Accident Analysis and Prevention 122 (2019) 207–214

Contents lists available at ScienceDirect

Accident Analysis and Prevention

journal homepage: www.elsevier.com/locate/aap

Computer-based hazard perception test scores are associated with the frequency of heavy braking in everyday driving

Andrew Hill*, Mark S. Horswill, John Whiting, Marcus O. Watson

School of Psychology, The University of Queensland, St Lucia, Brisbane QLD 4072, Australia

ARTICLE INFO

Keywords:
Hazard perception
Situation awareness
Instrumented vehicle
Driving
Event-triggered cameras
Dashcams

ABSTRACT

Computer-based hazard perception tests are used in a number of countries as part of the driver licensing processes, and hence evaluating the validity of such tests is crucial. One strategy for assessing the validity of the scores generated by a hazard perception test is to determine whether they can predict on-road driving performance. Only a few prior studies have attempted this, all relying on the subjective ratings of an examiner who was present during a single brief drive and was not blind to the driver's demographic characteristics, potentially contaminating the outcomes. Additionally, only one such study focused on the most relevant participant group with respect to the validity of tests used in licencing processes, namely young drivers. We sought to remedy this situation in the present project by measuring young drivers' performance over an extended period of everyday driving via g-force triggered video cameras ("dashcams") installed in their own vehicles. As a precursor to the dashcam study itself, we developed a new computerized hazard perception test and assessed the validity of its scores by more traditional means (Study 1). As expected, test scores distinguished between high-risk and lower-risk driver groups, and correlated with scores on an established hazard perception test previously shown to predict crash risk. In the subsequent dashcam study (Study 2), the frequency of heavy-braking events (controlling for distance driven) was used as a more objective measure of driving performance. Results indicated that drivers with higher rates of heavy braking had slower hazard perception response times, further supporting the use of these scores as a valid measure of drivers' ability to exercise hazard perception skill during real driving. More generally, this study also demonstrates the viability of using low-cost off-the-shelf dashcams to measure real-world driving behaviour.

1. Introduction

Computer-based hazard perception tests are currently being used in a number of countries as part of the driver licensing processes (Horswill, 2016a). For example, under the graduated licensing scheme operating in the state of Queensland, Australia, provisional licence holders need to pass a hazard perception test before they can proceed to the second, less restrictive stage of their provisional licence (Horswill et al., 2015; Wetton et al., 2011). These tests typically involve drivers being shown video clips of real traffic scenes, filmed from the driver's point of view (Horswill, 2016b). In the Queensland test, drivers are required to respond to each clip as soon as they anticipate a likely "traffic conflict", which is defined as a situation in which the camera car is on course to hit another road user if no evasive action is taken. The participant uses their computer mouse to click on the other road user involved in the traffic conflict, and their response time is measured from the earliest possible moment at which the conflict could plausibly have been predicted (i.e., the point at which the first predictive cue appears on screen). The overall test score is essentially an average of the participant's response times to a series of these traffic conflicts (with a standardisation process applied to ensure that each test item has an equal influence on the overall score). The purpose of such tests is to identify drivers who have poor hazard perception skill – and are therefore likely to have an elevated crash risk (Horswill et al., 2015) – so that their progress towards an unrestricted license can be delayed until they have achieved a mandated minimum standard in this skill (Pradhan and Crundall, 2017).

In order to justify the use of computer-based hazard perception tests as part of driver licensing processes, we need to be confident of their validity. That is, we need evidence indicating that scores on these tests

* Corresponding author.
E-mail address: a.hill@psy.uq.edu.au (A. Hill).

https://doi.org/10.1016/j.aap.2018.08.030
Received 10 March 2018; Received in revised form 25 August 2018; Accepted 31 August 2018
Available online 31 October 2018

actually reflect drivers' ability to exercise hazard perception skill during real driving.[1] One of the main strategies for gathering validity evidence has been to investigate the relationship between test scores and self-reported crash involvement (Horswill, 2016a). For instance, in one study, provisional drivers who predicted traffic conflicts earlier in the hazard perception test used for licencing purposes in Queensland had fewer self-reported crashes in both the year following the test and the years leading up to the test (Horswill et al., 2015). Similar crash relationships have also been found for other hazard perception tests (Boufous et al., 2011; Cheng et al., 2011; Darby et al., 2009; Horswill et al., 2010; McKenna and Horswill, 1999; Rosenbloom et al., 2011; Wells et al., 2008). Alternative strategies commonly used in validating hazard perception tests include determining whether test scores can differentiate between high-risk driver groups (e.g. young novices) and lower-risk driver groups (e.g. mid-age experienced drivers), and whether they correlate with scores on established hazard perception tests for which validity evidence has already been demonstrated (Wetton et al., 2011, 2010).

In the present paper, we will focus on a less frequently-used validation strategy for hazard perception tests, which is to investigate whether they can predict direct measures of on-road driving performance. We are aware of only three prior studies in which relationships between hazard perception test scores and on-road performance have been reported. These studies all involved participants completing a single drive (approx. 30–45 minutes) on public roads, adhering to a predetermined route. In all three studies, participants were accompanied by at least one examiner who rated their performance. Two of the studies involved older drivers (Ross et al., 2013; Wood et al., 2013) and the third involved younger drivers (Mills et al., 1998). All three studies found significant relationships between on-road performance ratings and hazard perception test scores.

The present paper aims to address four principal concerns with the three on-road studies described above. First, for the purposes of driver licensing, it is necessary to demonstrate test validity for novice drivers rather than older drivers. However, the single study in the literature that involved novices (Mills et al., 1998) was not peer-reviewed and did not report the magnitude of the relationship between on-road ratings and test scores (only that it was significant).

Second, in studies where an examiner rides in the vehicle with the participant, it is possible that the examiner's ratings will be influenced by assumptions made on the basis of the driver's appearance (Groeger, 2000). For example, if a driver appears particularly old and frail, the examiner might be biased towards presuming that their driving safety is likely to be compromised. When this same driver completes a computer-based hazard perception test, it is also possible that their performance could be suppressed relative to younger drivers with an equivalent level of hazard perception skill simply because they are less familiar with computers. In principle, this situation could therefore lead to a spurious correlation between on-road ratings and hazard perception test scores that did not result from the older, frailer drivers actually having poorer hazard perception skill and lower driving safety.

A third issue regarding typical on-road studies is that drivers may not drive as they usually do when they are acutely aware that their performance is being evaluated (Carsten et al., 2013). For instance, drivers are likely to vary in the extent to which they are susceptible to test anxiety. For those at the higher end of the spectrum, this anxiety might lead to substantially worse performance during both a one-off on-road drive in the presence of an expert evaluator and a computerized hazard perception test, thus creating an apparent correlation. However, it could be that, when drivers are not performing under test conditions, this correlation does not exist. That is, the relationship could be an artefact of the testing situation rather than a true reflection of whether the hazard perception test can predict drivers' everyday driving safety.

A fourth issue is that a single 30–45 minute drive is unlikely to yield a sufficient number of hazards to adequately inform examiners' judgements (Horswill, 2016b). For instance, when filming the video clips for the hazard perception test used in the present research, we estimate that it took about three hours of driving to encounter one hazard suitable for inclusion, even though the driver of the camera car deliberately avoided routes where traffic conflicts were less likely to occur. Hence, a 45 min drive may only have around a 25% chance of yielding a single encounter with such a hazard, assuming similar driving conditions.

An alternative approach to measuring on-road driving performance is to use instrumented vehicles. While this has traditionally been an expensive and elaborate endeavour involving specialist equipment (e.g. Klauer et al., 2006), it has recently become more affordable due to the increasing availability of low-cost consumer-orientated g-force activated in-car cameras, or "dashcams". These cameras, which can be mounted behind the windscreen, record video footage whenever the vehicle's electronics are turned on, but only save it to storage media if the g-forces acting on the camera exceed certain thresholds (e.g. due to the car suddenly decelerating because of a crash or heavy-braking event). In effect, the camera compiles a series of video clips documenting high g-force events, with each clip showing both the event that triggered the camera and its immediate aftermath. Once innocuous events have been removed, such as instances where the car passed over a speed bump, counts of these events can potentially be used as a measure of driving performance. Such an approach addresses many of the potential problems with supervised on-road assessment methods identified above. Specifically, because drivers can be monitored during all of their everyday driving over a period of weeks or months without the need for an observer to be present, test anxiety is less likely to influence the results. In addition, the measure itself is less likely to be contaminated by observer bias, and there is a greater chance of the driver encountering and having to deal with relatively rare hazardous events when on-road data are collected over a substantially longer period of time.

Simons-Morton et al. (2012) conducted a study in which teenage drivers' vehicles were fitted with an array of instrumentation, including accelerometers and cameras, for a period of 18 months. They reported a significant relationship between the frequency of elevated g-force events and participants' involvement in crashes or near-crashes. The frequency of sudden changes in acceleration has also been found to be associated with crash risk in other studies that employed elaborately instrumented vehicles (af Wåhlberg, 2007; Bagdadi and Várhelyi, 2011; Feng et al., 2017). These findings suggest that the frequency of g-force triggered events, as recorded by modern dashcams, could also potentially be a useful measure of driving safety.

The initial phase of the present project involved developing a new computer-based hazard perception test and evaluating its validity via traditional means, namely risk-related driver-group differences and correlations with established hazard perception tests (Study 1). However, the ultimate aim of the project was to provide more direct validity evidence for the new test by examining whether there was a significant relationship between young drivers' hazard perception scores and their frequency of heavy-braking events across an extended period of real driving, as measured using g-force triggered dashcams

[1] While the traditional conceptualisation of validity employed three mutually-exclusive categories (i.e. content validity, criterion validity, and construct validity), the contemporary unitary conceptualisation of validity regards all validity evidence (including evidence traditionally associated with "content validity" or "criterion validity") as contributing to "construct validity" (Downing, 2003). Given that these terms have been used differently over time, and continue to be used inconsistently by different authors, we avoid their use in the present paper in order to eliminate the potential for unnecessary confusion. Rather, across our two studies, we present a range of evidence to support the validity of scores derived from the new hazard perception test that we describe, including details of the content development and response process, as well as evidence of a consistent internal structure and relationships to several other relevant variables (see Downing, 2003).

installed in their own vehicles (Study 2). In so doing, we also aimed to assess the general viability of using these low-cost, off-the-shelf devices to measure real-world driving behaviour.

2. Study 1: preliminary validation of a new hazard perception test using high-risk (novice) and lower-risk (experienced) driver groups

As a precursor to the proposed dashcam study, we developed a pool of new hazard perception test items incorporating recently-recorded video stimuli. This was primarily to improve face validity because the traffic clips used in our existing tests (Horswill et al., 2008; Wetton et al., 2010) were starting to appear dated, both in terms of the appearance of the traffic environment and the quality of the video images. Therefore, the first aim of Study 1 was to verify that the new clips had similar measurement properties to our established clips. Hence, we replicated the preliminary validation process employed in our prior work (Wetton et al., 2011, 2010). First, we determined whether drivers' response times to traffic conflicts in the new pool of video clips could distinguish a high-risk driver group (younger novices) from a lower-risk driver group (mid-age experienced drivers), as in previous studies (Horswill, 2016b). Second, we examined the correlation between scores on the pool of new items and an established hazard perception test. This preliminary validation process is important because not all hazard perception tests have been found to pass basic validity checks in the past, and clip content has been proposed as the most likely distinguishing factor between effective and ineffective tests (Horswill, 2016b).

The second aim of Study 1 was to allow us to select the best 30 clips (in terms of their ability to distinguish between high-risk and lower-risk driver groups) for inclusion in a hazard perception test to be used in future research, including the proposed dashcam study (Study 2).

2.1. Method

2.1.1. Participants

The novice group comprised 36 younger drivers with no more than 3 years of driving experience since passing their on-road driving test, and the experienced group comprised 17 mid-age drivers who had been driving independently for at least 10 years (see Table 1 for participant characteristics). Each participant received course credit or AUD10 compensation for taking part.

2.1.2. Materials

2.1.2.1. Hazard perception test. We began by filming an extensive library of new high definition (1920 × 1080 pixels, 25 frames per second) traffic footage clips using a GoPro Hero 3 digital recording device mounted behind the windscreen of a medium-sized car and driven around locations in South East Queensland, Australia. The driver was an experienced hazard perception researcher, who flagged potentially useful clips for review by two other experienced hazard perception researchers. Of these, 57 clips were ultimately selected for inclusion in the new item-pool used in the present study.

The primary criteria for including a clip were that: (a) it contained a single unambiguous traffic conflict involving the camera car (i.e. the driver of the camera car had to slow down or change course to avoid a collision); and (b) the conflict was preceded by sufficient anticipatory cues of which drivers with good hazard perception skill could potentially take advantage (see Wetton et al., 2011, for further considerations). For example, in one clip, the camera car is travelling along a relatively narrow residential road with cars parked intermittently along either side. Far in the distance, a delivery truck can be seen negotiating the road in the opposite direction. By projecting forward the likely trajectories of both vehicles, and taking into account the width of the road and the locations of the parked cars, a driver with good hazard perception skill can quickly predict a potential traffic conflict with the truck. Specifically, it is probable that the truck will cross paths with the camera car at a point on the road where only one vehicle can fit through, due to the cars parked on both sides. In contrast, a driver with poor hazard perception skill might be unaware that a traffic conflict is likely until it becomes obvious that the vehicles are already on a collision course. In the video clip (as in all of the clips that were filmed especially for the test), the driver of the camera car did not take evasive action until a relatively late stage to avoid providing an extraneous cue to test participants. Descriptions of all items that were included in the final hazard perception test are available to researchers on request.

Once the clips to be included in the preliminary item-pool had been determined, each clip was subjected to further scrutiny by the reviewers to determine the point at which the first predictive cue to the traffic conflict appeared on screen (as the earliest point at which the conflict could conceivably be predicted by test participants), and a consensus was reached. These data were used to determine where the temporal scoring window for valid responses to each clip would begin.

The hazard perception test administered in Study 1 included all 57 items in the new item-pool, as well as all 22 items from an established test (Horswill et al., 2010; Wetton et al., 2010). In previous work, scores on the established hazard perception test were found to be associated with crash risk (Horswill et al., 2010). The clips were presented in a different random order for each participant, with new and established clips mixed together. As in the hazard perception test used for licensing purposes in Queensland (Horswill et al., 2015; Wetton et al., 2011), the temporal scoring window associated with each clip determined when an appropriate response (i.e. a mouse click on a road user likely to be involved in a traffic conflict) would be scored by the software to yield a hazard perception response time. The software also imposed a mandatory 30 s rest break (with an onscreen countdown) after every 16 clips.

The instructions for the hazard perception test took the form of a 5-minute video, similar to the one developed by Wetton et al. (2011), which defined and illustrated the concept of a "traffic conflict" and explained how to make valid responses in the test. Specifically, it instructed participants to use their computer mouse to click on any road user that was likely to become involved in a traffic conflict with the camera car (i.e. in situations in which they predicted that the driver of the camera car would have to slow down or change course to avoid a collision).

Table 1
Study 1 Participant Characteristics (N = 53).

Variable	Novice group (N = 36)	Experienced group (N = 17)
Age (years)	Mean = 19.17 SD = 1.76 Range = 17–26	Mean = 44.53 SD = 9.31 Range = 28–64
Sex	58.3% female 41.7% male	64.7% female 35.3% male
Years since passing on-road driving test	Mean = 0.94 SD = 0.74 Range = 0–2.5	Mean = 27.05 SD = 9.65 Range = 11.00–46.58
Distance driven per year (km)	Mean = 9,460.20 SD = 7,365.36 Range = 66–25,077[1]	Mean = 11,932.47 SD = 5,522.27 Range = 2,500–25,077[1]
Licence type	22% Learner Licence 78% Provisional Licence (allowed to drive unaccompanied with restrictions)	100% Open Licence (unrestricted)

[1] Both groups had the same irregular maximum value due to the slider response used for this survey item (however, participants did have the option of manually entering a higher number).

2.1.2.2. Mouse familiarization task and simple spatial reaction time test. The computer mouse familiarization task involved clicking on numbers (1–10) displayed around the screen. The *simple spatial reaction time test* (Poulsen et al., 2010) involved clicking on high contrast rectangles that appeared at random screen locations. This task was designed to control for individual differences in the hazard perception test response mode that could potentially be independent of drivers' actual hazard perception skill (i.e. their ability to manipulate a computer mouse and their reaction time).

2.1.2.3. Questionnaire. A brief questionnaire was used to collect the basic demographic information presented in the *Participants* section (see 2.1.1 and Table 1).

2.1.3. Procedure

Participants were tested (between 1 and 4 at a time) in a quiet research laboratory setting on a university campus. After consenting to take part in the study, each participant was seated at a PC equipped with a 1920 × 1080 monitor, and asked to complete a series of computerized tasks. First, they completed the computer mouse familiarization task, followed by the simple spatial reaction time test (see 2.1.2.2). Next, they watched the hazard perception test instruction video before completing the test itself (see 2.1.2.1). Finally, they completed the demographic questionnaire (see 2.1.2.3).

2.2. Results

For the initial analysis, the 57 new items were treated as a single hazard perception test to demonstrate the overall validity of the item-pool (before using the response data to reduce the number of items). As in previous work (Wetton et al., 2011), each participant's overall score was calculated via a multi-stage process. First, response times to each item were converted into z-scores (using the sample mean and SD of responses to that item), and averaged for each participant (excluding any items to which the participant did not respond). This was done because the length of the response window (in seconds) varied from clip-to-clip, depending on when the first predictive cue appeared; hence, the use of z-scores prevented clips with longer response windows from exerting an undue influence on the average. Next, each participant's mean z-score was re-standardized against the sample of average z-scores to yield an overall z-score for that participant (because an average of z-scores is not itself a z-score). Finally, each participant's overall z-score was converted back into an overall response time in seconds for ease of interpretation. To estimate the internal consistency of the full pool, misses were replaced with item means as a conservative strategy to allow alpha to be calculated (as per Wetton et al., 2010). Cronbach's α across all 57 items was 0.92.

Although the simple spatial reaction time test was included in the study for potential use as a covariate, participants' mean response times were found to have no significant correlation with their hazard perception test scores, $r(51) = -0.059$, $p = .675$. Hence, we concluded that inclusion of the covariate was unnecessary because individual differences in the response mode of the hazard perception test did not account for significant variance in hazard perception test scores.

On average, the experienced driver group responded to the 57 new traffic conflicts significantly faster ($M = 5.14$ s, $SD = 1.86$) than the novice driver group ($M = 6.37$ s, $SD = 1.58$), $t(51) = -2.51$, $p = .015$, $d = -0.74$. This indicates that the new item-pool can distinguish a high-risk driver group from a lower-risk group. In addition, mean scores on the 57 new items were also significantly correlated with mean scores on the established hazard perception test, $r(51) = 0.89$, $p < .001$, suggesting that the old and new items were likely to be measuring the same construct, as intended.

An "empirical criterion keying" process (Anastasi and Urbina, 1997; Cohen and Swerdlik, 1999; Hill et al., 2017) was used to select the best items for inclusion in the final version of the new hazard perception test for use in future research, including Study 2. Specifically, the 30 new clips that produced the largest novice/experienced differences were selected by conducting a separate between-groups *t*-test on participants' scores for each item, and selecting the items that yielded the largest *t*-values. The internal consistency reliability of this shortened test was 0.92 (replacing misses with item means as above). Overall scores were calculated for each participant in the same way as for the 57 item test. The novice/experienced difference remained significant for the new test, $t(51) = -4.09$, $p < .001$, $d = -1.20$ (novice drivers: $M = 7.19$, $SD = 1.75$; experienced drivers: $M = 4.94$, $SD = 2.14$), as did the correlation with the established test, $r(51) = 0.83$, $p < .001$.

2.3. Discussion

A new pool of hazard perception test items was created in accordance with published guidelines designed to ensure content validity (see Wetton et al., 2011). We found that the mean hazard perception response times from the new item-pool could discriminate between high-risk and lower-risk driver groups, yielding a large between-groups effect in which experienced drivers significantly out-performed novices, as expected. Hazard perception response times also correlated with scores on an established hazard perception test previously shown to predict crash risk (Horswill et al., 2010). These results support the validity of the new item-pool. Subsequently, we created a new hazard perception test for use in Study 2 by selecting the 30 individual items from the new item-pool that yielded the largest driver-group differences. Reanalysis of the data for these 30 items yielded a comparable pattern of results, consistent with the shortened test also being a valid measure of hazard perception skill. In addition, both the full pool of new items and the shortened test had excellent internal consistency.

3. Study 2: dashcam validation of the hazard perception test

The main aim of Study 2 was to investigate to what extent scores from the computer-based hazard perception test developed and evaluated in Study 1 would map on to the frequency of young drivers' heavy braking during real driving, controlling for distance travelled. Individuals with superior hazard perception skill are, by definition, better at anticipating potential traffic conflicts that may require them to take evasive action further down the road (Pradhan and Crundall, 2017; Wetton et al., 2011). When faced with these situations during real driving, this advantage should give these individuals more time to slow down or change course without resorting to heavy braking. Hence, we predicted that shorter hazard perception test response times would predict less frequent heavy-braking events. More broadly, this study was designed to provide insight into the ability of computerized driving tests to predict actual driving behaviour. We focussed on drivers aged under 25, given the high crash-risk associated with this age group (Elvik, 2010), the importance of ensuring that hazard perception tests used in licensing processes actually reflect young drivers' ability to exercise hazard perception during real driving, and the possible implications of the findings for future evaluations of training interventions designed to yield potentially life-saving improvements in the hazard perception skill of young drivers.

3.1. Method

3.1.1. Participants

We recruited 135 Australian driver's licence holders aged between 18 and 24 years using a psychology first-year participant pool, and via social media. To be eligible to participate, motorists had to be the sole driver of their car (though we made allowances for unexpected use of their vehicles by other drivers through the use of driving diaries). The present analyses included a final sample of 97 drivers after exclusions due to: (a) a broken dashcam ($n = 1$); (b) an incorrectly aligned dashcam (which had been knocked such that it faced the sky, rendering

Table 2
Study 2 Participant Characteristics (N = 97).

Variable	Descriptive statistics
Age (years)	Mean = 19.53 SD = 1.70 Range = 18–24
Sex	74.2% female 25.8% male
Years since passing on-road driving test	Mean = 2.45 SD = 1.74 Range = 0–8
Distance driven per year (km)	Mean = 13,426.36 SD = 7,300.33 Range = 2,418–45,000
Licence type	68% Provisional Licence (allowed to drive unaccompanied with restrictions) 32% Open Licence (unrestricted)
Vehicle type	61.9% hatchback 26.8% sedan 8.2% Sports Utility Vehicle 3.1% convertible

g-force measurements invalid, $n = 1$); (c) missing odometer data ($n = 1$); and (d) loss of hazard perception test data due to a computer hard drive failure ($n = 35$). Characteristics of the final sample participants are shown in Table 2.

3.1.2. Materials

3.1.2.1. G-force triggered dashcams. A small in-car video camera was mounted on the windscreen of each participant's usual vehicle near the rear-view mirror, and was connected to the vehicle's internal power socket. Two dashcam models with identical g-force sensors were used, the Lukas Pro LK-7700 and Ace LK-7900 (Qrontech Co. Ltd., Seoul, South Korea). The SD memory card inserted into each dashcam as video storage media was sealed in place with tamper-proof tape to alert the researchers to any attempted removal during the testing period.

Each dashcam continuously cached a recording of the forward view from the car when the vehicle's electronics were switched on. Whenever the g-forces registered by the camera exceeded a set threshold in any of three dimensions (longtitudinal 0.50 g; lateral 0.55 g; vertical 1.50 g), the camera saved the preceding 10 s and the following 20 s of footage to the memory card as a continuous 30-second video clip (hence capturing the event that caused the g-force trigger and its immediate aftermath). The g-force thresholds matched those used by McGehee et al. (2007), which were chosen to provide a balance between capturing as many safety-relevant events as possible while minimizing irrelevant events that could trigger the camera (e.g. passing over speed bumps). Footage was captured in high definition (1920 × 1080 pixels, 30 frames per second). No sound was recorded.

3.1.2.2. Hazard perception test and simple spatial reaction time test. Study 2 employed the 30-item hazard perception test developed in Study 1 (see 2.1.2.1 and 2.2), as well as the *simple spatial reaction time test* described there (see 2.1.2.2).

3.1.2.3. Questionnaire. A brief questionnaire was again used to collect basic demographic information, as presented in the *Participants* section (see 3.1.1 and Table 2).

3.1.2.4. Driving diary. This document allowed participants to record any instances in which someone else drove their vehicle while the camera was installed, including odometer readings before and after each instance. This was to facilitate the exclusion of any heavy-braking incidents that occurred while the participant was not driving, and the correction of our estimate of their total distance driven while the camera was installed.

3.1.3. Procedure

Study 2 participants attended two sessions at a university campus. In Session 1, the researcher obtained the participant's consent to take part in the study, installed a dashcam in the participant's car, recorded the vehicle's current odometer reading, administered the questionnaire (see 3.1.2.3), and issued the participant with a driving diary (see 3.1.2.4).

Session 2 took place 4.71–11.00 weeks ($M = 6.46$; $SD = 0.88$) after Session 1 (with most of the variability in the delay time attributable to the participants' availability). In this session, the researcher removed the dashcam from the participant's car, again recorded the vehicle's current odometer reading, and retrieved the completed driving diary. The researcher also presented the participant with a legal form allowing them to formally donate their dashcam footage to the university. Participants had been assured at the beginning of the study that the video recordings would remain their own property unless they subsequently chose to sign this form. This procedure was put in place to protect participants by allowing them to refuse to donate recordings if they believed that the camera may have captured potentially incriminating material. However, in practice, no participant refused to donate their clips.

The participant also visited a research laboratory on the same day that their camera was removed (usually directly afterwards). In the laboratory, they completed the simple spatial reaction time test and the hazard perception test (see 3.1.2.2) under the same conditions as Study 1 participants. Note that the hazard perception test was completed *after* the on-road measurement period in case completing the test influenced participants' subsequent on-road driving behaviour. Participants also completed additional measures that are not relevant to the present study. In these laboratory sessions, 1–4 participants were tested concurrently.

3.1.4. Dashcam video coding

Drive diaries were used to exclude dashcam video clips for which the participant was not the driver. Each remaining clip was independently coded by two researchers to determine what had caused one or more of the pre-determined g-force thresholds (see 3.1.2.1.) to be exceeded in that instance (causing the dashcam to automatically store a video clip of the incident). Specifically, the coders examined each clip and made a binary judgement as to whether it appeared to represent (a) a heavy-breaking event or (b) some other type of event that could trigger the dashcam (such as travelling over speed bumps, a pothole, or a driveway). To aid in making these judgements, the coders used special software (Lukas LK-7700 Viewer for PC) to view the video footage in combination with synchronised g-force graphs (see Fig. 1). Finally, the total number of heavy-braking events identified by each coder was calculated for each participant. Inter-rater reliability between the two coders' judgements was very high, *Spearman's Rho* $(95) = 0.99$, $p < .001$.

3.2. Results

Consistent with previous work on both elevated g-force events (Simons-Morton et al., 2012) and crash frequency counts (Wells et al., 2008; Horswill et al., 2015), negative binomial regression was used to determine the relationship with hazard perception test scores. (Note that the other option for count data, a Poisson distribution, was inappropriate as the variance exceeded the mean.) This analysis was conducted using a Generalized Linear Model (IBM SPSS version 25), with heavy-braking counts as the dependent variable and the hazard perception test score as the predictor. (Note that, for the purposes of this analysis, the two raters' counts were aggregated by summing them rather than averaging because the negative binomial distribution is specific to integer count data.) In addition, the total number of kilometres driven with the dashcam installed was included as an offset variable in the model to control for exposure. (This was calculated by

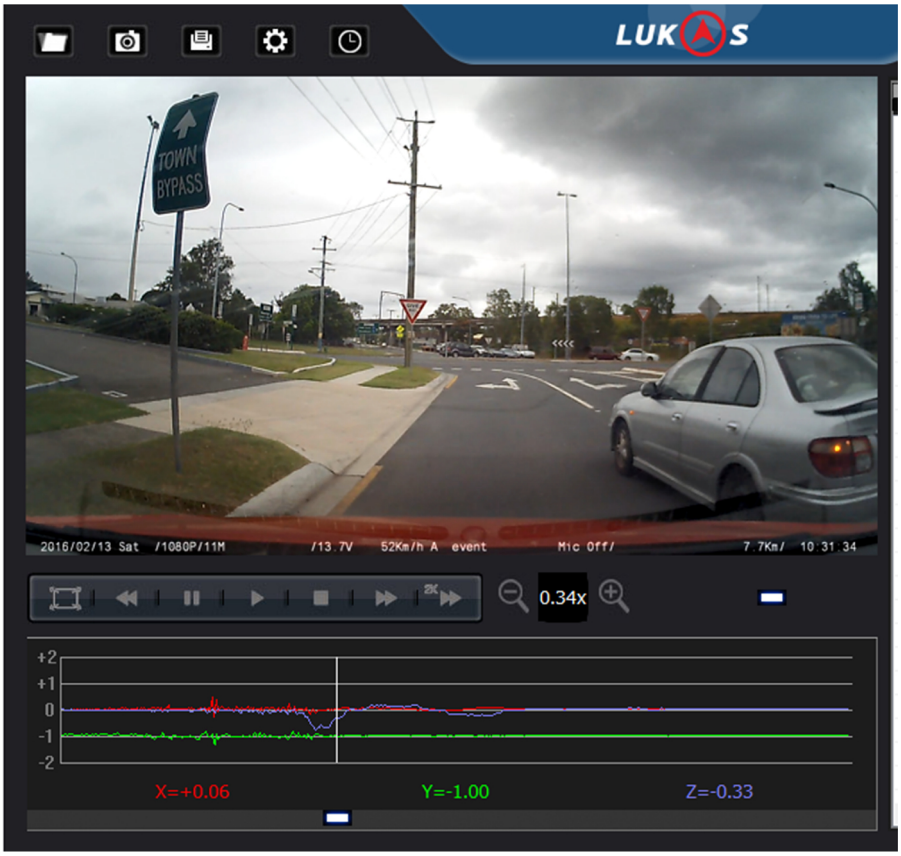

Fig. 1. Dashcam clip featuring a heavy-braking event, with g-force data for the lateral (x), vertical (y) and longitudinal (z) dimensions graphed beneath the video image (screenshot from Lukas LK-7700 Viewer for PC).

Table 3
Study 2 Descriptive Statistics (N = 97).

Variable	Mean	SD	Range
Count of heavy-braking events (averaged across raters)	9.62	14.46	0–86.5
Distance driven with dashcam installed (km)	1,618.05	1,095.32	355–7,884
Hazard perception test score (response time in seconds)	5.86	2.24	1.16–12.31
Simple spatial reaction time (seconds)	0.90	0.15	0.63–1.43

deducting the participant's odometer reading at Session 1 from their reading at Session 2, and adjusting for any use of the vehicle by another driver as disclosed in the driving diary.) The number of kilometres driven was entered as a natural logarithm as per the requirements for entering offset variables in this procedure (IBM Knowledge Center, 2012). Table 3 presents descriptive statistics for the heavy-braking counts, distance driven data, and hazard perception test scores.

Results indicated that hazard perception test scores significantly predicted the frequency of heavy-braking events (B = 0.15, *SE* B = 0.04, *Wald Chi-Square* = 11.08, p = .001). The odds ratio of this effect was 1.16 (95% CI = 1.06–1.26). This indicates that, for every 1 s that participants were slower in predicting traffic conflicts, they experienced a 16% increase in the number of heavy-braking events.

3.2.1. Additional analyses

To assess the robustness of the findings, the analysis was re-run separately on each individual coder's count data, yielding essentially the same result in each instance. Given that each coder tested a subset of the participants, there was a slim possibility that they may have implicitly remembered some risk-relevant characteristic of these drivers (e.g. gender) when viewing their videos, potentially biasing the coding. To check this, the analysis of each individual coder's counts was re-run, excluding all participants that the relevant coder had tested. This also had no meaningful effect on the outcome. In addition, a square-root transformation was performed on the hazard perception test data to reduce skew but, again, this did not affect the outcome (consequently, the untransformed data were retained for the main analysis reported above for ease of interpretation). Finally, simple spatial reaction time (see Table 3) was included as a covariate (log transformed to reduce skew), but this had no meaningful effect on the relationship between hazard perception test scores and heavy-braking counts. The p-values for these additional analyses ranged from < .001 to .002.

3.3. Discussion

As predicted, the results of Study 2 indicated that there was a significant relationship between the frequency of young drivers' heavy-braking events, as measured over a number of weeks using a g-force triggered dashcam, and their response times in a video-based computerized hazard perception test. This supports the proposal that scores on this test are a valid measure of actual driving hazard perception skill and that they are reflective of young drivers' real-world performance and safety, especially given the established link between elevated g-force events and crashes/near-crashes (Simons-Morton et al., 2012).

4. General discussion

The present research provides validity evidence for a computer-based hazard perception test among a sample of young drivers by measuring the relationship between test scores and a novel measure of on-road performance. Specifically, we measured the frequency of participants' heavy-braking events during real driving (controlling for

distance travelled) by coding videos captured using low-cost off-the-shelf consumer-level dashcams. As predicted, drivers with superior hazard perception skill (who, by definition, are better at predicting potential traffic conflicts that may require evasive action, such as braking or changing course) engaged in less-frequent heavy braking. We also obtained validity evidence for the test via two of the more traditional validation techniques, namely demonstrating that the scores can distinguish between driver groups known to differ in crash risk (Crundall, 2016), and that they correlate with scores on an established hazard perception test previously shown to predict crash risk (Horswill et al., 2010). Given the strong similarities between our new test and some of the hazard perception tests currently used in driver licensing (e.g. Horswill et al., 2015; Wetton et al., 2011), the present data provide further justification for the continued inclusion of this type of test in graduated licensing systems.

One advantage of using g-force triggered dashcams as a means of assessing the safety (or otherwise) of participants' driving is that the resulting measure directly reflects how participants behaved during the entirety of their everyday driving over an extended period of time (unless they also drove another vehicle), potentially yielding a relatively stable estimate of performance. In contrast, subjective ratings of safety made by examiners during a single brief supervised drive are potentially vulnerable to observer biases and transient aberrations in performance, such as those that may result from acute test anxiety. The dashcam method described in this paper also obviates the need to rely on self-reports of crash involvement, the accuracy of which has been questioned by some researchers (Chapman and Underwood, 1997). Arguably, it is also likely to be a much more sensitive method of measuring individuals' current on-road driving performance, since drivers typically experience actual crashes only around once per decade and crash risk is known to change with age (Evans, 1991).

One potential limitation of the dashcam method used in this paper is that participants may have maintained an awareness that they were being monitored while driving, which could potentially have affected the way that they drove. However, to some extent, reactivity potentially affects nearly all measures of driving behaviour and it is arguably likely to be less of an issue for the dashcam method used in the present paper because of the length of time over which the measurements were taken. That is, if drivers chose to drive more safely because of the presence of the camera, then they would have had to maintain this behaviour every time they drove over a period of many weeks.

An additional limitation of the present research is that we did not control for exposure variables apart from distance driven. Other relevant variables might include, for example, the type of roads being driven on and the average trip length, both of which could plausibly affect heavy-braking frequency. Similarly, it is likely that individual differences in risk-taking propensity may also have accounted for some of the variation in the heavy-braking data. In the present research, variables of this nature were assumed to contribute random noise to the data, but they could potentially be quantified in future work. Regarding the hazard perception test itself, one limitation that should be noted (which it shares with many other hazard perception tests used for research or licensing purposes) is that its reliance on forward-view road footage means that the test did not include traffic conflicts for which the anticipatory cues were available only via mirrors and/or shoulder checks.

Aside from providing validation evidence for the new hazard perception test, this research project served to demonstrate the general viability of using consumer-grade dashcams to measure driving behaviour. In particular, the chosen methodology ensured that the data collection was both ethical (i.e. the video donation procedure minimized the potential legal risks to participants associated with the presence of an in-car camera) and relatively low in cost (in contrast to the use of traditional instrumented vehicle rigs). Using this method, it is possible to evaluate a large number of drivers simultaneously, with the low unit cost of the devices making the technique a potentially viable option for large-scale behavioural studies (e.g. for the purposes of the present project, 50 dashcams were available, most of which were used sequentially by multiple participants).

Finally, it is worth noting that the dashcam measurement technique described in the present paper may also be useful for other types of research that require the quantification of driving behaviour. To give one example, it could be used for evaluating the impact of driver behaviour change interventions on actual driving. Intervention evaluations typically rely on self-report measures or performance in driving simulators, neither of which is likely to be an entirely accurate reflection of how participants behave during real driving. In contrast, if an intervention effect can be demonstrated using the dashcam technique over a sufficient period of time, then this could be argued to be more compelling evidence for the effectiveness of the intervention.

Declarations of interest

None.

Funding

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Acknowledgements

We would like to thank Mark Wetton for filming the video stimuli for the hazard perception test and Francine Smith, Genevieve Kieseker, and Estrella Paterson for testing some of the participants. We thank Paul Jackson for developing the hazard perception test software. Francine Smith and Genevieve Kieseker conducted the video coding.

References

af Wåhlberg, A.E., 2007. Aggregation of driver celeration behavior data: effects on stability and accident prediction. Saf. Sci. 45 (4), 487–500.
Anastasi, A., Urbina, S., 1997. Psychological Testing, 7th edition. Prentice-Hall, Upper Saddle River.
Bagdadi, O., Várhelyi, A., 2011. Jerky driving—an indicator of accident proneness? Accid. Anal. Prev. 43 (4), 1359–1363.
Boufous, S., Ivers, R., Senserrick, T., Stevenson, M., 2011. Attempts at the practical on-road driving test and the hazard perception test and the risk of traffic crashes in young drivers. Traffic Inj. Prev. 12 (5), 475–482. https://doi.org/10.1080/15389588.2011.591856.
Carsten, O., Kircher, K., Jamson, S., 2013. Vehicle-based studies of driving in the real world: the hard truth? Accid. Anal. Prev. 58, 162–174.
Chapman, P., Underwood, G., 1997. Reporting and forgetting accidents and other driving events. In: Grayson, G.B. (Ed.), Behavioural Research in Road Safety VII. Transport Research Laboratory, Crowthorne, UK.
Cheng, A.S.K., Ng, T.C.K., Lee, H.C., 2011. A comparison of the hazard perception ability of accident-involved and accident-free motorcycle riders. Accid. Anal. Prev. 43 (4), 1464–1471. https://doi.org/10.1016/j.aap.2011.02.024.
Cohen, R.J., Swerdlik, M.E., 1999. Psychological Testing and Assessment: An Introduction to Tests and Measurement. Mayfield Publishing Company, Mountain View.
Crundall, D., 2016. Hazard prediction discriminates between novice and experienced drivers. Accid. Anal. Prev. 86, 47–58. https://doi.org/10.1016/j.aap.2015.10.006.
Darby, P., Murray, W., Raeside, R., 2009. Applying online fleet driver assessment to help identify, target and reduce occupational road safety risks. Saf. Sci. 47 (3), 436–442. https://doi.org/10.1016/j.ssci.2008.05.004.
Downing, S.M., 2003. Validity: on the meaningful interpretation of assessment data. Med. Educ. 37, 830–837.
Elvik, R., 2010. Why some road safety problems are more difficult to solve than others. Accid. Anal. Prev. 42 (4), 1089–1096. https://doi.org/10.1016/j.aap.2009.12.020.
Evans, L., 1991. Traffic Safety and the Driver. Van Nostrand Reinhold, New York.
Feng, F., Bao, S., Sayer, J.R., Flannagan, C., Manser, M., Wunderlich, R., 2017. Can vehicle longitudinal jerk be used to identify aggressive drivers? An examination using naturalistic driving data. Accid. Anal. Prev. 104, 125–136. https://doi.org/10.1016/j.aap.2017.04.012.
Groeger, J.A., 2000. Understanding Driving. Psychology Press, Hove.
Hill, A., Horswill, M.S., Plooy, A.M., Watson, M.O., Rowlands, L.N., Wallis, G.M., Riek, S., Burgess-Limerick, R., Hewett, D.G., 2017. Assessment of polyp recognition skill: development and validation of an objective test. Surg. Endosc. 31, 2426–2436.
Horswill, M.S., 2016a. Hazard perception in driving. Curr. Dir. Psychol. Sci. 25 (6), 425–430.
Horswill, M.S., 2016b. Hazard perception tests. In: Fisher, D.L., Caird, J.K., Horrey, W.,

Trick, L. (Eds.), The Handbook of Teen and Novice Drivers. CRC Press, Boca Raton, FL, pp. 439–450.
Horswill, M.S., Marrington, S.A., McCullough, C.M., Wood, J., Pachana, N.A., McWilliam, J., Raikos, M.K., 2008. The hazard perception ability of older drivers. J. Gerontol. B Psychol. Sci. Soc. Sci. 63 (4), 212–218. https://doi.org/10.1093/geronb/63.4.P212.
Horswill, M.S., Anstey, K.J., Hatherly, C.G., Wood, J., 2010. The crash involvement of older drivers is associated with their hazard perception latencies. J. Int. Neuropsychol. Soc. 16 (5), 939–944. https://doi.org/10.1017/S135561771000055X.
Horswill, M.S., Hill, A., Wetton, M., 2015. Can a video-based hazard perception test used for driver licensing predict crash involvement? Accid. Anal. Prev. 82, 213–219. https://doi.org/10.1016/j.aap.2015.05.019.
IBM Knowledge Center, 2012. Weight and Offset (Generalized Linear Model). Retrieved from https://www.ibm.com/support/knowledgecenter/en/SS3RA7_15.0.0/com.ibm.spss.modeler.help/idh_glmm_weight_offset.htm. .
Klauer, S.G., Dingus, T.A., Neale, V.L., Sudweeks, J.D., Ramsey, D.J., 2006. The Impact of Driver Inattention on Near-Crash/Crash Risk: An Analysis Using the 100-Car Naturalistic Driving Study Data (DOT HS 810 594). Retrieved from https://www.nhtsa.gov/sites/nhtsa.dot.gov/files/810594.pdf. .
McGehee, D.V., Raby, M., Carney, C., Lee, J.D., Reyes, M.L., 2007. Extending parental mentoring using an event-triggered video intervention in rural teen drivers. J. Safety Res. 38, 215–227. https://doi.org/10.1016/j.jsr.2007.02.009.
McKenna, F.P., Horswill, M.S., 1999. Hazard perception and its relevance for driver licensing. J. Int. Assoc. Traffic Saf. Sci. 23 (1), 26–41.
Mills, K.L., Hall, R.D., McDonald, M., Rolls, G.W.P., 1998. The Effects of Hazard Perception Training on the Development of Novice Driver Skills. Department for Transport (UK), London.
Poulsen, A.A., Horswill, M.S., Wetton, M.A., Hill, A., Lim, S.M., 2010. A brief office-based hazard perception intervention for drivers with ADHD symptoms. Aust. N. Z. J. Psychiatry 44 (6), 528–534. https://doi.org/10.3109/00048671003596048.
Pradhan, A.K., Crundall, D., 2017. Hazard avoidance in young novice drivers: definitions and a framework. Handbook of Teen and Novice Drivers. CRC Press, pp. 61–74.
Rosenbloom, T., Perlman, A., Pereg, A., 2011. Hazard perception of motorcyclists and car drivers. Accid. Anal. Prev. 43 (3), 601–604.
Ross, R.W., Scialfa, C., Cordazzo, S., Bubric, K., 2013. Predicting older adults' on-road driving performance. Paper Presented at the 7th International Driving Symposium on Human Factors in Driver Assessment, Training, and Vehicle Design.
Simons-Morton, B.G., Zhang, Z., Jackson, J.C., Albert, P.S., 2012. Do elevated gravitational-force events while driving predict crashes and near crashes? Am. J. Epidemiol. 175 (10), 1075–1079.
Wells, P., Tong, S., Sexton, B., Grayson, G., Jones, E., 2008. Cohort II: A Study of Learner and New Drivers. Retrieved from http://www.dft.gov.uk/publications/cohort-ii-a-study-of-learner-and-new-drivers/. .
Wetton, M.A., Horswill, M.S., Hatherly, C., Wood, J.M., Pachana, N.A., Anstey, K.J., 2010. The development and validation of two complementary measures of drivers' hazard perception ability. Accid. Anal. Prev. 42 (4), 1232–1239. https://doi.org/10.1016/j.aap.2010.01.017.
Wetton, M.A., Hill, A., Horswill, M.S., 2011. The development and validation of a hazard perception test for use in driver licensing. Accid. Anal. Prev. 43 (5), 1759–1770. https://doi.org/10.1016/j.aap.2011.04.007.
Wood, J.M., Anstey, K.J., Horswill, M.S., Lacherez, P.F., 2013. Evaluation of screening tests for predicting older driver performance and safety assessed by an on-road test. Accid. Anal. Prev. 50, 1161–1168. https://doi.org/10.1016/j.aap.2012.09.009.