methods, data, analyses | 2022, pp. 1-29 DOI: 10.12758/mda.2022.10

Measuring Students‘ Reading Behavior with an Ambulatory Assessment – A Field Report on a Smartphone-Based Reading Diary Study

Franziska Maria Locher[1], *Verena Angelika Schnabel*[2], *Valentin Unger*[1] *& Maximilian Pfost*[2]

[1] *Institute for Educational Assessment, St. Gallen University of Teacher Education*

[2] *Department of Educational Research, University of Bamberg*

Abstract

In prior research, reading behavior was predominantly measured using either a questionnaire, which is economical and easy to implement but imprecise, or paper-pencil diaries that document reading behavior quite accurately, but which are time consuming and costly. The present study aims to introduce and evaluate a precise and easy to implement measure of reading behavior, namely a reading diary app in which participants can record their reading behavior on a smartphone. To evaluate the development procedure, the first research question asked whether data gathered with the app is of high quality (e.g., reliability). The second research question asked how reading time recorded via the app is related to reading time assessed via different retrospective questionnaires. $n = 31$ German university students recorded their reading activities for 14 days. Different approaches were applied to estimate the data quality and reliability and yielded satisfactory results. Participants reported more time spent reading daily on the retrospective questionnaire than when recording their reading time using the app. The correlation between reading diary app data and questionnaire data was medium in size. Our findings are discussed in the light of future directions for reading research and the use of ambulatory assessments.

Keywords: ambulatory assessment, reading diary, reading behavior, smartphone app, field report

Being able to effectively process written information is essential for cultural, social, and economic participation in our society. Reading facilitates self-exploration and self-enrichment. Therefore, reading competence is a central skill for today's society (e.g., Alexander, 2005; Artelt et al., 2001; Becker-Mrotzek et al., 2015; Marshall, 2000). The PIRLS 2021 study defines reading literacy as a functional construct capturing readers' ability to process written language in order to achieve personal or socially defined goals. It includes reading to learn from texts, reading to participate in society and reading for enjoyment (e.g., Mullis & Martin, 2019, see also OECD, 2019, for the PISA framework). The reading literacy construct encompasses both cognitive (knowledge and skills) and affective-motivational aspects of reading.

Reading behavior, defined as the sum of all activities related to reading (i.e., time spent reading, amount of reading, or being read to aloud in early childhood), is an important predictor of reading skill development. Many studies have provided convincing evidence of the positive relation between reading skills and reading behavior across the life course (e.g., Burgess, Hecht, & Lonigan, 2002; Bus, van IJzendoorn, & Pellegrini, 1995; Guthrie et al., 1999; Locher & Pfost, 2020; Mol & Bus, 2011; Pfost, Dörfler, & Artelt, 2013). However, beyond the well-replicated general finding of a positive relation between time spent reading and reading skills, there are still large areas of uncharted territory. For instance, there is scare evidence on what kind of reading material (e.g., with respect to text difficulty, content, type of text, writing style) individuals should read to facilitate the optimal development of their reading skills and reading motivation (e.g., Troyer et al., 2018). Thus, gaining deeper knowledge about the nature of people's reading development is of major interest to researchers and practitioners. This concerns above all reading behavior, which, as described above, is one of the most important predictors of reading skills. People at all stages of reading literacy development can face difficulties while reading: for example, while beginning readers might struggle to decode letters, advanced readers might struggle to extract information and construct meaning from the text (e.g., Chall, 1983; Kutner et al., 2007; OECD, 2021). Thus, it is important that research on reading does not end in adolescence. The better researchers

Acknowledgements

This research was supported by grants from the Bamberg Graduate School of Social Sciences (BAGSS) and the German Research Foundation (DFG: Grant number PF 840/2-2). We have no conflicts of interest to disclose.

We want to thank Peter Kuntner for his technical support in developing the reading diary app.

Direct correspondence to

Franziska Maria Locher, Institute for Educational Assessment,
St. Gallen University of Teacher Education, 9000 St.Gallen, Switzerland
E-mail: franziska.locher@phsg.ch

understand reading at different stages of individual development, the better interventions or instructional materials practitioners can develop to support readers in facing the challenges they encounter in later stages (e.g., Alexander, 2005).

Currently, in reading research as well as in psychology in general, most studies use global retrospective self-report data from questionnaires (e.g., an evaluation of average reading time per week; Fahrenberg et al., 2007b). Research questions such as the one above, however, can only be answered by taking a closer look at individuals' "real" reading activities, rather than merely relying on global retrospective measures that only provide information about average trends.

Continuous assessments of people's reading behavior (e.g., daily reading diaries) are seldom used because daily logs tend to be very time-consuming and can be a huge burden for participants, especially in paper-pencil studies. Therefore, the goal of this study was to develop a reading diary app for participants to record their reading behavior (e.g., reading time, reading material) and reading motivation with a smartphone in an economical way. In addition, the present study explores the reading behavior data collected via this smartphone app. The data on reading motivation will be analyzed in a further research project. The first aim was to examine the quality of the data (e.g., reliability) gathered with the reading diary app (ambulatory assessment). The second aim was to investigate how reading time assessed with the reading diary app is related to reading time assessed with global and retrospective questionnaire measures.

Theoretical Background

Conceptualization of Reading Behavior and How to Measure it

Reading behavior can be defined as the sum of all activities related to reading. As this definition can potentially include a wide range of reading-related activities, previous studies have operationalized reading behavior in many different ways. In order to clarify the concept of reading behavior, it seems worthwhile to differentiate between the quantitative aspects ("How much do people read?") and qualitative aspects ("What do people read?") of reading (Locher, Becker, & Pfost, 2019a). Quantitative aspects of reading behavior refer to the amount or volume of reading (e.g., number of books read in the last month) or time spent reading. Qualitative aspects of reading behavior are multifaceted. They comprise information about the nature of the reading material (e.g., type of text, text difficulty, text content, or medium, i.e., print or digital). Common ways to measure the quantitative aspect of reading behavior are global and retrospective self-reports of reading time (e.g., "About how much time do you usually spend reading outside of school?") such as those used in PISA (Programme for International Student Assessment; OECD,

2010), one of the largest and perhaps most well-established international large-scale comparison studies in the field of education.

However, in addition to these global self-report scales of time spent reading, recent research has provided evidence of differential effects of reading different types of texts on variables such as reading motivation or reading skills (e.g., Locher et al., 2019a; Jerrim & Moss, 2019; McGeown et al., 2015; McGeown et al., 2016; Pfost et al. 2013). For example, reading traditional fiction books (e.g., novels, short stories or tales) has been found to be more important for reading skill development than reading comics and newspapers or online media (e.g., Pfost et al., 2013). The finding that the type of text moderates the relation between reading behavior and reading skills was further supported by Jerrim and Moss (2019) in an analysis of PISA data: students who frequently read fiction books had better reading skills than their peers who do not read fiction. The authors did not find such an effect for other text types, such as magazines or non-fiction. Furthermore, with respect to reading motivation, recent research provides first evidence that reading classic literature, especially in comparison to modern fiction books, negatively relates to intrinsic situational reading motivation (Locher et al., 2019a). In addition, within the school context, students who read more difficult books were less motivated to read (Locher et al., 2019a). These results illustrate that more detailed insight into reading behavior is desirable.

Exploring qualitative aspects of students' reading behavior has often been neglected, probably because assessing such information comes at a high cost. Therefore, measures capturing the amount of time people spend reading different types of texts are a good complement to the global evaluation (Locher & Pfost, 2019b). Beyond the quantitative aspect of reading time, these measures provide at least some additional information about the average amount of time individuals spend reading fiction books, nonfiction books, newspapers, or other text types. Nevertheless, whether global retrospective self-reports from questionnaires (global and text type-specific measures of people's reading time) accurately capture individuals' behavior is doubtful, because this methodology "records mental representations rather than the actual experience and behavior" (Fahrenberg et al., 2007b, p. 207), and several potential biases might occur (Fahrenberg et al., 2007a). Data can be affected by cognitive schemata, response tendencies, judgment heuristics, or memory effects (Fahrenberg et al., 2007a; Gershuny, 2012). For instance, when determining daily reading time, some people might use the last week as a reference, whereas others might use the past month. Another source of bias might arise when the days selected are not representative with respect to the behavior being assessed (Kan & Pudney, 2008). The experience of time is also subjective, as persons perceive time use differently (Juster, Ono, & Stafford, 2003). This means that individuals' responses to a question about their normal daily reading time might be based on different heuristics. Moreover, individuals might only remember lon-

ger reading activities (e.g., reading a book for 3 hours on the weekend) and fail to recall brief reading activities (e.g., reading a newspaper for 5 minutes in a waiting room). The fact that people might not consider all of their reading activities might lead to biases in response behavior (memory effect). Another issue is that recalling all reading activities correctly and assigning them to the appropriate text category listed in the questionnaire can be a very difficult task, especially for children and young adolescents (Locher & Pfost, 2019b).

Thus, alternative approaches are required to obtain data that better captures a person's actual reading behavior. Options used in research fields such as psychology include the experience sampling method, which asks about experiences in the moment (e.g., Csikszentmihalyi & Larson, 2014; Hektner, Schmidt, & Csikszentmihalyi, 2007; Shumow, Schmidt, & Kackar, 2008; Zirkel, Garcia, & Murphy, 2015), as well as the day reconstruction method, in which participants reflect on their activities that day (e.g., Kahneman et al., 2004; Lucas et al., 2019). A third way to gather information that more closely approximates people's "real" reading behavior and is less error-prone is to use reading diaries in which people document their reading activities, namely how long and which books, magazines, newspapers, or other texts they read (e.g. Akbar et al., 2015; Anderson, Wilson, & Fielding, 1988; Nieuwenboom, 2008; Stoffelsma, 2018). Reading diaries are often seen as a kind of gold standard because they offer a quite precise documentation of people's reading behavior, provide concrete information about the books or texts the person read, and yield information that can be used in further analyses.

Paper-Pencil versus Digital Diaries – Using an Ambulatory Assessment

Most daily diary studies in psychology and educational research have relied on paper-pencil methods (Akbar et al., 2015; Bolger, Davis, & Rafaeli, 2003; Fahrenberg et al., 2007b; Wilhelm, Perrez, & Pawlik, 2012). However, this method is quite time- and space-consuming and can also be a huge burden for participants (Bolger et al., 2003). For example, people have to carry their documents/diaries with them at all times, or might not have their diary available when they need it due to the cumbersome nature of the paper documents. In the worst case, this results in missing information. Another possibility is that people enter their reading activities later. However, filling out the diary after too much time has passed increases the risk of a retrospective bias, as people have to estimate their activities (Bolger et al., 2003; Wilhelm et al., 2012). This drawback also applies to the day reconstruction method and so-called "end-of-day diaries", in which all activities are documented once a day. Therefore, although reading diaries are seen as the gold standard, they are seldom used as a method for continuously assessing people's reading behavior. One way to deal with the issues associated with the paper-pencil method is

electronic documentation of reading behavior via an ambulatory assessment. This promising and innovative method "refers to the use of computer-assisted methodology for self-reports, behavior records, or physiological measurements, while the participant undergoes normal daily activities" (Fahrenberg et al., 2007b, p. 206). In other words, ambulatory assessments aim to conduct research (i.e., monitoring people's psychological, emotional, behavioral, or biological processes) in daily life and in people's natural environment with digital assistance (Trull & Ebner-Priemer, 2014). In particular, smartphones have become more and more important for ambulatory assessments because a large amount of data can be collected very economically and easily via apps. In addition, there is no need to carry around anything extra, like a paper-pencil diary (Conner & Lehman, 2012; Miller, 2012; Trull & Ebner-Priemer, 2014; Zhang et al., 2018). Due to these advantages, apps have been used to measure behavior in various disciplines, for example in the field of health (e.g., Ahram, 2019; Ebner-Priemer et al., 2007; Glomann et al., 2019; McLaws et al., 1990).

In summary, an ambulatory assessment to assess reading behavior (e.g., a specific reading diary app) has numerous advantages over paper-pencil diaries. First, most adolescents and young adults use smartphones and therefore already carry one with them at all times (Lampert, Sygusch, & Schlack, 2007). This means that data can be collected in daily life at low cost, and participants do not need extra materials or extra recording devices or computers (Fahrenberg et al., 2007b). Second, participants can easily fill out the reading diary whenever and wherever they want. This means that it is much more convenient and requires less effort for participants to document their activities, presumably leading to better data quality. Third, questions and questionnaires can be adapted based on people's responses, opening up a broader range of possibilities and creating flexibility (Fahrenberg et al., 2007b). For example, participants can only be given information and questions that are relevant for them. This approach reduces the text load in digital reading diaries and may result in less workload, as everything "unimportant" can be hidden and data entry is made easier. Fourth, with paper-pencil diaries, researchers would likely never know if participants filled out the whole diary retrospectively at the very end of the study. In ambulatory assessment/reading diary app, researchers have better control over the timing and reliability of the entries, meaning that the use of electronic data should result in higher compliance (Bolger et al., 2003; Fahrenberg et al., 2007b). And fifth, it might be easier to recruit more study participants for studies using smartphone apps than paper-pencil diaries (Zhang et al., 2018). This could directly affect the generalizability of the diary data collected.

Of course, the use of diaries also comes with several challenges. Depending on the type of research question, software and programming costs can occur (Conner & Lehman, 2012). Thus, developing an app could be more costly and would probably require more resources than developing and implementing a questionnaire

or a paper-pencil reading diary. In addition, completing a digital diary requires certain technical skills that are not necessarily present in all individuals within society and thus may lead to sampling bias. Finally, it should be noted that apps can always produce errors (e.g., in data storage or data transmission).

Despite the huge potential of ambulatory assessments, we do not know of any studies that have used this approach to collect data on people's reading behavior in daily life. At least two exploratory studies used electronic diaries to assess reading behavior. Keller (2010) had 12 college students document their reading behavior over three days by taking digital photographs of their reading activities (e.g., pictures of books). Raith (2008) had ninth-grade students use weblogs to document their reading behavior. This qualitative study compared paper-pencil and weblog reading dairies, finding that students who documented their reading behavior with weblogs reflected on the content of the book better than students who documented the content with paper-pencil diaries. However, both studies used desktop computers and did not leverage the possible advantages of ambulatory assessment using personal smartphones to monitor daily reading behavior (e.g., flexibility, diary always readily available, thus yielding a large amount of precisely documented data).

Measurment and Data Quality in Ambulatory Assessments

A sufficient measurement and data quality is an important precondition that is needed for further analyses and the correct interpretion of results. Therefore, as for any empirical measure, in ambulatory assessment studies it is important to evaluate the quality of the collected diary data and measures used (Calamia, 2019). To date, only a small number of studies using ambulatory assessments have reported quality criteria such as the reliability of the self-reported data (Calamia, 2019). For instance, well-regarded diary studies such as Greaney and Hegarty (1987) or Allen, Cipielewski, and Stanovich (1992) lack information on measurement and data quality. This might be because there are no clear standards for evaluating measures and data quality within ambulatory assessments like there are for survey scales or test development.

One way of evaluating data quality in ambulatory assessments is to use postmonitoring interviews (e.g., reaction quesionnaires: Nieuwenboom, 2008; Stone et al., 2003). In such questionnaires, which are conducted after the monitoring period, participants answer questions about, for instance, whether they found the ambulatory assessment to be a huge burden, whether they think they behaved differently in some situations because of the ambulatory assessment, or whether they think their behavior differed from their average behavior throughout the ambulatory assessment. If this is the case (i.e., the majority of participants agree with the statements), the researcher must conclude that the data quality is not acceptable. Postmonitoring

questionnaires therefore provide some indication of the reliability and validity (e.g., does the fact of observation change the nature of the phenomenon being measured) or generalisability of a given assessment.

Another method for demonstrating measurement and data quality in reading diaries or ambulatory assessments (e.g., Anderson et al., 1988; McLaws et al., 1990; Nieuwenboom, 2008) is to demonstrate the reliability of the reading diary app by examining correlations in the goal construct (e.g., reading time) between the observed weeks as well as between even- and odd-numbered days. This approach is similar to the idea of split-half reliability, where a test is divided into two halves, scores on which should be correlated with one another. Furthermore, reliability can be demonstrated by examining internal consistency (e.g., Cronbach's alpha values for the amount of time spent reading per day across all days).

Study Aims and Research Questions

Two research questions were formulated.

(a) First: Does the ambulatory assessment of reading behavior via a smartphone-based diary app have satisfactory measurement and data quality? For an optimal quality check, existing approaches used in previous studies were combined. Measurement and data quality can be assumed to be sufficient when two conditions hold: 1) postmonitoring questionnaire results reveal no issues, and 2) the reading diary app measure is sufficiently reliable. The first condition is achieved if only a small proportion of study participants report irregularities and problems with app use (e.g. Category 4: "I very often forgot to make an entry in the diary"). With respect to the second condition, in accordance with previous research examining the reliability of diary measures (e.g., Anderson et al., 1988; McLaws et al., 1990; Nieuwenboom, 2008), the reading diary app measure can be assumed to be reliable if the correlation in reading time between Week 1 and Week 2 and between even- and odd-numbered days is around $r = .70$ or higher. Finally, our reading diary app measure can be assumed to be reliable if Cronbach's alpha (for the amount of time spent reading per day across all days) exceeds $\alpha = .80$.

(b) Second: How is reading time measured via the ambulatory assessment related to reading time measured via different global retrospective questionnaire measures? An important criterion for the quality of an instrument is construct validity, and convergent validity is one way to check for construct validity. Convergent validity refers to overlap in the results of different tests for the same or similar constructs (Moosbrugger & Kelava, 2012). Thus, high correlations between the results of two tests or measures of the same construct reflect high convergent validity (Pospeschill, 2010). Global retrospective questionnaire measures are widely used and therefore can be considered well established; thus, they seem to be well-suited

as a criterion for testing the quality of our reading diary app data. According to the literature, however, data obtained from reading diaries most closely reflect “real” reading behavior. Therefore, low correlations between these two measures may not necessarily indicate low validity of the reading diary app data, but may also indicate that the two instruments partially measure different constructs. Consequently, this second research question is examined in an exploratory manner.

Method

Participants

All analyses rely on data from a convenience sample of $n = 31$[1] German university students (77% women) with a mean age of 20.71 ($SD = 2.60$) years. The university students were in their third semester of higher education on average ($M = 3.01$); 23% had an immigrant background, meaning that at least one parent was born abroad. Nearly all students ($n= 28$) had taken courses in the fields of psychology and education science. The participants had received an average grade of 2 (10-12 points) in German language arts on their secondary school completion exams (Abitur), which reflects “good” performance according to the German grading system.

Study Design

The study included three measurement points.

First Measurement Point (M1). The first measurement point involved a one-hour session in small groups of three to nine study participants. A reading achievement test was administered at the beginning of the session. Afterwards, participants filled out a questionnaire with different reading behavior measures, which lasted about 15 minutes. Afterwards, the reading diary app was introduced and explained. The app was installed on the smartphones the participants had brought with them, and the app’s functions were tested. Finally, the further study procedure was explained and participants were instructed on how to use the app (see measures section for further details). The first session followed a standardized script.

Second Measurement Point (M2). The second measurement point represents the ambulatory assessment period via the reading diary app. For the ambulatory

1 The study began with a total of 35 participants. Three participants could not install the app on their smartphone and could thus not further participate in the study. It later turned out that the reading diary app did not work on devices with an older Android operating system. Of the 32 participants who used the reading diary app during the two-week survey period, one person did not fill out the final questionnaire and therefore was not considered either in the further analyses.

assessment, we followed an event-based design in which every reading activity was documented immediately after participants completed a reading event. All participants recorded their reading activities for 14 days. Based on arguments made by Foasberg (2014), the start of the ambulatory assessment period was placed in the middle of the semester so that there would be no bias due to final exam stress.

Third Measurement Point (M3). The day after the reading diary app was used for the last time (Day 15), participants received a link via e-mail to an online questionnaire created using SoSci Survey (Leiner, 2019). The questionnaire included different reading behavior measures as well as the postmonitoring questionnaire. The participants had four days to complete the online questionnaire, which lasted about ten minutes. The incentives for complete participation were made available the following week. We elected to conduct an online survey at M3. Participants were not required to return to the lab, reducing the effort required. This was of high importance in order to avoid non-response and missing data in the postmonitoring questionnaire.

The study was advertised and participants were recruited in university seminars, lectures and via the student council email list at the University of Bamberg. For technical reasons, only students with Android devices could participate in the study. For complete participation in the study, students received a 15€ voucher for a local bookstore. Bachelor's degree students in psychology could alternatively opt to receive four credit hours for their participation.

Measures

Reading diary app data

The reading diary app was developed by the authors of the present paper at the Department of Educational Research at the University of Bamberg for the purpose of the present study. The app has a clear structure and can be used intuitively[2]. As previously mentioned, the study followed an event-based design, meaning that participants were to document their reading behavior directly after each reading activity occurred. Personal communication such as emails and text messages were not to be taken into account. No restrictions were applied concerning text type, medium (print or digital device), or whether the reading activity was for enjoyment or for one's studies, and participants were requested to document all reading activities. Furthermore, browsing and looking things up on the internet was not explicitly excluded.

2 At M3, participants answered questions about user friendliness (e.g. "Scanning new books worked without problems", "The entries for reading were uncomplicated"). The feedback was good. On a four-point Likert scale ("strongly disagree" to "strongly agree"), participants gave each statement an average rating of $M = 3.0$.

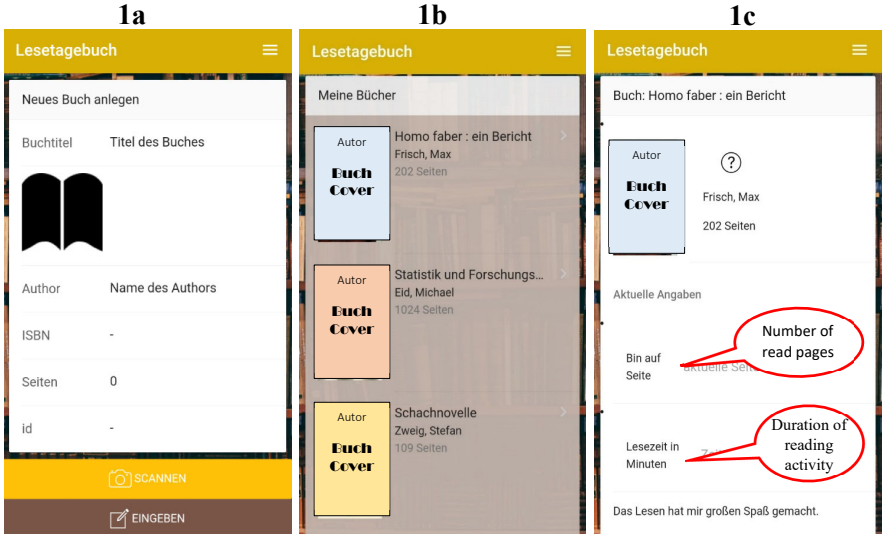

Figure 1 Screenshots showing the interface of the reading diary app

See Figure 1 for an illustration of what the reading diary app looked like. The key element of the reading diary app is each participant's personal library. Participants were able to add books/texts to their personal libraries by scanning the barcode on the back of the book with an app tool (see Figure 1a). This automatically entered book-related data such as book title, author's name, ISBN number, and the number of pages in the app, as the app was able to pull this information from the German National Library. Participants also had the opportunity to type in the title and author of the book manually. This also applied to all other texts (e.g., newspapers, digital articles, magazines) where no barcode was available. Then, every time the participant opened the reading diary app, his or her personal library appeared (see Figure 1b), encompassing all previously added books and texts (labeled "reading projects"). To measure the time participants spent reading, they were asked to make an entry every time they completed a reading event/activity. To make such an entry, participants first chose the reading project (i.e., the book or text they had read) from their library, and then they answered a short question about the duration of the reading event (in minutes) and the number of pages they had read (see Figure 1c). Participants were also asked to answer four short questions regarding aspects of situational reading motivation every time they completed a reading event. Log data provided additional information about the date and time of day when participants indicated they had read something (i.e., when they made an entry). All elements in the library (books or other texts) were recorded and visible to the participants until they had completed a reading project. When participants indicated that they had

finished a reading project (as well as at the end of the two weeks of using the app), they had to conclude the reading project. To do so, they had to answer questions about the reading project in general, such as the reading purpose.

Reading Behavior Measures from the Paper-pencil Questionnaire

Global evaluation of reading time. The global evaluation of reading time was captured in a manner comparable to the PISA study (Hertel, Hochweber, Mildner, Steinert, & Jude, 2014) by asking participants to answer the following question: "How much time do you normally spend reading per day?" A 5-point Likert scale was used (1 = *never*, 2 = *up to 30 min*, 3 = *between half an hour and 1 hour*, 4 = *1 to 2 hours*, 5 = *more than 2 hours*). The global evaluation of reading time was measured in the M1 and M3 questionnaires.

Evaluation of reading time for different types of texts. Equivalent to the global evaluation, the evaluation of reading time with respect to different types of texts was captured with the item: "How much time do you spend per day reading the following text types?," again on a 5-point Likert scale (1 = *never* to 5 = *more than 2 hours*). This time, however, participants were asked to indicate how much time they spent per day reading different types of texts. Similarly to the PISA study, this study asked about the following categories: (a) fiction books, (b) nonfiction books, (c) newspapers, (d) magazines, and (d) comic books. This variable was also measured at M1 and M3. Although more text types exist than the five categories mentioned, a recent study by Locher & Pfost (2019b) showed that a too fine-grained differentiation between text types tends to become counterproductive. Therefore, these broad text categories were used.

Comparative Reading Habits (CRH)

The CRH is a measure of reading habits developed by Acheson, Wells, and MacDonald (2008). Participants were asked to rate their reading habits in comparison with their peers (e.g., reading time: "Compared to other college students, how much time do you spend reading all types of materials?" or reading speed: "Compared to other college students, how fast do you normally read?"). For each of the five questions on the CRH, participants chose a number on a scale ranging from 1 to 7, with higher numbers indicating greater amounts of the quantity in question (e.g., reading time, speed). Similarly to the study by Acheson et al. (2008), a 7-point Likert scale was used to ensure sufficient variance in responses. The CRH was measured at M1 only.

Postmonitoring Questionnaire

Conscientiousness. To check whether participants regularly documented their reading activities, we asked participants at M3: "How regularly did you make entries

in the app after reading?" They were asked to rate this question on a 4-point Likert scale (1 = *forgot very often*, 2 = *sometimes forgot*, 3 = *regularly documented*, 4 = *always documented*).

Generalizability. To check whether participants' reading activities during the 2 weeks of using the reading diary app were comparable to their normal daily reading habits, participants were asked the following question: "Do you think you spent more or less time reading during the 2 weeks of using the reading diary app than you normally do?" A 5-point Likert scale (1 = *much less*, 2 = *a bit less*, 3 = *exactly the same*, 4 = *a bit more*, 5 = *much more*) was used to ensure sufficient differentiation.

Reaction. Three items were used to check for possible reaction effects caused by the reading diary app. The first item refers to boredom effects ("Filling out the reading diary app was boring"), the second item to the burden ("Filling out the reading diary app was a burden in everyday life"), and the third item to an unintentional intervention effect ("Filling out the reading diary app influenced my usual reading behavior"). All three items were rated on a 4-point Likert scale (1 = *disagree*, 2 = *somewhat disagree*, 3 = *somewhat agree*, 4 = *agree*).

Analysis Strategy

To explore the correlation between reading time in Weeks 1 and 2 and on even- and odd-numbered days, and to examine internal consistency, the total duration of all reading events each person documented throughout the 14 days was aggregated. In so doing, information was collected about the amount of time participants spent reading on each individual day as well as during each of the 2 weeks. In general, internal consistency measures whether different items that aim to measure the same construct produce similar scores. To compute the internal consistency of reading time as measured in the reading diary, the reading time on each day was treated as a single "item" measuring the same construct, namely, the amount of time spent reading.

For the second research question regarding the relation between reading time measured via the reading diary app and reading time measured via the questionnaire, the reading time data from the ambulatory assessment was transformed. In a first step, to differentiate between the different types of texts, each reading project from a participant's personal library was assigned to one of the categories from the questionnaire: fiction books, nonfiction books, newspapers, magazines, or comics. Some reading projects could not clearly be assigned to one of the text categories (e.g., lecture notes from university courses). These titles formed the category "other books" or the category "other texts." While categorizing the reading projects, it became apparent that the app failed to transfer title names from a substantial number of manually added reading projects to the server. Due to this technical problem,

these reading projects could not be categorized. These titles formed the category "texts with missing title." In a second step, daily reading time in minutes from the reading diary app was classified into one of the five response categories from the paper-pencil questionnaire: 1 = *never*, 2 = *up to 30 min*, 3 = *between half an hour and 1 hour*, 4 = *1 to 2 hours*, 5 = *more than 2 hours*. This made it possible to compute the average daily reading time across the 2 weeks on a categorical level. This was also done separately for each text type. After preparing the data in this manner, repeated-measures ANOVAs and correlation analyses were computed in SPSS (IBM-Corporation, 2012).

Results

Before presenting the results for the two research questions, some descriptive results will be highlighted to provide a first impression of the information we were able to collect with the reading diary app.

Descriptive Results for the Reading Diary Data

A total of 416 event-based entries were made during the ambulatory assessment. This means the 31 participants indicated that they had engaged in reading activities 416 times during the 2-week period. Figure 2 shows the times of day that were the most popular reading times. Most reading events were documented between 7 pm and 12 am, meaning that people mostly indicated spending time reading in the evening. Most reading and most entries were made on Mondays. Other than that, peak reading time was rather equally distributed (see Figures A and B from the electronic supplement for further information). On average, participants documented one reading event per day ($M = 0.96$, $SD = 0.66$). These reading events lasted an average of $M = 31.78$ min ($SD = 16.10$). The duration of a reading event ranged from 5 to 240 min. Approximately 17% of all reading events lasted 10 min or less, while 16% of all reading events lasted 1 hour or longer.

During the 2 weeks of the ambulatory assessment, participants added an average of four books ($M = 3.61$, $SD = 4.19$) and two additional texts, meaning newspapers, magazines, online articles, and so forth ($M = 2.10$, $SD = 2.17$), to their library. On average, four reading events per book ($M = 4.23$, $SD = 3.90$, Min = 1, Max = 19) and two reading events per text ($M = 2.28$, $SD = 1.61$, Min = 1, Max = 16) were recorded. Some participants did not document any reading time for some of the reading projects they entered into their library, meaning they did not read every book/text they entered. Therefore, Table 1 shows the number of reading projects overall and the number of reading projects with valid reading times. One possible explanation for this is that participants added some books they planned to read into

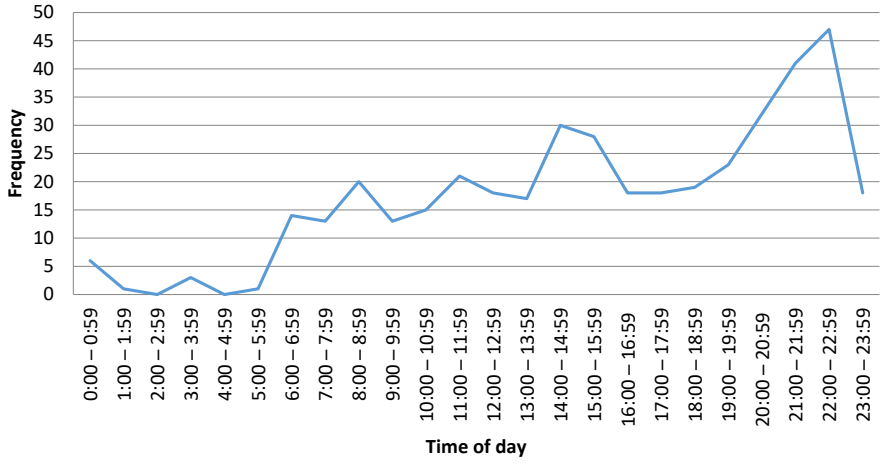

Figure 2 Frequencies of entries in the reading diary app per time of day

their library in advance, but then they did not actually spend time reading them. In-depth analyses revealed that it was often nonfiction books that were added to the library but had no reading time (see Table 1). Furthermore, the students added a higher number of books than other texts (e.g., newspapers) during the 2 weeks.

Table 1 Numbers of Reading Projects by Type of Text

Type of reading project	Type of text	*N*	n_T
Books	Fiction	46	42
	Nonfiction	54	36
	Other books	12	11
	Sum of books	112	89
Texts	Newspapers and magazines	15	13
	Other texts	15	15
	Texts with missing titles	35	33
	Sum of texts	65	61
Sum of reading projects		177	150

Note. N = Number of reading projects overall, n_T = Number of reading projects with valid reading time.

Table 2 Average Daily Reading Time in Minutes By Text Type Measured via the Reading Diary App

	n	*M*	*SD*	Min	Max
Fiction	42	16.82	25.37	0.00	119.29
Nonfiction	36	3.88	11.92	0.00	66.43
Newspapers	4	0.72	2.18	0.00	9.29
Magazines	9	1.61	5.81	0.00	32.14
Comics	1	0.03	0.19	0.00	1.07
Other books	11	1.81	4.35	0.00	16.43
Other texts	14	1.12	2.96	0.00	15.00
Texts with missing title	33	5.06	9.91	0.00	49.86
All	150	31.05	32.20	0.00	139.29

Note. $N = 31$ participants. n = number of reading projects to which this reading time can be subsumed.

Table 2 shows the average time people spent reading across all types of texts as well as differentiated by type of text. First, the results showed that on average, participants spent about half an hour a day ($M = 31.05$, $SD = 32.20$) reading. Differentiated by type of text, participants predominantly spent time reading fiction ($M = 16.82$, $SD = 25.37$) and nonfiction books ($M = 3.88$, $SD = 11.92$), whereas other types of texts such as newspapers and magazines were only read for a few minutes a day. On average, participants spent more time (in minutes) reading in the first week ($M = 34.18$, $SD = 37.90$) than in the second week ($M = 27.91$, $SD = 32.45$). However, this difference was not significant, $t(30) = 1.21$, $p > .05$, $r = .67$. Individual differences between participants in average reading time were large, as seen in the large standard deviation.

Measurement and Data Quality

Detailed results of the postmonitoring questionnaire can be found in Table 3. With respect to conscientiousness, only one person indicated that he or she often forgot to make entries in the reading diary app (= response option 1), whereas more than 75% stated that they regularly or almost always made entries (response option 3 or 4). Once again, the categories were 1 = "*forgot very often*", 2 = "*sometimes forgot*", 3 = "*regularly documented*", 4 = "*always documented*". Regarding the generalizability of the documented reading activities, the results of the postmonitoring questionnaire were satisfactory. Participants indicated how their reading activities during the ambulatory assessment compared to their normal daily reading habits

Table 3 Descriptive Statistics from the Postmonitoring Questionnaire (M3)

	M	*SD*	Cat 1 %	Cat 2 %	Cat 3 %	Cat 4 %	Cat 5 %
Conscientiousness	3.26	0.89	3.2	19.4	25.8	51.6	-
Generalizability	2.97	0.91	3.2	32.3	29.0	35.5	0.0
Reaction 1: Boredom	2.26	0.82	16.1	48.4	29.0	6.5	-
Reaction 2: Burden	1.68	0.65	41.9	48.4	9.7	0.0	-
Reaction 3: Intervention	2.23	0.81	19.4	41.9	35.5	3.2	-

Note. Data were collected from $N = 31$ participants. Cat in %= percentage of people selecting this category. Conscientiousness: Category 1 = *forgot very often* to Category 4 = *always documented*; Generalizability: Category 1 = *much less* to Category 5 = *much more*; Reaction: Category 1 = *disagree* to Category 4 = *agree.*

with the following categories: 1 = "*much less"*, 2 = "*somewhat less"*, 3 = "*exactly the same"*, 4 = "*a bit more"*, 5 = "*much more*". About 29% of participants indicated that they spent exactly the same amount of time reading during these 2 weeks compared with their usual reading time. A total of 32% stated that they usually read slightly less and 3% much less than in the 2 weeks of data collection. On the other hand, 36% of participants indicated that they read slightly more than they usually read. Consequently, participants' reading time during the 2 weeks of data collection seemed to be comparable on average to participants' usual reading activities. With respect to the reaction items (categories: 1 = "*disagree"*, 2 = "*rather disagree"*, 3 = "*rather agree"*, 4 = "*agree")*, only 7% of participants indicated that they got bored (Reaction 1) while using the reading diary app, whereas 65% stated that they were not bored or only slightly bored. About 90% of participants disagreed or somewhat disagreed that the task of monitoring their reading behavior with the reading diary app every day was a burden (Reaction 2). Furthermore, only one participant indicated that the ambulatory assessment influenced his or her daily reading behavior (i.e., he or she read a lot more than average). About 60% of participants stated that they did not think they changed their reading behavior due to the ambulatory assessment (Reaction 3).

As an additional quality measure, we calculated the internal consistency for reading time across all days. The reading diary app data had satisfactory internal consistence ($\alpha = .87$). This was also supported by the correlations between the sum total daily reading time on even- and odd-numbered days ($r = .81$, $p < .01$) and between the sum total of reading time in Weeks 1 and 2 ($r = .67$, $p < .01$). Both results serve as indicators of reliability.

Comparing Different Reading Behavior Measures

To compare the reading diary app data with the global retrospective questionnaire data, we used information about daily reading time from the event-based entries during the 2 assessment weeks. This data was transformed to match the response categories for the questionnaire items. Because the number of reading projects in the categories of newspapers and magazines was too small to analyze separately, a joint category was built for both the app and questionnaire data. No analyses were conducted regarding comic books because only one comic book was mentioned in the reading diary app.

Comparing the average amount of reading time per week measured as a global evaluation on the questionnaire and the average weekly reading time measured via the reading diary app (Table 5 and Figure 3) yielded a significantly lower average for the reading diary app data ($M = 2.06$, $SD = 0.70$) compared to the questionnaire data (M1: $M = 3.29$, $SD = 0.97$; M3: $M = 2.94$, $SD = 0.85$).

Furthermore, there were significant differences in the global retrospective questionnaire measure before (M1) and after (M3) the 2 weeks of reading behavior documentation, with participants indicating less time spent reading after the ambulatory assessment period. Comparable results were found when differentiating between text types (Table 4). For all text categories, the evaluation of reading time before using the reading diary app was descriptively but not significantly higher than the evaluation after the ambulatory assessment period. Turning to global reading time (i.e., summing up all reading activities across all types of texts), no participants indicated that they spent no time reading when asked in the questionnaire. However, according to the reading diary data, 23% of participants fell into the lowest category, which meant that on most days during the 2 survey weeks, they did not spend any time reading. The results of the correlational analyses in Table 5 revealed that the reading diary app data were significantly associated with the global retrospective evaluation from the questionnaire (M1: $r = .39$, $p < .05$ and M3: $r = .58$, $p < .01$). The global retrospective evaluation *after* the 2 weeks of data collection was more strongly related to the reading diary app data than the global retrospective evaluation *before* the 2 weeks of data collection. Comparable results were found when differentiating by type of text, with the exception of the newspapers and magazines category. Table 5 also shows the correlations with the CRH scale. One can see that reading time measured via the reading diary app was significantly correlated with the CRH ($r = .38$, $p < .05$). As the internal consistency of the CRH was quite low ($\alpha = .49$), and the CRH also includes items that refer to reading skills and reading speed, we additionally computed correlations with the item referring to reading time only ("Compared to other college students, how much time do you spend reading all types of materials?"). This item had a considerably stronger correlation with the reading diary app data ($r = .51$, $p < .01$).

Table 4 Average Time Spent Reading (Questionnaire and App Data) in General and by Type of Text

Type of text		*M*	*SD*	M1- M2	M2-M3	M1-M3	Never (%)	> 2 hr (%)
Fiction ($n = 42$)	M1: Questionnaire	2.84	0.93	$p < .01$	$p < .01$	*ns*	0.0	6.5
	M2: Reading diary app	1.59	0.64				51.6	0.0
	M3: Questionnaire	2.68	0.83				3.2	0.0
Nonfiction ($n = 36$)	M1: Questionnaire	2.21	0.90	$p < .01$	$p < .01$	*ns*	19.4	3.2
	M2: Reading diary app	1.13	0.27				96.8	0.0
	M3: Questionnaire	2.07	0.53				19.4	0.0
Newspapers and Magazines ($n = 13$)	M1: Questionnaire	1.60	0.46	$p < .01$	$p < .01$	*ns*	19.4	0.0
	M2: Reading diary app	1.05	0.11				100.0	0.0
	M3: Questionnaire	1.48	0.38				29.0	0.0
Global/All text types ($n = 150$)	M1: Questionnaire	3.29	0.97	$p < .01$	$p < .01$	$p < .05$	0.0	12.9
	M2: Reading diary app	2.06	0.70				22.6	0.0
	M3: Questionnaire	2.94	0.85				0.0	3.2

Note. $N = 31$ participants. n = entries/events. Reading time measured with the reading diary app was transformed into the same response categories used in the questionnaire. Analyses regarding differences between M1, M2, and M3 were computed using post hoc comparisons after computing repeated-measures ANOVAs. 5-point Likert scale: 1 = *I never read* to 5 = *I read for more than 2 hr.*

Table 5 Pearson Correlations for Reading Behavior Measures in General and by Type of Text

	Fiction			Nonfiction			Newspapers and magazines			Global (all text types)		
	1)	2)	3)	1)	2)	3)	1)	2)	3)	1)	2)	3)
1) M1: Questionnaire	-			-			-			-		
2) M2: Reading diary app	.43*			.45*			.18			.39*		
3) M3:Questionnaire	.66**	.58**		.64**	.48**		.64**	.23		.67**	.58**	
4) M1: CRH scale	.35	.26	.54**	.20	.32	.33	-.24	-.28	-.20	.38*	.38*	.31
5) M1: CRH (1 item)	.48**	.51**	.68**	.03	.22	.27	-.25	-.17	.08	.49**	.51**	.43*

Note. $N = 31$ participants. Cronbach's alpha of CRH was $\alpha = .49$. Correlation M1: CRH Scale and M1: CRH (1st item: "Compared with other college students, how much time do you spend reading all types of materials?"): $r = .65$**. *$p < .05$. **$p < .01$.

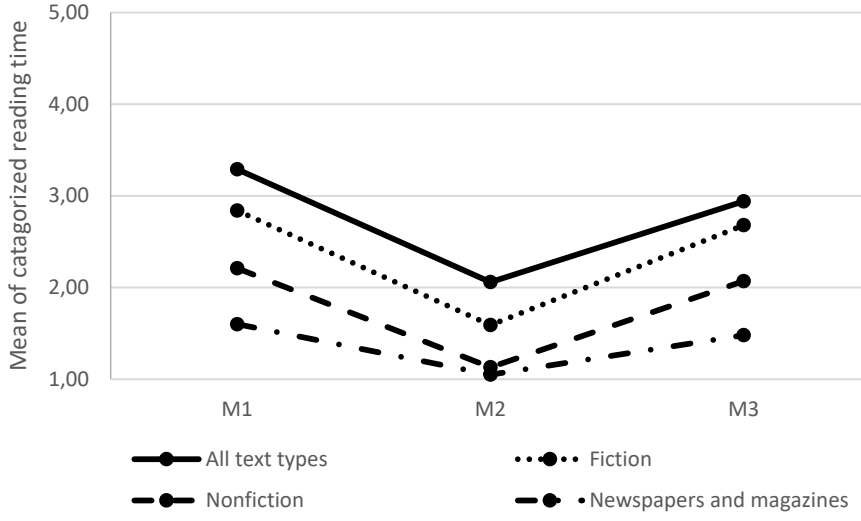

Figure 3 Presentation of differences in reading time between the three measurement points: M1 = prequestionnaire, M2 = reading diary app, M3 = postquestionnaire

Discussion

The main interest of this study was to gain deeper insight into people's "true" reading behavior using an ambulatory assessment. Therefore, a reading diary app was developed to monitor people's reading behavior in daily life. In this field report, we explored whether the ambulatory assessment had satisfactory measurement and data quality, which is an important precondition for further analyses. In addition, it was explored how the ambulatory assessment data were related to data from global retrospective questionnaire measures.

First, a reading diary app/ambulatory assessment seems to be an appropriate method for collecting data of satisfactory quality about reading behavior in daily life. As smartphone use allows for a relatively easy and economical data collection process, it seems feasible for participants to document their reading behavior continuously rather than, for instance, just once a day (e.g., as often the case for end-of-day diaries). Therefore, quite precise and detailed information about individuals' daily reading behavior can be obtained. The results showed that internal consistency was very good, and the correlations of time spent reading in Weeks 1 and 2 as well as between even- and odd-numbered days, another reliability indicator, were strong and in the expected direction (e.g., Anderson et al., 1988; McLaws et al., 1990; Nieuwenboom, 2008). Moreover, most participants reported that docu-

menting their reading behavior with the reading diary app was not a burden for them and that they made regular entries. However, the generalizability of this finding requires further research. For example, some persons (e.g., older people) may be less familiar with smartphone apps than university students, making the use of electronic reading diaries more difficult for this population (Conner & Lehman, 2012). Furthermore, some people may have privacy concerns related to an app that collects information on their personal behavior, including their reading behavior. It is also important to consider that participants only monitored their reading behavior for 2 weeks. Therefore, it might be argued that participants' reading behavior during those 2 weeks was not representative of their reading behavior in general. Nevertheless, two weeks are a common and sometimes recommended time period for diary studies (Conner & Lehman, 2012). Moreover, we chose a time period for our diary study in the middle of the semester when no exams had to be taken, as exam stress could affect college students' reading behavior. Furthermore, in the postmonitoring questionnaire, nearly all participants indicated that the amount of time they spent reading did not deviate significantly from their general reading behavior.

Second, the results showed that the reading diary app data and global retrospective questionnaire data (collected before and after the ambulatory assessment period) were closely related. However, despite the significant correlations, there were substantial differences between the reading time data collected via the reading diary app and the questionnaires. The average daily time spent reading was significantly lower in the ambulatory assessment compared to the questionnaire self-report scales. One possible explanation for this is that participants tend to overestimate the amount of time they spend reading each day when asked to make a global retrospective self-report on a questionnaire. This is in line with Nieuwenboom (2008), who found in a sample of third to fifth graders that students tended to overestimate their reading time in a questionnaire compared to a paper-pencil reading diary. Nevertheless, it must be noted that both the questionnaire and reading diary measures rely on self-reports. A third, independent source of data might be helpful in order to confirm whether participants really overestimated their reading time in the retrospective questionnaire or whether reading diaries tend to underestimate reading time. Although we implicitly assume that participants continuously documented their reading activities during the ambulatory assessment, which should result in less bias due to memory effects, some participants might have forgotten to document their reading time and then did not respond honestly to the conscientiousness question.

Finally, our results found significant differences in the global retrospective self-report before and after the 2 weeks of ambulatory assessment. Furthermore, the correlation between the reading diary app data and the global evaluation of reading time from the questionnaire was stronger after the ambulatory assessment.

One possible explanation is that after participants monitored their own reading behavior for 2 weeks, they revised their reading time estimates, leading to different and possibly more precise responses to the global retrospective question. This change in participants' responses could again be interpreted as a sign that global retrospective questionnaires are influenced by aspects such as heuristics and memory effects. Nevertheless, it should be kept in mind that all results are based on a rather small sample. Therefore, in order to confirm the results regarding the quality of the reading diary app as well as the relations with retrospective questions, the app would need to be applied in a larger sample and complemented with further sources of information, such as interview data. Consequently, future research should try to replicate the study with a larger and more heterogeneous sample with persons at different stages of life, from school students to older adults. Furthermore, data collection periods of varying length might be explored. Whereas longer time periods would help the diary data better capture habitual behavior, longer time periods might also lead to more missing data, measurement error, unwillingness to participate in the study and boredom effects (Bolger et al., 2003). Therefore, the effects of shorter data collection periods might also be examined.

Limitations of the Study

The present study also has some limitations. First, a small convenience sample of college students in psychology and educational science was used. Due to sample selectivity, the results may not generalize to the general population. Second, there were some unexpected technical problems with the reading diary app. The biggest issue was that the titles of 33 reading projects were not transferred to the server correctly. This information about the title/type of text was necessary for the differential analyses by type of text. Because it was not possible to restore this missing information, reading projects lacking information about the title/type of text had to be excluded from these analyses. This led to a reduction in the sample size in these categories and to a relatively small sample size in the newspapers and magazines category, which could be an explanation for the nonsignificant correlation between the app and questionnaire data. Third, it was not possible to determine whether participants made fake entries in the diaries for reasons such as social desirability (Carels et al., 2006; Gershuny, 2012). For example, social desirability effects have been found when parents report reading times with their children, with parents often exaggerating this reading time (Hofferth, 2006). Accordingly, participants might have indicated spending more time reading in the app than they actually spent reading. However, it might be seen as less likely for a participant to continuously make invalid statements for several entries across a two-week period compared to a single questionnaire response. Nevertheless, to address this limitation, it

would be useful for future research to examine a third source of information, such as interview data. Finally, there may be individual differences in data accuracy due an imprecise definition of the construct "reading event". Therefore, some study participants might have recorded reading events in the diary that other participants did not record. Hence, future studies should develop a more precise working definition of the term reading event and communicate this to study participants in order to improve data accuracy.

Conclusion

The present study is among the first to use an ambulatory assessment in the form of a reading diary smartphone app to examine people's reading behavior. In doing so, this study addresses the often-discussed necessity to use more innovative methods to study behavior in daily life (e.g., Fahrenberg et al., 2007b), and the need to obtain new and deeper insights into people's common reading activities and qualitative aspects of reading behavior (e.g., Troyer et al., 2018). Global retrospective measures often do not provide information that would allow for such insights because they only reflect average trends and tendencies rather than concrete information about the books and texts a person has actually read.

The present field report illustrates that a reading diary app is a promising method for economically collecting detailed data about people's reading behavior in daily life. However, ambulant assessment via a smartphone app also involves many challenges (e.g. susceptibility to technical problems, relatively large effort required to develop the app), as shown in this study and documented in this field report. While this study has taken a small first step in the direction of resolving these challenges, and there is a lot of work still to do and improvements to be made. The present study clearly illustrates that reading behavior is a much more complex construct than just the average time spent reading as measured in global retrospective questionnaires. The results showed that the amount of time individuals spent reading each day varied substantially across days. Furthermore, there is great variation in participants' reading material, which typically remains invisible in global retrospective data – except with respect to very general types of texts. But perhaps it is not just reading a lot but reading diverse books and texts (varying in content, complexity, and writing styles) that makes a competent reader (Kirsch et al., 2002). Given that existing evidence on the relation between reading behavior and reading skills or reading motivation is predominantly based on studies using global retrospective questionnaire data (e.g., Locher et al., 2020; Pfost et al. 2013; Troyer et al., 2018), future research should examine whether these findings can be replicated with more fine-grained measures of reading behavior. It might also be fruitful to further develop the reading diary app to promote increased reading behavior, e.g.,

by using a token system for the amount of reading time reached (see Robinson, Newby, & Ganzell, 1981). Akbar et al. (2015), for example, found that reading apps can help to improve reading speed. Such interventions might be a further perspective for future research with reading diary apps.

References

Acheson, D. J., Wells, J. B., & MacDonald, M. C. (2008). New and updated tests of print exposure and reading abilities in college students. *Behavior Research Methods, 40*(1), 278-289. https://doi.org/10.3758/BRM.40.1.278

Ahram, T. (Ed.). (2019). *Advances in artificial intelligence, software and systems engineering.* Cham: Springer International Publishing.

Akbar, R. S., Taqi, H. A., Dashti, A. A., & Sadeq, T. M. (2015). Does e-reading enhance reading fluency? *English Language Teaching, 8*(5), 195-207. https://doi.org/10.5539/elt.v8n5p195

Alexander, P. A. (2005). The Path to Competence: A Lifespan Developmental Perspective on Reading. *Journal of Literacy Research, 37*(4), 413–436. https://doi.org/10.1207/s15548430jlr3704_1

Allen, L., Cipielewski, J., & Stanovich, K. E. (1992). Multiple indicators of children's reading habits and attitudes: construct validity and cognitive correlates. *Journal of Educational Psychology, 84*(4), 489-503.

Anderson, R. C., Wilson, P. T., & Fielding, L. G. (1988). Growth in reading and how children spend their time outside of school. *Reading Research Quarterly, 23*(3), 285-303. Retrieved from http://www.jstor.org/stable/748043

Artelt, C., Stanat, P., Schneider, W. & Schiefele, U. (2001). Lesekompetenz: Testkonzeption und Ergebnisse. In J. Baumert, E. Klieme, M. Neubrand, M. Prenzel, U. Schiefele, W. Schneider et al. (eds.), *PISA 2000* (Basiskompetenzen von Schülerinnen und Schülern im internationalen Vergleich, pp. 69-137). Opladen: Leske + Budrich.

Becker-Mrotzek, M., Brinkhaus, M., Grabowski, J., Hennecke, V., Jost, J., Knopp, M., Schmitt, M., Weinzierl, C. & Wilmsmeier, S. (2015). Kohärenzherstellung und Perspektivübernahme als Teilkomponenten der Schreibkompetenz. Von der diagnostischen Absicherung zur didaktischen Implementierung. In A. Redder, J. Naumann & R. Tracy (eds.), *Forschungsinitiative Sprachdiagnostik und Sprachförderung – Ergebnisse*. (pp. 177–205). Münster: Waxmann.

Bolger, N., Davis, A., & Rafaeli, E. (2003). Diary methods: Capturing life as it is lived. *Annual Review of Psychology, 54*(1), 579-616. https://doi.org/10.1146/annurev.psych.54.101601.145030

Burgess, S. R., Hecht, S. A., & Lonigan, C. J. (2002). Relations of the home literacy environment (HLE) to the development of reading related abilities: A one year longitudinal study. *Reading Research Quarterly, 37*(4), 408-426. https://doi.org/10.1598/RRQ.37.4.4

Bus, A. G., van IJzendoorn, M., & Pellegrini, A. (1995). Joint book reading makes for success in learning to read: A meta-analysis on intergenerational transmission of literacy. *Review of Educational Research, 65*(1), 1-21. https://doi.org/10.3102/00346543065001001

Calamia, M. (2019). Practical considerations for evaluating reliability in ambulatory assessment studies. *Psychological assessment, 31*(3), 285. https://doi.org/10.1037/pas0000599

Carels, R. A., Cacciapaglia, H. M., Rydin, S., Douglass, O. M., & Harper, J. (2006). Can social desirability interfere with success in a behavioral weight loss program? *Psychology & Health, 21*(1), 65-78. https://doi.org/10.1080/14768320500102277

Chall, J. S. (1983). *Stages of reading development.* New York, NY: McGraw-Hill.

Conner, T. S., & Lehman, B. J. (2012). Getting started: Launching a study in daily life *Handbook of research methods for studying daily life*. (pp. 89-107). New York, NY, US: The Guilford Press.

Csikszentmihalyi, M., & Larson, R. (2014). Validity and Reliability of the Experience-Sampling Method. In *Flow and the Foundations of Positive Psychology* (pp. 35-54). Springer, Dordrecht. https://doi.org/10.1007/978-94-017-9088-8_3

Ebner-Priemer, U. W., Kuo, J., Kleindienst, N., Welch, S. S., Reisch, T., Reinhard, I. et al. (2007). State affective instability in borderline personality disorder assessed by ambulatory monitoring. *Psychological Medicine*, 37(7), 961–970. https://doi.org/10.1017/S0033291706009706

Fahrenberg, J., Myrtek, M., Pawlik, K., & Perrez, M. (2007a). Ambulantes Assessment - Verhalten im Alltagskontext erfassen [Ambulatory assessment - recording behaviour in an everyday context]. *Psychologische Rundschau, 58*(1), 12-23. https://doi.org/10.1026/0033-3042.58.1.12

Fahrenberg, J., Myrtek, M., Pawlik, K., & Perrez, M. (2007b). Ambulatory assessment-Monitoring behavior in daily life settings: A behavioral-scientific challenge for psychology. *European Journal of Psychological Assessment, 23*(4), 206. https://doi.org/10.1027//1015-5759.23.4.206

Foasberg, N. M. (2014). Student reading practices in print and electronic media. *College & Research Libraries*, *75*(5), 705–723. https://doi.org/10.5860/crl.75.5.705

Gershuny, J. (2012). Too many zeros: A method for estimating long-term time-use from short diaries. *Annals of Economics and Statistics* (105/106), 247-270. https://doi.org/10.2307/23646464

Glomann, L., Hager, V., Lukas, C. A. & Berking, M. (2019). Patient-centered design of an e-mental health app. In T. Ahram (Ed.), *Advances in artificial intelligence, software and systems engineering* (pp. 264–271). Cham: Springer International Publishing.

Greaney, V., & Hegarty, M. (1987). Correlates of leisure-time reading. *Journal of Research in Reading, 10*(1), 3-20. https://doi.org/10.1111/j.1467-9817.1987.tb00278.x

Guthrie, J. T., Wigfield, A., Metsala, J. L., & Cox, K. E. (1999). Motivational and cognitive predictors of text comprehension and reading amount. *Scientific Studies of Reading, 3*(3), 231-256. https://doi.org/10.1207/s1532799xssr0303_3

Hektner, J. M., Schmidt, J. A., & Csikszentmihalyi, M. (2007). *Experience Sampling Method*. Thousand Oaks, London, New Delhi: Sage Publications

Hertel, S., Hochweber, J., Mildner, D., Steinert, B., & Jude, N. (2014). *PISA 2009 Skalenhandbuch [PISA 2009 scaling handbook]*. Münster; New York: Waxmann.

Hofferth, S. L. (2006). Response Bias in a Popular Indicator of Reading to Children. *Sociological Methodology, 36*(1), 301–315. https://doi.org/10.1111/j.1467-9531.2006.00182.x

IBM-Corporation. (2012). IBM SPSS Bootstrappin 21. Retrieved from ftp://public.dhe.ibm.com/software/analytics/spss/documentation/statistics/21.0/en/client/Manuals/IBM_SPSS_Bootstrapping.pdf

Jerrim, J. & Moss, G. (2019), The link between fiction and teenagers' reading skills: International evidence from the OECD PISA study. *British Educational Research Journal*, 45: 181-200. https://doi.org/10.1002/berj.3498

Juster, F. T., Ono, H., & Stafford, F. P. (2003). An Assessment of Alternative Measures of Time Use. *Sociological Methodology, 33*, 19–54.

Kahneman, D., Krueger, A. B., Schkade, D. A., Schwarz, N., & Stone, A. A. (2004). A Survey Method for Characterizing Daily Life Experience: The Day Reconstruction Method. *Science, 306*(5702), 1776-1780.

Kan, M. Y., & Pudney, S. (2008). 2. Measurement Error in Stylized and Diary Data on Time Use. *Sociological Methodology, 38*(1), 101–132. https://doi.org/10.1111/j.1467-9531.2008.00197.x.

Keller, A. (2010). *Einsatz von digitalen Foto-Lesetagebüchern zur Erforschung des Leseverhaltens von Studierenden. [Use of digital photo reading diaries to research the reading behaviour of students].* Paper presented at the 5. Konferenz der Zentralbibliothek, Forschungszentrum Jülich.

Kirsch, I., de Jong, J., Lafontaine, D., McQueen, J., Mendelovits, J., & Monseur, C. (2002). *Reading for change. Performance and engagement across countries. Results from PISA 2000.* Paris: OECD.

Kutner, M., Greenberg, E., Jin, Y., Boyle, B., Hsu, Y., & Dunleavy, E. (2007). *Literacy in everyday life: Results from the 2003 National Assessment of Adult Literacy (NCES 2007–480).* Washington, DC: National Center for Education Research.

Lampert, T., Sygusch, R., & Schlack, R. (2007). Nutzung elektronischer Medien im Jugendalter [Use of electronic media in youth]. *Bundesgesundheitsblatt - Gesundheitsforschung - Gesundheitsschutz, 50*(5), 643-652. https://doi.org/10.1007/s00103-007-0225-7

Leiner, D. J. (2019). SoSci Survey (Version 3.1.06) [Computer software]. Available at https://www.soscisurvey.de

Lucas, R. E., Wallsworth, C., Anusic, I., & Donnellan, B. (2019). *A Direct Comparison of the Day Reconstruction Method and the Experience Sampling Method*. https://doi.org/10.31234/osf.io/cv73u

Locher, F. M., Becker, S., & Pfost, M. (2019a). The Relation Between Students' Intrinsic Reading Motivation and Book Reading in Recreational and School Contexts. *AERA Open*, 5(2), 1-14. https://doi.org/10.1177/2332858419852041

Locher, F. M., & Pfost, M. (2019b). Erfassung des Lesevolumens in Large-Scale Studien. Ein Vergleich von Globalurteil und textspezifischem Urteil. [Measuring reading volume in Large-Scale Assessments: A comparison of an overall evaluation and a differentiated evaluation relating different text types.]. *Diagnostica(1)*, 26-36. https://doi.org/10.1026/0012-1924/a000203

Locher, F., & Pfost, M. (2020) The relation between time spent reading and reading comprehension throughout the life course. *Journal of Research in Reading*, 43: 57– 77. https://doi.org/10.1111/1467-9817.12289.

Marshall, J. (2000). Research response to literature. In M. L. Kamil, P. B. Mosenthal, P. D. Pearson, & R. Barr (Eds.), *Handbook of reading research: Vol. III* (pp. 381-402). Mahwah, NJ: Lawrence Erlbaum Associates.

McGeown, S. P., Duncan, L. G., Griffiths, Y. M., & Stothard, S. E. (2015). Exploring the relationship between adolescent's reading skills, reading motivation and reading habits. *Reading and Writing, 28*(4), 545-569. https://doi.org/10.1007/s11145-014-9537-9

McGeown, S. P., Osborne, C., Warhurst, A., Norgate, R., & Duncan, L. G. (2016). Understanding children's reading activities: Reading motivation, skill and child characteristics as predictors. *Journal of Research in Reading, 39*(1), 109-125. https://doi.org/10.1111/1467-9817.12060

McLaws, M. L., Oldenburg, B., Ross, M. W., & Cooper, D. A. (1990). Sexual behaviour in aids-related research: Reliability and validity of recall and diary measures. *The Journal of Sex Research, 27*(2), 265-281. https://doi.org/10.1080/00224499009551556

Miller, G. (2012). The smartphone psychology manifesto. *Perspectives on Psychological Science, 7*(3), 221-237. https://doi.org/10.1177/1745691612441215

Mol, S. E., & Bus, A. G. (2011). To read or not to read: A meta-analysis of print exposure from infancy to early adulthood. *Psychological Bulletin, 137*(2), 267-296. https://doi.org/10.1037/a0021890

Moosbrugger, H. & Kelava, A. (Hrsg.). (2012). *Testtheorie und Fragebogenkonstruktion* (2.Aufl.). Berlin, Heidelberg: Springer.

Mullis, I. V. S. & Martin, M. O. (2019). PIRLS 2021 reading assessment framework. In I. V. S. Mullis & M. O. Martin (eds.): *PIRLS 2021 Assessment Frameworks.* (pp. 5-25). Retrieved from Boston College, TIMSS & PIRLS International Study Center website: https://timssandpirls.bc.edu/pirls2021/frameworks/

Nieuwenboom, J. W. (2008). *Wie viel lesen Kinder? Die Erfassung von Leseaktivitäten mit Hilfe von strukturierten Tagebüchern-eine methodologische Studie [How much do kids read? The recording of reading activities using structured diaries - a methodological study]*: Tectum Verlag.

OECD (2010). *PISA 2009 results: Learning to learn – Student engagement, strategies and practices. Volume III.* https://doi.org/10.1787/9789264083943-en

OECD (2019). *PISA 2018 Assessment and Analytical Framework*, PISA, OECD Publishing, Paris. https://doi.org/10.1787/b25efab8-en.

OECD (2021), 21st-Century Readers: Developing Literacy Skills in a Digital World, PISA, OECD Publishing, Paris. https://doi.org/10.1787/a83d84cb-en

Pfost, M., Dörfler, T., & Artelt, C. (2013). Students' extracurricular reading behavior and the development of vocabulary and reading comprehension. *Learning and Individual Differences*, 26, 89-102. https://doi.org/10.1016/j.lindif.2013.04.008

Pospeschill, M. (2010). *Testtheorie, Testkonstruktion, Testevaluation*. Munchen: Reinhardt.

Raith, T. (2008). Weblogs als Lesetagebücher im aufgabenorientierten Fremdsprachenunterricht–Ergebnisse einer Vergleichsstudie [Weblogs as reading diaries in task-oriented foreign language teaching - results of a comparative study]. *Aufgabenorientiertes Lernen und Lehren mit Medien: Ansätze, Erfahrungen, Perspektiven in der Fremdsprachendidaktik, 15*, 297.

Robinson, P. W., Newby, T. J., & Ganzell, S. L. (1981). A token system for a class of underachieving hyperactive children. *Journal of Applied Behavior Analysis, 14*(3), 307-315. https://doi.org/10.1901/jaba.1981.14-307

Shumow, L., Schmidt, J. A., & Kackar, H. (2008). Reading In Class & Out of Class: An Experience Sampling Method Study. *Middle Grades Research Journal, 3*(3), 97-120.

Stoffelsma, L. (2018). Short-term gains, long-term losses? A diary study on literacy practices in Ghana. *Journal of Research in Reading, 41*(S1), S66-S84. https://doi.org/10.1111/1467-9817.12136

Stone, A., Broderick, J., Schwartz, J., Shiffman, S., Litcher-Kelly, L., & Calvanese, P. (2003). Intensive momentary reporting of pain with an electronic diary: reactivity, compliance, and patient satisfaction. *Pain, 104*(1), 343-351. https://doi.org/10.1016/S0304-3959(03)00040-X

Troyer, M., Kim, J., Hale, E., Wantchekon, K., & Armstrong, C. (2018). Relations among intrinsic and extrinsic reading motivation, reading amount, and comprehension: a conceptual replication. *Reading and Writing*. https://doi.org/10.1007/s11145-018-9907-9

Trull, T. J., & Ebner-Priemer, U. (2014). The role of ambulatory assessment in psychological science. *Current Directions in Psychological Science, 23*(6), 466-470. https://doi.org/10.1177/0963721414550706

Wilhelm, P., Perrez, M., & Pawlik, K. (2012). Conducting research in daily life: A historical review *Handbook of research methods for studying daily life.* (pp. 62-86). New York, NY, US: The Guilford Press.

Zhang, J., Calabrese, C., Ding, J., Liu, M., & Zhang, B. (2018). Advantages and challenges in using mobile apps for field experiments: A systematic review and a case study. *Mobile Media & Communication, 6*(2), 179–196. https://doi.org/10.1177/2050157917725550

Zirkel, S., Garcia, J. A., & Murphy, M. C. (2015). Experience-Sampling Research Methods and Their Potential for Education Research. *Educational Researcher, 44*(1), 7-16. https://doi.org/10.3102/0013189X14566879

Date of online first publication: 2022-05-03