Is it finally time to kill self-report outcome measures in road safety? An investigation of common method variance in three surveys of a cohort of young drivers that included the Driver Behaviour Questionnaire.

Harrison, W.A.
Eastern Professional Services Pty Ltd

Abstract

In his challenging book on road safety research methods, Anders af Wåhlberg (2009) makes a strong case that there are significant, unaddressed problems in the use of self-report measures in road safety research. These include the effects of social desirability and response patterns that contribute to the common method variance largely ignored by road safety researchers. His argument, perhaps overstated, is that the problems associated with self-report measures such as survey questionnaires and the Driver Behaviour Questionnaire (DBQ) are so great that they should not be used in individual-differences research in road safety. This paper reports on an investigation of common method variance in relation to the DBQ, with a particular focus on social desirability effects. Survey data collected from three survey waves of a young driver cohort were analysed to investigate the relationships between measures such as self-reported behaviour, self-reported crash involvement, the DBQ, and a measure of social desirability. The results of these analyses were then used to draw some conclusions about the validity of af Wåhlberg's concerns and the prognosis for self-report methods.

Keywords

Social desirability, Common method variance, Survey, Self-report measures, Driver Behaviour Questionnaire

Introduction

A fundamental challenge in road safety research is how best to define and operationalize measures that provide information relevant to safety. Statistical power considerations make it difficult to use police-reported crash involvement as an outcome measure in many evaluations, and it is rarely possible to use police-reported crash data in research concerning psychological and social factors that may contribute to crash involvement.

Road safety researchers often choose measures that are one step or more removed from official records of crash involvement, and are therefore confronted by a number of issues concerning the reliability, validity, and relevance of potential measures. It is common for researchers to make use of self-report data in road safety research, relying on surveys or other forms of data collection that seek information directly from research participants.

It is my opinion that survey methods that collect self-report data are significantly flawed for a number of reasons and should therefore only be used with caution. Researchers appear to have a high level of optimism bias in relation to survey methods and their often unstated assumptions. Some of the questionable assumptions underlying the use of surveys are:

- Survey samples can be representative of the road user population of interest. It is almost always assumed that there are no systematic biases in sampling that are related to the safety issue of interest in the survey. This is often assessed in terms of relationships between sample demographics and population data without considering more-subtle biases that are likely to limit the generalizability of survey results to the population of interest. Indeed there have been instances where Australian researchers have so failed to understand the effect of sampling biases that they have argued that a numerically large sample is sufficient to minimise concerns about self-selection and other sources of bias.

- Survey participants are able to provide the information sought in the survey. It is assumed that this information is available for conscious processing at the time the survey is conducted. It is also assumed that responses to survey items reflect the causal mechanisms that underlie real-world behaviours – it is not unusual for survey items to seek information about attitudes, perceptions, and intentions and then to assume without justification that survey responses somehow represent factors that influence behaviour in the real world at some point in the future or in the past.

- Information recalled by participants is unbiased and accurate. Surveys assume that recall mechanisms are reliable and unaffected by the specific demand characteristics of the survey. There has been little research concerning the reliability of self-report measures in road safety except in relation to self-reported crash involvement and offence histories – where differences between official records and survey responses are often interpreted as a reflection of unreliable official data. The review published by af Wåhlberg (2009) demonstrates that this attribution is almost certainly incorrect – recall (even of crash involvement) appears to be inaccurate and inconsistent.

- Relationships between responses to items in a survey reflect underlying relationships between the things being measured – that there are no underlying, systematic biases in response patterns that would contribute to measured relationships between survey items. The assumption that there is some reasonable level of construct validity in survey items is critical – conducting a survey where there is some doubt about this would be pointless, but construct validity is rarely assessed.

Some of these issues are discussed in this paper in relation to a series of three surveys of a cohort of young drivers. These surveys were conducted as part of a larger, ongoing project that will be reported in greater detail in future.

Method

The surveys were delivered primarily as on-line surveys, with telephone surveys (with the same items) used as a supplementary method to improve the response rate. The first of three surveys using the same cohort of participants included items relating to demographic information; learner experiences (including supervision arrangements during the learner period, estimated hours of driving experience as a learner, the pattern of learner driving experience, factors that interfered with gaining experience, and the number of professional lessons obtained); driving exposure in the most recent two days; self-assessment of driving skill compared to peers in relation to hazard detection, safety, and driving in different driving contexts; the 27-item Driver Behaviour Questionnaire (DBQ); self-reported risk-related behaviours in the ten most recent driving trips; driving offences detected by Police; and crash involvement.

The second and third surveys (each conducted about six months after the preceding survey) included most of the items from the first survey, and an additional group of items to assess "social desirability" effects on responses.

A database of potential survey participants was drawn randomly from the VicRoads licensing database. It was composed of 2,500 probationary licence holders. All potential cohort members were licensed in the most recent six-month period available in the database (between 4 November 2007 and 3 May 2007). Sample membership was restricted to those who, amongst other things, obtained their learner permit at 16 years or older, were 18 years or older when they got their probationary licence, were resident in Victoria, and had known sex.

Recruitment into the survey involved initial contact and follow-up by post, and ultimately a follow-up telephone call (after various data sources were used to identify telephone numbers if these were not available in the licence database). An initial invitation letter was sent to all potential survey participants. In the first cohort survey these letters were sent to all potential sample members. In the second and third cohort surveys, the invitation letter was sent only to those who had completed the preceding survey. The letter was personally addressed, explained the purpose of the survey and how to participate. It included

the URL for the on-line survey and the participant's participant number. It also included a telephone contact number for potential participants without internet access.

In all the surveys participants were offered a small reward (a gift card from a major retail chain) for participating in the survey, and were also given the opportunity to participate in a prize draw for a higher-value gift card. Reminder letters and ultimately telephone reminders were used to encourage participants to complete the survey. The on-line survey was used by telephone interviewers if participants agreed to do the survey over the telephone during the reminder calls.

The identity and confidentiality of participants were preserved by ensuring the author had no access to the original database or to information that could link participant identity to their participant numbers.

The three cohort surveys each produced a data file with a number of response fields that should have remained constant between surveys. Each survey included the participant number (a unique number issued to participants and printed on their invitation letter along with the survey URL address) and the participant's sex and birth month and year. Matching of the two sets of survey data was not a simple matter as any of these variables could be entered incorrectly during completion of the survey. The result of the matching process and attrition between surveys was a final cohort with fully matched data for 676 participants.

Results and Discussion

Sample

The response rate at the end of the third survey wave was 27 percent. There was a 49 percent response rate for the first survey, a 67 percent response rate in the second, and an 82 percent response rate in the most recent survey. The analyses of the cohort data reported here are based on those participants with matched data across all three surveys (N = 676).

There were 381 female participants and 295 male participants. The age distribution of participants who completed the most recent survey did not differ significantly from the age distribution of those who did not ($F_{(1,1195)} = 0.57$, $p = .45$). Similarly, males and females who completed the first survey were equally likely to complete the third survey ($\chi^2_{(1)} = 2.09$, $p = .14$), and the socio-economic status of participants who completed the third survey (using an ABS relative advantage/disadvantage measure) did not differ significantly from those in the original cohort who did not complete the third survey ($F_{(1,1195)} = 2.82$, $p = .09$).

Sampling Bias

Surveys generally make use of a two-stage sampling process. The first stage involves selecting potential sample members from the larger population, and the second stage involves recruiting from potential sample members into the survey sample itself. There are therefore two points at which sampling biases can occur:

- The processes implemented to define the population of interest and to select potential participants for subsequent recruitment can bias the sample. Ideally it would be possible to select a random sample of the population for recruitment, but this is rare. It is more common to use techniques that involve sampling based on random telephone numbers, geographic areas, or existing databases of telephone numbers. These techniques are flawed because they introduce biases into the resulting survey sample – they ensure that potential survey participants who will be contacted about participation in the survey are not fully representative of the target population.

- There are self-selection and related biases that have their effect at the time potential participants are invited to complete the survey. These can relate to psychological characteristics such as motivational factors, but are also likely to be strongly influenced by practical factors such as time demands and access to facilities to complete the survey (such as the internet). These biases are always present in survey samples.

The cohort used in the current survey was sampled based on a randomly sampled subgroup of the population of young novice drivers. It therefore avoids the first of the two sources of sampling bias, but was still subject to the second source. Factors likely to lead to sampling bias in this series of surveys include the use of mail to contact potential participants and interpersonal differences in how young people respond to an invitation delivered in this way, interpersonal variation in responses to the incentives, differences in reading skill, differences in access to the internet (both in relation to access itself and then in relation to the type of access), interpersonal differences in availability for contact by telephone, differences in attitudes to government authorities and surveys, etc.

Most of these sources of sampling bias are not quantifiable in a way that allows comparison between responders and non-responders, and many of them are likely to correlate with safety-related behaviours and outcomes. The effect of these interpersonal differences cannot therefore be treated as simple sources of random error or noise in survey responses – they are likely to interfere with generalisation of the survey results to the population of interest.

There are some challenges in assessing sampling biases. No data are available for the non-responders, and there are some difficulties comparing sample and population demographics that need to be considered. Population data (census data, for example) provide a blunt instrument for assessing sampling bias and do little to limit concerns about the effects of underlying biases. A survey sample may appear to be representative in relation to age, sex, education level, and income – but this does not rule out the possibility that within these broad representative categories there are self-selection and other biases that are unrelated to blunt demographic characteristics but are related to key survey measures. This is a fundamental flaw with many survey-based studies in road safety – authors are often inclined to believe that their data are representative of the population based on blunt demographic comparisons.

Despite the use of a randomly selected group of potential participants and the broad similarities between sample and census characteristics, the current survey data suggest that there are sampling biases that limit the generalisability of the results. It was possible to compare respondents who completed the on-line survey and the telephone survey (in the first survey wave). Any differences between these groups might hint at potential biases in one or both survey method that could limit the generalisability of results.

There were significant differences between the on-line and telephone samples in relation to employment status ($\chi^2_{(1)} = 24.9$, $p = .00$) and educational status ($\chi^2_{(3)} = 25.9$, $p = .00$). On-line survey respondents were less likely to be in full-time employment (28 percent compared to 39 percent) and were more likely to be in part-time employment (51 percent compared to 43 percent). On-line survey participants were more likely to be studying at university (42 percent compared to 31 percent) and were less likely to report that they were not studying at the time of the survey (35 percent compared to 41 percent). There were significant relationships between the on-line and telephone samples in relation to their self-reported risky driving behaviours (with on-line survey participants consistently more likely to give socially desirable responses). Employment and educational status were also related to driving exposure and some risk-related driving behaviours.

These results hint at problems for one or both of the survey methods (on-line or telephone) used in the current project. There were differences between the demographic characteristics of participants who used each method, and these differences were in turn related to outcome measures used in the two surveys. This means that the sampling biases present in this survey are not independent of the survey outcome measures, and that there should be some concerns about generalising the results of the surveys to the broader population of young drivers.

Recall

It was suggested earlier that surveys make an assumption that survey participants are able to recall information relevant to survey items. The use of the same set of survey items over three surveys provides an opportunity to assess the reliability of participants' memory in relation to a typical young driver survey item. It is common in the context of graduated licensing systems to include survey items concerning the amount of experience accrued as a learner driver. Participants in the current survey series were asked how much supervised experience they had accrued as learner drivers. Their responses suggest significant reliability problems.

Only 21 percent of participants gave the same experience estimates in the first and second surveys, and 19 percent gave the same estimates in the first and third surveys. Only 61 participants (10 percent) gave the same experience estimates in all three surveys. The mean absolute difference between the survey 1 estimates and the estimates in surveys 2 and 3 was about 40 percent of the first experience estimate .

Figure 1 shows the considerable variability in responses to the learner experience survey item in Surveys 1 and 3. The majority of those who gave the same response in the two surveys gave responses of 80, 100, or 120 hours in the first survey. There were relatively few participants apart from these who gave the same response in the two surveys, and as noted earlier the differences in responses were sometimes large. The correlation between responses in Surveys 1 and 3 was r = .65 – suggesting that Survey 1 responses accounted for about 40 percent of the variation in Survey 3 responses.

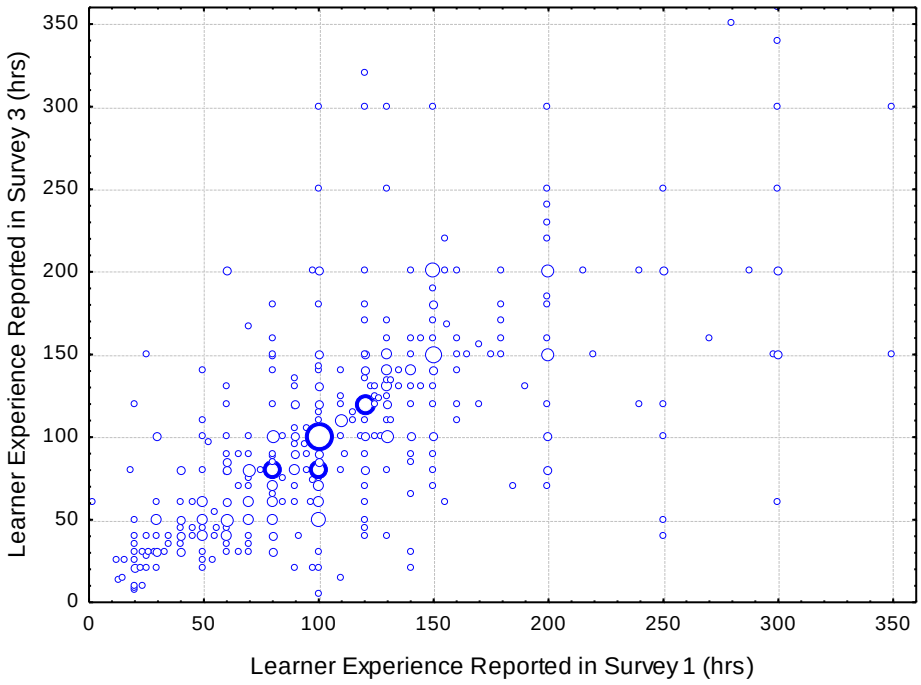

Figure 1: Scatterplot of recalled learner experience in Surveys 1 and 3. Large icons represent multiple participants at that point on the scatterplot

Responses to the learner experience item in the current series of surveys suggest a significant problem with the reliability of responses to this common item. Responses were generally inconsistent unless an easily memorable approximation was given in the first survey, and the variability between surveys was substantial. It is uncertain whether this reliability problem relates to poor recall or some other factor – but it does suggest that it would be unsafe to rely on the accuracy of responses to this type of item.

Common Method Variance and Social Desirability

Common method variance is the variation in survey item responses that originates in consistent intrapersonal response patterns across all survey items. If a survey is used to assess the relationships between attitudes, intentions, and self-reported behaviour, for example, individual participants are likely to approach the survey with specific response patterns that relate in part to their tendency to make use of response scales in different ways (some participants will more-readily use extreme scale values compared to other participants) and that relate in part to broad biases in response patterns (such as those that may relate to motivations concerning social desirability, responses to experimenter effects, and so on).

Common method variance is a problem because these personal response patterns are present across the survey and may result in specious correlations between variables. Where a single survey instrument collects data on behavioural intentions and self-reported behaviour, for example, correlations between these variables are at least in part the result of common method variance.

Social researchers have had concerns about common method variance and common method bias for some time. Doty and Glick (1998) were able to assess the effect of these problems using publications over a twelve-year period and concluded that common method variance and common method bias accounted for 58 percent of the variation in measures – leaving relatively little variation attributable to the underlying constructs.

There has been little interest in common method variance in road safety research, despite our increasing reliance on surveys and other self-report research methods. The challenging book on road safety methods by af Wåhlberg (2009) makes a strong case that there are significant problems resulting from common method variance and that this problem is a serious weakness in road safety research as it is currently conducted. Sullman and Taylor (2010) attempted to assess the contribution of social desirability to some common road safety questionnaires and concluded that the DBQ was not vulnerable to social desirability – although this paper did rely on a method that appears to assume that social desirability effects depend strongly on the location in which a survey is completed and did not actually measure social desirability effects.

The present survey included a set of items designed to assess social desirability effects and therefore provides an opportunity to assess the effects of this source of common method variance/bias on responses to safety-related responses.

Participants were asked to complete a series of ten items (in the third survey) to assess the extent to which they might give responses based on social desirability. These items form a subset of the Marlowe-Crowne Social Desirability Scale, which is widely used in social research.

Analysis of the ten items suggested a two-factor structure – five items loaded together on a factor that appears to reflect the willingness of participants to admit that they want to get their own way, and five items loaded together on a factor that appears to reflect people's willingness to admit that they are sometimes inconsistent, dishonest, or negative towards others.

Analysis of the social desirability factor scores suggests that social desirability played a role in responses to the survey items. There were small but statistically significant correlations between one or both social desirability factors and responses to items concerning experience accrued as a learner driver; the number of professional lessons taken; self confidence in relation to driving skills concerning hazards, unfamiliar roads, changing lanes, predicting others' behaviour, overtaking, and driving at high speeds; the incidence of risky driving behaviours including speeding, mobile phone use, and restraint use; crash involvement; and the DBQ Violations Scale and Lapses Scale. In all cases, the results suggest a motivation towards appearing "good" was associated with safer or better-behaved responses to the road safety items.

These simple correlations suggest that response patterns related to social desirability motivations may have an effect on responses to road safety related survey items. This is sufficient to warrant the inclusion of social desirability items in road safety surveys to act as a control for potential common method variance problems in future. On its own, however, this outcome is insufficient to discount survey methods in general. In the current study, for example, the common method variance relating to social desirability does not appear to have influenced the relationships between crash involvement and some potential predictors. This is shown in Table 1 – showing the relationships between self-reported crash involvement (all crashes) and scale scores for the four DBQ scales. The left side of the table shows the simple correlations between these variables, and the right side of the table shows the partial correlations controlling for the effect of the two social desirability factors. There is a consistent but trivial reduction in correlations between DBQ scale scores and crash involvement when variation due to social desirability is taken into account, but the key predictors of crash involvement (the Violations and Aggressive Violations Scales) remain as significant predictors.

This outcome suggests that social desirability related response patterns do contribute to the variation in responses to road safety survey items, but that in the present case the correlations were insufficient to

inflate the predictive relationship between the DBQ scale scores and crash involvement. This outcome suggests a concern that should be investigated further – indeed it may be appropriate to ensure that social desirability and other response patterns are taken into account in surveys as a safeguard against potential effects. The current survey made use of a broad social desirability instrument and may therefore be expected to demonstrate only weak links between social desirability and responses to other items. There are better instruments available – such as the Driver Social Desirability Scale (Lajunen et al, 1997) that focuses on social desirability in a road use context – and it might reasonably be expected that any common method variance problems relating to social desirability motivations will be more easily detected using a better instrument. Further surveys with the current cohort will make use of this instrument.

Table 1: Correlations between self-reported crash involvement and DBQ Scales, before and after controlling for social desirability factor scores. Shaded cells are non-significant correlations

	Correlations	Partial Correlations Controlling for Social Desirability
DBQ Errors	**.08**	.07
DBQ Violations	**.13**	**.12**
DBQ Lapses	.01	.00
DBQ Aggressive Violations	**.11**	**.10**

General Discussion

This paper has focused on some potential problems with the use of surveys in road safety research – problems concerning sampling biases, assumptions about the accuracy or reliability of survey responses, and common method variance associated with social desirability – with examples drawn from a recent series of surveys. Road safety research places a high level of reliance on data derived from self-report measures administered in surveys, and there is increasing cause for concern about the quality of these data.

Sampling bias is generally poorly addressed in survey research. The use of poor sampling methods (such as convenience samples) and increasing reliance on survey methods that specifically exclude participation by some members of the population (on-line surveys when internet access levels are still relatively low, and telephone survey methods when market researchers report increasing levels of refusal from potential participants) suggest that researchers have a remarkably poor understanding of sampling bias and its implications for the generalisation of survey results to the population. The reliance on blunt comparisons with census data to "demonstrate" how representative and therefore unbiased the sample is raises strong concerns about the optimism bias of researchers in this area.

In the current surveys, there are hints of sampling biases even with a sound method to select potential participants. It would be inappropriate to generalise from this sample to the broader population of young drivers, and it seems reasonable to suggest that it is inappropriate to generalise from almost all survey studies to the broader population because it is generally impossible to know how sampling biases are associated with the measures of interest in the survey.

The assumption that it is possible to rely on accurate responses in a survey was discussed with reference to young drivers' recollection of how much experience they accrued as learner drivers. The responses to this item in the three surveys were surprisingly inconsistent, and the variation from survey to survey was remarkably high – with the standard deviation of the difference between estimates from survey to survey being about 50 hours and estimates of experience in the first survey accounting for only forty percent of the variance in estimates in the third survey. This was an item of current interest, and although it might be unreasonable to expect young drivers to recall the exact number of hours experience accrued as a learner driver, the amount of variation from survey to survey is surprising.

The broad assumption that survey responses are accurate and reliable is fundamental to the use of surveys in road safety research. If they are not reliable, there seems little point collecting information in this way. The lack of consistency in estimates of learner driver experience in surveys separated by only six months suggests that researchers who make use of surveys for data collection should be careful in making assumptions about the accuracy and reliability of participants' responses to items.

The final issue addressed here related to the possibility that intrapersonal response patterns might result in common method variance/bias problems that could then result in specious correlations between variables. This is an issue of current interest in road safety research. The data from the current survey series suggest that there are small significant effects of social desirability throughout the survey, but that these relationships do not appear to inflate the relationship between crash involvement and some potential predictors. This issue needs further consideration, however, as the social desirability measure used in the third survey may not be the best measure for use in a road safety research context. It is intended to use a better instrument in subsequent surveys.

Despite the relatively positive result here, there are still some significant concerns about common method variance/bias that need to be addressed in road safety research that uses self-report measures. The substantial review of this area in af Wåhlberg (2009) raises a number of issues that may constrain the use of surveys as reliable sources of information in road safety research. At the very least it may be useful to include measures of social desirability in future surveys and to use responses to these measures as covariates in analyses.

I started this paper by suggesting that the use of survey methods and self-report measures is significantly flawed. The data derived from the young driver surveys discussed here is consistent with this broad concern – if survey methods deliver unreliable data, with the possibility of confounding within the data set, from samples of participants that cannot be reasonably viewed as representative of the target population, it is difficult to know what value can be put onto the method. There are good reasons to take a stronger interest in the value of continuing to use survey methods in road safety. The argument used by some researchers – which amounts to "it's better than nothing" – is a poor substitute for applying better scientific measurement principles to road safety.

References

af Wåhlberg, A. (2009) Driver Behaviour and Accident Research Methodology: Unresolved Problems. Farnham, UK: Ashgate.

Doty, D., & Glick, W. (1998) Common methods bias: Does common method variance really bias results. Organizational Research Methods, 1, 374-406

Lajunen, T., Corry, A., Summala, H., & Hartley, L. (1997). Impression management and self-deception in traffic behaviour inventories. Personality and Individual Differences, 22, 341–353.

Sullman, M., & Taylor, J. (2010) Social desirability and self-reported driving behaviours: Should we be worried? Transportation Research Part F, 13, 215–221