ESSnet Smart Surveys

Grant Agreement Number: 899365 - 2019-DE-SmartStat

Link to our CROS website

Workpackage 2

Smart Survey Pilots

Deliverable 2.1 – Time Use - WP2.2

Version 2.0, 07-04-2022

Prepared by:

Patrick Lusyne (Statistics Belgium, Belgium)
Justin Delloye (Statistics Belgium, Belgium)

Workpackage Leader:

Barry Schouten (CBS, Netherlands)
Jg.schouten@cbs.nl
telephone : +31 70 3374905

SUMMARY:

This document mainly reports on our experiences with organizing a small pilot on passenger mobility using MOTUS, a data platform developed by the Vrije Universiteit Brussels and Hbits. The pilot had three objectives: gaining as much experience as possible with organizing fieldwork with a tool like MOTUS, evaluating several usability issues, and assessing the feasibility to produce basic passenger mobility indicators. The study population was small; 39 persons participated, all recruited within statistics Belgium, but the research design elaborated. It consisted of several surveys, a slightly adapted diary, notifications, and a geofence.

We proved with this pilot that it's feasible to calculate basic mobility indicators by slightly adapting the diary functionality in MOTUS. The participants rated MOTUS usability high, but some had issues adjusting to this new collection method. Finally, organizing fieldwork revealed several operational challenges.

Throughout the entire experience, several general issues about smart statistics caught our attention and triggered insights. Although these insights are subjective, we address them briefly in the conclusions.

1. Introduction.

1. In 2019 more than 9 out of 10 young people aged 16 to 29 years used a mobile device to access the internet when away from home or work [1]. The implications of this figure for National Statistical Institutes are far-reaching. The clock is running. Within only a few years, a significant part of the population won't even be capable of imagining any other way of answering a survey than using a mobile device, their phone in particular. This prospect is both a challenge and an opportunity for statistical institutes, but a real gamechanger in any case.

2. In a recent paper, Riciatto e.a (source) speak of a paradigm shift [2]. They argue that the ability of mobile devices to combine active and passive data collection can lift the traditional survey (production)model onto an entirely new level. They also explicitly refer to the shifting interdependence between surveys' time and context, resulting in a 'context-aware' survey model. In that model, the timing and content of the questions are decided dynamically based on context information inferred from sensor data.

3. For National Statistical Institutes, harnessing this potential comes with challenges. There's still a need for certain developments on the technical side, but basically, the tools exist and are gaining maturity rapidly. There are many other facets, however, that still linger in the pioneering phase of the adoption process. Data protection and privacy matters often get the most attention, but also methodological and certainly operational issues remain. Even on the conceptual side, there remains work as a high-level key concept like usability still needs a workable translation before it can be practically made of value.

4. Pioneering offers the possibility to learn by experimentation, and in the last few years, compelling use cases have been realized [3,4]. However, a pitfall to experimentation and innovation is the tendency to overshoot key steps and consequently drain resources without achieving relevant results. The gap between the current survey production and the future might widen because of this, even from a psychological

perspective, as fellow production statisticians often react lukewarm to advanced technological solutions if they don't go hand in hand with mastering basic essential tasks.

5. Only a few National Statistical Institutes (NSIs hereafter) have field experience with smart surveys at this point. Therefore, in our vision, there's still a need for use cases that realize basic innovative survey layouts close to the existing surveys. By adding smart elements to these basic schemes, traditional survey methodology can gradually grow smart, and smart surveys can step by step achieve a place along with conventional surveys in NSIs.

6. The pilot we report on in this paper fits perfectly into that vision. Basically, it concerned organizing a small study with MOTUS, a data platform plus smart survey tool developed by the Free University Brussels. This work was part of work package 2 of the ESSnet Smart Statistics.

7. The setting of this research was purely experimental, and we recruited the few dozen participants in the pilot entirely among the staff of Statistics Belgium. We also translated and adapted most parts of the pilot to grant the colleagues of the other NSIs participating in the Essnet the opportunity to as well have a flavor of MOTUS[1]. However, as these repetitions took place well after the end of the original Belgian version, and for other practical reasons, we only report about the Belgian results in this report. Nonetheless, the valuable returns we've got from some colleagues added indirectly in a significant way to both particular insights as to our overall conclusions.

8. While this small study size is a drawback for the generalizability of the conclusions, it allowed us to deliberately increase the degree of experimentation beyond what would have been acceptable if actual respondents had been involved. The result was an elaborated research flow, arguably too elaborated, with several questionnaires, a diary, and innovative features like notifications and a geofence. More than the

[1] Seven German and eleven Italian colleagues grasp this opportunity and participated in partial copies of the Belgian pilot. These copies also included features like the notifications and the geofence.

number of research steps, the complicated sequence of tasks and the integration of advanced features stretched both the commitment of the participants and the application's capacities to their edge.

9. We had three research objectives. The first one was essentially gaining as much experience as possible with organizing fieldwork with a tool like MOTUS while exploring at least some of its more advanced features.

10. The second research objective focused on usability, both from the respondent's viewpoint and from an operational point of view, and we assessed several usability issues in more depth.

11. Finally, we turned to the domain of mobility research for the third research objective. While today seldom part of the core statistics NSI produces, the European Commission's new green deal implies that mobility and transport will only become more important as a research domain. Consequently, the need for comparable indicators to monitor progress in this field is significant, and it's clear that NSIs should anticipate. We explore if we can use MOTUS to collect data on people's mobility behavior and whether this might serve to produce basic passenger mobility indicators.

12. We elaborate on these objectives in the research design section, which precedes the result part of this report and the overall conclusion. As smart surveys are a relatively new field, however, there's still a need for conceptual elaboration. Therefore, we first elaborate on usability and related concepts in the next section. While preparing this study, we felt the need to concretize this concept.

13. To make the readers understand the genuine narrative of this report, we must explain what it isn't. To start with, this is no scientific-grade paper, nor by style, nor by content. The experimental setup makes that the results are anecdotal at best. It isn't a technical paper either. MOTUS has been extensively introduced elsewhere, and we strongly urge readers to consult these sources as well. Finally, readers should avoid reading this report as an evaluation report of MOTUS. MOTUS is continuously evolving, so this report only provides a snapshot, and the experimental scope of this research did not

lend itself to an in-depth evaluation. Additionally, and perhaps even more critical, was our lack of experience with organizing fieldwork with this new technology, making that many of the issues we discuss below trace back to this inexperience.

14. It's in the latter that the actual perspective of this report lies, in the experience. Basically, this report is an account. It describes a wide range of mainly practical and methodological issues when adopting a tool like MOTUS for the first time. Although these issues highly interact with technology, they don't necessarily necessitate an answer rooted in technology. However, the focus isn't entirely on the problems we encountered during this study. Perhaps most importantly, is that this report is mainly about opportunities. We hope that, even with this small pilot and despite the issues we touch on, readers will be convinced of the potential of this kind of technology, and MOTUS in particular. We certainly are.

2. Conceptual elaboration: functionality, usability, and deployability.

15. ISO defines usability as *"the extent to which a product can be used by specified users to achieve specified goals with effectiveness, efficiency, and satisfaction in a specified context of use"*[2]. [3]

16. Smart survey apps have a particular position in the mobile application landscape. Compared to many other mobile applications, the structural components of the ISO's definition can't be defined unambiguously for smart survey apps. The usability of a hiking app is all about enhancing hikers' experiences during hikes (the users, the goal, and the context). For smart surveys, the respondents are the prime users, but the goal is collecting data, making statisticians the ultimate users. Additionally, the contexts in which smart survey apps have to operate vary and depend on the domain and design features of a research.

17. This ambiguity leads to a dual perspective, the respondent's sense of how effective and efficient an app is and how satisfied he is with using it, versus the statistician's more professional viewpoint.

18. The focus regarding efficacy, efficiency, and satisfaction is somewhat different between both perspectives. A respondent will make an individual, more subjective assessment of using a survey app. The accent is slightly more on satisfaction. When statisticians use a survey instrument, the judgment is predominantly made from the statistical organization's angle, emphasizing efficiency.

19. Usability from the respondents' perspective aligns with the mainstream thinking on usability. There is a long tradition in usability research covering a wide range of issues about using all kinds of products by all sorts of users. Major research strands in that tradition concern product properties like *learnability*, how easy it is to learn to use a

[2] https://www.iso.org/obp/ui/#iso:std:iso:9241:-11:ed-2:v1:en

product, *accessibility*, if a product can be used by users with a wide range of capabilities, and *user experience* in general.

20. If we turn to the statistician's viewpoint on usability, some of these properties are directly applicable. *Learnability*, for example, translates effortlessly from the learning costs respondents endure to the learning curve statisticians and other staff of a statistical organization experience when they start organizing surveys with an application.

21. Although we said above that the statistician's perspective on usability is more on efficiency, this doesn't mean it's entirely void of subjective elements. In most organizations, change causes opposition. From a managerial and budgetary point of view, the benefits of smart survey apps, sensors, and new data sources are undeniably huge. Still, it's the motivation of the people on the floor that determines how smoothly new technology gets adopted. If using a new tool comes with severe practical and operational issues, or if that's the prevailing perception among statisticians, adoption will become much more complicated. We find this human factor somewhat underexposed in today's discussions on smart surveys and maybe even in the broader debate on innovation and modernization of statistical production in general.

22. It's not only a matter of getting through the initial learning and adoption phase. The costs and expertise involved in the long-term development and maintenance of apps remain high throughout their entire life cycle. In most statistical organizations covering this will take a fundamental reorientation of the workforce, implying recruitment and long-term engagement of very specialized IT profiles and critical investments in infrastructure. In reality, this is out of the league of most statistical organizations, basically leading to two, not excluding, options, outsourcing, and sharing statistical services between statistical organizations.

23. Regardless of the financial dimension, outsourcing needs to be balanced with the need for statistical organizations to govern their data's availability, security, and integrity. Outsourcing and governance aren't necessarily opposites, however, and outsourcing

a part or even the complete data collection process doesn't automatically imply a loss of governance for the statistical organization. Under some conditions, data may be even more securely governed by an external party if it's more knowledgeable and experienced in this field than the in-house IT department. Precise contractual terms are imperative, but technological developments, like containerization, enable new approaches to outsourcing.

24. There's another angle to outsourcing that, in our experience, tends to be somewhat neglected in the discussions on how to integrate smart surveys in today's statistical production. Although we believe that outsourcing the technology to have smart surveys may be the way forward, we also think that statistical organizations should avoid outsourcing the data collection as such.

25. It's not so much a matter of one thing or the other; it's subtle. If we consider the complete GSPBM life cycle of a survey, outsourcing may have a varying impact on the loss of grip on the survey process by the statistical organization. Therefore, modular, interoperable applications that offer a wide range of easily deployable services are preferable. Applications should have the capacity to take the burden of statisticians while simultaneously empowering them in critical parts of the survey process. We could refer to this as governability, the extent to which an application empowers statisticians to govern the survey process independently.

26. A comparison that may be far-fetched but valid, nevertheless, is the analogy with a package deal in the travel sector. These deals allow travelers to decide in complete freedom which services they wish to outsource and what they will provide for themselves while remaining in control of what the trip ultimately will look like.

27. The analogy with a package deal consisting of services fits smoothly with the GSPBM and CSPA. In CSPA[3] a statistical service performs one or more tasks in the statistical process, and services may operate at different levels of granularity. Small services cover small pieces of functionality; larger services may take on large parts of the

[3] https://statswiki.unece.org/display/CSPA

statistical production. From this point of view, a smart survey app is a system of integrated services that deal with particular (business) functionalities. These business functions are linked to the GSPBM business process model[4]. The wider the variety of services in the 'package' that an app can offer, the more phases of the GSPBM it will be able to cover.

28. GSPBM is not a rigid framework, however. It merely determines possible steps in the statistical business process and their interrelations. In reality, the model elements may occur in different orders in different projects, and subphases may have to be run through several times. This raises the need for flexibility and modularity in particular. Modularity means that an application should be able to deal with a segmented production chain broken down into subprocesses [4]. The principle of data abstraction is essential to modularity; modules should operate and be maintainable, independently of each other, except for their interconnections through the interface they share [4].

29. Features like modularity and shareability reduce the cost of development and maintenance. They determine the usability of an application from an organizational viewpoint. Because the angle is completely different, we propose to refer to this kind of usability as deployability.

30. If we apply the ISO's definition, deployability may serve as an overarching measure of an application's capability:
    - to be deployed by statistical organizations (the specified users in the ISO's definition)
    - to perform tasks of the GSPBM production cycle (the specified goals),
    - to conform with the CSPA recommendations about architecture, and to enhance modularity and shareability (efficacious and efficient, and usable in specified contexts),

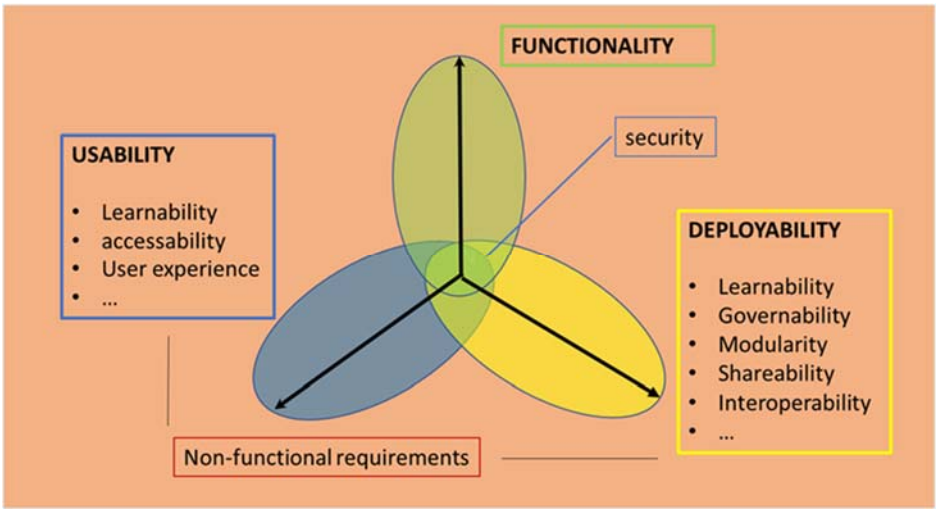
[4] https://ec.europa.eu/eurostat/cros/content/gsbpm-generic-statistical-business-process-model-theme_en

- and finally to make work easier and foster collaboration in and between statistical organizations (the with satisfaction element in the ISO's definition)

31. Deployability also relates to CSPA's principles about requirements. Applications have to meet both functional and non-functional requirements before statistical organizations can deploy them. Functional requirements concern the technical functioning of software or an application; non-functional requirements relate to how something works. In relation to the discussion here, functional requirements relate more to the functionality of an app, while non-functional requirements to its usability, and deployability in particular.

32. A 3-dimensional space emerges of this, with functionality, usability, and deployability as principal axis or spheres. While partially overlapping, each of these three stands for a distinct approach to the performance of an application or any other statistical service. They can also be considered as abstractions that group related features. Figure 1 visualizes that space, the three principal dimensions, and a tentative positioning of several other significant features in that coordinate system.

Figure 1: Representation of a three-dimensional construction formed by the functionality, usability, and deployability dimensions of applications or statistical services.



33. As the figure shows, some concepts are at the very heart of this system. Again, data security is a good example; it relates to all three spheres. The 'by design' extension that usually goes with it expresses this. This implies that to optimize the security of an application, all three perspectives need to be considered, not only the functional dimension. This is critical in choosing privacy-preserving technology; these techniques should not only do what they are made for, but they should also function in a way they don't become an obstacle, either for respondents or the statisticians organizing the data collection.

34. As always, a visualization simplifies reality. In reality, the features may counteract or reinforce each other between and within the main spheres. Certain functionalities may enhance deployability but impact usability negatively, or security may hamper shareability. Moreover, the same features can be rated differently according to the context of usage of an application and the organization adopting it.

35. The consequence is that the 3-dimensional space in figure 1 is itself an abstraction. Every statistical organization will have its version of this space and will calibrate spheres and features differently. Even organizational culture may play a role. Undeniable, this impedes shareability. If adopting a solution implies a change in an organization's DNA, resistance will be fierce.

36. We see modularity as a way forward to alleviate these kinds of tensions. An application, or a data platform, offering essential functional services, in combination with the ability to build modular flows on top of its technical kernel, grants statistical organizations the freedom to compose the data collection instrument they need, not a common denominator. If the core functionalities of this instrument are shared between statistical organizations, and within organizations between processes and domains, a high degree of shareability is still a viable option, however.

3. Research design.

37. Keeping the issues in mind that we raised above on deployability and governability in particular, we aimed at making our experiences with MOTUS in this pilot as hands-on as possible. We (Statistics Belgium) were able to operationalize the design of this study largely independently of the Free University Brussels, except for a limited number of interventions on their part. That said, the guidance they gave us, and Joeri Minnen in particular, was by times indispensable. We also highly appreciated the time they were willing to invest, for free, in educating us on MOTUS.

38. The overarching objective of this pilot was acquiring expertise about MOTUS, and the hands-on approach guaranteed us an in-depth experience with MOTUS, way beyond merely testing a survey. Testing a survey is evaluating the front end of the application. We partially worked with the backend and experienced setting up almost entire built, collect, and analysis phases of the GSBPM research cycle, regardless of the size of the pilot. The results of this objective are purely descriptive. We give an account of our experiences while preparing and organizing this pilot.

39. Within this overall objective, usability was the central theme. Usability is a vast field, and it would by no means be exaggerated to devote an entire research or even a project to usability issues in the context of smart surveys. In this pilot, we touch on three facets that relate to usability. First, we evaluate certain aspects of the usability of the short mobility survey. Secondly, at the end of the research, we appraise participants' perception of the overall experience and the use of MOTUS in particular. Thirdly, we also assess the use and functioning of notifications and the geofence. Although the research population is small, the results of both the analysis of the survey's usability and the overall user experience at the end are empirical. The results of the experiment with the notifications and the geofence are descriptive, on the other hand.

40. Lastly, the substantive objective of this pilot was assessing passenger mobility. Passenger mobility refers to the movement of people using any motorized, non-motorized, collective, or individual means of transportation[5]. Measuring passenger mobility is complicated, as there are many dimensions to cover. Consequently, we limited the scope in this pilot to an elementary proof of concept. We analyzed whether it was feasible to use MOTUS in its current stage of development to calculate a few of the principal indicators on passenger mobility.

41. An affirmative answer to that question has some value, in our opinion, for the existing efforts that are undertaken in the European Statistical System to innovate the collection of data on (passenger) mobility. However, the importance of that conclusion might even reach further. If this works for passenger mobility, it might also work for other domains, which supports the idea of using diary research and tools like MOTUS on a much broader front.

42. Table 1 gives an overview of all objectives in this study. For the Essnet on smart statistics, the overarching objective, the experiences we had with organizing this pilot, is the most interesting, followed by the usability assessments. However, we structure the research section the other way around. We start with reporting on passenger mobility, we then turn to the usability assessment, and we end with a description of our experiences with MOTUS.

[5] https://ec.europa.eu/eurostat/statistics-explained/index.php?title=Passenger_mobility_statistics#:~:text=Passenger%20mobility%20refers%20to%20the,or%20individual%20means%20of%20transportation.&text=One%20condition%20for%20receiving%20a,the%20end%20of%20this%20article).

Table 1: Research objectives.

STRAND	OBJECTIVE	RESULTS
mobility	1. feasibility calculation of principal passenger mobility indicators.	empirical
usability	2.1 usability mobility survey.	empirical
	2.2 overall user experience.	empirical
	2.3 Usability/functionality of the notifications and the geofence	descriptive
deployability	3.1 Experience with the build phase.	descriptive
	3.2 Experience with the collect phase.	descriptive

43. We strongly advocated a stepwise approach to developing smart surveys in the introduction. The research flow reflects that conviction. We build up gradually from questionnaires to the diary and then introduce the more innovative elements. Table 2 gives an overview of every step or component of the flow. Where applicable, the table also shows the relationship between each step and one or several of the research goals.

Table 2: Detailed overview of the research flow.

R1: Producing indicator passenger mobility

R2: Assessing Usability

R3: Gaining experience with MOTUS

[]	STEP/ELEMENT	TIMING	CONTENT	OBJECTIVE
0	Initializing communication recruitment	W-2	Invitations, recruitment, logins, briefing	R3
1	Mobility survey	W-1	Mobility	R1/R3
2	Usability survey	W-1	Specific usability issues survey	R2
3	Diary	W-0	Main diary – reference day – activity related questionnaire	R1/R3
4	Notifications	W-0	Parallel to main diary – 11AM – 3PM reference day	R2/R3
5	Geofence	W-0	Parallel to main diary – questionnaire	R2/R3
6	UX survey	W-0	Overall user experience survey	R2

44. The first two or three steps in table 2 correspond to the essential elements of organizing any survey. Like actual fieldwork, the flow kicks off with sending out the invitations and introducing the research. One of MOTUS's strengths is fully integrated functionality to organize and operationalize this. The survey on mobility [1] surveys the modes of transportation people have access to and other baseline information needed to calculate the mobility indicators. Peoples' mobility behavior as such, however, is assessed in the diary.

45. The first aim of the usability survey [2] is to probe some aspects of the respondents' experiences with completing the mobility survey. However, it also addresses the perception of issues like downloading and initializing MOTUS and communication. The

user experience (UX) survey [6] takes a much broader angle by questioning MOTUS's overall experience. It comes entirely at the end of the research and incorporates the entire experience, including the diary and working through the task list.

46. At this point, the diary functionality in MOTUS still has a format tailored to the needs of the Time Use Survey. However, the colleagues of the VUB intervened in the back end to enable us to tweak that format into a suitable design to assess mobility behavior [3]. This resulted in a 'lightweight' diary with basically only two activities, being somewhere versus being underway, without obligation to register a secondary activity. To additionally lower the response burden, we also reduced the reference period to a single day, Thursday, October 26, 2021.

47. As a diary implies by definition the registration of time and activities, diary research shifts the research focus to the interdependence between time, behavior, and context, especially when questions assessing contextual information on the activities are added. MOTUS enables this explicitly by featuring functionality to link contextual questions to the registration of a particular activity. In this pilot, every time a respondent registered a trip or an errand, MOTUS presented a set of additional questions on the purpose and mode of transportation.

48. In principle, 'mobile' diaries have a significant edge in countering common sources of error like recall bias. They allow to register activities instantaneously, not only when the respondent has access to a laptop or the paper version of the diary. Consequently, activity-registration spans may become much shorter, increasing the quality of both registrations and contextual data. However, the advantage of going mobile and having a short activity-registration span remains purely theoretical if respondents miss the reflex to register an activity shortly after ending it.

49. MOTUS features functionality that might help enforce that reflex, such as sending notifications to the participants. We sent several messages to the participants in which we urged them to record their trips and errands they had made during the diary reference day. We timed these notifications at 11 a.m., at three and six p.m.

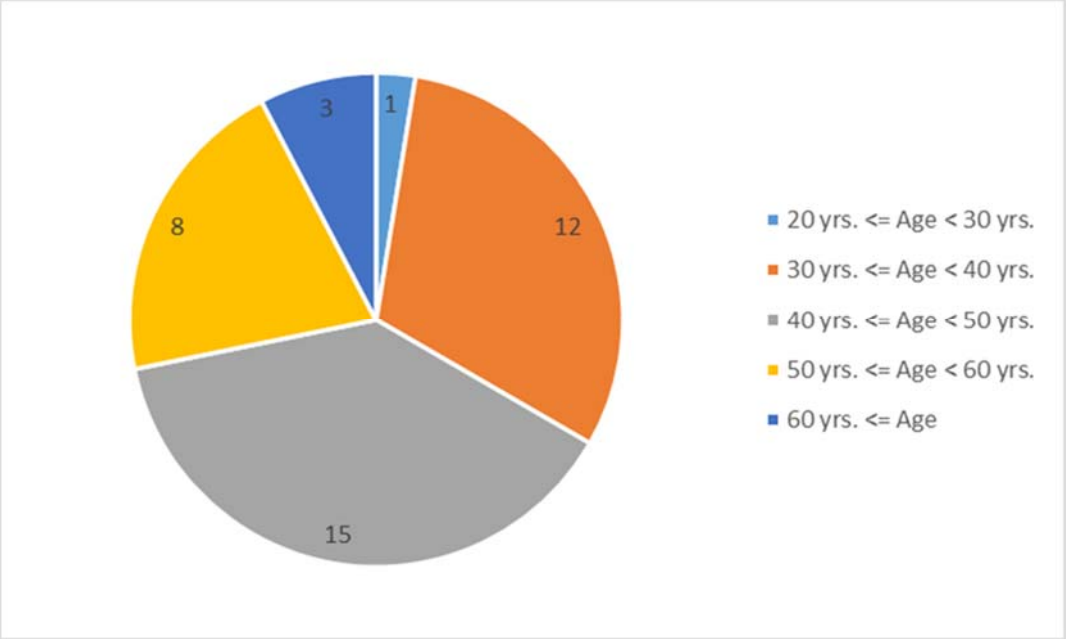
50. Being able to send out timed notifications is quite innovative compared to traditional data collection methods, but the geofence takes it a step further. A geofence triggers the application each time a respondent enters or leaves a predefined area. In this pilot, we operationalized a geofence around the premises of Statistics Belgium. Participating Statbel staff who showed up at the Brussels office during the diary reference day were urged to register their commute and automatically got a single question presented. The question had no substantive meaning to the mobility research as such; its only purpose was to evaluate the functioning of the geofence.

51. We recruited the study population of this pilot in two steps. We sent a general introduction by work mail to every 'Statbel' collaborator inviting them to express interest in participation. Exactly 50 of the plus-minus 350 employed by Statbel showed an initial interest to participate. In the next step, we invited these 50 persons to install MOTUS via a new email we sent using the application's communication functionalities. Thirty-nine colleagues finally did so and embarked on the experiment. Unfortunately, we experienced some dropout in the course of the pilot, especially in relation to the diary. Five participants had no entries in the diary; the information was of poor quality of three other participants. Therefore, we only have valid diary inputs for 31

participants. However, we recovered these participants for the end survey, which again concerns 39 respondents.

52. Despite the limited recruitment setting and the small study population, there was some diversity among these 39 participants. As Figure X shows, 11 of the 39 participants were at least 50 years of age, while those in their forties were the largest age group (15). Young people, or what could pass as young in this context, account for a third of the research population (minus 40 – 13 participants). There was also some variation in relation to the experience with data collection methods and familiarity with the use of apps. We report on these differences in the result section, however.

Figure 2: Age distribution of the 39 participants in the pilot.



4. Results.

4.1. Objective 1: Passenger Mobility.

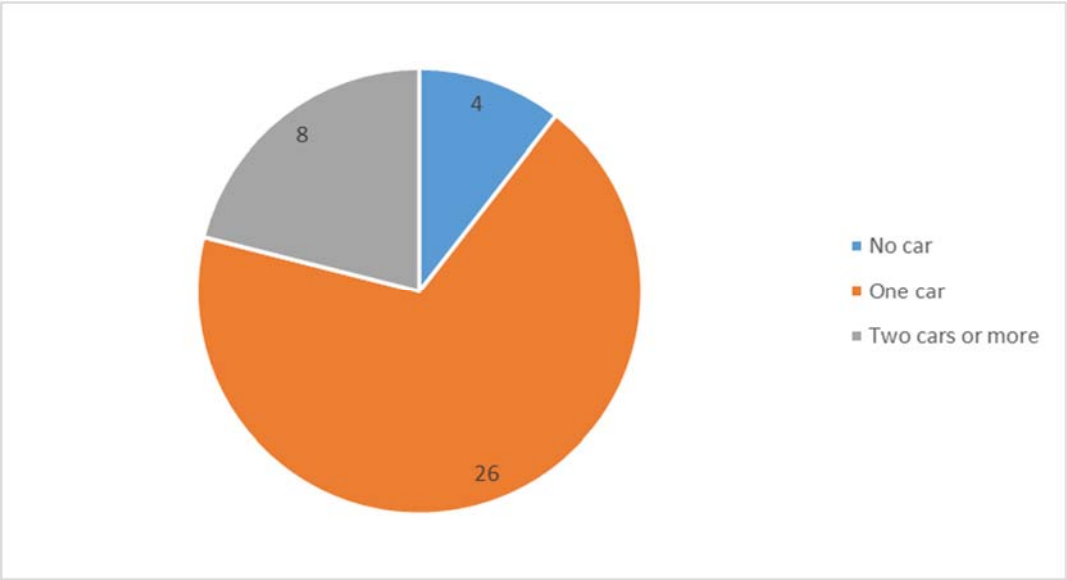
53. Passenger mobility is a vast domain, and there's by no means room in this document to cover even the basics. However, in the last decade, several initiatives and projects such as COST/SHANTI[6], and the Task Force on Passenger Mobility Statistics [7] resulted in excellent documentation on the subject.

54. Being a proof of concept, we focused in this pilot on basically two types of information; getting a view on the transportation means available to the respondents via the mobility survey and calculating several basic passenger mobility indicators based on the registration of trips in the diary. The indicators are far more interesting than the survey results, both substantively and methodologically. The survey is an introduction to the research and the theme. The experiences respondents had with answering these few questions also served as a reference for the questions on usability.

55. Although of lesser importance, the survey did a good job characterizing the availability of transportation means. The results don't surprise. Except for three, all respondents have a driving license, and car possession, or access as the question was formulated, is widespread. Figure 3.1 gives an overview. Only four respondents don't have access to a car or another four-wheeled motorized vehicle. Two-thirds (26) have one car available to their household. Seven had two, and one participant had even four cars available in his or her household.

[6] https://www.cost.eu/cost-action/survey-harmonisation-with-new-technologies-improvement-shanti/

Figure 3.1: Number of available (motorized 4-wheeled) vehicles to the household (1 case missing).



56. Only one respondent reports having a motorbike. Bikes, both traditional and electric, are standard, only one respondent indicates not to have a bike, and twenty participants have more than one bike in their household. Finally, being public servants, 36 participants have public transport passes.

57. As everywhere throughout Europe, the Covid-19 pandemic impacted mobility also in Belgium. The diary reference day (Thursday, October 26, 2021) fell within a period with firm Covid measures in place. Teleworking was strongly advised, and the presence at the office was limited to a maximum of one day a week. The choice for a Thursday as reference day was based on the observation that this was the day of the week with the most people present in our main office in Brussels. Eventually, eight participants turned up. Consequently, rather than commuting, our initial interest, the information collected in this pilot reflects how people spent their day working at home and how this new way of working impacts mobility.

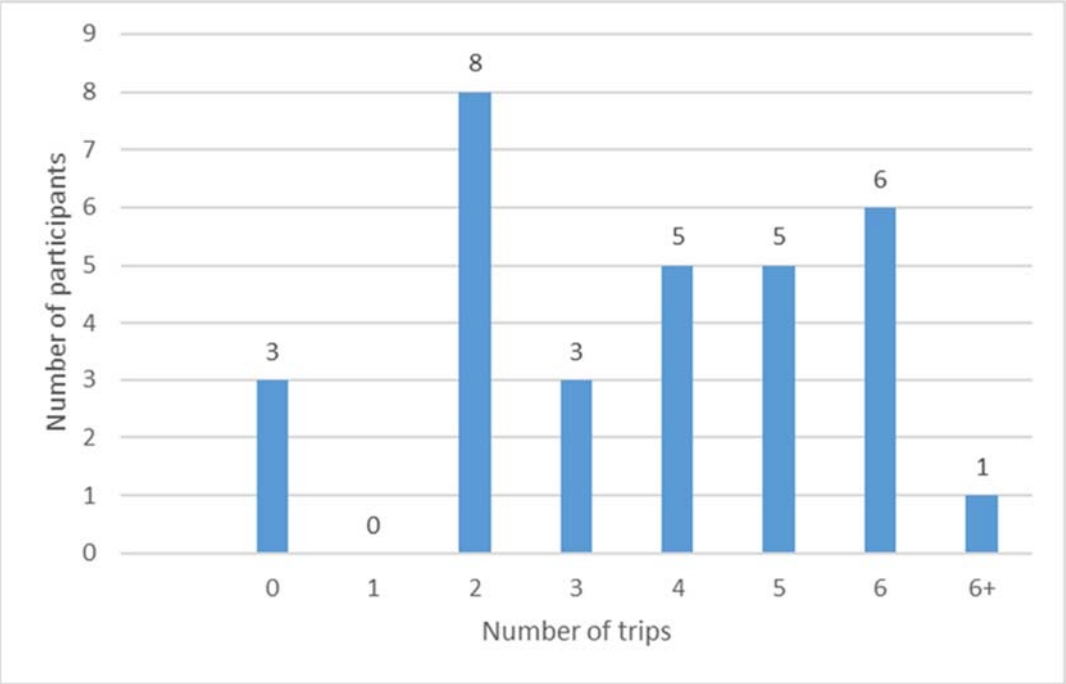
58. As we already mentioned, we lost eight participants between the end of the usability survey and the start of the diary. The quality of the diary entries of the remaining 31 participants was fine, except for one issue, widespread in mobility research, caused by misinterpretation of the definition of a trip.

59. We applied the conventional definition [7, p.8]; a trip is a movement from an origin to a destination. By this definition, dropping someone off and returning makes up for two trips. However, there's a tendency among respondents in mobility research to interpret such a roundtrip as one trip.

60. As this definition quickly gets complicated in practice, lengthy explanations and elaborate illustrations are needed to counteract this tendency. These are much harder to display on a mobile phone, at least in a conventional way, without new media. An additional mistake was that participants couldn't return to the screen with the information on the trip definition after being introduced to it at the start of the diary. This was a beginners' mistake, not an issue with MOTUS. However, regardless of this mistake, the observed trip definition errors were a stark reminder that using smartphones to organize surveys necessitates an entirely different, read a visual approach to communication.

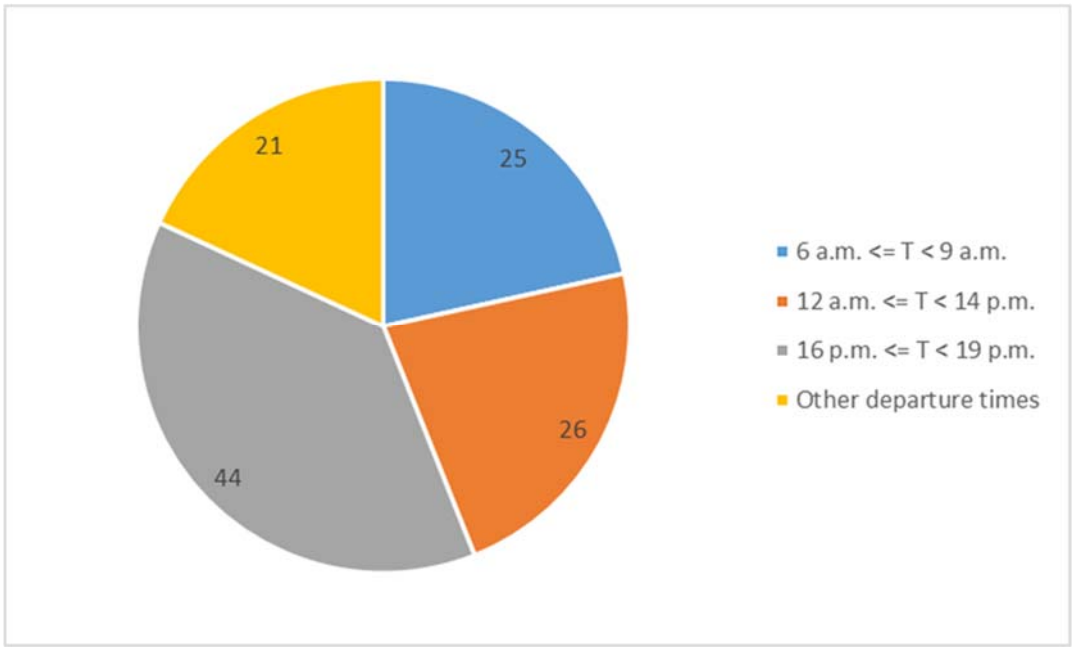
61. The 31 participants initially reported 93 trips during the diary reference day. To correct the above issue, we imputed 23 additional return trips, resulting in 116 trips based on information of 31 persons. Most participants made up to four trips during the diary reference day (mean 3.74, median 4.0). Even in this small group, there are marked individual differences. Three persons didn't make any trip during the reference day, while one person made as many as ten journeys.

Figure 3.2: Number of trips by participants N = 31.



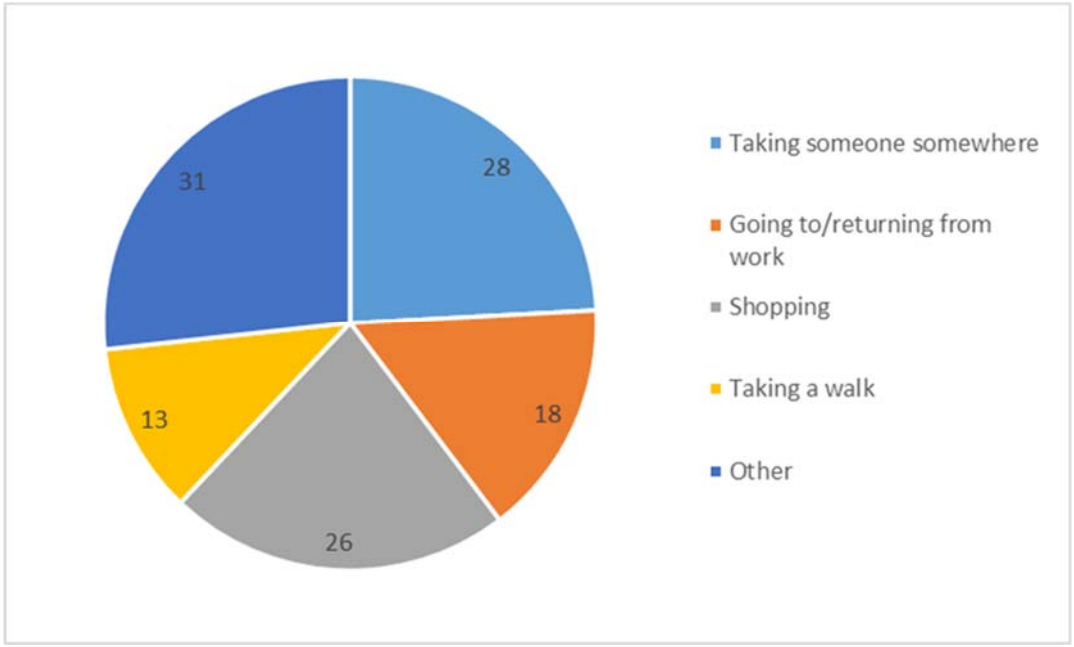
62. Although spread throughout the day from 6 a.m. to midnight, the number of trips reported peaked early in the morning, during the lunch break and early evening, as expected. A quarter of the trips were undertaken before nine a.m. (25 trips or 21,5%), another quarter between midday and two p.m. (26 trips or 22,4%), and a little more than a third of the trips from four to seven p.m. (44 trips – 37,9%).

Figure 3.3: Number of trips by departure hour N = 116.



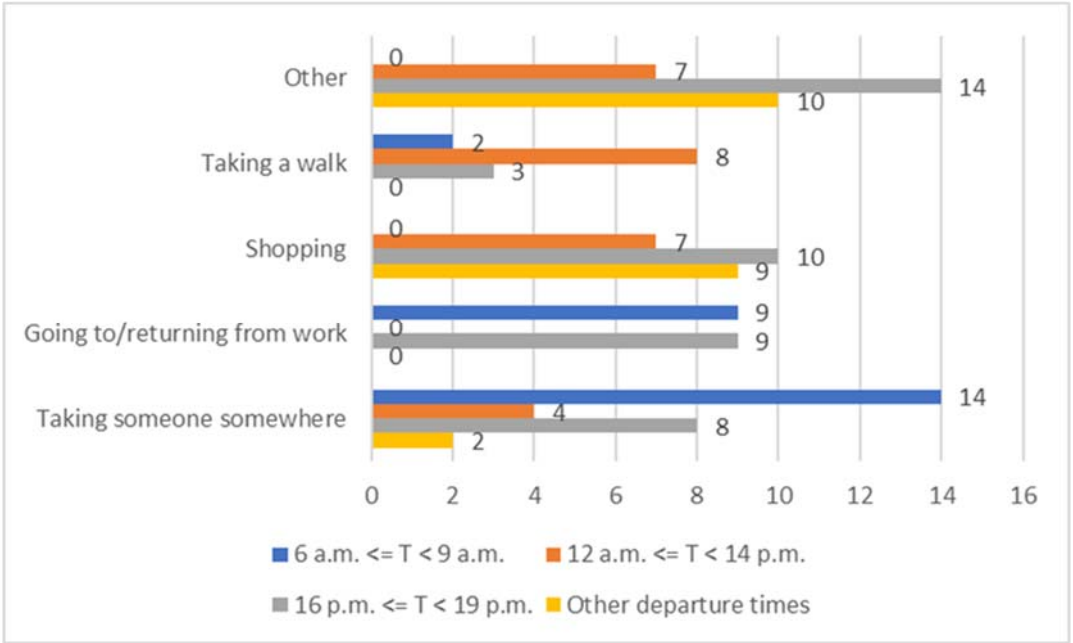
63. Also, trip purposes were in line with the expected movements on a (tele-) working day. Together, almost half of the trips were associated with taken someone somewhere (28 trips or 24%) and shopping (26 trips or 22%). Commuting accounted for 18 trips (15%). Considering the definition of a trip, not all of these trips were exact roundtrips. Going for a walk was the fourth most frequent trip purpose (13 trips). Twelve of the 31 remaining trips in the category other were associated with a private matter (e.g., a doctor visit), five with having a meal, two each with following a course and visiting family, and ten were classified by the respondents as having another purpose than the once already mentioned.

Figure 3.4: Number of trips by trip purpose N = 116.



64. In figure 3.5, the temporal distribution by trip purpose is given. Once more, the results do not surprise and show the expected association between time and purpose.

Figure 3.5: Number of trips by trip purpose and departure time N = 116.

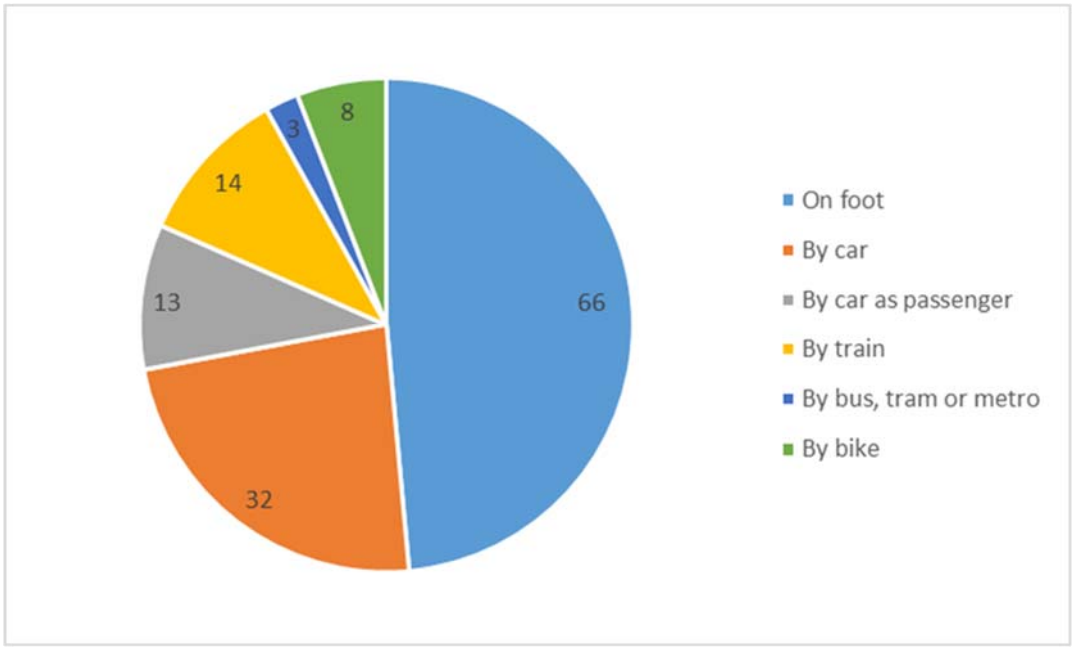


65. Exactly half of the 28 trips that involved taking someone somewhere took place before 9 a.m. Participants shopped during lunchbreak (7 out of 26 trips) and after work (10 out of 26). Finally, more than half of the walks were done during lunch break (noon to 2 p.m. - 8 out of 13 walks).

66. Looking at the other way around, we see that trips before 9 a.m. are almost entirely related to starting the day. More than half of the trips (14 out 25) involved bringing a household member somewhere, and nine trips concerned going to work. Except for the two walks, we didn't observe any trips for other purposes, not even shopping. The participants actively spent their lunch breaks with walking, shopping, and several other activities. Finally, we observed the most diversity in trip purposes in the hours immediately after the end of the working day (4 to 7 p.m). Many respondents were underway during these hours for diverse reasons.

67. Most of the trips were partially made on foot (66 of the 116 trips in total). Cars are widely used, which doesn't surprise, but the difference is marked compared to public transport and cycling. A vehicle was involved in 45 trips. In 13 of these, the respondent was a passenger. Fourteen trips were at least partially made by train, and bus, tram, or a metro were only used in three trips. Finally, only eight trips were made by bike. Figure X below gives an overview. We also observed some differences by trip purpose. Of the 28 trips to take someone somewhere, 15 involved using a car, and nine were done on foot. Shopping slightly shows another pattern. Only ten involved using a vehicle, and 15 were at least partially done on foot

Figure 3.6: Number of trips by transportation mean.



68. In figure 3.6, the sum of all modes doesn't add up to the total number of trips (116) simply because of the modal split, combining several transportation modes in one trip. This flexibility is essential in mobility studies, as it's seen as one of the key solutions to mobility problems.

69. In 24 trips (21%), several transportation means were combined. Walking was combined with another mode in all except one of these trips, and 14 of the 24 concerned commuting. This implies that of the 18 commuting trips in total, 14 involved a combination of walking and another mode of transportation. Not surprisingly, the train is that other mode of transportation. It was used in 12 of these 14 trips. We observed the combination of car and train, whether or not in combination with walking, only four times.

70. Although these results are highly anecdotal and all kinds of circumstances play a role, they trigger substantive questions. Even in this small study population, working from home diversifies mobility patterns. We observed people grasping opportunities that

probably weren't an option before; taking a household member somewhere in the morning, for example, or having walks during lunch breaks.

71. There's no explicit need to organize a study on mobility to infer these insights, as any regular use survey also covers this information. Unfortunately, the reality is that in a time use survey, the perspective on mobility tends to get watered down to avoid response burden and to balance it with other themes. This small pilot proves, in our vision, that it's feasible to collect very detailed information on peoples' mobility behavior by slightly adapting existing technology to organize time use surveys, and MOTUS' case in particular, by harnessing the modular capacities of the application.

4.2. Objective 2: Usability.

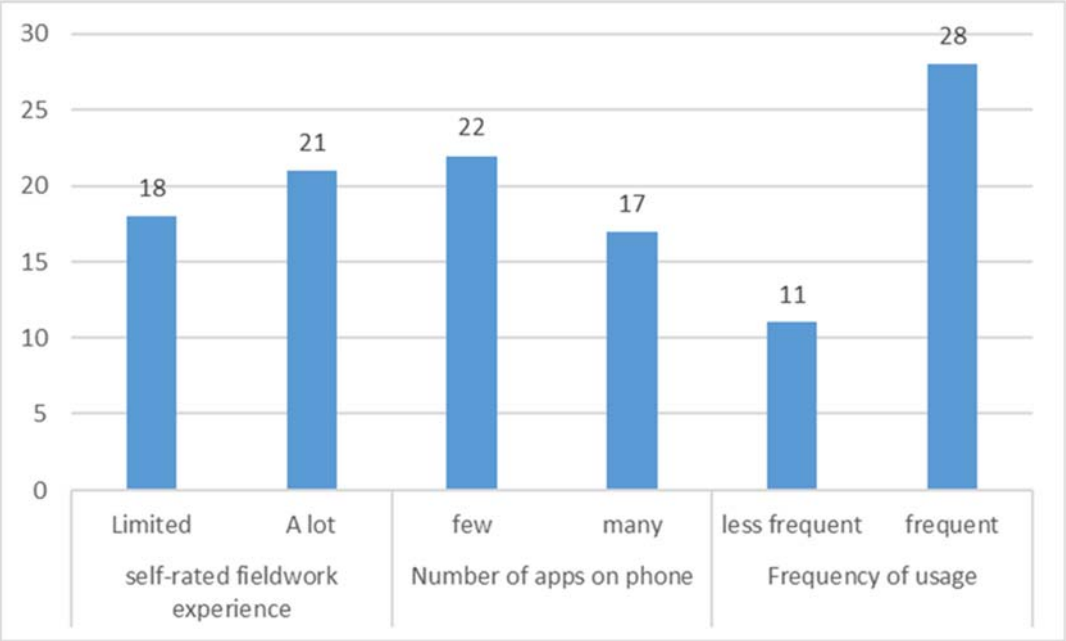
4.2.1. Usability assessments.

72. The extensive research flow gave the participants in this pilot the opportunity to become acquainted with many different aspects of MOTUS. As indicated in the section on the research design, we went into three of those aspects: (1) we analyzed how they perceived the usability of the mobility survey, (2) we examined their global perception at the end of the research, and finally, (3) we also assessed how the notifications and the geofence worked. As they are related, we cover the first two items together.

73. Actively participating in a survey, answering a questionnaire, is the most essential activity a respondent can have in a smart survey application. However, even for an activity as basic as that, many parameters will influence the rating of the experience. Features of the application, in the first place, but also the organization of the research, and individual characteristics of the respondents, will all have an impact.

74. Regarding the respondents' characteristics, the limited recruitment setting and the small study population made us expect the worse. Like xe already mentioned, however, at first sight, the 39 Belgian participants turned out to be somewhat diverse after all. We probed for several characteristics that could potentially influence their experience with MOTUS, including the extent to which they are involved in data collection, whether they have many various apps on the smartphone, and the extent to which they use them.

Figure 4.1: participants' experience with data collection, number of apps, and app usage frequency. N = 39.



75. Just over half of the respondents (21 of the 39) rated themselves as having a lot of experience with data collection and having a good view of the issues involved. A majority among them, 27, or almost seven out of the ten, had more than ten apps. Having many apps doesn't imply using them or using them often. As figure 4.1 shows, overall, only 17 respondents said they use many different apps, irrespective of the frequency of usage. In comparison, 28 respondents indicated that they use apps regularly (defined as at least several times a day). This suggests a tendency among the participants to use only a few apps but use them regularly. The number of only nine participants that indicated they use many apps regularly throughout a day reflects this

76. Despite the differences in their profiles, a short explanatory analysis revealed that these characteristics hardly influenced the participants' opinion on MOTUS, neither on their general rating of the application nor on the perception of distinct features. The most likely explanation is that this group was simply too small to find any trace of an effect.

77. Nonetheless, our exchanges with many participants during the pilot revealed another reason. Several participants with relatively advanced digital skills indicated that they also struggled with at least some parts of the pilot. Further enquiring revealed, however, that it was the sequence and the range of tasks that troubled them rather than the technical dimension. Consequently, as the task list was a little bit of a struggle for everyone, this might have prevented the appearance of the expected divide between those with high and less developed digital skills. Moreover, the observation that most participants experienced issues at some point in the pilot doesn't mean that the participants' experiences were negative overall, as the results below will demonstrate.

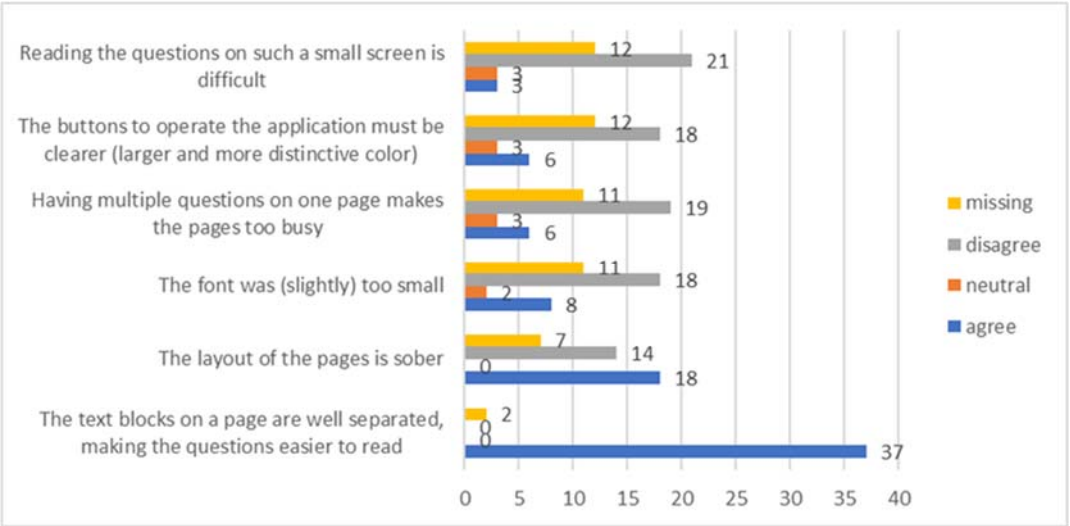
78. Let's first turn to the initialization of the research. We observe that almost nine out of ten respondents (35 of the 39) indicated they were sufficiently informed and had enough information to start the research. Thirty-four out of 39 had no issue finding the application in the app or play store. Nobody experienced problems during installation, and only two persons reported minor issues with opening the app for the first time and initializing the research (but eventually succeeded). Finally, 32 participants, or four out five, thought that the initializing mail sent to them from within MOTUS was attractive enough in terms of layout and structure. Interestingly, though, five of the seven colleagues who weren't convinced are directly involved in collecting data at Statistics Belgium.

79. We had six questions assessing respondents' perception of basic questionnaire features like layout, font sizes, spacing, and presentation of questions. Unfortunately, these questions suffered badly from partial non-response as no less than eleven respondents left all questions blank from the second question in the series onwards. However, anyone of them had an indifferent response behavior elsewhere in the survey. The reason for this deviation became soon apparent. These six questions were practically the only ones presented on a single page, implying that the respondents needed to scroll down to answer the second and following questions. However, the button to continue to the next page is immediately visible from the first question. Consequently, depending on screen sizes, the beginning of the second or third

questions wasn't visible, which misled the eleven respondents to immediately continue to the next page, skipping most of the questions. While this is an embarrassing mistake from the point of view of this report, it's a very valued learning experience in general.

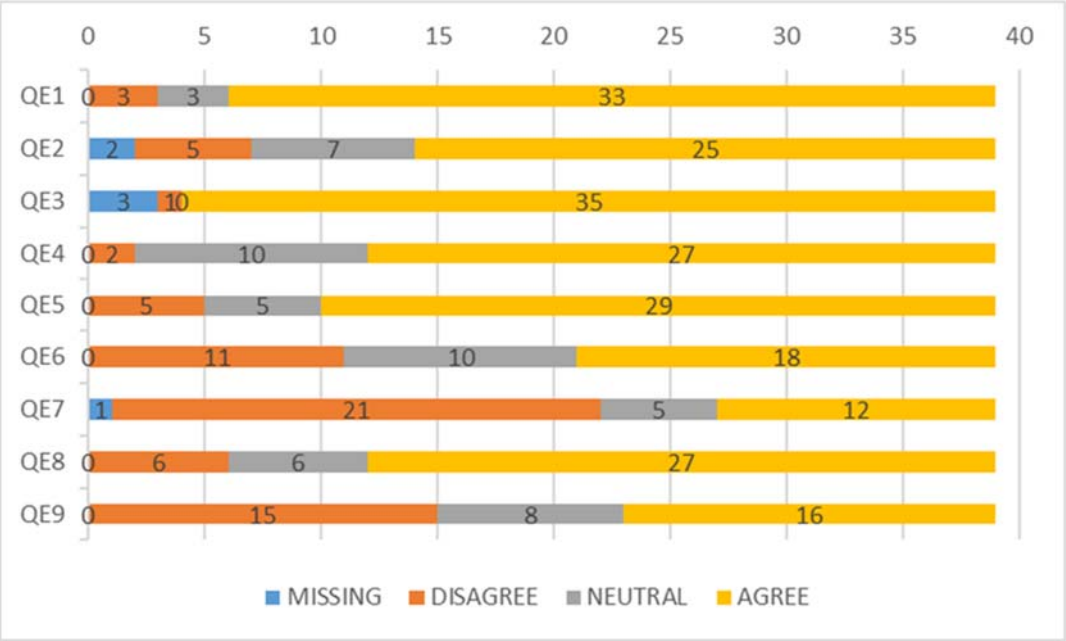
80. Figure 4.2 gives an overview of the distribution of the answers for each of the six questions. The initial 5-point scales (fully agree to fully disagree) were transformed into a 3-point scale. The table indicates that, varying from question to question, roughly a fifth to a fourth of the respondents, or even a third if the neutral category is included, expressed some concerns about particular features. For example, eight out of the remaining 28 agreed to the statement that font sizes were slightly too small, three found reading questions on such a small screen difficult, and six indicated that buttons could be made more visible. We also asked respondents if they had issues with working through the questionnaire, going from question to question, and sending the results after finalizing it. Very few respondents reported problems with this, however.

Figure 4.2: answers to six questions regarding usability experiences with survey completion. N = 39.



81. Without downplaying the issues and concerns reported in the questionnaire on usability following the initial mobility survey, the overall perception of the experiment and MOTUS didn't seem to be substantially impacted by them. As mentioned before, the more general questions on respondents' perceptions were presented at the end of the research, implying that they cover the whole experience and not only the survey. As some reported issues with working through the research flow, good appraisals become less likely, in principle. Nevertheless, the respondents generally rated both the pilot and the overall experience with MOTUS as reasonably positive. Figure X gives an overview of the answering patterns of nine questions covering the participants' general perception of MOTUS and this research.

Figure 4.3: answers to nine questions regarding overall user experience and rating. N = 39.



QE1	Overall I think that for a first test, the functioning of this app is good.
QE2	I prefer this way of participating in a survey over the traditional questionnaire methods.
QE3	I think this kind of technology has a future within our statistical production.
QE4	Many respondents (with a smartphone) prefer this way of participating in a survey over participation with an interviewer (CAPI).
QE5	I think that many respondents (with a smartphone) will prefer this way of participating in a survey to participation via laptop or PC (CAWI).
QE6	Many respondents will be more worried about their data security when they are questioned via smartphone than with traditional survey methods.
QE7	As a respondent, I would be more concerned about this.
QE8	Many respondents will experience problems with the technical side of participating via smartphone (downloading the app, filling out the survey, etc.).
QE9	I also found this not so easy.

82. Thirty-three of the 39 participants agreed or rather agreed with the statement that for a 'first' test, to use the exact wording of the question, MOTUS functioned overall well. Almost nine out of 10 participants (35) see a future for this technology as part of the Statbel data production. This is even ten out of ten for those involved in the organization of fieldwork (21 persons) (except one person who didn't answer the question). Additionally, precisely 32 participants indicated that they preferred this way of participating to research above the traditional way. Of the remaining seven, five hesitated, and only two expressed that they preferred conventional survey methods.

83. We also asked the participants in this pilot if they expect similar preferences among actual respondents in the surveys we organize as part of the current statistical production. Twenty-seven agreed to the statement that 'many respondents will prefer this way of participating above CAPI, but ten took an intermediate position and didn't agree, nor disagree. With 29 agreeing, a similar answering pattern emerged concerning CAWI. However, while only two participants disagreed with the statement about CAPI, five disagreed concerning CAWI.

84. Data protection is a vital issue to Statistics Belgium, so not surprisingly, most participants in the pilot showed a particular sensitivity for the matter. Therefore, 18 of the 39 agreed with the statement that 'respondents will be more worried about their data security when participating by phone compared to traditional data collection tools, and an additional ten took an intermediate position. When asked if they would also be more worried as a respondent in actual research, 12 answered affirmative, five hesitated (didn't agree, nor disagree), and 21 expressed no issues (they disagreed with the statement). Of these 21, however, ten indicated that they expected respondents, in general, to have worries concerning data security, and only four claimed this wouldn't be an issue. The other seven took an intermediate position. This result indicates that the participants in this pilot were conscious that their reference frame is somewhat different from the average respondent's.

85. The answers to the final two questions reflected that awareness. Asked if they rated their experience as challenging, 16 participants agreed, eight didn't agree nor disagree, and 15 disagreed. However, 27 participants agreed with the statement that it could be expected that ordinary respondents might encounter problems when participating in this kind of research. Of the 16 that rated this study as challenging, 14 think some respondents will also face problems. This figure is still half (7) among the 15 that indicated having no issues with this experiment.

86. The number of participants that predict issues for ordinary respondents is, at first sight, surprisingly high. This seems contradictory to the observation that the majority found that the application functioned fairly well for a first test and that this kind of technology has a place in future data production. Out of the exchanges that we had with many of the participants, it became clear that both observations rather match than contradict, however. The majority of the participants shared two conclusions. The first one is that it's not only about MOTUS, and by elaboration, not only about any application. Most participants clearly distinguished between the functioning of the tool and the challenges that arise with organizing fieldwork with this kind of technology. The second conclusion is that while most recognize the potential, they also realize that there's still a long road ahead before we'll be able to harness the full

potential of this kind of technology, especially in the context of complicated research designs.

4.2.2. Functioning of the notifications and the geofence.

87. In MOTUS, sending notifications or implementing a geofence can be part of any step of a research flow. However, as the focus in this pilot was on trips, both were tested as part of the diary stage of this research. This means that only the 31 participants that entered the diary stage were part of this evaluation.

88. To appraise the actual ability of notifications to support respondents' registration behavior, two primary conditions need to be met: respondents need to have received and noticed them, and there should be a reason to act on them, there should be activities to register. To get insight into the first condition, we presented them a question. In the research flow, this question followed immediately after finalizing the diary. It was only temporarily open as respondents were sent automatically to the next stage of the research flow within 24 hours after the diary day. Unfortunately, this led to additional non-response for six respondents. They had left the last activity in the diary open-ended and didn't get to see the question before getting sent to the next stage.

89. Somewhat surprisingly, of the remaining 25, as many as one in three (eight) indicated they didn't receive notifications. This can have many causes, ranging from simply not noticing them to even technical issues related to the phone type. Still, an informal inquiry learned that phone settings and preferences played a role. We urged respondents to allow the app to trace them, but we didn't insist on accepting that it sends them notifications. Several respondents testified they had refused to enable this functionality and most added they do so reflexively when installing apps. It's easy to imagine that this is a common practice. Consequently, one of the main conclusions is that while sending notifications to respondents is a great asset (on paper), harnessing their potential in practice may not be as straightforward as it seems.

90. To verify if there was a need to act on the notifications, we scrutinized the entries in the diary. The time between the end of a trip and the registration is of interest. Table 3 presents a detailed overview of this lag in relation to the departure time of the journey. The table only takes the 93 trips reported by the 31 participants into account, leaving out the 23 trips we imputed to correct for the interpretational bias.

Table 3: Time lag between departure time and registration N = 93.

Time lag between departure and registration	6 a.m. <= Dep Time < 9 a.m.	12 p.m. <= Dep Time < 14 p.m.	16 p.m. <= Dep Time < 19 p.m.	Other departure times	Total
[0 – 30 min.]	5	9	19	5	38
[30 min. – 1 hr.[	2	3	5	3	13
[1 hr. – 2 hrs.[	3	5	3	5	16
[2 hrs. – 3 hrs.[	1	1	4	2	8
[3 hrs. – 4 hrs.[	2	1	2	1	6
[4 hrs. – 5 hrs.[	1	2	1	0	4
>=5 hrs.	6	0	2	0	8

91. The participants were quite disciplined in registering their trips. 51 of the 93 trips (55%) were reported within the hour after the end of the trip. There is some variation associated with the moment of the trip. Most early morning trips were only reported more than an hour after ending the trip. There might well be a substantive explanation for this as mornings are probably more busy for most people than early afternoons, for example. However, the numbers in this pilot were too small to rule out other explanations, and traditional time use research is much better equipped to give insight into this.

92. The observation that most trips got promptly reported suggests that there were almost no trips to register when the reminders were received. Consequently, the notifications can hardly have had any impact. Figure support that hypothesis. The

figure shows the diary entries for each quarter of an hour of the diary day between eight a.m. and eight p.m. (irrespective of the starting time of the trips). The table represents only the registration of the seventeen respondents that said they had received notifications (44 of the 93 journeys initially reported).

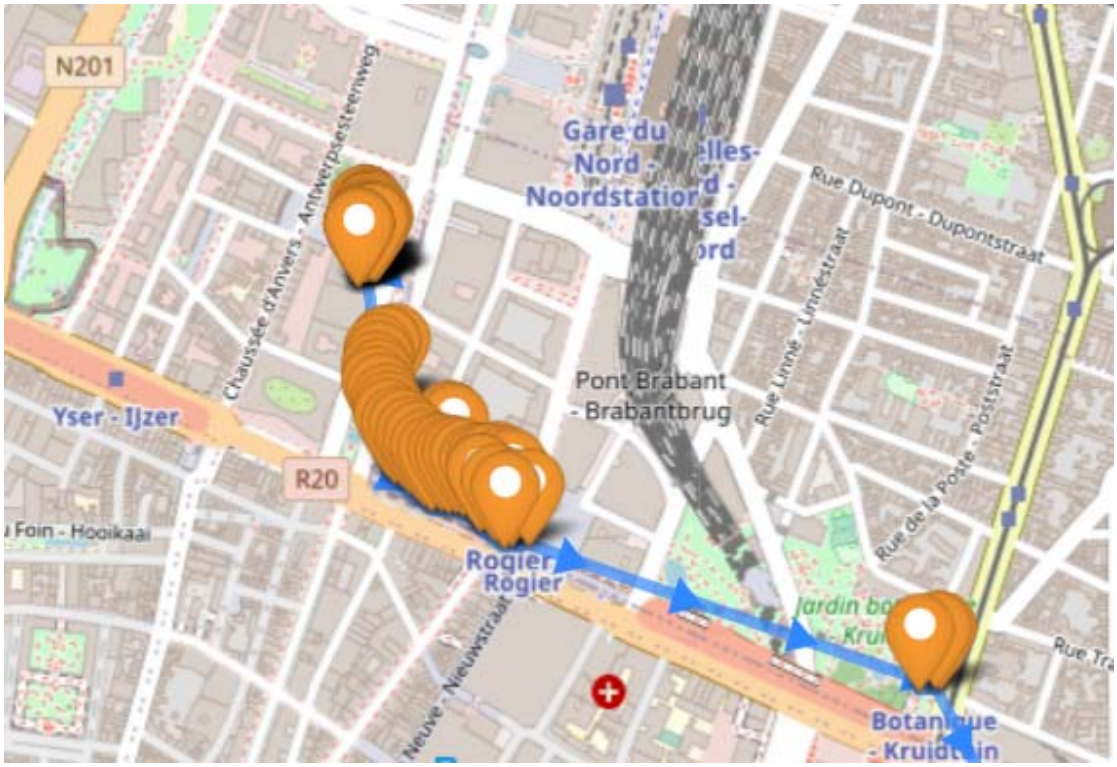
Figure 4.4: Number of trips by hour and quarter of registration. N = 44.

93. As expected, registrations tend to be spread out randomly over every quarter within an hour; 12 trips were reported during a first quarter of an hour, 6 during a second quarter, 18 during a third and finally 8 during a fourth quarter. If the notifications have an impact, we expect to see a peak in entries shortly after reception, at eleven a.m. and three and six p.m. sharp. There's no real trace of that in the figure, although the three trips that were reported shortly after 3 p.m., while no other trips got reported between 3 and 4 p.m., catch the eye. We scrutinized these registrations, and their pattern is very distinctive. They belong to one respondent and were reported within 10 minutes after receiving the 3 p.m. notification. Two of them were undertaken between 9 a.m. and 10 a.m., and the third one between 1 and 2 p.m. Assuming the impact of the message, we contacted the person in question and got confirmation that it was indeed the reminder that had given rise to these registrations.

94. Due to the Covid measures at the pilot time, we had no idea how many participants would turn up in Brussels during the diary reference day. Eventually, eight of the participating colleagues showed up at the office in Brussels.

95. Within this small group, we were able to evaluate the operation of the geofence directly, but operationalizing the geofence implied that we needed to trace all respondents throughout the entire reference day. Even while the research population in this pilot was tiny, this produced a significant amount of throughput data. Figure 4.5 visualizes part of a journey that one participant made, including the data points over this small distance. The figure is more than merely illustrative. We present it to give a slight indication of the scalability and privacy issues of implementing this technology. These issues, however, are elsewhere addressed in the Essnet.

Figure 4.5: Visualization of a part of a journey made by one of the participants.



96. As indicated above, we evaluated the functioning of the geofence using a single question that was presented to the respondent when the fence was activated. Judging by the results of that question, the test failed, or at least from a business perspective, because we have proof that the tracing functioned. We only got an answer back from three of the eight people who entered the fence that day.

97. A significant observation was that the respondents were utterly clueless about why the question didn't pop up on their phones. Some even unintentionally gave us information that made finding the cause more complicated. Additionally complicating is that examining even straightforward reasons for failure is complex if these causes can interact.

98. Ultimately, we found the most critical clue in the registrations of the activities. The three participants for whom the geofence had worked reported an activity while at work, implying that the app was activated from within the fence. This indicated that the way the app manages states transitions was at least part of the explanation.

99. Applications can be active or passive when running in the fore- or background. Operating systems automatically put apps in the background. In this case, MOTUS gets suspended when moving from the fore- to the background. This interrupted the tracing, only to be resumed when the respondent's application was reactivated. Unfortunately, every time we pretested the geofence, we reactivated the app from within the fence to check if it had worked. While an absolute rookie mistake, this demonstrates the degree of complexity that comes with the simplicity of testing an application. Or is it the other way around?

100. In principle, entering the geofence triggers a notification. Clicking this notification unlocks the phone, reactivates MOTUS (if not activated yet, a necessary condition in this case, like explained), and presents the associated (one-question) questionnaire to the respondent as part of his task list. Figure X shows the geofence in action (these

images are exact reconstructions of how the geofence was presented to the respondents on the diary reference day).

101. While there's as such no issue with this method, it brings challenges similar to those mentioned in the context of using notifications in general. Participants need to notice the notifications to act on them. This highly depends on phone settings and preferences that are partially decided when installing the application.

4.3. Objective 3: Gaining experience with MOTUS

102. Until now, we mainly informed on research results, not on our experiences with setting up the research. Organizing this pilot enabled us (Statistics Belgium) to accumulate fieldwork experiences with MOTUS leading to all kinds of observations. We discovered the power of even basic functionalities, encountered remaining technical and operational issues, and got a much more profound insight into MOTUS's potential as a data collection tool.

103. This section predominantly reports on significant features and issues that determined the overall experience. It's impossible to go into every detail. We also differentiate between the built and collect phases and address positive and negative experiences. To the first, we refer as strengths, to the latter as challenges.

104. The choice to use the term challenges instead of weaknesses isn't window dressing in favor of MOTUS. Throughout the study, the VUB-team insisted on feedback as this experimentation is also valuable. MOTUS is a rapidly evolving tool, and receiving feedback and input from users is crucial for the developing team to address the challenges ahead of them adequately. Moreover, some of the issues we raise below, and other more minor concerns, have been fixed in the meanwhile.

105. Additionally, and without downplaying the value of our work, we also have to put our conclusions a little bit in perspective. It's evident from the above results that this pilot is neither a complete nor a definitive analysis of MOTUS' capacities. Many misconceptions occurred, and the problems that we experienced, even those that seem at first sight technical in nature, are in no way by definition related to the functioning of MOTUS. The reverse may also be the case, however. If specific weaknesses didn't show up in this study, this doesn't imply that they are absent from MOTUS.

4.3.1 Built phase – strengths.

106. What we mean with the built phase essentially converges with subphase 3.1 of the GSPBM workflow, preparing the different data collection instruments and operationalizing the research flow. In this phase, we extensively used the so-called MOTUS builder, a GUI-like application in the MOTUS backend.

107. The MOTUS builder integrates a wide range of functionality in one very accessible environment. It manages all kinds of tasks such as the design of surveys and diaries, the operationalization of the research flow, the upload of respondents, the organization of logins and passwords, and almost everything concerning communication. The architecture of the MOTUS builder is more or less hierarchically organized. Most of its services are developed as separate applications, also called builders, that integrate into the overarching Motus builder. There's a survey and diary builder, a communication builder, a respondent builder, and a research builder, to only mention the most important ones we used in this pilot.

108. The latter, the research builder, is a crucial component of the overarching MOTUS builder. It operationalizes the research flow, allowing easy integration of the different research components (questionnaires, diaries, communications) in a more or less chronological flow (more or less because research steps can converge). Consequently, research flows become highly modular and can easily be tailored to specific research needs. Therefore, studies with the same components may have entirely different flows, while research elements are shareable between studies.

109. Adding to the builder's attractiveness of the builder and MOTUS, in general, is that research elements are also reusable within a single research flow. For example, the same questionnaire can serve as a standalone survey but can also be used to assess contextual information in a diary conditionally. An example is the small questionnaire on trip characteristics that we operationalized in the survey builder (trip distance, purpose, and mode of transportation) and integrated in the diary.

110. While this might not be special, the builder's ease of use is, in our opinion. It functions in a GUI-like manner, and almost all functionality is accessible by clicking buttons, and no technical knowledge is required.

111. The analogy with a cockpit comes to mind; The MOTUS builder enables the statistician to be in charge of the research independent of any other party like an internal or external IT department. Consequently, the focus of the work shifts away from technical issues towards methodological and domain-related matters, where it should be, in our opinion. The Motus builder also entirely reflects the intrinsic modular approach that MOTUS advocates, or better, it's the result of it.

4.3.2 Built phase - challenges.

112. While the capacity to assemble fully independent rich research designs is one of the great merits of the MOTUS builder, it inevitably increases the complexity of the designs. Designing research is no longer merely operationalizing a survey. It's also about incorporating surveys in diaries, integrating communications, operationalizing actions and events, and all kinds of other interventions that make a research design stronger but more complicated.

113. To deal with this complexity, especially if the purpose is to put it in the hands of statisticians and not IT staff, a tool's structure needs its roots in apparent business logic. The logic behind the Motus builder goes a long way already, but we see room for improvement, however.

114. Basically, and without going into too much detail, we felt the need for more clarity on certain key elements of the builder, like states, actions, and events. Especially the difference between what a state is, essentially a step in the research, and what actions or events are, what happens during a step, could be made more straightforward. It's not a matter of documentation; it's conceptual. In our opinion, the concept of a state should be kept as pure as possible. States should be empty building blocks that need filling with research elements such as surveys, diaries, and, e.g., notifications and geofences. For example, at this point, surveys can be a state (as a step in a research

flow) and an action (part of a step). The issue is not with the latter, it's great to be able to integrate surveys where needed, but then it's better to consider the steps as empty placeholders that need filling.

115. A similar line of reasoning can be applied to the reusability that MOTUS features; the logic behind this can be extended much further, in our opinion. At this moment, 'high-level' components like surveys and diaries are easily reusable, but answering categories are not, for example. Ideally, every single element should be reusable anywhere in MOTUS. We suppose this is 'low hanging fruits for the MOTUS developer team, as the same programming structures that serve to reuse higher-level elements probably also work to make lower-level components reusable.

116. We encountered several other issues that somewhat hampered working efficiently with the Motus builder. In many cases, these concerned details or issues that needed adaption from our side. The way the builder deals with translations is an exception to this, however. Motus easily handles several languages, but there isn't an upload service for surveys available at this point. Therefore, translations need to be integrated question by question, which is burdensome. We partially adapted by constructing and translating parts of the questionnaire first, to operationalize them simultaneously in MOTUS afterward. Nevertheless, it's clear that the development of such an upload service is a necessity.

4.3.3 Collect phase - strengths.

118. The importance of the MOTUS builder reaches beyond the design phase as it also enables micro-management of fieldwork. With micro-management, we mean follow-up of individual respondents as they work through the stages of the research and the ability to intervene in their task flow actively. Interventions like setting back a respondent to an earlier phase of the research flow, or making him advance through a task flow, are easy to perform. The result is a powerful functionality that highly facilitates dealing with respondent-specific bugs or issues. As it allows for swift interventions independent of a supporting party like an IT department, it's also once more an example of the empowering capacity that MOTUS has for statisticians or any other person managing fieldwork. Finally, this functionality also enhances the quality of the fieldwork as it enables a personal touch to helping respondents.

119. Not a must, but an interesting feature of MOTUS, nevertheless, is that it aims at operating entirely independently of interviewers. Consequently, communication with the respondents to start and manage the fieldwork is critical, requiring an array of communication functionalities.

120. Motus is very well equipped to that purpose, with even a separate 'builder' in the overarching Motus builder dedicated to communication. This builder enables different kinds of communications that serve a variety of goals. Editing and sending emails and messages to cover global communications and invitations is easy, and logins and passwords are integrated automatically. The same builder also makes it possible to edit the information supporting respondents working through the task list. Introductions to specific phases of the research, for example, or background information on definitions. This kind of information is not sent but integrated displayed in the front end on the respondents' device.

121. While the communication builder serves at organizing communication in general, some situations necessitate direct contact with individual respondents. MOTUS has deeply integrated functionality to that end. The internal database system managing the allocation of respondents to studies also features email and messaging services to contact individual respondents from within the application. More even, similar functionality is integrated into the component that serves to follow the progress participants make in the research flow. When an individual task flow needs an intervention, the service that makes intervening possible handles sending feedback on the intervention to the respondent. There is additional functionality in MOTUS we didn't use in this pilot that makes managing fieldwork even easier. For example, R scripts can be integrated to evaluate response rates in the analysis builder without leaving the MOTUS environment.

4.3.4 Collect phase - challenges.

122. Slightly expected, the fieldwork of this small study was not entirely free of technical flows. Most of these issues seemed to be occasional and limited. For example, we needed to intervene in the back end using the MOTUS builder to launch at least one respondent's research flow. Because it was not clear why, similar cases require continuous follow-up in future tests. The most widespread issue observed was a problem with the language setting in MOTUS. When logging in, the language setting needed to be reset every time again. Several respondents also got quite frustrated as they kept receiving tasks in one of the non-native languages, even after changing the language setting several times. The MOTUS developer team has since informed us that they identified this bug and fixed it.

123. The general feeling that prevails is that the fieldwork of this small study was much more of a challenge than expected. This perception is mainly unrelated to MOTUS, however. As already explained, MOTUS features an array of functionalities that help manage fieldwork. However, we also learned that even a minor technical flow, like the issue with the language setting, immediately has a damaging impact on fieldwork.

124. But it goes beyond technical flows. More thorough pretesting would undoubtedly have revealed these issues and made us avoid them. However, the real challenge is that quite some respondents felt a little bit out of their comfort zone, as we already reported in the section on usability. Some expressed a sentiment of feeling lost in the research or insecure about handling specific tasks in the app. There was a need for more guidance throughout the study.

125. The feeling of being somewhat out of one's comfort zone wasn't limited to the participants; we also felt disconnected from what we know from organizing traditional fieldwork. This kind of technology comes with new dependencies. We have extensively demonstrated the dependence on phone settings, but there are less visible issues such as the policies pursued by the app stores.

126. However, dealing with these dependencies is a challenge, not an insurmountable obstacle. New approaches, like respondent panels, can help overcome many of the problems we reported. Several respondents told us that they had become much more familiar with MOTUS by the end of the study. A more visual-based communication strategy is also inevitable. The objectives and the way of working wouldn't have been so obscure to many participants had we already used this in this pilot.

4. Conclusion and Open issues.

127. The prevailing thought during the organization of this pilot was that we felt privileged, especially about how we were able to embark on it. This very practical, first-hand experience deepens our views on both opportunities and challenges of smart surveys. It shed light on the road ahead. This section elaborates on several insights about this study and smart surveys in general.

128. Based on our overall experiences, the object of the principal research objective, we rate MOTUS highly. We didn't have a hassle-free experience from a technical perspective, and we also experienced minor annoyances while working with MOTUS. However, the weight of these issues is limited, and we are convinced that the MOTUS developer team will gradually overcome most of these shortcomings. Motus is already deployable today, or at least in its core capacities, such as its survey and diary functionalities. We also believe that the application is at the threshold of incorporating much more innovative features like geotracking and -fencing, and receipt scanning.

129. As a second research objective, we assessed several particular usability-related issues and how participants rated the overall experience of using MOTUS and participating in this experiment. About MOTUS, overall ratings are fairly high. This is also reflected in the observation that many believe this kind of technology will become an essential part of our future data production. However, many also expect issues when this is 'unleashed' on actual respondents. Moreover, most clearly sensed that this is only partly if at all, related to MOTUS and the technology. They realized that there's also a business challenge. How to organize fieldwork with an application like MOTUS?

130. The third research objective on mobility was of the least interest to the Essnet on smart statistics. Still, it has a substantive significance in relation to the challenges ahead that the Green Deal brings to NSIs. Despite its small size and limited scope, we believe that this pilot proves that there's an alternative for the highly fragmented production of passenger mobility statistics inside and outside the European Statistical System. By teaming up and by deploying a platform like MOTUS, it looks feasible to

organize a harmonized cross-national survey capable of measuring the most elementary mobility indicators in a harmonized way.

131. If we take some distance from this study, several more general issues about smart statistics caught our attention and triggered insights. We briefly address three of them here: [1] the issue of privacy (and trust), [2] the lack of a conceptual framework to test applications, and [3] the interaction between active and passive measurement and its consequences for the urther development of the field of smart statistics. We should warn, however, that the insights we express below are strongly inspired by our vision and background, and can't be considered as objective conclusions of the study, let alone the project.

132. [1] Even within this small study, with only colleagues participating, the privacy issues involved in transitioning to passive measurement were tangible. As geotracking is an invasive technology that gives insight into the exact whereabouts of people at any moment, the collected information vastly overshoots the research purpose. Therefore, even in fairly limited experience like this pilot there's a strong legitimation for the statistical community's awareness of privacy and protecting confidentiality. Failing to find deployable answers to this will certainly harm further progress in the field of smart statistics. Concretely, in the case of this research, the question is if we are truly at the point that we can 'unleash' tracking on genuine respondents?

133. As the tone of the question already indicates, the answer at this point is 'no', at least in relation to large field studies. However, this negative answer isn't unequivocal, and MOTUS already anticipates by featuring an array of legal and technical measures to deal with the privacy challenge. The written privacy policy[7], for example, describes the difference between foreground and background tracking, and users provide consent in a 2-stage approach, where in the last step, the platform asks for final approval for tracing the device (as a proxy to the user).

134. Additionally, respondents can switch off tracking from within the application's menu without turning to the general setting of the phone. The idea is to develop this further to snooze

[7] https://www.motusresearch.io/privacy/full/

specific periods within days, for example. And although still on the design table, the MOTUS developer team envisages the capacity to send whereabouts back to respondents on a map as tentative data to allow them to validate or erase the data if they feel the need. Finally, MOTUS' modular capacities are also helpful as the tracking can easily be confined to particular stages of the research.

135. MOTUS' way of dealing with the privacy challenge is only one approach, tailor-made to its design, capacities, and architecture. However, there are other approaches, and the broader issue of privacy (preserving techniques) is central to this Essnet and WP 3 in particular. Nevertheless, in line with the main message in this paper, we stress the usability dimension of the privacy issue in smart statistics.

136. Usability from a participant's viewpoint brings complexity to the privacy challenge. Firstly, measures should be adequate but shouldn't hamper the user's experience. However, there's also a profoundly subjective dimension to both assessments; and to what's considered adequate in particular. Few people will reference the privacy-preserving measures of an application to existing legislation or technology. Most will make a subjective assessment based on their personal beliefs about privacy, their understanding of the technology behind an application, and, maybe even more important, based on the trust they have in the organizer of the study. As all participants were colleagues, trust wasn't to much of an issue in this small pilot. However, as the results of the usability assessment in this study show, many participants indicated that, even based on this limited experience, it undeniable will be.

137. [2] If we turn back to the organization of the pilot itself, the lack of a conceptual framework to guide testing an application as MOTUS struck us particularly. There are frameworks and software to test apps in general, but these evaluate mainly the functional-technical dimension of applications. As far as we can see, there's no ready-to-use model available to assess the usability of smart surveys in-depth. For deployability this even seems completely absent. In recent years a lot of effort has been put into conceptualizing features like modularity and shareability of statistical services, and also in this Essnet there's work done on this. However, as far as we can

see, a transparent and integrated model or a framework that converts these concepts into indicators that evaluate mobile applications and smart surveys isn't existing yet.

138. [3] We see another conceptual ambiguity, relatively unimportant at first glance but with potentially far-reaching consequences at second thought. The adverb 'smart' in 'smart statistics' seems to imply a sharp separation between data collection methods that are smart and those that aren't. Passive data collection using sensors is usually considered a threshold to define an application or a process as smart. This definition is not the issue, but there's a tendency to neglect an entire field of data collection methods situated between smart and traditional, active data collection using smartphones in particular.

139. This study, the ESSnet from our point of view, was an eye-opener in that respect. Technically everything seems in place for rapid deployment of mobile surveys and diaries. We neither see scalability issue; we are convinced that we could have easily scaled up certain parts of this study. But as extensively demonstrated, the methodological and operational challenges are considerable, so NSIs will have to invest substantially in acquiring this expertise if they want to deploy this kind of technology.

140. Some may wonder if we have to make such an investment if active data collection becomes secondary or even irrelevant in the future. However, in our opinion, the premise of that question is wrong. We are convinced that the active component plays a much more critical role than the current storyline in smart statistics suggests.

141. The prime reason for this lies in the paradigm shift we referred to in the introduction. Smart surveys do change the time-context dimension of the traditional survey model and have high potential, therefore. However, to harness that potential, we have to give enough weight to the context.

142. In fact, without context, smart surveys are only watered-down versions of big data if we may use that term one more time. The mobility theme in this study is a speaking example of this. What's the added value of tracing a small sample of persons if

theoretically millions of phones can be traced in real-time? The answer is, of course, that the survey enables to assess the context of these movements. Why and how do people travel? But more importantly, do they have alternatives, and why aren't they considering these, for example? Policymakers need trustful answers, not hypotheses or deductions, to all of these questions.

143. As a consequence, instead of a relict, to put it with some exaggeration, active/mobile measurement becomes leverage for smart surveys. It's neither about one thing nor the other; it's about balance. Finding that balance between the active mobile and passive smart dimensions changes the perspective on smart surveys and perhaps even on how to innovate our statistical production altogether. Smart surveys become an almost natural endpoint of an evolution in which the current data production goes mobile first and then gradually becomes smarter with the addition of passive elements. In other words, To become smart, we need to get clever first. This approach, this perspective on smart statistics, and smart surveys in particular, might also be the clever way to convince or reassure legal departments throughout Europe.

144. Finally, it might also shift balances between stakeholders in the data collection process. Active measurement, or rather integrating passive measurement with active data collection, intensifies fieldwork highly, as this study shows. Consequently, statisticians remain key players in the smart data collection process, and any application or data platform must empower them to play that role. In other words, the premise should be to innovate old-school data collection methods out of statistics, not statisticians.

References

[1] Eurostat Statistics Explained. Being young in Europe today – digital world. (2020) https://ec.europa.eu/eurostat/statistics-explained/index.php?title=Being_young_in_Europe_today_-_digital_world#:~:text=In%202019%2C%2092%20%25%20of%20young,portable%20computer%20in%20this%20way.&text=This%20is%20one%20of%20a,Being%20young%20in%20Europe%20today.

[2] Ricciato, F., Giannakouris, K., Wirthmann, A., Hahn, M., Trusted Smart Surveys: a possible application of Privacy Enhancing Technologies in Official Statistics. https://ec.europa.eu/eurostat/cros/content/trusted-smart-surveys-possible-application-privacy-enhancing-technologies-official-statistics-short-paper-sis-2020_en

[3] McCool, D., Lugtig, P., Mussmann, O., Schouten, B., (2021) An App-Assisted Travel Survey in Official Statistics: Possibilities and Challenges. Journal of Official Statistics, 37 (1): 149-170.

[4] Keusch, F., Wenz, A., Conrad, F., (2021) Do you have a smartphone with you? Behavioral barriers for measuring everyday activities with smartphone sensors. Computers in Human Behavior, 127 (10).

[5] Bevan, N., e.a., (2016) New ISO Standards for Usability, Usability Reports and Usability Measures., International Conference on Human-Computer Interaction July 2016.

[6] Salgado, D., e.a., (2018) Data organisation and Process Design Based on Functional Modularity for a Standard Production Process. Journal of Official Statistics 34 (4):811-833.

[7] Eurostat Manuals and Guidelines, Eurostat guidelines on Passenger Mobility Statistics, EU Transport Statistics, December 2018.
https://ec.europa.eu/eurostat/documents/29567/3217334/Guidelines_on_Passenger_Mobility_Statistics+%282018_edition%29.pdf/f15955e3-d7b4-353b-7530-34c6c94d2ec1?t=1611654879518