Commentaries and Responses to "The Driver Behaviour Questionnaire as a predictor of accidents: A meta-analysis" [Commentaries lead by Anders af Wahlberg; Responses lead by J.C.F. de Winter]

Link to publication record in USC Research Bank:
http://research.usc.edu.au/vital/access/manager/Repository/

Document Version:
Author accepted manuscript (postprint)

Citation for published version:

Copyright Statement:
Copyright © 2012. This manuscript version is made available under the CC-BY-NC-ND 4.0 license http://creativecommons.org/licenses/by-nc-nd/4.0/

General Rights:
Copyright for the publications made accessible via the USC Research Bank is retained by the author(s) and / or the copyright owners and it is a condition of accessing these publications that users recognize and abide by the legal requirements associated with these rights.

Take down policy
The University of the Sunshine Coast has made every reasonable effort to ensure that USC Research Bank content complies with copyright legislation. If you believe that the public display of this file breaches copyright please contact research-repository@usc.edu.au providing details, and we will remove the work immediately and investigate your claim.
The following discussion is in response to a 2010 article published in the *Journal of Safety Research* by J.C. de Winter and D. Dodou entitled “The Driver Behaviour Questionnaire as a predictor of accidents: A meta-analysis” (Volume 41, Number 6, pp. 463-470, available on sciencedirect.com). The editors are pleased to provide a forum for this exchange and welcome further comments.

"The Driver Behaviour Questionnaire as a predictor of accidents: A meta-analysis"

Comments

Anders af Wåhlberg\(^a\), Lisa Dorn\(^b\)

\(^a\)Department of Psychology, Uppsala University, Uppsala, Sweden

\(^b\)Department of Systems Engineering and Human Factors, Cranfield University, Cranfield, United Kingdom

Abstract

A previously published meta-analysis of the predictive power of the Manchester Driver Behaviour Questionnaire (DBQ) versus road traffic crashes is criticized upon a number of counts, including the incomplete handling of common method variance, failure to control for exposure and how the correction for unreliability of the accident variable was undertaken. It is concluded that the results reported, and the conclusions drawn from these, are too favorable to the DBQ, despite the effect sizes being miniscule. In effect, the DBQ has no proven predictive power versus collisions.

Key words: DBQ, meta-analysis, crash, accident, self-report, common method variance
1. Introduction

In a recent issue of the Journal of Safety Research, a meta-analysis of the predictive power of the Manchester Driver Behaviour Questionnaire (DBQ) versus traffic crashes was published (de Winter & Dodou, 2010). As the DBQ is probably the most widely used driver self-report instrument in the world, and that this popularity to some degree would seem to rest upon its often stated ability to predict crash involvement, a meta-analysis of the published findings was indeed a welcome contribution to the traffic safety literature.

However, this meta-analysis also contained features and statements that we feel there is a need to comment upon. This does not necessarily mean that we disagree with de Winter and Dodou, but that some information is lacking.

2. Problematic features of the DBQ meta-analysis

2.1 Common method variance

The main problem associated with the interpretation of the DBQ/accidents association is common method variance (i.e., systematic biases in the self-reports used), which create part or all of the associations found (Podsakoff, Mackenzie, Lee, & Podsakoff, 2003; Chang, van Witteloostuijn, & Eden, 2010). Common method variance may be due to a number of different factors, and in some studies, substantial effects have been found (e.g., Hessing, Elffers, & Weigel, 1988; for reviews see Cote & Buckley, 1987; Podsakoff et al., 2003). This problem, however, was only partially treated by de Winter and Dodou, as only consistency motif and common scale anchors were discussed. Furthermore, when discussing the effects for one of the ways of circumventing consistency motif, using self-report data gathered on different occasions, the results from using this method in af Wåhlberg (2010), one of the publications included in the
meta-analysis, were left out. Here, the predictive power of the violation scale was halved when used to predict accidents not reported on the same occasion. It can also be noted that this cross-correlation test is a weak test of response bias, as it does not control for report biases that are stable over time, but are limited to occasion-specific effects.

Furthermore, the effect of social desirability was only mentioned in passing, although this was also reported upon in af Wåhlberg (2010). As it was found to strongly influence the violations/accidents association, this omission is of some importance.

In general, however, we agree with de Winter and Dodou in that the correlation between self-reported behavior and accidents can be, even is very likely to be, spurious.

2.2 Intercorrelations of different instruments

Another feature of the de Winter and Dodou paper is that inter-correlations with other self-report instruments seem to be interpreted as a positive feature of the DBQ; "...the DBQ errors and violations factors are strongly situated in a network of correlations with other questionnaires and tests..." (p. 463). In standard literature on psychometrics (e.g., Anastasi, 1988), this would usually be called a lack of divergent validity (a negative feature). It can be noted that the theoretical content of the self-report instruments mentioned (religious orientation, trait anxiety etc), does not seem to have much in common with the DBQ factors, and any correlation between such scales would therefore seem to be an indicator of common method variance (i.e., artefactual associations due to the method used).

However, whether the DBQ predicts other variables than accidents is really beside the point, if the point is to predict safety. As all known variables have very weak correlations with accidents, the use of a proxy variable is not a valid method for estimating the safety association,
only a way to find a strong correlation, especially if both variables are self-reported (af Wåhlberg, 2009).

2.3 Accident data validity

de Winter and Dodou also presented the standard argument that archived (actually police-recorded) accident data are not necessarily more valid than self-reports, apparently as a reason for downplaying the results of the only DBQ study that has used recorded data. This argument, we find irrelevant.

First, the study referred to, af Wåhlberg, Dorn, and Kline (2011), used company data, which presumably have much higher validity than state records, although this has not been extensively researched (af Wåhlberg, 2009). Such information was also given in the study regarding one of the data sets used, where the validity had been tested in various ways and found to be superior to self-reports of the same drivers. Second, as an example of bias in recorded data, the authors refer to older people more easily being injured, and that accidents involving senior drivers therefore are over-represented in official data. However, this injury bias argument is not valid for professional drivers, especially of heavy vehicles, as they are not so old, and seldom injured at all in their crashes. Another interesting feature of the age/injury argument is that it would lead to more variation in the accident variable for older drivers, which should boost the effect, and therefore make it easier to find. As noted by de Winter and Dodou, this does not seem to happen in the DBQ.

A final observation regarding the treatment by de Winter and Dodou of the af Wåhlberg, Dorn, and Kline study is that they did not comment upon the higher variance of the recorded accident data, as compared to the self-reports. As variance is crucial for bringing out an effect
when working with uncommon events like accidents; this difference loaded the comparison between criteria in favor of the records. Despite this advantage, the effect for archived data was much smaller than for the self-reports. Such an effect is predicted by the common method variance hypothesis, but is at odds with the interpretation of the DBQ as a valid instrument for accident prediction.

2.4 Exposure

Although the authors stated that exposure could be a confounder in this meta-analysis, exposure to risk of accident was apparently not controlled for in any way. Although in general, the coefficients for exposure versus the DBQ factors were very small (from -.09 to .17), this does not mean that this did not have a significant effect, as the overall correlations for the DBQ were also very small. There might also be a difference between studies, where some might be influenced by this, and others not, as we have noted in the British evaluation data (af Wåhlberg, 2010). As the correlation between a DBQ scale and accidents increases in different samples, so does the explained variance that is due to exposure (forthcoming publication), and it can therefore be suspected that the studies yielding larger DBQ/accidents correlations were more influenced by differences in exposure (reports).

2.5 Correcting for unreliability

At the end of the paper, the authors make a Spearman-Brown correction for unreliability of the measures, including accidents. The latter correction is unusual, and deserves some scrutiny. Referring to unpublished results, the authors state that accidents have a (mean) coefficient of stability of .11 (apparently a Pearson correlation). However, reliability of crashes actually differ
very strongly with the time period and/or the mean in the sample (or rather the variance; af Wåhlberg, 2009; af Wåhlberg & Dorn, 2009), and applying a single value from one sample would seem to be a somewhat crude method. At least, the time periods and/or means of the studies in the meta-analysis should have been compared to those of the sample used to calculate the coefficient of stability. Such a calculation would probably have yielded a higher correlation, and thus less of a correction factor because the time periods used for stability calculations by de Winter and Dodou were much shorter (maximum 1 year) than is usually the case for the DBQ (three to five years).

The corrected mean correlation of the errors and violations scales with accidents returned 'true' r values of .4 for both. This would seem to indicate that these two variables explained an astonishing 32% of the variance in accidents (unless they are strongly inter-correlated). Given that the DBQ does not incorporate medical problems, illegal drugs, cell phone use, psychiatric disorders, or information processing, all of which have been found to predict crash involvement, it would seem to be somewhat strange that such a strong effect could be found. If these other factors were also corrected for unreliability of the variables, the total amount of variance explained by all variables would probably exceed 100% (see Clarke & Robertson, 2005, for a similar claim of very large effects).

There appears to be only two possible explanations; either there is extensive overlap between the DBQ and these other factors, or one or more of them have inflated effect sizes, due to response bias. As it is difficult to see how the factors mentioned could possibly be measured by the DBQ, given its item content, the second explanation would seem to be the most viable.

2.6 Further references
Although we believe this to be of minor importance, we can also point out that the following references, which contain results for the association between DBQ factors and accidents, apparently were not included in the de Winter and Dodou study; Chapman, Roberts, and Underwood (2000); Parker (1999); Quimby, Maycock, Palmer, and Grayson (1999); Sümer and Özkan (2002); Stradling, Meadows, and Beatty (2004); West (1995); and West and Hall (1998). Elliott, Baughan, and Sexton (2007) reported values for motorcyclists and Diaz (2002) for pedestrians. Also could be mentioned Meadows, Stradling, and Lawson (1998), and Dobson, Brown, Ball, Powers, and McFadden (1999), which, however, could be using the same data as other studies that were included in the meta-analysis.

3. Conclusions

To summarize, we believe that the meta-analysis of de Winter and Dodou is somewhat too favorable to the DBQ by failing to make the above points. Yet we paradoxically agree with de Winter and Dodou in one of their conclusions and recommendations; more studies using other-source criteria are needed. The self-report-source only data are not reliable, and conclusions about the predictive power of the DBQ factors versus traffic safety are not yet possible to draw.

4. References


Response to commentary on “The Driver Behaviour Questionnaire as a predictor of accidents: A meta-analysis”

J.C.F. de Winter & D. Dodou

Department of BioMechanical Engineering, Faculty of Mechanical, Maritime and Materials Engineering, Delft University of Technology

Corresponding author:
Joost de Winter, Department of BioMechanical Engineering, Faculty of Mechanical, Maritime and Materials Engineering, Delft University of Technology, Mekelweg 2, 2628 CD Delft, The Netherlands
E-mail: j.c.f.dewinter@tudelft.nl

Abstract

A commentary on our previously published meta-analysis about the predictive validity of the Driver Behaviour Questionnaire (DBQ) raised a number of points. These points do not dispute the quantitative results as such, but suggest that our introduction and discussion overly favor the DBQ and are incomplete in a number of ways. The commentary targeted the following topics: common method variance, intercorrelations of different instruments, accident data validity, correcting for measurement error, correcting for exposure, and missing studies. Some of these points, such as the risk of common method variance (CMV) when self-reported data are intercorrelated, are valid albeit already dealt with in our article. For most of the points, however, we show that the commentary fails to effectively synthesize the existing evidence. Here we provide new empirical results on the importance of correcting for measurement error and show that the DBQ-accident correlation is robust to exposure. It is concluded that the commentary raised some valid points but changes nothing to our conclusions. We are pleased that af Wåhlberg and Dorn have kindly appraised our meta-analysis as a welcome contribution. Although they do not dispute our quantitative results and main conclusions as such, they do regard parts of our introduction and discussion as overly in favor of the DBQ and incomplete in a number of ways. Their commentary raises several points that we obviously agree with, since we have already presented them in our article. Their other points appear to be slanted against the
DBQ, in line with their earlier work in which they assert, “the world’s most popular driver questionnaire is unvalidated” (af Wåhlberg, 2009, p. 46). Below we respond to each comment.

1. **Common method variance**

Our meta-analysis provided an extensive discussion of validity threats, including ones not treated before in the DBQ literature, such as common scale anchors and publication bias. We see little added value in the commentary when it informs the readership of common method variance (CMV), as we clearly did this in our article. It is widely understood that CMV can account for a large share of the variance when self-reported data are intercorrelated (see Podsakoff, MacKenzie, Lee, & Podsakoff, 2003, for a highly cited review on CMV and recommended remedies). However, the position af Wåhlberg and Dorn take on regarding CMV is one-sided and fails to incorporate existing counterevidence.

Specifically, the commentary refers to temporal cross-correlation, a technique we used to test for CMV, and alerts us to the work of af Wåhlberg (2010), which showed that the correlation between DBQ violations and self-reported accidents was lower between two different questionnaires than within the same questionnaire ($r = .087$, $N = 5,667$ vs. $r = .140$, $N = 7,638$). Our meta-analysis presented similar results, but for 16 temporal combinations as part of a 36-month longitudinal study (Transport Research Laboratory [TRL], 2008; see also Wells, Tong, Sexton, Grayson, & Jones, 2008; $Ns$ between 1,718 and 9,138), yielding an average same-questionnaire violation-accident correlation of .12 and an average different-questionnaire correlation of .11 (see Table 2 of our meta-analysis). All violation-accident correlations were between .09 and .15 and so the results of af Wåhlberg fall within this range. The difference between both our results is likely due to measurement error: Even with a sample size as large as
5,000, the 95% confidence interval around a correlation of .11 is .08 to .14. While af Wåhlberg and Dorn concluded on the basis of their deduction that “the predictive power of the violation scale was halved when used to predict accidents not reported on the same occasion” (af Wåhlberg, commentary), we reported that “in some but not all cases, correlations seemed somewhat stronger when assessed within the same questionnaire as compared to correlations for time periods relatively far apart” (de Winter & Dodou, 2010, p. 466). By combining both af Wåhlberg’s and our results, it can be concluded that same-questionnaire correlations are probably higher than different-questionnaire correlations, but the difference is not so large as to invalidate the DBQ.

The commentary further points out that cross-correlation is a weak test of CMV, because it does not control for response biases that are stable over time. This comment is valid when read in isolation, but overlooks the notion that even without CMV, different-questionnaire correlations are expected to be lower than same-questionnaire correlations, because drivers’ risky behaviors are not stable over time. This is particularly true when drivers are in a training/developmental phase as was the case with both af Wåhlberg’s (2010) study and the TRL data (see also Jessor, Turbin, & Costa, 1997; Vassallo et al., 2010 for changes in risky driving behaviors among young drivers). As illustration, Table 1 presents correlations of DBQ violations across four time periods. The matrix corresponds to a simplex structure, having “the largest values near the diagonal (i.e., adjacent time periods) and declining values as one moves away from the diagonal (i.e., the correlations between variables further temporally removed from each other)” (Ackerman, 1987, p. 13). Summarized, attenuated temporal cross-correlations cannot be attributed to CMV as such, because individual traits are not stable over time.
The commentary states that we mention the effect of social desirability only in passing. Let us, therefore, clarify that several studies have investigated the effect of social desirability on the DBQ outcomes (Conner & Lai, 2005; Groeger & Grande, 1996; Harrison, 2009; Lajunen & Summala, 2003; Sullman & Taylor, 2010; Wickens, Toplak, & Wiesenthal, 2008), by calculating correlations with social desirability scales (e.g., lie scale from Eysenck Personality Inventory, Marlowe-Crowne Social Desirability Scale, and the Balanced Inventory of Desirable Responding), by partialling out the effect of the social desirability scores, or by administering the questionnaire in public versus private settings and calculating the differences in the DBQ responses. Each of these studies concluded that the effect of desirable responding on the DBQ factors and/or their correlations with external criterion variables such as accidents was small. The work of af Wåhlberg (2010) stands out by concluding that “the problem for traffic research is grave” (p. 99). In this study, af Wåhlberg used the Driver Impression Management (DIM) scale containing seven items on traffic violations, such as “I have never driven through a traffic light when it has just been turning red,” “I always keep sufficient distance from the car in front of my car,” and “If there were no police control, I would still obey the speed limits,” on a scale from “not true” (1) to “very true” (7). As Paulhus (1991) explained, impression management should only be controlled when it is “conceptually independent of the trait being assessed” (p. 23). Since all the DIM items capture the content of the construct of interest (i.e., violations), this scale should not have been used for assessing the responding style in the DBQ. Lajunen, Corry, Summala, and Hartley (1997) had already shown that the scale correlated .45 with self-reported traffic rule compliance. It is utterly unsurprising that the violation-accident correlation was attenuated when the DIM scale was partialled out, and af Wåhlberg’s position cannot be
justified. Instead, it can be concluded that up to now all valid evidence concurs: DBQ is robust to socially desirable responding.

2. Intercorrelations of different instruments

The commentary reacts to an introductory sentence in which we stated that the DBQ is strongly situated in a network of other questionnaires and tests (such as Trait Anxiety, Cognitive Failures Questionnaire, and Sensation Seeking Scale), by asserting that we seem to interpret such correlations as a positive feature of the DBQ and by pointing out that correlations between the DBQ and other self-reports may have arisen spuriously because of CMV.

We dispute the assertion that we regard correlations between self-reports necessarily as a positive feature, as our article reveals that we are well aware of CMV. In this exact line, we explained in the discussion of our paper that as the positive correlations between DBQ errors and violations reported by many are not expected from a theoretical viewpoint, and we attributed this phenomenon to CMV (common scale anchors, to be precise).

Furthermore, we disagree that the DBQ suffers from lack of discriminant (divergent) validity. The references accompanying the commented sentence of our introduction contain good demonstrations of such. Groeger and Grande (1996), for example, showed that errors, not violations, correlated with Broadbent’s Cognitive Failures Questionnaire investigating minor lapses and errors in everyday tasks ($r = -.32$ and $-.37$ vs. $-.03$, $N \approx 300$). DBQ violations correlated ($r = .29$) with the Sensation Seeking Scale, whereas DBQ errors did not ($r = -.04$ and $-.10$, $N = 101$; Schwebel et al., 2007). Similar were the findings by Rimmö and Åberg (1999) who found that sensation seeking explained 27%, 3%, 2%, and 3% of the variance ($N = 705$).
with violations, mistakes, inattention, and inexperience, respectively, and concluded that “this is interpreted as additional support for the suggested distinction between violations and other error types” (p. 162).

It is unfortunate that af Wåhlberg and Dorn tend to consider all non-zero correlations between self-reports as confirmations of the CMV hypothesis, whereas the literature contains dozens of studies spanning more than two decades that have linked DBQ correlates to a theoretical framework about driving behavior. The fact that women report more errors but fewer violations than men is just one example that provides interpretability to the violation and error factors and their distinct underlying psychological mechanisms (see Reason, Manstead, Stradling, Baxter, & Campbell, 1990 for seminal DBQ research and theoretical discussion; note that the gender variable is usually coded without error and the correlations are therefore unaffected by CMV). Similarly, the effects of anxiety on perceived risk, working memory, processing efficiency, and cognitive interference have been used to explain the correlations between trait anxiety and errors, whereas the relation of very low anxiety level with overconfidence has been proposed as the reason underlying the trait anxiety correlation with violations (Shahar, 2009; Stephens & Groeger, 2006). These findings run counter to the position of the commentary that trait anxiety does not seem to have much in common with the DBQ factors.

We were surprised by the statement that “the use of a proxy variable is not a valid method for estimating the safety association,” because accident prediction was not the purpose of our introductory statement about the network of correlates of the DBQ. Furthermore, the commentary downplays the meaning of proxy variables. Although correlations between the DBQ and surrogate safety measures alone do not prove that the DBQ predicts accidents, they are at
least highly suggestive. For example, Åberg and Wallén Warner (2008) found that DBQ violations predicted logged speeding in cars during field tests ($r = .43$, $N = 175$), while the error factors did not correlate (and inexperience errors even significantly negatively correlated) with logged speeding. Because logged speeding was determined objectively, the reported correlations are unaffected by CMV. Kinematic and vehicle dynamics analyses can easily prove that speed is a risk determinant, as it influences stopping distance and maneuverability. Indeed, speed has been reported to have a strong and causal effect on road safety (Elvik, Christensen, & Amundsen, 2004) and individual differences research (an area more prone to random measurement error than aggregated accident statistics) has also shown that registered speed and speeding are predictive of self-reported and registered accidents (Cooper, 1997; Maycock, Brocklebank, & Hall, 1998; Taylor, Lynam, & Baruya, 2000; Wasielewski, 1984; West, Elander, & French, 1992; for a review and other studies see Aarts & Van Schagen, 2006). The commentary appears insensitive to such findings and only emphasizes that CMV undermines the validity of the DBQ.

3. Accident data validity

The commentary rejects our position that not only self-reported accident data are susceptible to biases, but recorded data too. Note that our remark applied to all types of recorded accidents, including police reports, hospital data, insurance data, as well as fleet data from professional drivers, and not just company data as in the work by af Wåhlberg, Dorn, and Kline (2011), which is mentioned in the commentary. The literature discusses several sources of bias for recorded accidents that are less likely when using anonymous self-reports:
• Minor crashes do not come forward in recorded data, and even the more severe crashes are seriously underreported in hospital or police data (Derriks & Mak, 2007; Ward, Lyons, & Thoreau, 2006; see Elvik & Mysen, 1999 for a meta-analysis).

• Registering practices vary over time and region. Ranney (1994) discussed the inadequacy of public records, including different damage thresholds, resulting in large differences between individual states.

• Drivers may have personal interests in not reporting accidents to the police or to their insurance company, for instance, to avoid accountability, to hold on to claim-free years, or when the insurance does not cover the damage (O’Day, 1993). We agree with the commentary that the overrepresentation of injured older drivers in accident reports is less expressed amongst professional drivers. For nonprofessional drivers, however, the tendency to call the police to the scene or the likelihood to be hospitalized does depend on casualty age as well as on a variety of other variables, including drivers’ personality, number of vehicles involved, vehicle type, injury severity, and geographic region (e.g., Alsop & Langley, 2001; European Transport Safety Council, 2006), resulting in artifacts when establishing correlations with accidents. In other words, certain groups of drivers might be overrepresented for reasons unrelated to their actual accident risk (Elander, West, & French, 1993).

• Over-reporting also takes place. Green and Matterson (2010) found overreporting due to duplicate records. McKenna, Duncan, and Brown (1986) described a loose distinction between incidents and accidents in company data, with accident databases contaminated with such incidents as stones thrown through the window of a bus, attacks on staff, and passengers’ falls.

af Wåhlberg and Dorn point out that the variance is higher in recorded data, and that “variance is crucial for bringing out an effect when working with uncommon events like
accidents.” Although reliability is a necessary condition for validity, we do not see why the mere fact that more accidents are included in a database implies a higher validity. Whether an effect will be found depends on the types of incidents recorded in the database. The commentary also refers to a study by af Wåhlberg “where the validity [of recorded accident data] had been tested in various ways and found to be superior to self-reports of the same drivers” (af Wåhlberg, commentary). We are not clear which study this comment refers to, but content-wise the closest are presumably either af Wåhlberg (2002) or af Wåhlberg et al. (2011), which used both self-reported and recorded accident data. In either case, no empirical evidence was provided supporting the validity of the recorded accident data, and a definition of validity was not provided.

Summarizing, our point is that both self-report and recorded accidents should be treated with some caution when they are used as a criterion in differential accident involvement research. Perhaps automatic crash recording systems (e.g., Sul & Cho, 2009; Svensson & Wingård, 2003) are a good step towards providing trustworthy accident data for testing the validity of the DBQ in the future.

4. Correcting for unreliability

The commentary pointed out that our correction for attenuation is unusual, a surprising claim considering the established importance of correction for measurement error in theory testing (Liu & Salvendy, 2009; Schmidt & Hunter, 1999). The available DBQ research provided almost no information on measurement error, so we did not apply a correction as part of the meta-analysis. Instead, we applied one afterwards, based on the raw data of the largest DBQ study available,
and illustrated that observed DBQ-accident correlations are seriously attenuated. Below we go into more detail on how measurement error affects the DBQ-accident relationship.

From the TRL study (also used in our meta-analysis), an average accident stability of .11 was calculated (see Table 2 of this rebuttal for all correlation coefficients). This value was also mentioned in our meta-analysis and neatly corresponds to review data on accident stability (see af Wåhlberg & Dorn, 2009, Figure 1) for this average number (thus variance) of accidents (average of about 0.20 accidents per period). When using data from 1,138 subjects for whom DBQ and accident data were available for all four periods, the summed number of accidents in questionnaires 1 and 3 versus questionnaires 2 and 4 revealed an intercorrelation of .23, a clear increase in stability.

The principle of aggregation applies not only to accident data but also to DBQ data. Our meta-analysis averaged the DBQ-accident correlations amongst the subfactors. We explained in our work: “For example, when a correlation of .07 was described between errors and accidents and a correlation of .09 between lapses and accidents, then a correlation of .08 was noted between errors and accidents” (p. 465). This approach attenuates the DBQ-accident correlations in comparison to using composite score predictors (Barrick & Mount, 1991). From the TRL data we calculated a total violation score by summing the violations and aggressive violations scales, and a total error score by summing the errors, inexperience errors, and slips scales. By doing that, the average test-retest reliability amongst the four questionnaires increased from .70 to .74 (see Table 1 for raw data) for violations and from .65 to .73 for errors. As with the accident data, a further reduction of measurement error could be obtained by summing across questionnaires of different periods. For example, the summed score of questionnaires 1 and 3 versus the summed score of questionnaires 2 and 4 revealed a correlation of .88 for violations and of .87 for errors.
Having established that the reliability of accident data and DBQ data increases by aggregating across subfactors and across questionnaires administered months apart, it is now useful to investigate whether the DBQ-accident correlation also increases through aggregation. Using the same sample ($N = 1,138$), the non-aggregated average same-questionnaire violation-accident and error-accident correlations were both .11, in correspondence to our meta-analysis outcome. The correlation between the DBQ violation score summed across subfactors and the four periods, and the number of accidents summed across the four periods, was established at .22 (see Figure 1 for illustration of this correlation). Using the same method of aggregation, the correlation between summed DBQ errors and summed number of accidents was established at .18.

af Wåhlberg (2009) has previously noted that DBQ data aggregation is a good way to test for CMV. Specifically, he noted, “the scales of the questionnaires can be added and tested against the accident variables. If there are CMV effects, the resulting correlations should be lower than those within each questionnaire. If there is a true effect, the correlations should instead increase, because the mean of two (or more) measurements should be a better estimate of the true value, as random errors tend to cancel each other out between waves (p. 54).” af Wåhlberg (2010) concluded that the DBQ fails such a test, offering “further evidence of CMV” (p. 5), without providing quantitative data. The present results clearly show that the DBQ passes this CMV test.

Our estimate was that the DBQ-accident correlation could rise to .4 if all measures were perfectly reliable. This approximation is not unreasonable, given that the average summed number of accidents amongst all four surveys was still only $0.71 (SD = 1.03, N = 1,138)$, with 56% of the drivers reporting no accidents at all. We do not understand why the commentary
considers a correlation of .4, notably after correction for measurement error, “astonishing.”

Previous studies have also found that accidents can be predicted with correlations of .4 and higher. For example, Ball, Owsley, Sloane, Roenker, and Bruni (1993) found a correlation of .52 between Useful Field of View scores and at-fault crashes from state records ($N = 294$). The difference with the work of Ball et al. and many studies on accident prediction is that they oversampled for drivers having multiple crashes.

The commentary also pointed out that the DBQ does not incorporate medical problems, illegal drugs, cell phone use, psychiatric disorders, or information processing, all of which have been found to predict accident involvement. We think that this statement reveals a severe misunderstanding of the DBQ. The very purpose of the Driver Behaviour Questionnaire, as its name suggests, is to capture how a driver behaves on the roads. A person with latent conditions, such as medical problems or psychiatric disorders, is expected to score highly on the violation and error factors. For example, Reimer et al. (2005) showed that ADHD subjects had significantly higher scores in errors, lapses, and violations compared to controls. As explained by Sümer’s (2003) contextual mediated model, aberrant behaviors, including violations and errors, are proximal predictors of accident involvement and mediate the association of accidents with more distal elements such as psychological symptoms, risk taking, and sensation seeking. Furthermore, it may be noted that alcohol consumption was already included in the original DBQ (Reason et al., 1990) and that contemporary versions of DBQ have included cell phone use (Freeman, Davey, & Wishart, 2007).

5. Exposure
The commentary points out that our results were not corrected for exposure in any way. First of all, this statement is false, because our meta-analysis did include effect sizes corrected for exposure. We used a special moderator category for effect sizes other than zero-order correlations. These effect sizes were derived from regression analysis, often with mileage as one of the predictors.

Second, it may be noted that af Wåhlberg himself reported that the association between exposure and accidents is “usually weak” (af Wåhlberg, 2009, p. 88) when discussing the limitations of his own accident stability research that did not correct for exposure. Our meta-analysis showed that the correlation between mileage and DBQ factors is weak as well. It can be therefore be implied that the DBQ-accident correlation is also likely to be weak.

Third, as Evans (2004) explained, “There is no all-purpose definition of exposure; it always depends on the question being addressed” (p. 11). That said, we strongly doubt whether DBQ scores should be corrected for exposure, as it is the very purpose of the DBQ to assess how often a certain error or violation is committed by a driver. Correcting for exposure (e.g., number of trips) discards important information; therefore one can anticipate a (slightly) attenuated DBQ-accident relationship.

Either way, it is useful to put to the test what happens to the DBQ-accident relationship after applying correction for exposure. Above we showed that the aggregated DBQ-accident correlation was .22 and .18 for violations and errors, respectively ($N = 1,138$). Of these subjects, 975 provided mileage data in each of the four questionnaires, resulting in an average mileage amongst the four periods of 5,118 miles ($SD = 5,549$). Next, the average mileage was rank-transformed because of the highly skewed distribution. The correlations of mileage were .29 with
DBQ violations, .03 with DBQ errors, and .18 with the total number of self-reported accidents. The DBQ-accident correlation with mileage partialled out was determined at .20 and .20 for violations and errors, respectively ($N = 975$). The questionnaires also contained 11 questions on the number of trips, such as “In the last 6 months, how often did you drive in a busy town or city centre” and “In the last 6 months, how often did you drive on country roads,” all to be answered on a scale from 1 (never) to 6 (every day). Of the 1,138 subjects, 891 answered these 11 questions in all four questionnaires. The average score per person across all four periods was 3.35 ($SD = 0.63$) and correlated with .32 with violations, .00 with errors, .15 with self-reported accidents ($N = 891$), and .70 with mileage ($N = 782$). The DBQ-accident correlations with this average number of trips partialled out were .17 for violations and .16 for errors. Hence, the DBQ-accident correlation holds after correcting for two forms of exposure.

6. Further references

The commentary attended us to 11 studies supposedly not included in our meta-analysis. We appreciate af Wåhlberg and Dorn’s close scrutiny of our reference list for potential omissions on the DBQ-accident relationship. It may be noted that missing a small fraction of studies in a meta-analysis of this scope is almost inevitable.

The samples of four of the studies (Dobson, Brown, Ball, Powers & McFadden, 1999; Elliott, Baughan & Sexton, 2007; Parker, 1999; Stradling, Meadows, & Beatty, 2005) identified in the commentary were already included in our meta-analysis. Another study (Chapman, Roberts, & Underwood, 2000) was also included, but unfortunately, we omitted to cite it in our
reference list. We were aware of two studies (Díaz, 2002, also cited in our work; and Sümer & Özkan, 2002), but excluded them based on the criteria reported in our article.

Thus, it seems that not 11, but 4 studies were actually missing from our meta-analysis. The study by Meadows, Stradling, and Lawson (1998) analyzed a sample of young male offenders from a reform center, 48% driving without a license, and reported large correlations between DBQ and accident involvement ($r = .51$ for violations and .25 for errors, $N = 100$). The study by West and Hall (1998) reported a correlation of .26 between DBQ violations and accidents over the past 12 months in a sample of pre-licensed drivers ($N = 809$). West (1999) reported correlations between DBQ violations and accidents of .23 and for young ($N = 316$) and .19 for older drivers ($N = 376$), respectively. These three studies support a major result of our meta-regression: that the violation-accident correlation is strongest among young drivers.

Finally, the work of Quimby, Maycock, Palmer, and Grayson (1999) revealed significant correlations between DBQ violations and logged speed ($r = .41$, $N = 113$), providing further evidence that the DBQ predicts variables that are not affected by CMV. A correlation between DBQ violations and self-reported accidents was also described ($r = .09$). Finally, we conducted an additional literature search and retrieved five DBQ studies published since our meta-analysis, two of which providing correlations between the DBQ and accidents (Constantinou, Panayiotou, Konstantinou, Loutsiou-Ladd, & Kapardis, 2011; Feenstra, Ruiter, Schepers, Peters, & Kok, in press; Özkan, Lajunen, Doğruyol, Yıldırım, & Çoymak, in press; Rowden, Matthews, Watson, & Biggs, 2011; Sunday & Akintola, 2011). With the study of Sunday and Akintola being conducted in Nigeria, DBQ has now been applied in all continents.
Discussion

The purpose of a meta-analysis is to provide a quantitative summary of a metric of interest, correlations between the DBQ, and external criteria in our case. af Wåhlberg and Dorn’s commentary does not dispute the quantitative results in themselves (except with regard to the correction for exposure), but targets the qualitative introduction and discussion of our article.

The commentary raises some valid points, albeit points already discussed in our article. We too believe that common method variance can considerably distort the DBQ-accident correlations, and it is somewhat disconcerting to note the abundance of DBQ-accident research administering the DBQ and the accident item on the same piece of paper. We share the conclusion that more studies using other-source criteria are needed, although as pointed out in our article, quite some of such research already exists.

We appreciate af Wåhlberg and Dorn’s thorough knowledge of conducted research, and are grateful to them for bringing four additional references to our attention. However, we are not convinced by the way they synthesizes the available information. As this rebuttal demonstrates, there is considerable evidence—including temporal cross-correlations and correlations with objective criteria—showing that CMV effects may be small, certainly not large enough to invalidate the DBQ.

This rebuttal also provides new empirical evidence, in particular that the DBQ-accident correlation rises to about .20 when DBQ and accident data are aggregated across subfactors and separate questionnaires, a finding that runs counter to af Wåhlberg’s CMV hypothesis. Furthermore, we show that the DBQ-accident correlation is robust to correction for two metrics
of exposure. Summarizing, the commentary raised some valid points but changes nothing
germane to our conclusions.
References


Table 1

*Pearson correlations between DBQ violations score (sum of violation score and aggressive violation score) in longitudinal study consisting of four questionnaires taken at 6, 12, 24, and 36 months after passing the driving test (N = 1,138)*

<table>
<thead>
<tr>
<th>DBQ violations</th>
<th>6 months</th>
<th>12 months</th>
<th>24 months</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Accidents</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>12 months</td>
<td>.75</td>
<td></td>
<td></td>
</tr>
<tr>
<td>24 months</td>
<td>.68</td>
<td>.79</td>
<td></td>
</tr>
<tr>
<td>36 months</td>
<td>.66</td>
<td>.76</td>
<td>.81</td>
</tr>
</tbody>
</table>

*Note.* These results are based on the TRL data archive (2008). Only subjects without missing DBQ data and without missing accident data were included in the sample.

Table 2

*Pairwise Pearson correlations between number of accidents in longitudinal study consisting of four questionnaires taken at 6, 12, 24, and 36 months after passing the driving test (Spearman correlations in parentheses)*

<table>
<thead>
<tr>
<th></th>
<th>6 months</th>
<th>12 months</th>
<th>24 months</th>
</tr>
</thead>
<tbody>
<tr>
<td>6 months</td>
<td>.15</td>
<td></td>
<td></td>
</tr>
<tr>
<td>12 months</td>
<td>(.13)</td>
<td>.09</td>
<td>.13</td>
</tr>
<tr>
<td>24 months</td>
<td>(.06)</td>
<td>(.13)</td>
<td></td>
</tr>
<tr>
<td>36 months</td>
<td>(.05)</td>
<td>(.09)</td>
<td>(.15)</td>
</tr>
</tbody>
</table>

*Note.* Ns varied between 1,788 and 5,619. The average Pearson correlation was .11 and average Spearman correlation was .10. Average numbers of accidents were 0.24 (*SD = 0.53; N = 9,731,*
0.15 ($SD = 0.41; N = 7,189$), 0.19 ($SD = 0.46; N = 4,092$), and 0.19 ($SD = 0.45; N = 2,695$) for the respective questionnaires. When considering only the 1,138 subjects with DBQ and accident data for all four questionnaires, average Pearson and Spearman correlations were .12 and .11, respectively. These results are based on the TRL data archive (2008).

*Figure 1.* Number of self-reported accidents summed across the four periods as a function of DBQ violation score summed across subfactors (i.e., violations and aggressive violations) and the four periods. The 1,138 subjects were sorted by their total DBQ violation score, and automatically divided into eight groups based on percentiles (due to equal DBQ scores, the $Ns$ per group varied between 117 and 162). The horizontal axis shows the mean total DBQ violation score and the vertical axis shows the mean total number of accidents for the eight groups. Vertical lines are the 95% confidence intervals of the mean total number of accidents per group. These results are based on the data archive of TLR (2008).
Commentary on the rebuttal by de Winter and Dodou

Anders af Wåhlberg\textsuperscript{a}, Lisa Dorn\textsuperscript{b} & James Freeman\textsuperscript{c}

\textsuperscript{a} Department of Psychology, Uppsala University, Uppsala, Sweden

\textsuperscript{b} Department of Systems Engineering and Human Factors, Cranfield University, Cranfield, United Kingdom

\textsuperscript{c} Centre for Accident Research and Road Safety, Queensland University of Technology, Australia

In their rebuttal of our commentary to their meta-analysis, de Winter and Dodou have forwarded further evidence and arguments about the issues we raised. We would like to comment upon these and the conclusions drawn. In this further comment, we have followed the outline of the rebuttal, using the same headings, although this is to some degree misleading, as several subjects are discussed that should be considered under other headings.

1. Common method variance

de Winter and Dodou seem to agree that common method variance (CMV) does exist, and is a problem for self-report research in general. These authors also claim that they have covered all aspects of CMV in their meta-analysis. Readers can see for themselves whether this last claim is valid. However, what is interesting is that de Winter and Dodou refuses to acknowledge that CMV could be a problem for the DBQ, given the data available.
de Winter and Dodou first discuss the results found with temporal cross-correlation, and concludes from the available evidence that the differences found do not invalidate the DBQ. We will discuss the logic behind this conclusion in our general section.

First, however, we want to point out that the change in predictive power of the DBQ versus accidents between same and different questionnaire waves in their own data (mean rs of .12 and .11), actually constitutes a change in amount of explained variance of 16%.

de Winter and Dodou correctly points out that this difference between within and between waves scale/accident correlations in their study lies within the limits of measurement error. However, they fail to recognize that it might also go the other way. It is equally possible that the effect is larger than the difference indicates. What we currently have to work with are the figures found, while the size of the error is mainly an indication of the size of samples needed in future research.

Second, de Winter and Dodou discuss the principle of cross-correlation (same/different questionnaires) as a test of CMV. However, we cannot understand their position on this question. We originally claimed that this method is a weak test of CMV, which de Winter and Dodou do not seem to accept, or, alternatively, they do not think that cross-correlation is a test at all. Either way, they argue that individual traits are not stable over time, and that this explains lower inter-correlations between questionnaire waves (presumably of the trait variables), and that these are therefore not necessarily due to lack of CMV.

However, the correlations we discussed were between predictors and dependent variables, not repeated measures of the same predictor. The data presented by de Winter and
Dodou in their Table 1 are therefore not very relevant to CMV (and the 'accident' variable in the table is apparently erroneous), and in fact support our argument.

The problem we wanted to point out that can be tested with cross-correlations is; how can individual traits predict accident record if they are not stable over time? As accident record is usually measured over several years, the behaviors would need to be stable over a similar time to have predictive power. Also, if the time period between questionnaire waves is small as compared to the time period for which accidents are measured, this variable should change very little, because it is the same data that is (or should be) canvassed. If the correlations between predictors and crashes are higher within waves than between (and the time period between waves is short), then this indicates CMV.

Turning to the problem of effects of socially desirable responses in the DBQ, de Winter and Dodou refer to a number of studies, all of which they claim have found that such an effect is small. These studies will therefore be briefly discussed here, while the main problem (rejecting 'small' effects) will be covered in the general discussion.

Conner and Lai (2005), controlled for social desirability (Crowne-Marlowe scale) in the correlation between DBQ and accidents. They concluded that there were no effects, but reported no values. This was also the case for Groeger and Grande (1996). Given that de Winter and Dodou have refused to take into account the effects of social desirability reported in af Wåhlberg (2010), due to these not having been stated in detailed numbers, the use of such imprecise information as an argument in favor of the DBQ is peculiar.

The Lajunen and Summala (2003) and Sullman and Taylor (2010) studies are not relevant in the present discussion, as they were not about individual differences, but between
situations effects. As for Wickens, Toplak, and Wiesenthal (2008), they used psychology students as subjects, which is not a good population from which to sample if you are using a lie scale. The problems inherent in these studies have been discussed in more detail in af Wåhlberg (2009).

In the study by Harrison (2010), the differences in correlations between the DBQ scales and crash involvement were 'small' when controlling for responses to the Crowne-Marlowe lie scale. However, as pointed out above, in terms of differences in amount of explained variance, even a seemingly very small difference can actually account for a large part of the effect. In the case of Harrison's data, the variance explained by the lie scale was even larger than in de Winter and Dodou's data.

Finally, de Winter and Dodou criticize the use of the Driver Impression Management (DIM) scale in the control for social desirability, as it is conceptually similar to the items in the DBQ and other driver scales. However, we would like to point out that, in contrast to other lie scales, the DIM scale has been validated for traffic use (af Wåhlberg, Dorn, & Kline, 2010). In this study, it was found across several samples that the DIM and its sister scale correlated -.1 with self-reported crashes, and 0.01 with recorded accidents, a result that would be expected if self-reports of crashes and the lie scale are both influenced by CMV mechanisms.

Our conclusion, from the same results, is therefore rather the opposite from that of de Winter and Dodou; the DBQ is indeed susceptible to effects of CMV, although it is difficult to separate social desirability from other effects.

2. Intercorrelations of different instruments
de Winter and Dodou's position on the subject of the meaning of the associations between the DBQ scales and other self-report instruments was initially interpreted as a claim of such correlations as a positive feature by us. We have disagreed with this interpretation. However, in their rebuttal de Winter and Dodou consider this to be a misunderstanding, and attribute the non-theoretically expected correlations between violations and errors to common scale anchors (a CMV effect). This position we could in principle accept, although we do wonder why only this specific type of association could be due to common scale anchors, and not other ones.

Turning to the question of divergent validity, de Winter and Dodou mainly seem to discuss the differences in correlation pattern between violations and errors. This we find to be not relevant to the point we were making; that if the questionnaires named by de Winter and Dodou correlates with the DBQ, then what is its unique contribution?

de Winter and Dodou consider it to be 'unfortunate' that we consider non-zero correlations as confirmations of the CMV hypothesis. Although we consider it equally unfortunate that they only pay lip service to the possibility of CMV effects, we would like to point out that our position is that all of the results based solely on self-report data might be contaminated by CMV. We do not consider such correlations as evidence of anything, only as unreliable data.

Finally, proxy road safety variables are discussed. Here, the standard arguments for proxy variable use are repeated. This question has been treated at length in af Wåhlberg (2009), and we therefore refer to this source, with one exception.

de Winter and Dodou refer to the study by Åberg and Wallén-Warner (2008) as evidence in favor of the DBQ, saying that the violation factor correlated .43 with logged speeding. What
they fail to mention, though, is that self-reported speeding correlated .63 with violations. Such a
difference we interpret as a sizeable effect of CMV, although a methodological problem makes it
impossible to draw firm conclusions (see the discussion in af Wåhlberg, 2009).

3. Accident data validity

de Winter and Dodou argue again that recorded data tends to have various shortcomings, and
conclude that we need to be cautious with these as well as self-reported crashes. We agree with
this position and would respectfully advice that de Winter and Dodou follow their own advice
concerning self-reports of crashes. It can also be pointed out that the shortcomings of recorded
data do not tend to yield spurious results, as CMV effects are usually not present.

However, de Winter and Dodou do not seem to have understood the point we were trying
to make in our initial commentary. The point is that the general argument that recorded data can
be biased is not appropriate for the Swedish data in the study af Wåhlberg, Dorn, and Kline
(2011), as validity for these records has been established by more than a decade of close
collaboration with the bus company involved, and tested in af Wåhlberg (2002; see also af
Wåhlberg, 2011). Amongst the features found was that the records contained more incidents than
the drivers could report for the same time period (three years), despite the fact that they
themselves had written the reports found in the archive. Therefore, the af Wåhlberg et al. study
reports results that are difficult to explain by any mechanism other than CMV. de Winter and
Dodou claim that no test of validity has been undertaken for these data. This might be true in
some sort of absolute sense, as this is practically impossible. However, what was clearly shown
in these studies was that the validity of the recorded data was higher than that for the self-reports
for the same time period, unless, of course, de Winter and Dodou would like to suggest that more than 20% of the reports had been faked to begin with, and a high volume of vehicle damages would have gone unreported by the drivers and unnoticed by the bus company (but reported by the drivers when asked about them years later).

de Winter and Dodou also point out that no definition of validity was given in our commentary. This we found to be superfluous, as we use the word in the rather common sense of the reflection of the measured variable as a measure of reality (i.e., how well it does measure, what it is supposed to measure). Concerning traffic accidents, this refers to unintended traffic events that lead to some sort of damage to objects or living creatures. We note that de Winter and Dodou did not themselves provide a definition of validity in their original meta-analysis.

It can also be noted that although the validity of the recorded data from the United Kingdom and Canada in af Wåhlberg et al. (2011) was not tested, it is nevertheless somewhat strange that the effect of the DBQ can be found with self-reported, but not recorded data, given that both sources can be believed to be unreliable. This feature has not been commented upon by de Winter and Dodou, although it is possible that this is due to their belief that self-reports in general are more valid than records.

As for variance, this is a separate feature from validity. It is the statistical truth that restricted variance in one of the variables restricts the size of the effect (see Peck, 1993). Our point here is that the lack of correlations between the DBQ and recorded accidents in af Wåhlberg et al. (2011) cannot be due to low variance, as the recorded data we used had higher variance than the self-reports of crashes.
4. Correcting for unreliability

For this point, it can only be said that de Winter and Dodou, despite their lengthy reply, do not mention the main problem involved; that they have over-corrected the original DBQ-accident correlation, by using a very low reliability figure for accidents on data sets that probably have higher reliabilities.

As this point was not referred to in their rebuttal, we offer instead the example of correcting for the reliability of a four-item scale in eight-item scale data. Usually, the longer version of a scale has better reliability. Using the value for a shorter version of the same scale will therefore overcorrect, and a spuriously strong effect will result. Again, our point is that the .11 reliability figure found by de Winter and Dodou cannot be applied to the studies they have included in their meta-analysis, because these probably have higher reliabilities (as estimated from the time periods involved).

Turning to the question of what happens to accident correlations when DBQ data are aggregated, we note that de Winter and Dodou have not understood the logic of this method, and that their test does not address the question of CMV. The suggested method is to measure the scale and accidents repeatedly, and correlate these within each wave. Thereafter, the questionnaire is averaged over occasions, and used to separately predict the accidents in each wave. If CMV is at play these latter correlations will be weaker than the first ones, while if not, they will be stronger (due to higher reliability of the questionnaire). Similarly, the scale in wave one can be used to predict the accidents within wave two, and vice versa, with the prediction being that these correlations will be even lower than the values for the aggregated scale.
de Winter and Dodou, on the other hand, aggregated the accidents and predicted them with the aggregated DBQ. This method will of course yield a higher correlation than the average of the within-waves correlations, because CMV has not been removed or weakened. Instead, variance in the accident variable has been increased, thus increasing the correlation.

It can also be noted that apparently de Winter and Dodou used accident reports for separate time periods, while the suggested method pre-supposes that the reports are for strongly overlapping time periods, and the questionnaire waves are not more than a few months apart. If these required circumstances are not fulfilled, the effects found will to some degree be due to real changes in behavior, which is not the subject of the test.

de Winter and Dodou also claim that quantitative data were not provided for this kind of specific CMV test, which was undertaken in af Wåhlberg (2010). This is to some degree true, as, due to space limitations, these calculations were mainly described in text (but the same kind of effect can be calculated from Table 4). It can be noted, however, that de Winter and Dodou do not seem to have a problem using qualitative information when it supports the DBQ, as in the cases of Conner and Lai (2005) and Groeger and Grande (1996).

However, to clear up this matter, the exact figures for the effects reported in af Wåhlberg (2010) are presented in Table 1. It can be noted that N for the Young Driver Scheme sample is somewhat larger than that reported upon previously, as data was still being gathered when the 2010 paper was written.

The most important results are for waves 1 and 2 versus the dependent variables in wave 1 (second and third rows). These waves were distributed only about a month apart, and the predictive power of the DBQ scale should therefore be almost identical, if no CMV was
influencing the results, while here it is more than halved. For the other calculations, the effects are even stronger, but what is due to CMV and what is a natural effect of the accidents being canvassed for different time periods cannot be disentangled.

Table 1 about here

As for the strength of this test for CMV effects, it can be noted that it does not capture persistent report effects, like those of social desirability. The strongest test is comparing the correlations for self-reported and recorded crashes (or other dependent variables that can be objectively measured). If the recorded data have good validity, the difference between correlations will be a measure of all CMV effects.

5. Exposure

In our first commentary, we stated that the DBQ/accident correlations included in their meta-analysis had not been corrected for exposure, while de Winter and Dodou in their rebuttal claims that this had indeed been undertaken. Upon repeated inspection, we find that apparently, exposure was included in the 'Other' category of Table 1, although we cannot find that this was explicitly stated. Furthermore, not all studies can have controlled for exposure, as the number of studies predicting accidents is much larger than the numbers under the categories 'Mileage' and 'Hours driven per week'. Also, it would seem to be the zero-order values that were used for calculating "...the true DBQ-accident correlations..." (p. 468) of .4, although it was not stated which values were used.
In a short passage, de Winter and Dodou appear to be implying that a similar error to their own was committed in af Wåhlberg (2009). However, the difference is that in the section referred to (p. 88), the lack of exposure control in the data used is specifically pointed out as a problem.

The position taken by de Winter and Dodou on exposure control as not necessary was somewhat unexpected, offering a very different interpretation of the DBQ. Usually, it is agreed within traffic safety research that exposure should be controlled for, because individual differences mean differences between individuals under equal environmental circumstances. In terms of the DBQ violation scale, this would mean that drivers would tend to differ in their number of violations when driving equal numbers of miles, in the same environment.

de Winter and Dodou, however, would seem to suggest that what the violation scale is measuring is the total number of violations committed, and that exposure therefore should not be controlled for. This, of course, is an acceptable position, if stated when the DBQ is used. However, it could be asked whether this interpretation was the intention of the originators of the DBQ. Also, the scales should be re-labeled (the 'violations and amount of driving' scale), to more truthfully reflect their meaning. It can also be pointed out that if the DBQ is to be compared with other variables, which are used to predict accident involvement, they need to be compared on an equal basis with respect to exposure.

The data example provided by de Winter and Dodou in their rebuttal was very instructive, as was their conclusion from their calculations that 'the DBQ-accident correlation holds after correcting for two forms of exposure.' Actually, what happened was that three out of four correlations shrank considerably, when calculated as the change in variance explained.
6. General conclusions and arguments

de Winter and Dodou claim to accept that there exist a number of validity threats against the DBQ-accident correlations (social desirability, common scale anchors, exposure, etc.), but dismiss each of these with the argument that the effect is small, and thus does not invalidate the DBQ. Thereafter, this small effect is totally discounted when the overall conclusion is drawn.

There are two problems with this position; first, de Winter and Dodou are using double standards for what is considered 'small,' or rather of practical significance, when it comes to effect sizes. The effect size of the DBQ versus accidents is thus rather small to begin with, and it is against this size that the sizes of the validity threats should be compared, not against some general standard (which they have not stated). Rejecting an effect that alters the DBQ-accidents correlation by tens of percent because it is 'small' by some standards is therefore incorrect.

Second, de Winter and Dodou only consider each possible CMV effect in isolation, rejecting each as insignificant. If, on the other hand, all possible effects were controlled for (as when another source of data for the dependent variable is used), the summed effects would probably be rather large. The available evidence (as reviewed in af Wåhlberg, 2009) would seem to say that the effect for predicting traffic safety variables with self-reports is at least halved when different data sources are used, as compared with self-reports of the same variables within the same wave.

The current debate would very much seem to be due to basic differences in how evidence is viewed, or whether the glass is half full or half empty. de Winter and Dodou seems to accept the correlations between self-reported accidents and DBQ scales as valid evidence until (a lot of) evidence is forwarded that this is unreliable data. We, on the other hand, believe that the studies
based solely upon self-report data are unreliable, as they could be contaminated by CMV, and that a fair amount of evidence exists that this is indeed the case.

As de Winter and Dodou seem to be rather keen on exact figures, we would like to ask for an exact figure on the subject of DBQ-accident correlations, as guidance in further debate. de Winter and Dodou repeatedly states that the effects found in what few studies that have tested the validity of the DBQ does not invalidate it. However, it can then be asked what evidence they would accept as proof, or rather at what level of accident correlation they would see the DBQ as being invalidated, or not useful? As it is, they have found actual, zero-order, correlations around .1. We claim that this is an over-estimation, due to CMV effects. If it could be shown that, for example, half of the explained variance is due to CMV, would the remaining correlation (.007) be strong enough to 'validate' the DBQ?

The reason for asking this question is that social science researchers might, in principle, accept any correlation as evidence of 'validity,' as we do not have any guidelines for this. Therefore, any amount of evidence for CMV in the DBQ could be presented, and de Winter and Dodou could still claim that it did not invalidate the DBQ, because some miniscule true correlation would always remain. We therefore respectfully ask for their view on the lowest acceptable figure, and the reasons for this choice.

As for ourselves, we believe that at least half of the .1 correlations found by de Winter and Dodou is indeed due to CMV. There might exist a true, very small association between these self-reports and actual accidents, but we believe this to be too small to be of any practical or theoretical significance.
As a further step in the DBQ discussion, we would also like to ask researchers and anyone else who has used the DBQ, and who have included accidents and/or violations in their measures, but have not published these data, to contact us. Although de Winter and Dodou found no evidence of publication bias, we still believe this is a source of error that requires further investigation.

Finally, we agree with de Winter and Dodou that more research is needed on the validity of the DBQ and that the use of objective correlates can resolve the current debate. Time and data will tell whether the glass is half full, or half empty, or just empty.
7. References


Table 1: The correlations between the DBQ violation scale and self-reported accidents and penalty points in the Young Driver Scheme (YDS) sample, as described in af Wåhlberg (2010). DBQ-V was canvassed in all waves, while the dependent variables were measured in the first wave and third waves. First, calculations for the prediction of accidents in the first wave by the DBQ in the first and second waves are shown (second and third rows). Second, calculations for the prediction accidents in both the first and last waves by the DBQ in all three waves are shown.

<table>
<thead>
<tr>
<th>DBQ-V</th>
<th>N</th>
<th>Crashes, wave 1</th>
<th>Penalty points, wave 1</th>
<th>Crashes, wave 3</th>
<th>Penalty points, wave 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Wave 1</td>
<td>8376</td>
<td>.125***</td>
<td>.042***</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Wave 2</td>
<td>8376</td>
<td>.079***</td>
<td>.025*</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Wave 3</td>
<td>1186</td>
<td>.133***</td>
<td>.093**</td>
<td>-.044</td>
<td>.047</td>
</tr>
<tr>
<td>Wave 1</td>
<td>1186</td>
<td>.055</td>
<td>.025</td>
<td>.068*</td>
<td>.001</td>
</tr>
<tr>
<td>Wave 3</td>
<td>1186</td>
<td>.062*</td>
<td>.015</td>
<td>.072*</td>
<td>.054</td>
</tr>
</tbody>
</table>
Response to second commentary on “The Driver Behaviour Questionnaire as a predictor of accidents: A meta-analysis”

J.C.F. de Winter & D. Dodou

Department of BioMechanical Engineering, Faculty of Mechanical, Maritime and Materials Engineering, Delft
University of Technology

Corresponding author:
Joost de Winter, Department of BioMechanical Engineering, Faculty of Mechanical, Maritime and Materials Engineering, Delft University of Technology, Mekelweg 2, 2628 CD Delft, The Netherlands
E-mail: j.c.f.dewinter@tudelft.nl

Where we agree
Before we respond to the commentators’ criticism, it would be useful to elucidate where we agree and disagree. The commentators appear to agree with the meta-analysis in itself. That is, their dispute does not concern the meta-analytic procedures or numeric outcomes per se. The commentators also agree that the correlation between the DBQ and self-reported accidents is susceptible to bias having a random and systematic component (but they disagree about the possible direction and magnitude of the bias, see below). Furthermore, they agree that the DBQ has no proven predictive power of recorded accidents, as our meta-analysis showed by a 95% confidence interval of −.03 to .13.
Worthy of repetition is our position that common method variance can be a severe source of systematic bias in questionnaire research. Research in other domains has shown that in some cases 50% or more of the variance of relationships between variables can be attributed to CMV. Even two completely unrelated constructs could reveal a positive correlation, merely because of common method effects. The amount of CMV varies according to research design and context. For example, CMV is reported to be higher when the research involves abstract constructs (e.g., affective states) rather than concrete behavior (Podsakoff & Organ, 1986). CMV can inflate relationships, but at the same token, it can deflate relationships (Cote & Buckley, 1987). The latter possibility, ignored by the commentators, will occur when “the correlation between the methods is lower than the observed correlation between the measures with method effects removed” (Podsakoff, MacKenzie, Lee, & Podsakoff, 2003, p. 881).

Where we disagree

Our meta-analysis determined that the DBQ predicts self-reported accidents with correlations of .10 (errors) and .13 (violations), both correlations without correction for attenuation. The commentators’ main disagreement with us concerns whether these correlations are representative of correlations with real accidents or whether they arise from common method variance. There is no definitive answer to this; after all, no DBQ research with real accidents (e.g., recorded by an automatic crash record system) has ever been conducted. However, empirical research is available that allows to make an informed judgment about the DBQ-real accidents correlation.

The position of the commentators is that the observed DBQ-accident correlation is “very likely to be spurious.” They also state that the correlations are at least halved when particular
CMV sources are removed, and suggest that this reduction is even larger when multiple CMV sources are combined. Elsewhere, they point out that it is “not true” (af Wåhlberg, 2010a, p. 101) that the DBQ predicts accident involvement, that “the association with (self-reported) crashes is due to common method variance” (af Wåhlberg, 2009, pp. 46–47), that “the DBQ scale only predicts self-reported accidents, not recorded crashes” (af Wåhlberg, Dorn, & Kline, 2011, p. 66), and that “the DBQ can weakly predict self-reported accidents, but not company or state-recorded (i.e., objective) data” (af Wåhlberg et al., 2011, p. 77; emphasis added). We gather that the commentators’ position is that the DBQ-real accident correlation is probably close to .00, or, in their words, “too small to be of any practical or theoretical significance.”

We argue that the commentators’ proposition is implausible and that the DBQ-real accidents correlation is positive. To respond to the commentators’ request for a numeric threshold, our estimate is that the DBQ-real accident correlation is greater than .05, which becomes about .2 after correction for random measurement error. More likely, however, the correlation is around .1 (.4 after correction), similar to the correlation between the DBQ and self-reported accidents or, to borrow the commentators’ metaphor, it is a completely full glass.

Below we discuss in detail issues not mentioned the earlier commentaries, including the hypothetical case that the commentators’ model is correct, that is, that the DBQ-accident correlation is close to zero. We show why this hypothesis is unlikely to be true and even inconsistent with their own research findings.

**Temporal cross-correlations**
Our meta-analysis showed that DBQ violations and errors are consistently predictive for accidents that were self-reported 30 months before and 30 months after administering the DBQ, which indicates that for this large sample dataset, CMV due to cognitively accessible constructs and transient factors such as mood is small. Our interpretation is strengthened in light of the fact that the correlations were established in a group of novice drivers, who were still maturing, learning to drive (see the learning curve of the mean number of accidents per year in Table 2), and had unsettled lifestyles.

In their second commentary, af Wåhlberg et al. provide an extension of their previously published data, presenting somewhat larger temporal shrinkages than we found (from .13 to .08, \( N = 8,376 \)). The commentators state that the predictive power of their violation factor should be almost identical between waves 1 and 2 if no CMV were involved, because these waves were separated by only about a month. Curiously, however, af Wåhlberg himself had earlier characterized the DBQ violation factor of the same Young Driver Scheme sample as “highly unstable over even such short time” (af Wåhlberg, 2010b, p. 336). Indeed, a closer inspection reveals that the wave 1-wave 2 correlation of their violation scores was only .47; \( N = 5,658 \) (af Wåhlberg, 2010b). In Table 1, we summarize all test-retest reliabilities of the DBQ violations we could retrieve from the literature. It can be seen that af Wåhlberg’ correlation is an outlier compared to the stabilities that others have reported. Their low stability could be the result of an online course in safe driving, which was administered in between waves 1 and 2, or a consequence of the fact that the respondents were taking part in a police scheme that targeted young drivers who had committed traffic violations, making the DBQ and accident data susceptible to regression towards the mean (cf., Elvik, 2002). Either way, if repeated measures of the same predictor reveal such an unusually low stability that cannot be explained by random
error, then this indicates that the drivers have modified their behavior and it is to be expected that
the different-questionnaire correlations between predictor and criterion would be lower than the
same-questionnaire correlations.

Furthermore, since this is a discussion about CMV, it would have been prudent of the
commentators to remind the reader that this e-learning course and survey were used as an
alternative to court prosecution, fine, or penalty points, and that these threats were present in
waves 1 and 2, but not in wave 3 (af Wåhlberg, 2010a). This masks the relationships between
wave 3 and the other two waves, and virtually guarantees shrinkage of the temporal cross-
correlations. Because of these methodological issues, the cross-correlations provided by af
Wåhlberg et al. (2011) do not allow for generalization in the CMV-DBQ debate. The temporal
separation test should preferably be applied in a group of experienced drivers, without training
interventions or conditional sampling.

The commentators argue that our previous analysis of the presence and size of CMV in
self-reported data is incorrect. In Tables 2 and 3, we report results according to their proposed
methodology. A slight increase is observable in the aggregated DBQ-accident correlation, from
.12 to .13 (.11 to .13 for Spearman correlations) for DBQ violations, and from .10 to .11 (.09 to
.09) for DBQ errors, supporting the no-CMV hypothesis formulated by the commentators: “If
CMV is at play these latter correlations will be weaker than the first ones, while if not, they will
be stronger.”

We agree with the commentators that temporal separation does not control for more
enduring response tendencies (e.g., due to overconfidence or other personality traits); no
The commentators consider the work of Wickens, Toplak, and Wiesenthal (2008) invalid because psychology students would know about SD and would therefore avoid this behavior. However, the opposite may be true instead, namely that “psychology students are more likely to respond in a socially desirable way because they are most familiar with attitude assessments” (Leite & Cooper, 2010, p. 288).

Other researchers have also pointed out that SD influencing the DBQ-accident correlation may act in a direction opposite to that assumed by the commentators. For example, Parker and Manstead (1996) noted that: “It seems reasonable to argue that this [violations-accident] relationship would be weakened rather than enhanced by the operation of social desirability biases, since those who regard their driving violations as a mark of their superior driving ability (and may therefore be inclined to overestimate the number of violations they commit) would presumably also be inclined to under-report the number of accidents in which they are involved” (p. 201; see also Lawton, Parker, Stradling, & Manstead, 1997 for similar remarks).

The commentators further contend that the robustness of the DBQ to public-private manipulation is “not relevant, as they were not about individual differences,” despite this technique being well established (e.g., Lautenschlager & Flaherty, 1990; Paulhus, 1984). If DBQ scores are invariant to such manipulation at the item level, it is plausible that DBQ-accident correlations are also invariant to this manipulation. Moreover, the commentators’ claim that
Conner and Lai (2005) “reported no values” is inaccurate, as correlations between DBQ factors and SD are presented in that report.

It would have been useful if the commentators had clarified what correlations with SD scales mean to them. There is much controversy on this topic and the commentators did not discuss the possibility that partialling out SD scales may suppress the relationships of interest (cf., Ganster, Hennessey, & Luthans, 1983; McCrae & Costa, 1983). This clearly happened in af Wåhlberg’s work (2010a), where he partialled out a violations construct from the DBQ violations-accident correlation, thus engaging in what has been called the “partialling fallacy” (Gordon, 1968). As early as 1960, the SD research of Crowne and Marlowe warned against the “problem of overlapping meanings” (p. 353).

The commentators suggest that various CMV sources and validity threats should be “summed.” Obviously, CMV effects are not additive, as CMV sources are interacting and have common causes. It would have been more informative if the commentators had provided the readership with a framework explaining the various CMV sources, including the direction of the effects and their temporal granularities, as well as some information on how CMV could be combined and remedied.

Correcting for unreliability

af Wåhlberg et al. (2011) assert that the overall DBQ-accident correlations of .10 (errors) and .13 (violations) of our meta-analysis are “rather small.” They also argue against proxy variables because these have “very weak correlations with accidents.” The commentators seem to invoke the small-correlation argument at their own convenience. For example, elsewhere they
(correctly) point out that low correlations between accidents result from “artefacts of the short time periods and low-risk populations utilised” (af Wåhlberg & Dorn, 2009, p. 88).

Indeed, correlations with accidents are necessarily small because accidents are rare events following a discrete probability distribution. We illustrated the principle of aggregation: the DBQ-accident correlation rose to .22 when the results of four separate questionnaires were combined. Using Spearman’s correction for attenuation, we showed that a correlation of about .40 is plausible if predictor and criterion were perfectly reliable.

af Wåhlberg et al. (2011) suggest that we made an overcorrection because other data “probably have higher reliabilities.” In Wells, Tong, Sexton, Grayson, and Jones (2008), to date the largest DBQ study available, the respondents reported an average of 0.19 accidents per period, which is not unusually low compared to other DBQ studies. In the second largest DBQ study (Elliott, Baughan, & Sexton, 2007), 89% of the drivers reported no accidents at all. Perhaps the third author of this commentary should refer to his own DBQ research, which states that 86% of the drivers reported zero accidents (Freeman, Wishart, Davey, Rowland, & Williams, 2009; with 4,792 respondents, hence also receiving a large weight in the meta-analysis). Note as well that our correction did not consider range restriction, an artifact that may lead to underestimations of true validities as high as 20–30% (Hunter, Schmidt, & Le, 2006). For example, several DBQ studies focused on specific groups of drivers, such as exclusively women or drivers from a specific region, and therefore these data have lower standard deviations than the total driving population.

We believe that the criticism of overcorrection is out of place, if only because af Wåhlberg’s (2009) own tabular overviews nicely show that registered data from insurance
companies and state records report stabilities are also around .1 and even somewhat lower for about the same number of accidents per driver (e.g., Daigneault, Joly, & Frigon, 2002, $r = .09–.11$; Gebers & Peck, 1994, $r = .07–.13$; Hauer, Persaud, Smiley, & Duncan, 1991, $r = .09$; Peck, McBride, & Coppin, 1971, $r = .03–.06$; all $Ns > 50,000$).

**Self-reported versus registered accidents**

The commentators invoke the argument that their registered data are valid because they have been working with the bus company for 10 years. Using the same rhetoric one could argue that the DBQ must be more than doubly valid, because researchers have been working with the DBQ for 22 years.

We agree with the commentators that registered data (i.e., derived from insurance company, fleet company, state records, or police data), in principle, are less susceptible to CMV than self-reported data. However, as the commentators explain, their bus drivers reported accidents themselves (with some safeguards in place such as vehicle inspections). It can be hypothesized that due to work stress, fear of losing a job or unanimity, such registered company data are even more susceptible to CMV than self-reported data from anonymous questionnaires, which are usually administered by scientific institutions with whom the respondents are not affiliated. We also regard the characterization “objectively recorded data” (af Wåhlberg et al., 2011, p. 76) as somewhat misleading, because data were not recorded in a mind-independent manner, as they would have been if measured by an automatic crash record system for example.

The commentators’ main argument about validity of registered data seems to rest on the number of accidents in their database as compared to the number of self-reported accidents. We
feel that the commentators make a fundamental error here. Under the appropriate use of the word ‘validity,’ accident data should be compared to real accidents. Instead, the commentators compare self-reported accident data to registered accident data and label the former invalid. In other words, their definition of validity carries with it the implicit assumption that registered accidents are superior. This may also explain why they used both a negative and a positive difference to claim that their registered accidents are more valid. For example, Dorn, Stephen, af Wåhlberg, and Gandolfi (2010) have stated previously, “unfortunately, the use of recorded crashes makes the statistical power of the analysis weaker, due to low variance” (p. 1430), whereas in their second commentary af Wåhlberg et al. claim the very opposite, namely that “the lack of correlations between the DBQ and recorded accidents in af Wåhlberg et al. (2011) cannot be due to low variance, as the recorded data we used had higher variance than the self-reports of crashes.”

In conclusion, we see no reason why self-reported accident data should be regarded as inferior or superior to registered accidents. Self-reported and registered criteria are susceptible to different sorts of systematic bias, and therefore have a complementary value. Obtaining a definitive answer on the DBQ-real accident correlation requires more than merely “time and data” (af Wåhlberg et al., second commentary); it also requires a valid criterion.

**What if the commentators’ model holds?**

Suppose indeed that DBQ violations and errors are not predictive of accidents. This implies that drivers who report a large number of violations and errors are not involved in more accidents than drivers who report no or only few violations and errors, a position that defies common sense
and is at odds with the observation that human fallibility is a contributing factor in more than 90% of traffic accidents (Evans, 1996). We prefer to adhere to Laplace’s principle that extraordinary claims require extraordinary evidence. However, the commentators adhere only to some questionable CMV evidence and do not forward any behavioral or physical (vehicle-dynamics) mechanisms to support the claim that the DBQ does not correlate with real accidents. For example, do they believe that DBQ factors do not relate to real violations and errors because drivers are completely unable to assess what they do in real traffic due to cognitive biases? Is it that violation-prone or error-prone drivers somehow have superior alertness or crash-avoidance abilities when a collision is impending, or do people who report making many violations and errors generally drive in safer cars in terms of maneuverability or mass? In order to be convinced that the DBQ does not predict real accidents, we would require stronger evidence than the provided CMV argumentation.

The commentators attach no value to the fact that the DBQ factors correlate interpretably with an array of other questionnaires. They attribute all such correlations to CMV, without considering the magnitude and direction of these correlations, and simply “do not consider such correlations as evidence of anything, only as unreliable data.” We agree that many of the observed correlations between self-reports may be biased, sometimes extensively, for example due to acquiescence when the same item format is used. Yet, the unduly draconian position presented in the second commentary is difficult to reconcile with an abundance of traffic safety research, including many of its second and third authors’ own publications reporting correlations between self-reports (e.g., stress in Driver Stress Inventory and Driver Behaviour Inventory, stress coping in Driver Coping Questionnaire, driving attitude in Driver Attitude Questionnaire,
work attitude in Safety Climate Questionnaire, DBQ itself) and drawing meaningful conclusions from them.

The commentators attach no value to the fact that the DBQ predicts numerous accident predictors that are likely to be unaffected by CMV, including gender, age, speeding recorded by in-vehicle recording systems, unobtrusive radar measurements, on-board observers, and speeding infringements issued by the police (Åberg & Wallén-Warner, 2008; de Angeli et al., 1996; Palamara & Stevenson, 2003; Quimby, Maycock, Palmer, & Buttress, 1999; Quimby, Maycock, Palmer, & Grayson, 1999), reaction times (Sümer, Ayvaşık, & Er, 2005), ADHD (Fried et al., 2006), behavior in driving simulators (e.g., Neimer & Mohellebi, 2009; see also original meta-analysis), knowledge of road rules (Tay, 2010), peripheral motion detection thresholds (Henderson, Gagnon, Bélanger, Tabone, & Collin, 2010), and so forth. The commentators seem to ignore research such as that of Cooper (1997) which linked driver record information with insurance-claim data, and showed that drivers who were not caught with excessive speed had on average 0.25 accidents ($N = 1,031,861$), whereas drivers who were caught once and twice had an average number of 0.49 ($N = 60,391$) and 0.80 ($N = 6,238$) accidents, respectively. The commentators simply suggest that using a proxy variable, including speeding, is not a valid method for estimating a safety association.

Although the commentators attach no value to proxy variables for road safety, af Wåhlberg advocates the message that accident liability is itself predictable and a “very stable” (af Wåhlberg, 2009, p. 81) individual characteristic; he also proposes reviving the term “accident

---

1 Åberg and Wallén-Warner (2008) found that DBQ violations predicted registered speeding ($r = .43$), a finding which the authors themselves found “impressive” because at least five reasons necessarily deflated this correlation (regional and temporal differences of drivers, Intelligent Speed Adaptation-device (ISA) errors, mismatching of drivers, range restriction due to volunteer sampling, and range restriction because of the effects of the ISA). More than a year passed between the DBQ and the observations of actual speeding behavior. Self-reported speeding and self-reported violations intercorrelated .63, an expected number because speeding has high factor loadings on the violations score (de Winter & Dodou, in press).
proneness” as a general trait. This leaves us wondering which construct may then predict accidents. af Wåhlberg (2006) himself has kindly answered this question. His “driver celeration theory” proposes that celeration—a metric of speed variability correlating strongly (> .7) with mean speed—is “superior to all other variables as a predictor of individual traffic accident involvement” (p. 43). The celeration theory (which, actually, we regard as potentially valid) is impossible to reconcile with the commentators’ debunking of the speed-accident correlations.

The first two authors of the commentary have also found that self-reported stress and stress coping (surveyed with items from the Driver Stress Inventory and the Driver Coping Questionnaire combined into a Bus Driver Risk Index) are valid accident predictors of recorded accidents (Dorn et al., 2010; while including, incidentally, a few potshots at the DBQ). Validation of their questionnaire was based on a number of significant correlations (.09 at maximum) with registered accidents. Dorn et al. (2010) also validated their coping questionnaire by using celeration in a driving simulator, while af Wåhlberg (2009) devoted a book chapter to pointing out that “driving simulator variables have very little backing as safety proxies for individual differences” (p. 200). Curiously, in their article, Dorn et al. (2010) applied Matthews’ transactional model of driver stress, and they (correctly) explain that “violations” and “driver error” (p. 1420) are amongst the variables that mediate the relationship between stress and accidents. In the same vein, Matthews explained that the impact of stress on road safety may “be mediated by behaviors including cognitive lapses, errors, and intentional traffic violations” (Rowden, Matthews, Watson, & Biggs, 2011, p. 1332). It is precisely these constructs that the DBQ aims to capture. Thus it is unclear why the commentators are so against the DBQ as a valid predictor of accidents, when they simultaneously adhere to DBQ constructs as mediators between their own questionnaire and accidents.
The commentators leave us with an internally inconsistent and isolated picture of the world.

**Conclusion**

In this second rebuttal, we argued that the hypothesis “correlation between the DBQ and real accidents is greater than .05 (or greater than .20 for the disattenuated correlations)” is more plausible than the commentators’ alternative. The discussion of our meta-analysis highlighted diverse positive and negative features of the DBQ, providing a more nuanced picture than the commentators’ exclusively negative view, a view that only emphasizes that the DBQ is invalid and does not even recognize the possibility that method effects can also deflate observed correlations.

The commentators adhere to isolated arguments and appear to attach no weight at all to other sorts of accumulated evidence, such as correlations between the DBQ and recorded speeding, or a recent case control study demonstrating a DBQ-accident association (Jayatilleke, Poudel, Dharmaratne, Jayatilleke, & Jimba, 2010). A more constructive approach would be to consider the well-established nomological net of the DBQ, and acknowledge the likelihood that the DBQ and accidents share a meaningful correlation.

Finally, we note that af Wåhlberg et al. (2011) have previously concluded that “the DBQ literature is characterized by much less coherence than is usually implied by the researchers active in this field, including the power to predict accidents” (p. 77). Apparently, they did not consider that basing conclusions on individual significance tests is known to yield systematically erroneous conclusions (Schmidt, 1992). Our meta-analysis (de Winter & Dodou, 2010) estimated
overall correlations by pooling a large number of DBQ studies involving over 45,000 respondents. We revealed that DBQ research is markedly homogenous with almost all DBQ-accident correlations falling within the 95% confidence interval of a fixed-effects model.\textsuperscript{2} Utilizing meta-analysis, we were able to see the forest for the trees and have contributed cumulative knowledge to the field.

\textsuperscript{2} As mentioned in the meta-analysis, an outlier was Freeman et al. (2009) reporting a relative large correlation between errors and self-reported accidents, which fell outside the 95% confidence interval.
References


<table>
<thead>
<tr>
<th>Reference</th>
<th>Time period</th>
<th>Correlation per violation factor</th>
<th>Number of subjects</th>
<th>Number of items</th>
<th>Mean age</th>
<th>Intervention</th>
</tr>
</thead>
<tbody>
<tr>
<td>af Wåhlberg (2010b, YDS)</td>
<td>2 months</td>
<td>.47</td>
<td>5,658</td>
<td>7</td>
<td>21.9</td>
<td>Yes(^a)</td>
</tr>
<tr>
<td>af Wåhlberg (2010b, Control)</td>
<td>2 months</td>
<td>.67</td>
<td>133</td>
<td>7</td>
<td>22.4</td>
<td>No</td>
</tr>
<tr>
<td>Burgess and Webley (1999)</td>
<td>6 months</td>
<td>.69</td>
<td>608</td>
<td>8</td>
<td></td>
<td>About 20% between 17–21 Yes(^a)</td>
</tr>
<tr>
<td>Harrison (2009)</td>
<td>6 months</td>
<td>.75 &amp; .72(^b)</td>
<td>822</td>
<td>6 &amp; 4</td>
<td>20.8</td>
<td>No</td>
</tr>
<tr>
<td>Özkan, Lajunen, and Summala (2006)</td>
<td>36 months</td>
<td>.76</td>
<td>622</td>
<td>9</td>
<td>43.5</td>
<td>No</td>
</tr>
<tr>
<td>Palamara and Stevenson (2003)</td>
<td>12 months</td>
<td>.54</td>
<td>995</td>
<td>8</td>
<td></td>
<td>Only 17-years old drivers</td>
</tr>
<tr>
<td>Parker, Reason, Manstead, and Stradling (1995)</td>
<td>7 months</td>
<td>.81</td>
<td>54</td>
<td>8</td>
<td></td>
<td>About 40</td>
</tr>
<tr>
<td>Rimmö (1999)</td>
<td>6 months</td>
<td>.84</td>
<td>65</td>
<td>8</td>
<td>41</td>
<td>No</td>
</tr>
<tr>
<td>Wells et al. (2008)</td>
<td>6–30 months</td>
<td>.70 &amp; .68(^c)</td>
<td>Between 1,758 and 5,490</td>
<td>6 &amp; 6</td>
<td>22.8</td>
<td>No</td>
</tr>
</tbody>
</table>

\(^a\)Course as alternative to court prosecution, fine, or penalty points

\(^b\)Ordinary violations and aggressive violations, respectively
Violations and aggressive violations, respectively. The average of the six combinations between the four waves is reported; .73 when the aggressive violations and violations subfactors are first summed.
Table 2. Descriptive statistics of self-reported number of accidents and DBQ violations score (sum of violations score and aggressive violations score) in longitudinal study consisting of four questionnaires taken at 6, 12, 24, and 36 months after passing the driving test ($N = 1,138$).

<table>
<thead>
<tr>
<th></th>
<th>$M (SD)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Accidents Q1</td>
<td>0.23 (0.53)</td>
</tr>
<tr>
<td>Accidents Q2</td>
<td>0.13 (0.38)</td>
</tr>
<tr>
<td>Accidents Q3</td>
<td>0.18 (0.43)</td>
</tr>
<tr>
<td>Accidents Q4</td>
<td>0.17 (0.42)</td>
</tr>
<tr>
<td>DBQ-V Q1</td>
<td>2.70 (0.75)</td>
</tr>
<tr>
<td>DBQ-V Q2</td>
<td>2.82 (0.81)</td>
</tr>
<tr>
<td>DBQ-V Q3</td>
<td>2.97 (0.92)</td>
</tr>
<tr>
<td>DBQ-V Q4</td>
<td>3.05 (0.95)</td>
</tr>
<tr>
<td>DBQ-V Q1–4</td>
<td>11.55 (3.08)</td>
</tr>
</tbody>
</table>

Note. These results are based on the Transport Research Laboratory (TRL) data archive (2008). Only subjects without missing DBQ data and without missing accident data were included in the sample. Q1 and Q2 covered the past 6 months, whereas Q3 and Q4 covered the past 12 months. The mean number of accidents per year can be therefore approximated as 0.46, 0.26, 0.18, 0.17, indicating a learning curve.
Table 3. Pearson correlations between DBQ violations score (sum of violations score and aggressive violations score) and self-reported number of accidents in longitudinal study consisting of four questionnaires taken at 6, 12, 24, and 36 months after passing the driving test ($N = 1,138$). Spearman correlations between parentheses.

<table>
<thead>
<tr>
<th>Violations</th>
<th>Accidents</th>
<th>$r$ (rho)</th>
</tr>
</thead>
<tbody>
<tr>
<td>DBQ-V Q1</td>
<td>Accidents Q1</td>
<td>.15 (.15)</td>
</tr>
<tr>
<td>DBQ-V Q2</td>
<td>Accidents Q2</td>
<td>.13 (.11)</td>
</tr>
<tr>
<td>DBQ-V Q3</td>
<td>Accidents Q3</td>
<td>.10 (.10)</td>
</tr>
<tr>
<td>DBQ-V Q4</td>
<td>Accidents Q4</td>
<td>.10 (.09)</td>
</tr>
<tr>
<td><strong>Average</strong></td>
<td></td>
<td><strong>.12 (.11)</strong></td>
</tr>
<tr>
<td>DBQ-V Q1–4</td>
<td>Accidents Q1</td>
<td>.14 (.15)</td>
</tr>
<tr>
<td>DBQ-V Q1–4</td>
<td>Accidents Q2</td>
<td>.15 (.14)</td>
</tr>
<tr>
<td>DBQ-V Q1–4</td>
<td>Accidents Q3</td>
<td>.11 (.11)</td>
</tr>
<tr>
<td>DBQ-V Q1–4</td>
<td>Accidents Q4</td>
<td>.12 (.11)</td>
</tr>
<tr>
<td><strong>Average</strong></td>
<td></td>
<td><strong>.13 (.13)</strong></td>
</tr>
</tbody>
</table>

*Note.* These results are based on the Transport Research Laboratory (TRL) data archive (2008). Only subjects without missing DBQ data and without missing accident data were included in the sample.
Final response to the DBQ meta-analysis discussion

Anders af Wåhlberg, Lisa Dorn & James Freeman

It is with great interest that we have read the counter-arguments forwarded by de Winter and Dodou in this exchange about the validity of the DBQ, and would like to thank them for their very thorough study of the relevant literature. However, we now feel that we have reached a point where little new information is being added to this discussion. We therefore refrain to comment upon the last points raised, which has strayed away from the main questions we posed, centering instead upon the consistency of various papers of ours. We do not claim to be free from having committed errors, or to being totally consistent over time. Science is a winding road upon which we all travel without knowing what will turn up around the next bend.

We would also like to thank the editors of JSR for this opportunity to discuss a question that is so very central to research into individual differences in traffic safety research (as this is not just about the DBQ, but about self-reports and common method variance in general). This type of discussion would seem to be extremely rare within traffic safety, which is a pity, as we feel that our exchange with de Winter and Dodou has greatly added to the clarification of views, highlighted important data sources and their possible interpretations, and identified what kind of research is needed.
Final response to the commentaries on “The Driver Behaviour Questionnaire as a predictor of accidents: A meta-analysis”

J.C.F. de Winter & D. Dodou

Department of BioMechanical Engineering, Faculty of Mechanical, Maritime and Materials Engineering, Delft

University of Technology

Corresponding author:

Joost de Winter, Department of BioMechanical Engineering, Faculty of Mechanical, Maritime and Materials Engineering, Delft University of Technology, Mekelweg 2, 2628 CD Delft, The Netherlands

E-mail: j.c.f.dewinter@tudelft.nl

We thank af Wåhlberg, Dorn, and Freeman for their thoughtful comments on our meta-analysis. Although we disagree with their negative view of the validity of the Driver Behaviour Questionnaire (DBQ), we do agree that it is important to control for common method effects whenever possible. One precautionary measure against common method variance is to obtain the criterion variable from a different source, for example from spouses, unobtrusive observers, or objective measurement systems, such as driving simulators and in-vehicle dataloggers. Another measure, providing a lower level of evidence, is to introduce a time delay between the self-reported predictor and criterion variables.

We too feel that this type of discussion is rare within the traffic safety research community, and we are grateful to have had the opportunity to go into some depth about the
validity threats of the DBQ. These commentaries and replies could serve as general literature citations when discussing the validity of self-reports within traffic safety research.