H.-I., Lee, U., Kim, Y. J. (2016). A validation study of a smartphone-based finger tapping: Application for quantitative assessment of bradykinesia in Parkinson's disease. *PLoS ONE*, *11*, e0158852. https://doi.org/10.1371/journal.pone.0158852

Levine, R., & Bartlett, K. (1984). Pace of life, punctuality and coronary heart disease in six countries. *Journal of Cross-Cultural Psychology*, *15*, 233–255.

Levine, R. V., & Norenzayan, A. (1999). The pace of life in 31 countries. *Journal of Cross-Cultural Psychology*, *30*, 178–205. https://doi.org/10.1177/0022022199030002003

Mehl, M. R., Pennebaker, J. W., Crow, D. M., Dabbs, J., & Price, J. H. (2001). The Electronically Activated Recorder (EAR): A device for sampling naturalistic daily activities and conversations. *Behavior Research Methods, Instruments, & Computers*, *33*, 517–523. https://doi.org/10.3758/BF03195410

Miller, G. (2012). The smartphone psychology manifesto. *Perspectives on Psychological Science*, *7*, 221–237. https://doi.org/10.1177/1745691612441215

Raento, M., Oulasvirta, A., & Eagle, N. (2009). Smartphones: An emerging tool for social scientists. *Sociological Methods and Research*, *37*, 426–454. https://doi.org/10.1177/0049124108330005

Reips, U.-D., & Funke, F. (2008). Interval-level measurement with visual analogue scales in Internet-based research: VAS generator. *Behavior Research Methods*, *40*, 699–704. https://doi.org/10.3758/BRM.40.3.699

Rosa, H. (2003). Social acceleration: Ethical and political consequences of a desynchronized high-speed society. *Constellations*, *10*, 3–33.

Schwarz, S., & Reips, U.-D. (2001). CGI versus JavaScript: A Web experiment on the reversed hindsight bias. In U.-D. Reips & M. Bosnjak (Eds.), Dimensions of Internet science (pp. 75–90). Lengerich, Germany: Pabst.

Stieger, S., Göritz, A. S., & Voracek, M. (2011). Handle with care: The impact of using Java applets in web-based studies on dropout and sample composition. *Cyberpsychology, Behavior, and Social Networking*, *14*, 327–330.

Stieger, S., & Reips, U.-D. (2010). What are participants doing while filling in an online questionnaire: A paradata collection tool and an empirical study. *Computers in Human Behavior*, *26*, 1488–1495.

Stisen, A., Blunck, H., Bhattacharya, S., Prentow, T. S., Kjærgaard, M. B., Dey, A., . . . Jensen, M. M. (2015). Smart devices are different: Assessing and mitigating mobile sensing heterogeneities for activity recognition. In J. Song, T. Abdelzahar, & C. Mascolo (Eds.), Proceedings of the 13th ACM Conference on Embedded Networked Sensor Systems (SenSys) (pp. 127–140). New York, NY: Association for Computing Machinery. https://doi.org/10.1145/2809695.2809718

Stone, A. A., Shiffman, S., Schwartz, J. E., Broderick, J. E., & Hufford, M. R. (2002). Patient noncompliance with paper diaries. *British Medical Journal*, *324*, 1193–1194. https://doi.org/10.1136/bmj.324.7347.1193

Stroop, J. R. (1935). Studies of interference in serial verbal reactions. *Journal of Experimental Psychology*, *18*, 643–662. https://doi.org/10.1037/0096-3445.121.1.15

Witt, S. T., Laird, A. R., & Meyerand, M. E. (2008). Functional neuroimaging correlates of finger-tapping task variations: An ALE meta-analysis. *NeuroImage*, *42*, 343–356. https://doi.org/10.1016/j.neuroimage.2008.04.025

Wrzus, C., & Mehl, M. R. (2015). Lab and/or field? measuring personality processes and their social consequences. *European Journal of Personality*, *29*, 250–271. https://doi.org/10.1002/per.1986