# How Reliable Are Drivers' Statements About Their Engagement in Non-Driving-Related Tasks?

**Tibor Petzoldt,** TU Dresden, Germany, *tibor.petzoldt@tu-dresden.de*, **Daniel Eisele,** TU Dresden, Germany, **Sophie Feinauer,** Audi AG, Germany

**ABSTRACT**

When investigating road user behaviour that has the potential to increase risk (e.g., engagement in non-driving-related tasks; NDRTs), we often must rely on self-reported data. However, for a variety of reasons, the reliability of such self-reports can be called into question. To compare self-reported and actual engagement in NDRTs, we utilised a dataset collected as part of the SHRP 2 large-scale naturalistic driving study, which contains video footage of drivers and their activities while driving. A subset of 144 drivers in three distinct age groups was selected for analysis. In each age group, there were 12 drivers each who reported that over the past 12 months, they had either (a) never, (b) rarely, (c) sometimes, or (d) often engaged in potentially distracting activities such as texting, eating, or smoking. For each driver, 120 short episodes of driving were randomly selected, and any observable NDRT engagement was annotated. The analysis of the data revealed a significant association between self-reported and observed frequency of NDRT engagement. However, it also showed a considerable degree of within-group variance. The predictive value of the self-reports was moderate overall. Differences between the age groups emerged with regard to both the extent and the type of observed NDRT engagement. The results indicate that such self-reports, while reasonably accurate at the group level, should be taken with a grain of salt when used to predict individual behaviour.

**Keywords:** Distraction, Self-Reports, Naturalistic Driving, Validation, Driver Age.



## 1 INTRODUCTION

When investigating road user behaviour, especially potentially risky behaviour, we often rely on road users' self-reports, as collected through questionnaires or interviews. Self-reports are comparatively easy to obtain, and, for some forms of behaviour, might be the only approach that allows us to gather any information at all. Prominent examples, such as the Driving Behaviour Questionnaire (DBQ; Reason et al., 1990) or the Multidimensional Driving Style Inventory (MDSI; Taubman-Ben-Ari, Mikulincer, & Gillath, 2004) are grounded in theory and, to a certain extent, have been validated against actual on-road behaviour (e.g., Rowe et al., 2015). In many cases, however, the actual validity of such self-reports is difficult to gauge, simply because reliable data for external validation is hard to obtain.

But why would there be a need for validation in the first place? Why should there be a disconnect between the response to a seemingly simple question such as "How often do you do X while driving?" and actual behaviour? Strack and Leonard (1987) explain that in a survey situation, respondents are

confronted with a variety of potentially complex cognitive tasks and must make several judgements. They need to interpret the question, which requires them to have a clear understanding of the terms used and their meaning in the context of the survey. They need to access their memory and recall behaviour. They must form a judgement based on their recollections and potentially need to fit that judgement into predefined response categories. Additionally, they might feel the need to edit their response to better comply with perceived social norms. Each of these steps along the way to a response is prone to biases and subject to context effects.

One behaviour that has received continuous attention over the past 30 years is driver distraction, or, more neutrally phrased, the engagement in non-driving-related tasks (NDRTs). Beginning with the advent of in-vehicle information systems, countless studies have examined and often confirmed the potential risk posed by NDRT engagement. However, reliable data about frequencies of engagement proves difficult to obtain. Observations from outside the vehicle (e.g., Kathmann et al., 2019; National Center for Statistics and Analysis, 2024) are limited in what can be observed, and restricted to suitable locations for such observation. Observations inside the vehicle (through so-called naturalistic driving studies; for a definition see van Schagen et al., 2011) are highly expensive, and typically (though not always) limited to smaller, non-representative samples. Using self-reports of NDRT engagement appears to be the obvious solution. Among the more structured approaches to collecting representative self-reported data on driver distraction are the works of Huemer and Vollrath (2011), Kreusslein, Schleinitz and Schumacher (2024), or McEvoy and Stevenson (2006).

There have been previous attempts to validate self-reported NDRT engagement using observation data. Kreusslein, Schleinitz and Schumacher (2019) collected behavioural and interview data from 94 drivers, analysing two trips per driver. Simply asking whether or not drivers had engaged in a particular NDRT, the authors found sensitivities (a metric for how many observed instances of engagement were actually reported) ranging from 100% (smoking) to 8.95% (changing clothes). The authors acknowledge that the usefulness of the self-reports – even when it is merely about whether a certain activity was carried out while driving a few minutes prior – is limited at best, and dependent on the particular type of activity. Petzoldt and Utesch (2016) interviewed 15 drivers immediately after a drive, and asked them whether or not, and for how long they had engaged in certain activities. During the interviews, it became apparent that drivers struggled to answer the questions for a variety of reasons, ultimately leading the authors to conclude that "subjective accounts of secondary task engagement might provide information about what drivers believe they are doing, but should not be understood as a means to actually quantify driver distraction" (p. 216). However, it should be acknowledged that the rather small sets of observation data available for comparison against the self-reported behaviour that the authors had available in these studies certainly did not work in their favour.

One of the largest sets of naturalistic driving data is the so-called SHRP2 naturalistic driving dataset. This dataset was collected between 2010 and 2013, and contains observational data on everyday driving behaviour of about 3,400 US drivers in age categories from 16 to 99 years, most of which were followed for a year or two (Hankey, Perez, & McClafferty, 2016). The resulting dataset contains about

50 million miles of driving and about one million driving hours of driving time, for which about two petabytes of video footage (incl. drivers' faces and hands) were captured (Dingus et al., 2015).

Crucially, the dataset also contains information from a variety of participant questionnaires, including self-reported data on the frequency with which they engaged in potentially distracting activities and their perceptions of the risk associated with such activities. Therefore, the aim of this study was to utilise that dataset and test whether drivers' self-reported behaviour is consistent with observation data. More precisely, we sought to determine the degree to which the reported frequency of engagement in NDRTs matches the observed frequency.

# 2 METHOD

## 2.1 Dataset

To address our research question, we utilised the SHRP2 naturalistic driving dataset. Our analysis focused on drivers' NDRT engagement, so we made extensive use of the available video data. For the self-reported frequency of engagement, we relied on the following questionnaire item: "In the past 12 months while driving, how often did you do other things while driving, like use cell phone, eat or drink, put on makeup, read things, or smoke cigarettes?". Participants could respond that they either "Never", "Rarely", "Sometimes", or "Often" engaged in such activities. An additional item asked participants to rate the risk associated with such activities on a seven-point Likert scale.

## 2.2 Participant sample

In the first step, we selected a sample of drivers (from the total dataset) based on their self-reported engagement in non-driving-related activities. We selected drivers in three age groups: younger (16-19), middle-aged (35-49), and older (65-84). Within each group, gender and self-reported frequency of engagement in non-driving-related tasks were balanced across the sample where feasible. The resulting sample can be found in Table 1. Overall, we selected 48 drivers per age group, and 36 per category of self-reported frequency of NDRT engagement, resulting in a total of 144 drivers in our dataset.

**Table 1 – Sample of drivers analysed**

| **Self-reported NDRT engagement** | **Distribution per age group** | | | **Total** |
|---|---|---|---|---|
| | **Younger (16-19)** | **Middle-aged (35-49)** | **Older (65-84)** | |
| Never | 6m/6f | 6m/6f | 6m/6f | 36 |
| Rarely | 6m/6f | 6m/6f | 6m/6f | 36 |
| Sometimes | 6m/6f | 6m/6f | 6m/6f | 36 |
| Often | 6m/6f | 6m/6f | 7m/5f | 36 |
| Total | 48 | 48 | 48 | 144 |

### 2.3 Episode sample and annotation

Once the drivers had been chosen, we needed to select our sample of driving episodes for analysis. The first 20 trips that were recorded for each driver were excluded from the sample (to allow the drivers to "settle in"), as well as any recording with a duration of less than five minutes (based on previous experience, many of the shorter recordings contain no driving on public roads, or no driving at all). For the remaining recordings, we randomly generated timestamps for episodes within these recordings to analyse. Any timestamps that brought up non-driving situations (e.g., a parked vehicle with the engine running) or driving on private property (e.g., parking lots) were discarded. We followed the SHRP 2 approach of analysing "baseline events" by annotating six-second episodes of driving. We annotated all cases of NDRT engagement, particularly those that corresponded to the tasks mentioned in the screening item (cell phone use, eating or drinking, putting on makeup, and smoking), following the SHRP 2 Researcher Dictionary for Safety Critical Event Video Reduction Data (VTTI, 2015). In total, we annotated 120 episodes per driver, i.e., a total of 17,280 episodes.

## 3 RESULTS

Observed NDRT engagement ranged from 0 to 85 instances per driver (with multiple occurrences in some analysed episodes), with a skewed distribution showing a concentration of values at the lower end (across self-reported frequencies and age groups). As depicted in Figure 1, drivers who reported often engaging in NDRTs were indeed observed engaging most frequently (total *n* = 956, per-driver *MD* = 21.5, IQR = 27.5), followed by those who reported engaging sometimes (total *n* = 832, per-driver *MD* = 19.0, IQR = 25.2). A notable drop was observed in drivers who reported rarely engaging in NDRTs (total *n* = 412, per-driver *MD* = 6.0, IQR = 18.0). Interestingly, those who reported never engaging in NDRTs were descriptively close to the rarely engaging group (total *n* = 359, per-driver *MD* = 6.5, IQR = 11.2) and noticeably above zero. In fact, only nine out of the 36 drivers who reported never engaging in NDRTs were actually observed to abstain completely in the analysed clips. The highest observed count in this group was 41.

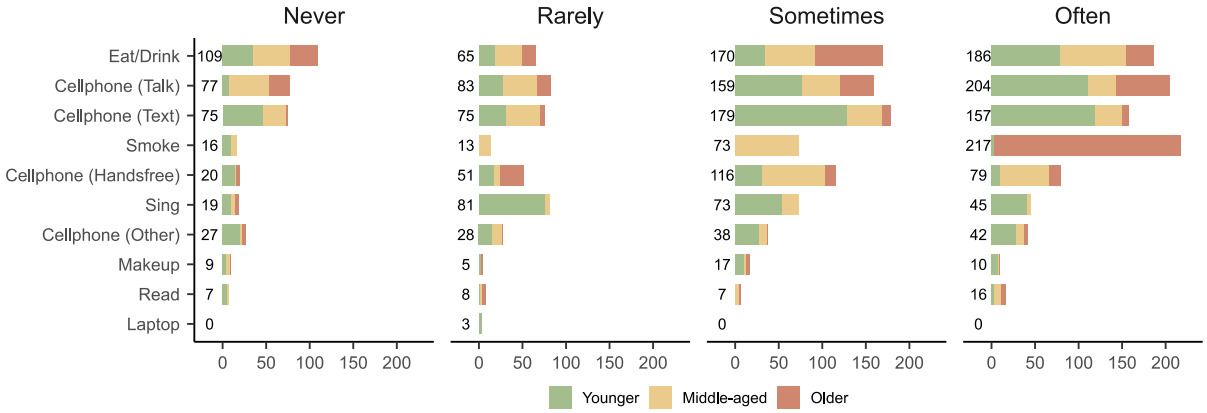


**Figure 1 – Observed NDRT Engagement by Self-Reported Frequency and Age Group**

While the overall average number of episodes with NDRT engagement was M = 17.8 out of 120 (SD = 17.8, MD = 12, IQR = 22), a further breakdown revealed that mean and median counts were highest in the younger age group and decreased noticeably with age, as shown in Figure 2. Additionally, Figure 1 illustrates substantial differences in the types of NDRTs across age groups (e.g., older drivers who reported often engaging in NDRTs mostly restricted their engagement to smoking, while hand-held cellphone use was mostly the domain of younger drivers).

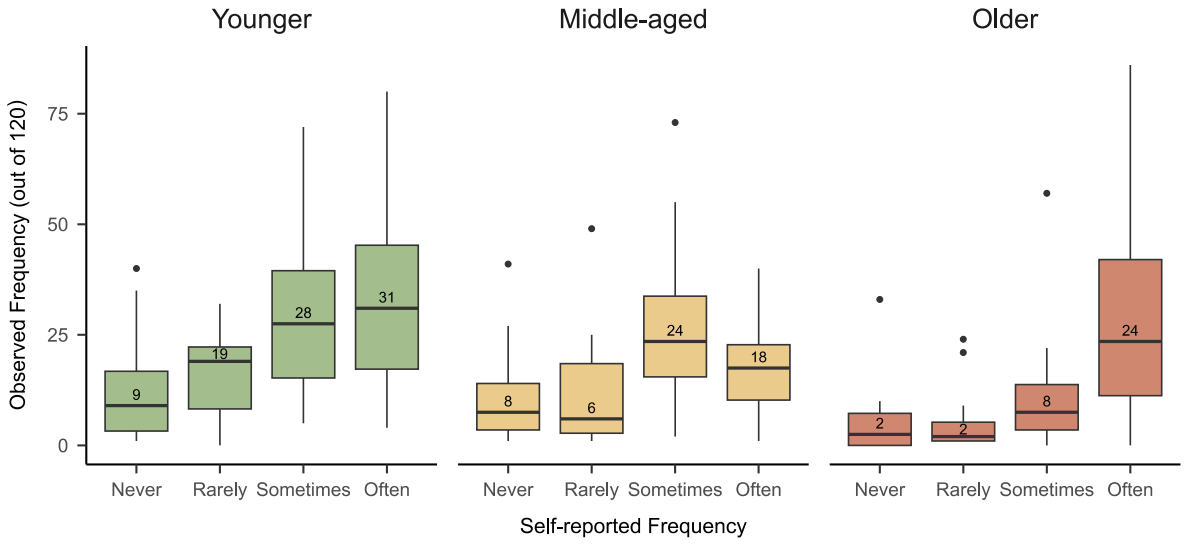


**Figure 2 – Observed NDRT Engagements by Self-Reported Frequency and Age Group**

To examine these patterns statistically, we fitted a linear model (using OLS estimation) to predict observed NDRT engagement at the per-episode level, based on self-reported frequency and age group. The model explained a statistically significant moderate proportion of variance (*adj.* $R^2$ = 0.22, $F(5, 138) = 7.73$, $p < .001$). Self-reported frequency accounted for partial $R^2$ = .173, while age contributed a partial $R^2$ = .066. The model intercept, representing the expected count for a young driver who reported no NDRT engagement, was 15.35 ($t(138) = 4.69$, $p < .001$). Within the model, self-reported rare engagement was not a significant predictor ($\beta = 1.47$, $t(138) = 0.39$, $p = 0.698$). However, reporting to sometimes ($\beta = 13.14$, $t(138) = 3.47$, $p < .001$) and often engaging in NDRTs ($\beta = 16.58$, $t(138) = 4.39$, $p < .001$) significantly increased the predicted number of observable NDRTs. While being middle-aged did not predict a significant decrease in NDRTs ($\beta = -5.94$, $t(138) = -1.81$, $p = .072$), being older did ($\beta = -10.19$, $t(138) = -3.11$, $p = .002$) .

Further iterations of the model revealed no significant interactions between self-reported frequency and age, nor any significant effect of gender. A polynomial model did not improve model fit. Validation of the model predictions using non-parametric Kruskal-Wallis tests, followed by Wilcoxon pairwise comparisons, reproduced the exact pattern of findings, reinforcing the robustness of the results.

Overall, self-reported frequency was moderately correlated with observed NDRT engagement ($\rho = .40$, $p < .001$). The correlation was noticeably weaker in the middle-aged ($\rho = .30$, $p = .039$) than the

younger (ρ = .48, $p < .001$) and older group (ρ = .49, $p < .001$). Correlation strength further varied by task type.

Regarding the perceived risk of NDRT engagement, a strong negative correlation was found with self-reported frequency (ρ = -.49, $p < .001$). A corresponding negative correlation, though smaller in magnitude, was found with observed frequencies ($\rho$ = -.25, $p$ = .003). The difference between these two correlations was significant, with a medium effect size across all tests provided by the cocor package (all $p \leq .003$; Diedenhofen & Musch, 2015).

# 4 DISCUSSION

The aim of the analysis presented in this paper was to find out how closely self-reported frequency and actual frequency of engagement in so-called non-driving-related tasks (NDRTs) correspond. The results were mixed. Overall, drivers who reported more frequent engagement in NDRTs did, indeed, engage more often on average. The correlation between these metrics, however, was only moderate, as was the predictive value of the self-report for observed engagement. Given that both metrics are, in theory, supposed to measure one and the same behaviour, this is not an ideal outcome. When examining the different age groups, the picture did not change substantially, although correlations in the younger and older subgroups were slightly higher than in the middle-aged group.

It should be noted that age appeared to play a role in NDRT engagement, albeit a statistically minor one. The descriptive data, however, highlight that, beyond the small differences in engagement frequency, it is the type of NDRT drivers engage in that differs considerably between age groups. At the same time, given that the dataset is now more than a decade old, the question arises as to whether these are indeed effects of age (in the literal sense), or rather differences between generational cohorts. Is it reasonable to assume that today's younger drivers, who will be the seniors of tomorrow, abstain from using the cell phones while driving when they reach a certain age? Or will they continue engaging in the same behaviours – just as current generation of senior drivers probably do the same things they did 30 years ago? In the context of the analysis presented in this paper, it appears that we were, to an extent, comparing apples and oranges when assessing drivers of different age groups who reported engaging in non-specific NDRTs to varying degrees.

It is evident that the single item used in SHRP 2 to quantify NDRT engagement is highly problematic It groups together various types of tasks into a single question, leaves ambiguity as to which activities should be included (by providing a non-exhaustive list of examples), employs an extremely vague scale, and references a 12-month timeframe that makes accurate memory retrieval nearly impossible. Approaches specifically designed to assess driver distraction (Huemer & Vollrath, 2011; Kreusslein et al., 2024; McEvoy & Stevenson, 2006) are clearly superior in their design and may help mitigate some of the inherent issues associated with the SHRP 2 item.

To be fair, within the context of SHRP 2, the item was never intended to provide a precise measure of NDRT engagement. Instead, it was part of a 32-item "Risk Taking Questionnaire," where all questions followed a standardised format ("In the past 12 months while driving, how often…"). However, it could

be argued that our findings somewhat undermine the validity of the overall questionnaire, whose data has been used in subsequent analyses of driver behaviour (e.g., Khakzar et al., 2021; Richard et al., 2020).

Overall, however, self-reports of driver behaviour remain a valuable and often indispensable source of information. They offer useful insights and reasonable-quality data. It just seems that, at times, this data quality could be improved if more care were taken in question formulation. Ultimately, though, it remains imperative to validate such instruments against behavioural data whenever possible. Otherwise, questions will always remain, whether well-founded or not.

# 5 ACKNOWLEDGEMENTS

The data used in this analysis was provided by the Virginia Tech Transportation Institute (VTTI). The findings and conclusions of this paper are those of the authors and do not necessarily represent the views of VTTI, the Transportation Research Board, the National Academies, or the Federal Highway Administration.

# REFERENCES

Diedenhofen, B., & Musch, J. (2015). cocor: A comprehensive solution for the statistical comparison of correlations. *PLOS ONE*, *10*(4), e0121945. https://doi.org/10.1371/journal.pone.0121945

Dingus, T. A., Hankey, J. M., Antin, J. F., Lee, S. E., Eichelberger, L., Stulce, K. E., … Stowe, L. (2015). *Naturalistic Driving Study: Technical Coordination and Quality Control*. Washington, D.C.: Transportation Research Board.

Hankey, J. M., Perez, M. A., & McClafferty, J. A. (2016). *Description of the SHRP 2 naturalistic database and the crash, near-crash, and baseline data sets*. Virginia Tech Transportation Institute.

Huemer, A. K., & Vollrath, M. (2011). Driver secondary tasks in Germany: Using interviews to estimate prevalence. *Accident Analysis & Prevention, 43*(5), 1703-1712. https://doi.org/10.1016/j.aap.2011.03.029

Kathmann, T., Scotti, C., Huemer, A. K., Mennecke, M., & Vollrath, M. (2019). *Konzept für eine regelmäßige Erhebung der Nutzungshäufigkeit von Smartphones bei Pkw-Fahrern.* Berichte der Bundesanstalt für Straßenwesen, Heft M 287.

Khakzar, M., Bond, A., Rakotonirainy, A., Trespalacios, O. O., & Dehkordi, S. G. (2021). Driver influence on vehicle trajectory prediction. *Accident Analysis & Prevention, 157*, 106165. https://doi.org/10.1016/j.aap.2021.106165

Kreusslein, M., Schleinitz, K., & Schumacher, M. (2019). What you see is what you get? Correspondence of video and interview data on secondary task engagement while driving-a naturalistic driving study. In *Proceedings of the Tenth International Driving Symposium on Human Factors in Driver Assessment, Training and Vehicle Design* (pp. 196-202). University of Iowa.

Kreusslein, M., Schleinitz, K., & Schumacher, M. (2024). Sociodemographic, contextual and psychological factors predicting secondary task engagement: A nationwide interview study among car drivers in Germany. *Transportation Research Part F: Traffic Psychology and Behaviour, 103*, 387-403. https://doi.org/10.1016/j.trf.2024.04.008

McEvoy, S. P., Stevenson, M. R., & Woodward, M. (2006). The impact of driver distraction on road safety: Results from a representative survey in two Australian states. *Injury Prevention, 12*(4), 242-247. https://doi.org/10.1136/ip.2006.012336

National Center for Statistics and Analysis (2024). *Driver electronic device use in 2023* (Traffic Safety Facts Research Note. Report No. DOT HS 813 660). National Highway Traffic Safety Administration.

Petzoldt, T., & Utesch, F. (2016). Trying to validate subjective reports with naturalistic data - A case against questionnaires and surveys to quantify driver distraction. In A. Morris & L. Mendoza (Eds.). *Proceedings of the European Conference on Human Centred Design for Intelligent Transport Systems* (pp. 208-218).

Reason, J., Manstead, A., Stradling, S., Baxter, J., & Campbell, K. (1990). Errors and violations on the roads: a real distinction?. *Ergonomics, 33*(10-11), 1315-1332. https://doi.org/10.1080/00140139008925335

Richard, C. M., Lee, J., Brown, J. L., & Landgraf, A. (2020). *Analysis of SHRP2 speeding data (No. DOT HS 812 858)*. United States. Department of Transportation. National Highway Traffic Safety Administration.

Rowe, R., Roman, G. D., McKenna, F. P., Barker, E., & Poulter, D. (2015). Measuring errors and violations on the road: A bifactor modeling approach to the Driver Behavior Questionnaire. *Accident Analysis & Prevention, 74*, 118-125. https://doi.org/10.1016/j.aap.2014.10.012

Strack, F., & Martin, L. L. (1987). Thinking, judging, and communicating: A process account of context effects in attitude surveys. In *Social information processing and survey methodology* (pp. 123-148). New York, NY: Springer New York.

Taubman-Ben-Ari, O., Mikulincer, M., & Gillath, O. (2004). The multidimensional driving style inventory—scale construct and validation. *Accident Analysis & Prevention, 36*(3), 323-332. https://doi.org/10.1016/S0001-4575(03)00010-1

Van Schagen, I., Welsh, R., Backer-Grondal, A., Hoedmaker, M., Lotan, T., Morris, A., … Winkelbauer, M. (2011). *Towards a large scale European Naturalistic Driving study: final report of PROLOGUE.* PROLOGUE Deliverable D4.2. SWOV Institute for Road Safety Research, Leidschendam, The Netherlands.

VTTI (2015). Researcher Dictionary for Safety Critical Event Video Reduction Data, Version 4.1. Blacksburg, VA: Virginia Tech Transportation Institute.