Accident Analysis and Prevention 106 (2017) 31–43

Contents lists available at ScienceDirect

Accident Analysis and Prevention

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

A meta-analysis of in-vehicle and nomadic voice-recognition system interaction and driving performance

Sarah M. Simmons[a], Jeff K. Caird[b,*], Piers Steel[c]

[a] Department of Psychology, University of Calgary, Calgary, Alberta, Canada
[b] Departments of Psychology and Community Health Sciences, University of Calgary, Calgary, Alberta, Canada
[c] Department of Human Resources & Organizational Dynamics, University of Calgary, Calgary, Alberta, Canada

ARTICLE INFO

Keywords:
Driver distraction
Driving performance
Voice-recognition
Speech-to-text
Meta-analysis

ABSTRACT

Driver distraction is a growing and pervasive issue that requires multiple solutions. Voice-recognition (V-R) systems may decrease the visual-manual (V-M) demands of a wide range of in-vehicle system and smartphone interactions. However, the degree that V-R systems integrated into vehicles or available in mobile phone applications affect driver distraction is incompletely understood. A comprehensive meta-analysis of experimental studies was conducted to address this knowledge gap. To meet study inclusion criteria, drivers had to interact with a V-R system while driving and doing everyday V-R tasks such as dialing, initiating a call, texting, emailing, destination entry or music selection. Coded dependent variables included detection, reaction time, lateral position, speed and headway. Comparisons of V-R systems with baseline driving and/or a V-M condition were also coded. Of 817 identified citations, 43 studies involving 2000 drivers and 183 effect sizes (*r*) were analyzed in the meta-analysis. Compared to baseline, driving while interacting with a V-R system is associated with increases in reaction time and lane positioning, and decreases in detection. When V-M systems were compared to V-R systems, drivers had slightly better performance with the latter system on reaction time, lane positioning and headway. Although V-R systems have some driving performance advantages over V-M systems, they have a distraction cost relative to driving without any system at all. The pattern of results indicates that V-R systems impose moderate distraction costs on driving. In addition, drivers minimally engage in compensatory performance adjustments such as reducing speed and increasing headway while using V-R systems. Implications of the results for theory, design guidelines and future research are discussed.

Driver distraction is a growing and pervasive issue with tragic consequences that requires multiple solutions. Over the past ten years, fatalities associated with distracted driving have been increasing (Wilson and Stimpson, 2010). In 2013, 10% of fatal crashes, 18% of injury crashes, and 16% of all police-reported crashes involved driver distraction in the United States (NHTSA, 2015). Evidence from naturalistic and driving performance literature indicates that a number of visual-manual and cognitive tasks increase crash risk and degrade driving performance (Caird et al., 2008, 2014; Dingus et al., 2016; Klauer et al., 2014; Simmons et al., 2016). As a countermeasure, safe driving legislation that targets handheld cell phone use has been implemented in almost all US states, all Canadian provinces and numerous countries (CCMTA, 2013; GHSA, 2015; WHO, 2015). The limited effectiveness of legislation and other interventions on the use of nomadic and integrated systems by adult drivers in general and younger drivers in particular is concerning (Caird and Horrey, 2017; Klauer et al., 2014; McCartt et al., 2014). Research shows that the dangers of distracted driving are appreciated by the general public, but awareness of the dangers associated with distracted driving does not necessarily translate into behavioral modifications among drivers (Atchley et al., 2011; Hamilton et al., 2013).

The demand for communication and connectivity behind the wheel, in conjunction with the implementation of safe driving legislation targeting handheld mobile phones, has led to the limited adoption of hands-free technologies (Pickrell and KC, 2015). Notably, hands-free technologies allow drivers to have cell phone conversations without having to physically hold a cell phone to the ear; the phone may be replaced with a headset, or the phone may be paired with a vehicle's built-in microphone and speaker system. However, hands-free systems can also support non-conversational secondary tasks, such as dialing or text messaging, by using speech-to-text (voice-recognition) technology. These systems operate by recognizing the user's spoken words and converting these into commands that can be used by the system to accomplish various tasks. In some cases, hands-free systems may rely on

[*] Corresponding author.
E-mail addresses: sarah.simmons@ucalgary.ca (S.M. Simmons), jkcaird@ucalgary.ca (J.K. Caird), piers.steel@haskayne.ucalgary.ca (P. Steel).

http://dx.doi.org/10.1016/j.aap.2017.05.013
Received 17 November 2016; Received in revised form 27 April 2017; Accepted 17 May 2017
Available online 29 May 2017

voice-recognition technology to initiate a task, such as dialing a phone number or searching for an address. In other cases, the entire task may rely on voice-recognition, such as the dictation of a text message to be sent to a friend. In most cases, the manual manipulation of a keyboard or input device is minimized (some systems require the push of a button or a touchscreen press to start the system), and the user's eyes are available, in theory, to focus on driving. Additionally, hands-free systems are also more likely to comply with legislation targeting handheld cell phones.

Some hands-free systems that use voice-recognition (V-R) are integrated into the vehicle to control not only entertainment and climate, but also hands-free mobile phone use. Most vehicle manufacturers offer integrated in-vehicle communication and entertainment systems that support indirect, hands-free interactions with a mobile phone (Mehler et al., 2015c; Reimer et al., 2016; Strayer et al., 2015a, 2015b). In addition, smartphones carried by a driver into a vehicle are often equipped with personal assistant software capable of voice-recognition, opening up the potential for these devices to function as nomadic in-vehicle communication and entertainment systems in place of more expensive integrated systems. A number of large technology corporations offer V-R based personal assistants that support tasks such as initiating calls with contacts, and sending and reading text messages. However, these systems can also be used to provide navigation, find and read emails, create events and reminders, set alarms and timers, complete web searches, play selected music, take notes, perform calculations, identify songs by sound, and even report nearby cinema show times and make reservations at restaurants (Klein, 2015). In summary, V-R systems have become much more accessible through smartphone applications and vehicle integration, and the variety of non-driving tasks that a driver may engage in behind the wheel is quite extensive.

The development and release of new hands-free technologies, and the sheer variety of ways they can be used behind the wheel, offers a possible solution to visual-manual (V-M) distraction but also poses a potential cognitive distraction problem. The dangers of using a handheld mobile or smartphone that take the eyes off the road are well demonstrated within distracted driving research (Caird et al., 2014; Dingus et al., 2016; Klauer et al., 2014; Simmons et al., 2016). Although V-R systems differ in design, with many systems offering auditory feedback, visual feedback, or both, the intuitive benefit of hands-free nomadic and integrated cell phones or other in-vehicle convenience technologies is that they should allow the driver to minimize their eyes-off-road time. Consequently, V-R systems should interfere less with the visual aspects of driving than V-M systems. However, it is not entirely clear whether minimization of eyes-off-road time necessarily translates into safety benefits.

Based on reviews of research, V-R systems are not necessarily distraction-free and may exert cognitive and visual distraction costs. Barón and Green (2006) reviewed the results of 15 early studies using a vote counting system of analysis that tallied the number of studies that used different methods and measures. The summary of results concluded that, "people drove at least as well as manual systems, if not better (less lane variation, speed was steadier), when using speech interfaces than manual interfaces" (p. i), but participants drove worse with speech systems than without them. Tijerina (2016) reviewed 12 selected studies from 1998 to 2014 that included a variety of V-R systems and methodological approaches. Across reviewed studies, researchers arrived at different conclusions regarding the impact of V-R systems on distraction and safety. Some researchers found that V-R systems negatively impacted driving (Cooper et al., 2014; Strayer et al., 2015a, 2015b), whereas others concluded that V-R systems are a benefit to drivers over the use of V-M systems (Shutko et al., 2009). Still others conclude that not enough is known about the veridical effects of V-R systems on driver behavior and crash risk (Fitch et al., 2013). It was also noted that drivers may moderate the negative effects of V-R system interaction by compensating for using a system by decreasing speed and increasing headway. A previous meta-analysis found that conversations using hands-free phones increased reaction time to stimuli and events compared to baseline driving, but drivers did not engage in compensation by increasing headway and decreasing speed (Caird et al., 2008). Based on this research, V-R system interactions are likely to have a similar pattern of effects on driving performance.

Against this background of differences in V-R system results and interpretation, the purpose of this meta-analysis is to provide a clear and comprehensive understanding of how V-R systems affect driving performance compared to both baseline driving and V-M systems. The advantages of a meta-analysis include a comprehensive search for studies, quantitative precision, focused hypotheses, identification of moderators, and a higher level of results evidence (Rosenthal and DiMatteo, 2001). Based on previous research, V-R interaction was hypothesized to be associated with lower detection rates and slower reaction times relative to baseline driving, and higher detection rates and faster reaction times compared to the same task using a V-M system. It was also hypothesized that drivers would not necessarily compensate when interacting with a V-R system by increasing their headway or decreasing their speed relative to baseline driving. In addition, drivers were hypothesized to maintain lateral positioning or lane keeping while interacting with a V-R system because drivers would direct visual attention to the forward roadway while steering. V-R systems were expected to outperform visual-manual systems on lateral positioning and longitudinal control measures.

1. Method

This paper is formatted in accord with PRISMA guidelines and checklist for the reporting of meta-analyses, which includes title, structured abstract, search and coding methods, bias and meta-analytic results and discussion of evidence and limitations (Moher et al., 2009).

1.1. *Eligibility criteria*

1.1.1. *Participants*

Studies with participants of all age groups, nationalities and gender were sought for inclusion in the meta-analysis. Except for one study (Reimer et al., 2010), none of the included studies reported selecting participants with clinical conditions. The Reimer et al. (2010) study involved two groups – a group of young drivers with ADHD, and a group of controls. Only the group of controls was included in this meta-analysis. Participant characteristics such as age, sex and nationality were coded. Descriptions of participants represented in the meta-analysis appear in Table 1.

1.1.2. *Independent variables*

Tasks involving interactions with systems employing voice-recognition – where the driver states a command (or series of commands) and the system interprets and acts on these commands – were required in order for the study to meet inclusion criteria. General examples of V-R interactions of interest include dialing or initiating a call, texting or emailing, destination entry with a navigation system or selecting music. Two conditions had to be met to satisfy this criterion. First, the interaction with the V-R system, which could be real or simulated, needed to be structured so that participants provided verbal input that the system's actions were either contingent upon, or were at least assumed to be contingent upon. Tasks and/or studies where drivers merely provided vocal responses that were not assumed to be useable system inputs were excluded. Second, the V-R task was required to be representative of texting, emailing, navigation entry, or some other interaction type that might reasonably be engaged in with a V-R system during normal driving in everyday life.

Tasks that involved isolated conversation were excluded because meta-analytic reviews of cell phone conversation on driving perfor-

Table 1
Overview of included studies.

Study	Setting	*N*	Age (SD) and Sex	Voice-recognition system type	Voice-recognition interaction tasks	Performance measures
Angell et al., 2006 (Laboratory)	Simulator	50	Age: 7 in 20s, 10 in 30s, 9 in 40s, 8 in 50s, 9 in 60s, 7 in 70s; Sex: 26 M	Real V-R (Cell Phone)	Number dialing.	Detection, RT, SDLP.
Angell et al., 2006 (Interstate Highway)	On-road	101	Age: 17 in 20s, 17 in 30s, 16 in 40s, 19 in 50s, 18 in 60s, 14 in 70s; Sex: 49 M	Real V-R (Cell Phone)	Number dialing.	Detection, Headway, RT, SDLP, Speed.
Angell et al., 2006 (Test Track)	Test-track	64	Age: 11 in 20s, 11 in 30s, 12 in 40s, 11 in 50s, 11 in 60s, 8 in 70s; Sex: 31 M	Real V-R (Cell Phone)	Number dialing.	Detection, Headway, RT, SDLP, Speed.
Beckers et al., 2014	Simulator	24	Age: *M* = 25.0 (2.6); Sex: 12 M	Real V-R (Cell Phone)	Navigation input.	Detection, RT, SDLP.
Bruyas et al., 2009	Simulator	30	Age: *M* = 34 (11), 18–50; Sex: 15 M	Simulated V-R (Wizard of Oz)	Using a hands-free "answerphone."	Detection, RT.
Carter and Graham, 2000	Simulator	32	Age: *M* = 29.3 (8 young males), *M* = 66.4 (8 old males), *M* = 30.0 (8 young females), *M* = 59.4 (8 old females); Sex: 16 M	Real V-R (In-Vehicle Speech Recognizer)	Music selection, climate control, phone functions.	RT, Root Mean Squared Error of Lane Position (analogous to SDLP).
Cooper et al., 2014	On-road	36	*M* = 28.1 (3.89), 22–36; Sex: 18 M	Real V-R (Infotainment Systems)	Number dialing, contact dialing, music selection.	RT.
Cuřín et al., 2011	Simulator	28	Age: 18–55; Sex: 14 M	Real V-R System (Prototype)	Text messaging.	MDLP, RT, SDLP.
Graham and Carter, 2001	Simulator	48	Age: *M* = 35.2, 20–50; Sex: 27 M	Real V-R System (Prototype)	Number dialing.	RT, Root Mean Squared Error of Lane Position (analogous to SDLP).
Greenberg et al., 2003	Simulator	63	Age: 25–66 (48 participants), 16–18 (15 participants); Sex: 32 M	Real V-R for Dialing Task (Hands-Free Headset), Simulated V-R for Voice-Mail Task (Wizard of Oz)	Number dialing, voice-mail interaction.	Detection, Headway.
Harbluk et al., 2013	Simulator	16	Age: *M* = 29.4, 21–46; Sex: 8 M	Real V-R (Cell Phone)	Asking intelligent personal assistant predetermined questions, and to read texts and make calendar appointments.	RT.
Harbluk et al., 2007	Simulator	32	Age: *M* = 34, 24–58 (16 visual-manual interface users), *M* = 33, 21–48 (speech-based interface users); Sex: 13 M (16 visual-manual users), 13 M (speech-based interface users)	Real V-R (Navigation System)	Navigation input.	MDLP.
He et al., 2014	Simulator	35	Age: *M* = 21.6 (3.67); Sex: 11 M	Real V-R (Cell Phone)	Text messaging (of phone numbers).	Headway, Mean Lane Position, RT, SD Headway, SDLP, SD Speed, Speed.
He et al., 2015	Simulator	25	Age: *M* = 20.48 (2.14), 18–25; Sex: 12 M	Real V-R (Cell Phone)	Text messaging.	Detection, Headway, Mean Lane Position, RT, SD Headway, SDLP, SD Speed, Speed.
Itoh et al., 2004	Simulator	11	Age: *M* = 35.1, 25–47; Sex: 9 M	Real V-R (Prototype)	Music selection, navigation input.	SDLP.
Lee et al., 2001	Simulator	24	Age: 18–24; Sex: Not Reported.	Simulated V-R (Wizard of Oz)	Emailing.	RT.
Maciej and Vollrath, 2009	Simulator	29	Age: *M* = 33.2 (11.9), 19–59; Sex: 16 M	Real V-R (In-Vehicle Information Systems)	Contact dialing, music selection, navigation input.	MDLP, RT, SDLP.
McCallum et al., 2004	Simulator	24	Age: *M* = 22.8, 18–35; Sex: 12 M	Simulated V-R (Wizard of Oz)	Emailing, internet activities.	RT.
McWilliams et al., 2015	Simulator	40	Age: *M* = 24.6 (2.8), 20–29 (20 younger); *M* = 61.6 (3.4), 55–69 (20 older); Sex: 20 M	Real V-R (Cell Phones)	Navigation input.	SDLP, Speed.
Mehler et al., 2015a (Corolla)	On-Road	48	Age: *M* = 39.8 (17), 20–69 for 24 females, *M* = 40.3 (16.7), 20–67 for 24 males; Sex: 24 M	Real V-R (Infotainment System)	Contact dialing, navigation input.	SD Speed, Speed.
Mehler et al., 2015b (Impala)	On-Road	48	Age: *M* = 41.8 (16.6), 22–68 for 24 females, *M* = 39.6 (16.4), 21–68 for 24 males; Sex: 24 M	Real V-R (Infotainment System)	Contact calling, navigation input.	SD Speed, Speed.
Mehler et al., 2014	On-Road	64	Age: *M* = 22.12 (1.1), 20–24 for 8 females, *M* = 22.00 (1.3), 20–23 for 8 males, *M* = 33.00 (5.9), 25–41 for 8 females, *M* = 29.38 (5.3), 20–36 for 8 males, *M* = 46.75 (3.0), 41–50 for 8 females, *M* = 46.62 (3.7), 42–54 for 8 males, *M* = 58.25 (1.8), 56–61 for 8 females, *M* = 59.50 (4.2), 55–69 for 8 males; Sex: 32 M	Real V-R (Infotainment System)	Music selection, navigation input.	Speed, SD Speed.
Mehler et al., 2015c (CLA)	On-Road	48	Age: *M* = 38.9 (15.6), 20–65 for 24 females, *M* = 39.8 (15.3), 20–67 for 24 males; Sex: 24 M	Real V-R (Infotainment System)	Contact calling, navigation input.	SD Speed, Speed.
Munger et al., 2014	Simulator	24	Age: 20–24 (12 younger), 55 and over (12 older); Sex: 12 M	Real V-R (Cell Phone)	Navigation input.	Detection, RT, SDLP.
Neurauter et al., 2012	Test-Track	24	Age: 18–30 (younger), 45–55 (older); Sex: 12 M	Real V-R	Navigation input, text messaging.	Speed Variance.

(continued on next page)

Table 1 *(continued)*

Study	Setting	*N*	Age (SD) and Sex	Voice-recognition system type	Voice-recognition interaction tasks	Performance measures
				(Hands-Free Cell Phone Kit)		
Ranney et al., 2005	Test-track	21	Age: *M* = 40.3 (13.9), 22–67; Sex: 10 M	Real V-R (Infotainment System)	"Baseline" tasks (dialing phone numbers, tuning radio stations), "Simple tasks" (opening message containing list of items, creating voice memo) and "complex tasks" (opening message, opening phone book, autodialing automated system, retrieving information, creating voice memo).	Detection, RT, SDLP.
Reimer et al., 2011	Simulator	37	Age: *M* = 20.7 (0.9), 19–23 (18 younger), *M* = 56.3 (4.5), 51–66 (19 late middle age); Sex: 19 M	Simulated V-R (Wizard of Oz)	Number dialing, interacting with automated phone trees and voicemail systems.	Speed, SD Speed.
Reimer et al., 2010	Simulator	35	Age: *M* = 20.65 (1.89); Sex: 20 M	Simulated V-R (Wizard of Oz)	Number dialing, interacting with automated phone trees and voicemail systems.	Speed.
Reimer et al., 2013	On-Road	60	Age: *M* = 24.73 (3.0), 20–29 for 15 younger females, *M* = 24.00 (2.7), 20–29 for 15 younger males, *M* = 64.13 (3.0), 60–68 for 15 older females, *M* = 66.20 (2.9), 60–69 for 15 older males; Sex: 30 M	Real V-R (Infotainment System)	Contact dialing, music selection, navigation input.	SD Speed, Speed.
Reimer et al., 2015 (ACC)	On-Road	24	Age: *M* = 63.2 (2.6), 61–68 for 6 older females, *M* = 64.8 (2.8), 60–68 for 6 older males, *M* = 24.2 (3.2), 20–28 for 6 younger females, *M* = 24.2 (1.8), 21–26 for 6 younger males; Sex: 12 M	Real V-R (Infotainment System)	Contact dialing.	SD Speed, Speed.
Reimer et al., 2016	On-Road	80	Age: *M* = 40.4 for females, *M* = 40.3 for males; Sex: 40 M	Real V-R (Infotainment Systems)	Contact dialing.	SD Speed, Speed.
Salvucci, 2001	Simulator	11	Age: *M* = 25, 19–32; Sex: 6 M	Simulated V-R (Wizard of Oz)	Number dialing, contact dialing.	Root Mean Squared Error of Lane Position (analogous to SDLP).
Salvucci and Macuga, 2002	Simulator	7	Age: 18–40; Sex: Not Reported	Real V-R (Cell Phone)	Contact dialing.	Root Mean Squared Error of Lane Position (analogous to SDLP), Root Mean Squared Error of Speed (analogous to SD Speed).
Schreiner et al., 2004	Test-track	37	Age: *M* = 23.4 (18 younger), *M* = 56.6 (19 older); Sex: 9 M (younger), 9 M (older)	Real V-R (Prototype)	Number dialing.	Detection, RT.
Schreiner, 2006	Simulator	12	Age: *M* = 36.9, 22–53; Sex: 4 M	Real V-R (Cell Phones)	Number dialing, contact dialing.	Lane Position Variance, Speed Variance.
Serafin et al., 1993	Simulator	12	Age: *M* = 24, 20–35 (younger), *M* = 70, 60–76 (older); Sex: 6 M	Unclear	Number dialing.	SDLP.
Strayer et al., 2015a (Smartphone Experiment 1)	On-road	31	Age: *M* = 42, 21–68; Sex: 16 M	Real V-R (Cell Phones)	Number dialing, contact dialing, music selection.	RT, Speed.
Strayer et al., 2015a (Smartphone Experiment 2)	On-road	34	Age: *M* = 42.5, 22–68; Sex: 19 M	Real V-R (Cell Phones)	Text messaging.	RT, Speed.
Strayer et al., 2013 (Experiment 2)	Simulator	32	Age: *M* = 23.5, 19–36; Sex: 22 M	Simulated V-R (Wizard of Oz)	Text messaging, emailing.	Headway, RT.
Strayer et al., 2013 (Experiment 3)	On-road	32	Age: *M* = 23.5, 18–33; Sex: 12 M	Simulated V-R (Wizard of Oz)	Text messaging, emailing.	RT.
Strayer et al., 2015b	On-road	257	Age: *M* = 44, 21–70; Sex: 127 M	Real V-R (Infotainment Systems)	Number dialing, contact dialing, music selection.	RT.
Strayer et al., 2014 (Experiment 2)	Simulator	41	Age: *M* = 25.2, 18–40; Sex: 21 M	Simulated V-R (Wizard of Oz)	Climate control, emailing, text messaging, navigation entry.	Headway, RT.
Strayer et al., 2014 (Experiment 3)	On-road	40	Age: *M* = 26.1, 20–19. Sex: 23 M	Simulated V-R (Wizard of Oz)	Climate control, emailing, text messaging, navigation entry.	RT.
Terken et al., 2011	Simulator	41	Age: 20–29; Sex: Not Reported	Real V-R (Prototype)	Emailing, text messaging.	Headway, Speed, SD Headway, SD Speed.
Törnros and Bolling, 2005	Simulator	23	Age: Unclear (*M* is approximately 34, range is approximately 24 to 54); Sex: Unclear (11 or 12 M)	Real V-R (Cell Phone)	Number dialing.	Detection, RT, SDLP, Speed.
Truschin et al., 2014	Simulator	98	Age: Unclear (*M* is approximately 23, *SD* is approximately 5); Sex: Unclear (between 74 and 88 M)	Real V-R (Prototype)	Emailing.	MDLP.

(continued on next page)

Table 1 (*continued*)

Study	Setting	*N*	Age (SD) and Sex	Voice-recognition system type	Voice-recognition interaction tasks	Performance measures
Tsimhoni et al., 2004	Simulator	24	Age: *M* = 24, 20–29 (12 younger), *M* = 69, 65–72 (12 older); Sex: 12 M	Simulated V-R (Wizard of Oz)	Navigation input.	SDLP, SD Headway, Speed.
Yager, 2013	Test-track	43	Age: 2 aged 16–17, 16 aged 18–24, 4 aged 25–29, 3 aged 30–39, 10 aged 40–49, 7 aged 50–59, 1 aged 60 and over; Sex: 20 M	Real V-R (Cell Phones)	Text messaging.	Detection, RT, Speed.

mance have been published elsewhere (see Caird et al., 2008). Tasks that did not have face validity for texting, emailing, navigation entry, or some other interaction type that might reasonably be engaged in with a V-R system during normal driving were also excluded. Studies that did not meet these inclusion criteria typically used cognitive tasks such as verbal or spatial working memory tasks. All excluded studies, and their rationale for exclusion, are shown in Fig. 1.

1.1.3. *Dependent variables*

To be included in the meta-analysis, a measure of driving performance was required. Driving performance can be measured in many different ways but are generally categorized into response time measures, discrete outcome measures (e.g. lane and roadway departures, speed violations), and continuous outcome measures with means and standard deviations (e.g. lateral position, speed and headway) (SAE International, 2015). Specifically, studies that measured reaction time, speed, headway and lateral position were targeted. These four measures are commonly reported in meta-analyses of driving performance but were chosen primarily in order to allow for comparisons specifically with Caird et al. (2008) and Caird et al. (2014), which meta-analyzed the effects of conversation and texting, respectively, on each of these dependent measures.

1.1.4. *Study design*

Studies were required to be experimental in nature. As in any experimental design, both an experimental and a comparison condition were required. For each experimental condition (i.e., the "voice-recognition" or V-R condition), participants must engage in a secondary task with a V-R system while simultaneously engaging in a driving task. Two types of comparison conditions were targeted: visual-manual (V-M) comparison, and baseline (BL) driving. In a V-M comparison condition, participants engage in an analogous secondary task with a V-M system (i.e., a traditional handheld phone or integrated system utilizing touchscreens, buttons or knobs) instead of a V-R system. Including this condition allows inferences to be made about whether voice-recognition systems have any driving performance benefit over traditional handheld phones or integrated in-vehicle systems that lack V-R capabilities. A baseline driving condition had participants complete comparison driving roadway segments without engaging in the secondary task. Including this condition allows inferences to be made about the impact of V-R systems relative to just driving.

1.2. *Information sources*

A subject librarian was consulted to develop a targeted list of electronic databases to search for studies. Search terms were entered into PsycINFO, SPORTDiscus, Academic Search Complete, PubMed, Medline, TRID and Scopus with no limitations on publication year. A preliminary search was conducted using Google Scholar, which was searched with combinations of two to three of the search terms listed below. Electronic searches were completed in January 2016 and included studies up to this date. In addition, personal collections of studies were searched. If a review paper reported a relevant study, the original study was searched for and obtained. Reference lists were also crosschecked for new papers. Authors and experts were contacted to identify additional studies when they appeared as first authors on two or more studies selected for inclusion in the meta-analysis.

1.3. *Search strategy*

The search terms "driver," "driving," "performance," "behavior," "behavior," "voice recognition," "voice-recognition," "speech recognition," "speech-recognition," "voice to text," "voice-to-text," "speech to text," "speech-to-text," "handsfree" and "hands-free" were used. Search terms were combined with Boolean operators.

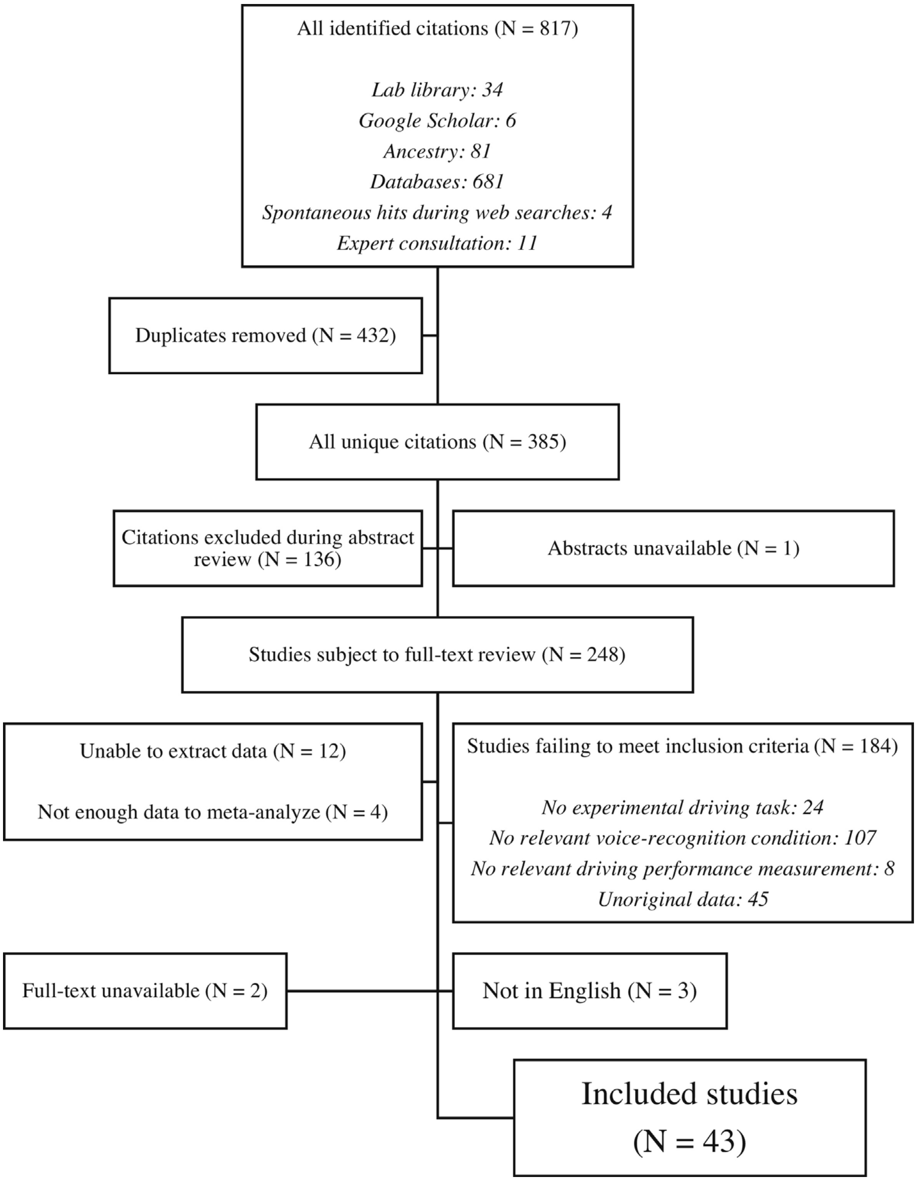
All identified citations (N = 817)
Lab library: 34
Google Scholar: 6
Ancestry: 81
Databases: 681
Spontaneous hits during web searches: 4
Expert consultation: 11
Duplicates removed (N = 432)
All unique citations (N = 385)
Citations excluded during abstract review (N = 136)
Abstracts unavailable (N = 1)
Studies subject to full-text review (N = 248)
Unable to extract data (N = 12)
Not enough data to meta-analyze (N = 4)
Studies failing to meet inclusion criteria (N = 184)
No experimental driving task: 24
No relevant voice-recognition condition: 107
No relevant driving performance measurement: 8
Unoriginal data: 45
Full-text unavailable (N = 2)
Not in English (N = 3)
Included studies
(N = 43)

Fig. 1. Study selection.

1.4. Study selection

All identified studies were imported into Mendeley Desktop (v. 1.16.1), and duplicate citations were removed. All remaining abstracts were screened against the inclusion criteria, which are listed above. Studies that passed the abstract screening process were then subjected to full-text review against the inclusion criteria (see Fig. 1).

Studies that reported unique, original data were identified. When multiple references seemed to contain the same data, the most complete reference was retained. For example, when a conference paper and a journal publication contained the same data, typically the later journal publication was included and the conference paper excluded. Similarly, when an exhaustive technical report and a conference paper were both identified with the same data, the most complete or most interpretable data was included. Finally, when participant data appeared to be reanalyzed and/or re-reported in separate papers, care was taken to extract relevant data only once. In cases where it was unclear whether two papers reported on the same participant sample or the same experiment, authors were contacted to determine if the papers used the same data multiple times.

1.5. Data collection

All data collection was completed by one coder and added to an electronic coding database consisting of spreadsheets. When difficulty or uncertainty was encountered during coding, the item in question was discussed with a second coder (JKC) until a consensus of interpretation could be achieved.

1.6. Data items

First, general publication and demographic information was collected. General publication information included the title of the study, the year of publication, the lead author, and the source (i.e., journal article, conference proceeding, etc.). Demographic information included the sample size, the number of males and females in the sample, the age of the sample (including mean, standard deviation and range), the source of the sample (i.e., convenience sample, employees, etc.) and the nationality of the sample (see Table 1). These items were used for study identification purposes, as well as to provide context for the source of the statistical data extracted.

Next, methodological information was collected. These items were extracted to compute effect sizes for V-R, V-M or baseline conditions, to assess study quality, and to conduct moderator analyses. Brief descriptions of the baseline driving task, the V-R task and comparison task were collected. Descriptions included the type of V-R system (integrated in-vehicle or mobile phone), the type of V-M system (integrated in-vehicle or mobile phone), the research setting (simulator, test-track or on-road), the experimental design (within- and/or between-subjects factors, including random assignment and/or counterbalancing), and whether the participant was allowed to choose when to interact with a V-R system. Because studies varied in the quality of reporting, confidence ratings (percentages accompanied by notes) were assigned. The purposes of the confidence ratings were to identify weaknesses in study quality and to identify potential moderator analyses based on study characteristics.

Descriptions of V-R systems were collected and coded as a basis for moderator analyses. First, the interaction type was coded, including whether it involved music selection, climate control, navigation, emailing, text messaging, or some other type of interaction that might reasonably be expected to occur with the use of a voice-recognition system (see Table 1). Next, input types were coded, including whether the system used true speech recognition or whether a Wizard-of-Oz approach was taken (i.e., a person filled in for the system), and whether a button press was required to start the system (see Table 1). Finally, output types were coded, including whether the system provided auditory feedback, visual feedback, a combination of auditory and visual feedback, or no feedback.

In addition, standardized mean difference effect sizes were computed from means, standard deviations, *F* values, *t* values and *p* values extracted from included studies and stored in an electronic coding database (Lakens, 2013). When possible, effect sizes were computed from reported means and standard deviations. At times, standard deviations were calculated from standard errors and confidence intervals (Cochrane Collaboration, 2011b). These standardized mean difference effect sizes were then converted into *r* effect sizes (Cooper et al., 2009, p. 234). When only *F* or *t* values were available, *r* was computed directly (Schmidt and Hunter, 2014). Results are reported for both effect size forms in Table 2.

In cases where means were available but standard deviations were not given, missing standard deviations were replaced using regression imputation (The Cochrane Collaboration, 2011b; Cooper et al., 2009, pp. 407–408). Standard deviations were regressed against their associated means using the Missing Value Analysis function in SPSS (v. 23). For performance measures where regression imputation was used, sensitivity analyses were conducted (Cochrane Collaboration, 2011a). In these analyses, the magnitude of the effect sizes were inspected for appreciable differences arising from the inclusion or exclusion of effect sizes computed with imputed standard deviations. For this study, an appreciable difference was operationalized using a cut-off point, $r = .10$. If inclusion of effect sizes computed with imputed standard deviations inflated *r* by .10 or more, these effect sizes were excluded. This was based on the rationale that a “weak” effect size is first observed when $r = .10$ (Cohen, 1992). Thus, changes of .09 or less were deemed below the threshold an appreciable effect.

1.7. Data synthesis

Effect sizes in the form of *r*, converted from standardized mean differences (Cohen's *d*), were meta-analyzed using Meta-Excel (Steel, 2016, personal communication). Sensitivity analyses were conducted concurrently using Meta-Excel. Relevant output provided by Meta-Excel includes effect sizes (both *r* and *d*), confidence intervals, credibility intervals and funnel plots. Additional meta-regression analyses were conducted using IBM SPSS Statistics (v. 23). Planned analyses included testing whether sample size was predictive of effect size (publication bias) and whether study setting was predictive of effect size.

2. Results

2.1. Included studies

Of 817 identified citations, 59 studies met inclusion criteria. Of these 59 studies, 43 were ultimately included, and the remaining 16 were excluded. For 12 of these excluded studies, there were insufficient descriptions of materials, methods or statistical analyses such that no data could be extracted for the purposes of a meta-analysis. The remaining 4 excluded studies reported only binary outcome performance measures, specifically lane errors and speed errors. There were not enough studies to meta-analyze these performance measures. A brief overview of each included study is presented in Table 1.

2.2. Publication bias

Determining whether a collection of identified studies contains publication bias is an essential step in meta-analysis (Sutton, 2009). Publication bias was also assessed with weighted least squares regression, setting *N* as a predictor and *r* as the criterion. Publication bias refers to the tendency to favor publishing studies with larger effects (Borenstein et al., 2009). Studies with fewer participants tend to have more sampling error variability, which may lead to a variety of effect size magnitudes. Larger studies tend to be more precise. When publication bias occurs, smaller effect sizes from studies with smaller sample sizes are typically absent, leading to a negative correlation between *N* and *r* (Borenstein et al., 2009).

Across all performance variables with ten or more studies, there were no indications of publication bias [F (1, 27 = 1.725, $p = .200$ for RT, VR v. BL; F (1, 16) = .632, $p = .438$ for RT, VR v. VM; F (1, 17) = 5.119, $p = .037$ for Detection, VR v. BL; F (1, 11) = 1.737, $p = .214$ for SDLP, VR v. BL; F (1, 15) = .758, $p = .398$ for SDLP, VR v. VM; F (1, 16) = 8.289, $p = .011$ for Speed, VR v. BL; F (1, 10) = .009, $p = .926$ for Speed, VM v. VR; F (1, 9) = 8.414, $p = .018$ for Speed Variability, VR v. VM] after adjusting for multiple comparisons ($p = .00625$).

2.3. Meta-analysis

In total, 207 effect sizes were extracted, 183 of which were ultimately included after outliers and some imputed data were removed. If outlier inclusion or imputed data inclusion was found to cause appreciable fluctuations in the magnitude of the meta-analyzed effect size (specifically, an increase or decrease in *r* by 0.10; Cohen, 1992) they were removed. Sensitivity analyses were conducted to test for these fluctuations. The results of the meta-analysis of the 183 effect sizes, for each comparison type (i.e. voice-recognition v. baseline, and voice-recognition v. visual manual) and grouped by each dependent variable, are presented in Table 2.

For each meta-analyzed effect size reported in Table 2, both a confidence interval and a credibility interval are reported. Confidence intervals are a measure of the accuracy of the effect size estimate; they are produced using the mean effect size's standard error and reflect the range within which the average should occur (Whitener, 1990). Credibility intervals reflect the residual variation *after* considering sampling error, usually attributed to moderator effects that can increase or decrease effects sizes (Whitener, 1990). In meta-analysis, moderators are always expected to occur because differences across studies are always present. The wider the credibility interval, the more that heterogeneity between studies is present and the mean is less likely to be representative of any particular study. If these credibility intervals cross zero, it indicates a lack of generalizability.

2.3.1. Detection

Detection refers to the number of visual targets or events that participants respond to during a detection task. For example, some tasks involved responding to peripheral targets such as flashing lights, and

Table 2
Results of meta-analysis.

Variable	*k*	*N*	*r*	*d*	95% CI		95% CrdI	
					L	*U*	*L*	*U*
Detection								
Voice recognition v. baseline	19	948	−.41	−.90	−.50	−.32	−.71	−.11
Voice recognition v. visual-manual	5	136	.21	.42	−.06	.48	−.27	.69
Reaction time (RT)								
Voice recognition v. baseline	29	1005	.55	1.32	.48	.62	.26	.84
Voice recognition v. visual-manual	18	430	−.29	−.60	−.38	−.20	−.29	−.29
Standard deviation of lane position (SDLP)								
Voice recognition v. baseline	13	227	.20	.40	.05	.35	−.09	.48
Voice recognition v. visual-manual	17	356	−.39	−.84	−.48	−.30	−.39	−.39
Mean deviation of lane position (MDLP)								
Voice recognition v. baseline	9	229	.26	.53	.14	.38	.26	.26
Voice recognition v. visual-manual	7	160	−.18	−.36	−.34	−.02	−.29	−.07
Headway								
Voice recognition v. baseline	8	410	.16	.32	.06	.25	.16	.16
Voice recognition v. visual-manual	3	123	−.18	−.37	−.36	−.01	−.18	−.18
Headway variability								
Voice recognition v. baseline	4	125	.19	.39	.02	.36	.19	.19
Voice recognition v. visual-manual	4	125	−.11	−.22	−.28	.07	−.11	−.11
Speed								
Voice recognition v. baseline	18	761	−.13	−.26	−.20	−.06	−.13	−.13
Voice recognition v. visual-manual	12	576	.09	.18	.01	.17	.09	.09
Speed variability								
Voice recognition v. baseline	9	265	−.05	−.11	−.23	.12	−.44	.34
Voice recognition v. visual-manual	11	372	.08	.15	−.06	.21	−.21	.37

In Table 2, *k* represents the number of effect sizes contributing to the overall meta-analyzed effect size, *N* indicates the total number of participants represented across all included studies per meta-analyzed effect size, *r* is the meta-analyzed effect size converted from *d, d* is the standardized mean difference between a voice-recognition system interaction and either baseline driving or a visual-manual system interaction, and *L* and *U* are lower and upper limits, respectively, for both the 95% confidence interval and the 95% credibility interval. Positive *r* and *d* values indicate that voice-recognition system interactions have a larger effect in the dependent variable than the second listed condition, and negative *r* and *d* values indicate that voice-recognition system interactions have a smaller effect in the dependent variable than the second listed condition. An *r* value of .10 indicates a small effect size, .30 indicates a medium effect size, and .50 indicates a large effect size (Cohen, 1992).

others involved responding to simulated hazardous scenarios, such as a lead car suddenly decelerating. Subgroup analysis indicated that detection tasks involving peripheral targets and detection tasks involving simulated hazardous scenarios were similar; thus, effect sizes from both types of detection tasks were pooled.

Compared to baseline driving, drivers who interacted with a voice-recognition system while driving detected moderately fewer targets (*r* = −.41, 95% CI [−.50, −.32]). It appears likely that moderators are present and in operation in this effect (95% CrdI [−.71, −.11]). Interestingly, there does not appear to be a detection advantage during interactions with voice-recognition systems compared to visual-manual systems (*r* = .21, 95% CI [−.06, .48]). Again, moderators appear to be in operation (95% CrdI [−.27, .69]).

2.3.2. Reaction time

Whereas detection refers to the number of targets that a participant provides responses to, reaction time (RT) refers to the time interval between the onset of a stimulus and the participant's first observable response to that stimulus (see SAE International, 2015 for complete measurement descriptions). Larger effect sizes reflect larger time intervals, and larger time intervals indicate greater impairment. Compared to baseline driving, driving while interacting with a voice-recognition system is associated with increases in reaction time to stimuli and events (*r* = .55, 95% CI [.48, .62]). Credibility intervals indicate that moderators are likely in operation (95% CrdI [.26, .84]). Interactions with a V-R system had a smaller reaction time effect compared to interactions with V-M systems. The negative effect size indicates a moderate reaction time advantage (i.e. a decrease in reaction time) associated with V-R systems compared to interactions with V-M systems (*r* = −.29, 95% CI [−.38, −.20]).

2.3.3. Standard deviation of lane position (SDLP)

SDLP is a measure of lateral position maintenance, with larger effect sizes indicating increased within-lane oscillations (impaired lateral position maintenance ability). Compared to baseline driving, V-R systems are associated with a small increase in SDLP (*r* = .20; 95% CI [.05, .35]). Credibility intervals indicated the potential presence of moderators (95% CrdI [−.09, .48]). Interactions with V-R systems were associated with a moderate decrease in SDLP versus interactions with V-M systems (*r* = −.39, 95% CI [−.48, −.30]).

2.3.4. Mean deviation of lane position (MDLP)

MDLP, also known as “offset,” refers to the mean difference between the participant's average lateral position and a reference position. Larger values for MDLP indicate increased distance from the reference position. Compared to baseline driving, driving while using a V-R system led to a small increase in MDLP (*r* = .26, 95% CI [.14, .38]). Compared to V-M systems, V-R systems led to small decreases in MDLP (*r* = −.18, 95% CI [−.34, −.02]). Credibility intervals indicated the potential presence of moderators (95% CrdI [−.29, −.07]).

2.3.5. Headway

Compared to baseline driving, driving while using a V-R system is associated with small increases in following distance or headway (*r* = .16, 95% CI [.06, .25]). Compared to V-M systems, V-R systems led to small decreases in headway (*r* = −.18, 95% CI [−.36, −.01]).

2.3.6. Headway variability

Compared to baseline driving, driving while using a V-R system leads to small increases in headway variability (*r* = .19, 95% CI [.02, .36]). V-R systems did not significantly impact headway variability compared to V-M systems (*r* = −.11, 95% CI [−.28, .07]).

2.3.7. Speed

Compared to baseline driving, driving while using a V-R system is associated with a small decrease in speed (*r* = −.13, 95% CI [−.20,

− .06]). Compared to V-M systems, V-R systems do not have an appreciable influence on speed (*r* = .09, 95% CI [.01, .17]).

2.3.8. Speed variability

Compared to baseline driving, driving while using a V-R system has no effect on speed variability (*r* = −.05, 95% CI [−.23, .12]). However, caution should be taken in interpreting this result. This meta-analyzed effect size is highly unstable according to the sensitivity analyses, and credibility intervals indicate the potential presence of sizeable moderators (95% CrdI [−.44, .34]). Compared to V-M systems, V-R systems had no effect on speed variability (*r* = .08, 95% CI [−.06, .21]). However, credibility intervals indicate that moderators are likely present (95% CrdI [−.21, .37]).

2.4. Meta-regression

Although meta-regression was planned to test whether effect sizes differed based on experimental setting, a limitation of the current study is small sample sizes per performance measure. Thus, meta-regression was only feasible in a small subset of performance measures (i.e., performance measures with more than ten studies per covariate) (Cochrane Collaboration, 2011c). The results should be interpreted in accord with this limitation. For each set of meta-regressions conducted, Bonferroni corrections were used to adjust for multiple comparisons.

2.4.1. Experimental setting

In several previous meta-analyses (Caird et al., 2008; Horrey and Wickens, 2006), no differences in effect size magnitudes were observed across experimental settings. Simulator, on-road and test-track studies were dummy coded and entered into a WLS regression model, weighted using inverse sampling error (Steel and Kammeyer-Mueller, 2002), with *r* specified as the criterion. For dependent performance measures where meta-regression was feasible, experimental setting was not predictive of effect size [*F* (2, 26) = 1.267, *p* = .298 for RT, V-R v. BL; *F* (2, 15) = 1.978, *p* = .173 for RT, V-M v. V-R; *F* (2, 16) = 4.812, *p* = .023 for Detection, V-R v. BL; *F* (1, 15) = 5.000, *p* = .041 for SDLP, V-M v. VR; *F* (2, 15) = .734, *p* = .497 for Speed, V-R v. BL; *F* (2, 9) = .776, *p* = .489 for Speed, V-M v. V-R; *F* (2, 8) = 3.182, *p* = .096 for Speed Variability, V-R v. V-M], adjusted for multiple comparisons (*p*-value = .007). Effect size magnitudes do not appear to vary between simulator, on-road and test-track settings.

2.4.2. Voice-recognition accuracy

During coding, voice-recognition accuracy became a suspected moderator. Some studies used true V-R systems, such as commercially available in-vehicle and personal assistant systems. Other studies used a "Wizard of Oz" where the voice-recognition system was simulated by a team of researchers. Participants' vocal commands were interpreted by research assistants who then manually controlled the system's next action. The benefit of a Wizard of Oz approach is that the system can be made to operate with perfect voice recognition accuracy. However, perfect system accuracy removes the variance associated with individual differences in proficiency in using a real voice-recognition system.

For dependent performance measures where meta-regression was feasible, voice-recognition accuracy was not significantly predictive of effect size [*F* (1, 26) = .163, *p* = .690 for RT, V-R v. BL; *F* (1, 17) = .516, *p* = .482 for Detection, V-R v. BL; *F* (1, 11) = .319, *p* = .583 for SDLP, V-R v. BL; *F* (1, 16) = .035, *p* = .854 for Speed, V-R v. BL] after adjusting for multiple comparisons (*p* = .0125). V-R systems that operated with perfect accuracy (i.e. Wizard of Oz systems), when compared to those assumed to operate with imperfect accuracy (true voice-recognition systems and Wizard of Oz systems specified as operating with imperfect accuracy), had a similar impact on driving performance.

2.4.3. Year of publication

The date of publication was identified as a possible moderator. Over time V-R systems may have improved in accuracy and functionality. Time was not significantly predictive of effect size for reaction time [*F* (1, 18) = 3.434, *p* = .080, V-R v. BL], for speed [*F* (1, 12) = .075, *p* = .788, V-R v. BL] or for SDLP [*F* (1, 10) = .323, *p* = .583 V-R v. BL] after adjusting for multiple comparisons (*p* = .0125). However, time was significantly predictive for detection [*F* (1, 11) = 13.475, *p* = .004 for Detection, V-R v. BL]. A positive relationship between year and effect size (*r* = .742) indicates that higher rates of detection are associated with later study publication dates.

3. Discussion

The primary purpose of this meta-analysis was to synthesize the effects of using a voice-recognition (V-R) system to perform everyday tasks while driving. A secondary purpose was to meta-analyze the differences in driver performance between using visual-manual (V-M) systems and V-R systems. An exhaustive search was conducted to identify all studies that measured certain dependent variables while using a V-R system while driving with comparisons to either a baseline driving condition and/or with a V-M system. A meta-analysis provides a higher level of evidence over a more complete set of measures than results from individual studies, which may be underpowered or cannot be replicated.

Driving while interacting with a V-R system is associated with increases in reaction time and lane positioning, and decreases in detection compared to driving without a system. Fig. 2 depicts the pattern of effect sizes for these variables grouped as hazard perception, longitudinal control and lateral position (also see, Table 2). In the figure, small, medium and large effect sizes correspond to *r* = .10, *r* = .30 and *r* = .50, respectively (Cohen, 1992). Measures of hazard perception, reaction time and detection had large and medium effect sizes, respectively. While using a V-R system, reaction time was increased and detection was decreased compared to baseline driving. Measures of driving performance that may indicate compensatory adjustments to system interactions associated with using a V-R system had small effect sizes (see Fig. 2). Speed, speed variability, headway and headway variability had small effects while using a V-R system compared to driving without a system. Thus, slowing down or increasing following distance were minimally affected from the synthesis of

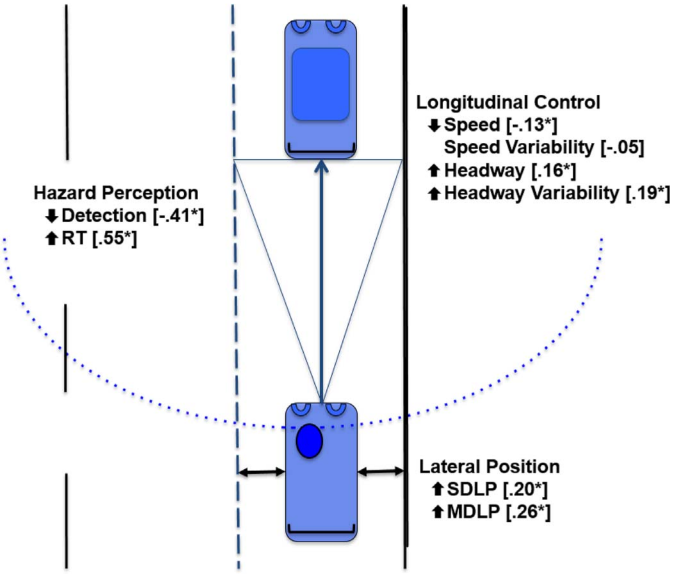

Fig. 2. Voice recognition (V-R) system effect sizes [*r*=] for analyzed measures with the direction of effects from baseline driving indicated by arrows. Asterisks indicate statistical significance based on 95% confidence intervals. See text for additional details.

effects across studies.

Lane variability (SDLP) and lane offset (MDLP) had small effect sizes while using V-R systems compared to baseline driving, which was not expected. One explanation for these effects is that drivers may have looked at the V-R system to confirm spoken commands. Visual system checking, even in the presence of auditory confirmation, likely affects lateral vehicle position as visual steering information is temporarily not available or in peripheral vision while the eyes are off the road (Caird et al., 2014; Cuřín et al., 2011; Green, 2016; Land and Lee, 1994; Strayer et al., 2015b). The accuracy of a V-R system may affect glance behavior during learning to use a system and may carry through to long-term use patterns, which requires additional study (Strayer et al., 2016). The present meta-analysis did not analyze eye movements, which would provide a stronger confirmation of this explanation. During coding across studies, different eye movement variables were used and a limited number of studies measured eye tracking. In general, Strayer et al. (2016) concludes that V-R system interactions are not necessarily "eyes-free" activities due to confirmation checking and incidental interactions with systems.

When V-M systems were compared to V-R systems, drivers had moderately better performance with V-R systems on reaction time, SDLP and headway. The pattern of effect sizes across variables is shown in Fig. 3 (also see Table 2). Each effect size represents the advantage of V-R over V-M systems. Speed, speed variability and headway variability, which are interpreted as compensatory variables, showed no differences between V-R and V-M systems. Of particular interest though, detection did not show an advantage for V-R systems. However, because the credibility interval crosses 0, interpretation of this effect size is problematic and is not necessarily generalizable.

Compared to V-M systems, a number of studies reported that drivers who used V-R systems spent more time looking at the road, had shorter glance times away from the road, and had fewer and shorter glances toward the visual display (Cuřín et al., 2011; Itoh et al., 2004; Maciej and Vollrath, 2009; Reimer et al., 2016). The small decrease in offset (MDLP) and the moderate decrease in lane variability (SDLP) observed for V-R system use, compared to V-M system use, likely reflects the minimization of eyes-off-road time. Overall, V-R systems show modest advantages over V-M systems for reaction time and lane keeping.

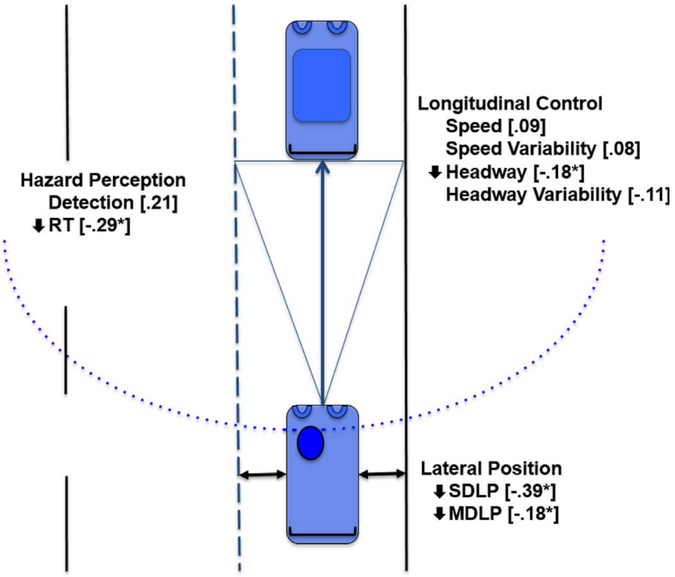

Fig. 3. Voice recognition (V-R) system effect [$r=$] for analyzed measures with the direction of effects from baseline driving indicated by arrows. Asterisks indicate statistical significance based on 95% confidence intervals. See text for additional details.

3.1. Theoretical implications

The Multiple Resource Theory (MRT) predicts that separate visual and auditory inputs permit easier task sharing, whereas dual-task interference worsens when two simultaneous tasks compete for the same input modality (Wickens, 2002, 2008; Wickens et al., 2013). Voice-recognition systems are expected to minimize diversion of the driver's eyes from the roadway because interactions can be completed using verbal inputs. In contrast to visual-manual input, verbal inputs would not necessarily interfere with the visual tasks of vehicle control and hazard perception, which are predominately visual. Thus, driving while completing V-R tasks would not theoretically use the same visual input modality compared to driving while completing the same V-M system tasks.

Despite these hopeful expectations, actual V-R system use involves some competing visual inputs that likely affect driving performance measures. Looking at the roadway for steering information and potential hazards competes with visual V-R system interactions. A number of the studies included in this meta-analysis used V-R systems that provided visual feedback to confirm system actions. For example, Cuřín et al. (2011) noted that drivers frequently looked at visual displays for feedback even though the same information was available through acoustic feedback. Strayer and colleagues (2015b) also observed that participants routinely glanced at in-vehicle displays during interactions with V-R systems. In another study, eye movements to the button of a speech-based interface were reported even though feedback was not provided by the system (Neurauter et al., 2012). Visual checking of a V-R system for feedback about verbal inputs does not seem to decline with short-term practice (Strayer et al., 2016). Visual checking of a V-R system for feedback or looking at a system to complete incidental interactions compete for the visual input modality with hazard perception and lane keeping. From this meta-analysis, decreased detection and increased RT and lateral position measures likely reflect, in part, the effects of this visual competition.

MRT also predicts declines in performance when tasks overlap based on spatial and verbal codes across processing stages (Wickens, 2002). Given the variation of spatial and verbal codes across dialing, initiating a call, texting, emailing, destination entry or music selection tasks, visual and auditory overlap and crosstalk across codes may also contribute to performance decrements. When high overall demands are placed on working memory from multiple tasks, the slight advantage of the separation of auditory and visual modalities associated with spoken interactions and driving may be overwhelmed (Wickens et al., 2013). Properties of V-R systems such as low system intuitiveness, high task complexity, recall from long-term memory, comparisons within working memory, requirements to develop task switching strategies and long task engagements may also contribute to oscillations in cognitive demand that may result in driving performance decrements (Strayer et al., 2013, 2014, 2015a, 2015b). Delays in RT and missed detections, from this meta-analysis, may also indicate variations in cognitive demands on working memory that affect these measures. For example, drivers may look directly at a target during a detection task, yet fail to sufficiently process and respond to that target due to insufficient attention resources (White and Caird, 2010).

3.2. Implications for drivers and designers

Voice-recognition systems are advertised as an alternative to less-safe visual manual systems. Drivers may perceive V-R systems as distraction-free, in part, because manufacturers market these systems to consumers as a possible solution to driver distraction. Minimization of eyes-off-road time appears to be a large selling point for these systems. However, the results of this meta-analysis question whether these systems are necessarily 'risk-free' for several reasons. First, hands-free systems that use voice recognition can still interfere with hazard detection, reaction time and lane keeping. Drivers should know that V-

R systems might take a driver's eyes or mind off the road. Although V-R systems appear to offer slight advantages over V-M systems, advocacy for voice recognition systems must be interpreted cautiously. Second, drivers may use V-R systems more frequently than comparable V-M systems under the belief that the former is safe and the latter is unsafe. Similarly more frequent use of V-R systems may also occur in more dangerous environmental contexts such as at intersections that will raise the overall distraction risk exposure of the driver.

V-R systems are in use now and use is likely to increase in the future. Designers and system evaluators need to develop additional ways to minimize the potential for visual and cognitive distraction. For example, designers may want to explore ways of maximizing the benefits of V-R systems by testing a variety of designs and their effects on driving performance such as the form and content of feedback to drivers. Visual confirmations may tempt drivers to divert their eyes from the roadway. Conversely, purely auditory feedback may impose working memory costs that are time and sound dependent. However, if a driver dictates a text message but receives *no* confirmation about the message, the driver has no way of knowing whether the dictated message was processed correctly or reached an intended recipient. Lack of feedback may result in a frustrating driver experience and may require additional steps to confirm task completion.

National Highway Transportation Safety Administration's (NHTSA) visual-manual guidelines for in-vehicle and portable devices, which are voluntary, recommend that systems allow drivers be able to maintain eyes forward on the roadway most of the time, to keep at least one hand on the wheel at all time, and to be able interrupt and control the pace of a task (NHTSA, 2012, 2014). Based on the results of the current meta-analysis, voice-recognition system designs that allow the driver to maintain their eyes on the forward roadway most of the time may not be sufficient to reduce cognitive and visual distraction, which may be addressed in subsequent NHTSA auditory-visual guidelines.

3.3. Limitations and future research

There are a number of limitations to the current study. Foremost, the role of moderators needs to be considered. A number of credibility intervals indicated the potential presence of moderators, which complicates interpretation. The more variance contributed by moderators, the wider the credibility interval (see Table 2). For a given study, a number of potential moderators could not be coded and tested using meta-regression including driving environment (urban, rural, suburban), recognition accuracy (90%, 95%, 100%), sex (male, female), age (teen, adult, older), language spoken (English, French, etc.), technology proficiency (low, medium, high), length of use (short, long), task types (email, texting, music selection, etc.), and task completion time (short, medium, long) among others. Each of these variables represents contributions to heterogeneity of variance when multiple studies are combined together. Although wider credibility intervals pose a challenge in that they indicate uncertainty in the meta-analyzed effect size owing to heterogeneity, credibility intervals do indicate the range of possible values (i.e. effect sizes), which contain the realized results of any particular study, such as everyday task variability across V-R system use.

Secondly, small sample sizes (i.e., fewer than ten studies) for dependent driving performance variables also complicated moderator analyses that used WLS regression. Notably, this complicated statistical tests for publication bias. Although no moderator effects (including publication bias) were detected in any of the dependent performance measures tested, meta-regression analyses in the case of the current study are likely underpowered and other factors that could not be coded likely contributed to wide credibility intervals that suggested additional moderators.

Across studies, a lack of consistency in reporting of methods and results leads to problems of data extraction. About 20% of the studies that met inclusion criteria could not be used to compute effect sizes, primarily due to insufficient descriptions of materials, methods and statistical analyses (see Fig. 1). Many of the included studies did not sufficiently report their statistical information for all measures such that effect sizes could be extracted (also see, Barón and Green, 2006; Caird et al., 2014). In many cases, relevant pairwise comparisons were not reported, so descriptive statistics were required to compute effect sizes. However, descriptive statistics were often unavailable as well, especially for null results. Problems were also encountered when figures were used to visually represent results. Error bars (i.e., representations of standard error or standard deviation) were often unlabeled or absent. In other cases, error bars were present, but they were adjusted for within-subjects designs (i.e., see Loftus and Masson, 1994). These adjusted error bars are quite useful for making visual inferences about significant differences, but effect sizes cannot be computed with this information alone. Descriptive statistics are needed to supplement these types of figures in order to compute effect sizes. From the perspective of a meta-analyst or peer reviewer, future researchers are encouraged to make some specific changes. First, tables of descriptive data, including means and standard errors or standard deviations, should be reported. This data is valuable in computing effect sizes and making comparisons beyond those reported in omnibus test results. Second, figures should be labeled clearly to aid interpretation.

In addition, several problems were encountered while coding study methods. In many cases, it was unclear what form of feedback was provided to the driver for each experimental task. The exact steps taken to interact with a voice recognition system were often insufficiently specified. Interaction often requires a number of visual, cognitive and manual steps that impose different costs to ongoing driving. If limited by word count, inclusion of important information in the form of appendices should be considered. It is also possible that several studies that may be otherwise eligible for inclusion may not have been identified for inclusion in the present study. Although this may occur with any meta-analysis, it is highly likely that there is a body of dark literature, which is only available to those within certain corporations, related to this topic. For example, studies conducted by automotive and technology companies may be protected by nondisclosure agreements, complicating identification and access. Additionally, studies conducted by automotive companies may not necessarily be written in English, but potentially in German, Japanese and other languages.

Finally, it is important to note that although the current study demonstrates that voice-recognition systems have a distraction cost, it is unclear whether these translate into increased crash risk. Epidemiological and naturalistic research on crash risk associated with these systems is still needed.

Acknowledgements

We would like to thank Laura Koltutsky for assisting us with the development of key word and electronic database search strategies and an anonymous reviewer for focusing our results and discussion. We would also like to express our gratitude to all of the authors who provided additional study information and data upon request. An abstract of this paper was presented at the 5th International Conference on Driver Distraction and Inattention in Paris.

References[1]

*Angell, L., Auflick, J., Austria, P.A., Kochhar, D., Tijerina, L., Biever, W., … Kiger, S., 2006. Driver Workload Metrics Project: Task 2 Final Report (Report No. DOT HS 810 635). National Highway Traffic Safety Administration, Washington, DC.

Atchley, P., Atwood, S., Boulton, A., 2011. The choice of text and drive in younger drivers: behavior may shape attitude. Accident Anal. Prevent. 43, 134–142.

Barón, A., Green, P., 2006. Safety and usability of speech interfaces for in-vehicle task while driving: A brief literature rewiew (Tech. Rep. No. UMTRI-2006-5). University of Michigan Transportation Research Institute, Ann Arbor, MI.

[1] Note: * indicates studies included in the meta-analysis.

*Beckers, N., Schreiner, S., Bertrand, P., Reimer, B., Mehler, B., Munger, D., Dobres, J., 2014. Comparing the demands of destination entry using Google Glass and the Samsung Galaxy S4. Proceedings of the Human Factors and Ergonomics Society Annual Meeting 58, 2156–2160. http://dx.doi.org/10.1177/1541931214581453.

Borenstein, M., Hedges, L.V., Higgins, J.P.T., Rothstein, H.R., 2009. Introduction to Meta-Analysis. Wiley, West Sussex, U.K.

*Bruyas, M.-P., Brusque, C., Debailleux, S., Duraz, M., Aillerie, I., 2009. Does making a conversation asynchronous reduce the negative impact of phone call on driving? Transport. Res. Part F: Traffic Psychol. Behav. 12 (1), 12–20. http://dx.doi.org/10.1016/j.trf.2008.06.002.

Caird, J.K., Horrey, W., 2017. A review of novice and teen driver distraction. In: Fisher, D.L., Caird, J.K., Horrey, W.J., Trick, L. (Eds.), The Handbook of Teen and Novice Drivers: Research, Practice, Policy and Directions. CRC Press, Boca Raton, FL.

Caird, J.K., Johnston, K., Willness, C., Asbridge, M., Steel, P., 2014. A meta-analysis of the effects of texting on driving. Accident Anal. Prevent. 71, 311–318.

Caird, J.K., Willness, C.R., Steel, P., Scialfa, C., 2008. A meta-analysis of the effects of cell phones on driver performance. Accident Anal. Prevent. 40 (4), 1282–1293. http://dx.doi.org/10.1016/j.aap.2008.01.009.

Canadian Council on Motor Transport Administrators (CCMTA), 2013. Canadian fines and demerit points for: Hand-held cell phone or electronic communication device use - August 2013. Retrieved September 27, 2015, from http://ccmta.ca/images/publications/pdf//Canadian-fines-and-demerit-points-cell-phones-ensligh.pdf.

*Carter, C., Graham, R., 2000. Experimental comparison of manual and voice controls for the operation of in-vehicle systems. Proceedings of the Human Factors and Ergonomics Society Annual Meeting, Vol. 58 2156–2160. http://dx.doi.org/10.1177/154193120004402016.

Cochrane Collaboration, 2011a. 16.1.3.1 Imputing standard deviations. In: Higgins, J.P.T., Green, S. (Eds.), Cochrane Handbook for Systematic Reviews of Interventions Version 5.1.0 [updated March 2011], . Retrieved from www.cochrane-handbook.org.

Cochrane Collaboration, 2011b. 7.7.3.2 Obtaining standard deviations from standard errors and confidence intervals for group means. In: Higgins, J.P.T., Green, S. (Eds.), Cochrane Handbook for Systematic Reviews of Interventions Version 5.1.0 [updated March 2011], . Retrieved from www.cochrane-handbook.org.

Cochrane Collaboration, 2011c. 9.6.4 Meta-regression. In: Higgins, J.P.T., Green, S. (Eds.), Cochrane Handbook for Systematic Reviews of Interventions Version 5.1.0 [updated March 2011], . Retrieved from www.cochrane-handbook.org.

Cohen, J., 1992. A power primer. Psychol. Bull. 112 (1), 155–159. http://dx.doi.org/10.1037/0033-2909.112.1.155.

Cooper, H., Hedges, L.V., Valentine, J.C. (Eds.), 2009. The Handbook of Research Synthesis and Meta-Analysis, 2nd ed. Russel Sage Foundation, New York, New York, USA.

*Cooper, J.M., Ingebretsen, H., Strayer, D.L., 2014. Mental Workload of Common Voice-Based Vehicle Interactions across Six Different Vehicle Systems. AAA Foundation for Traffic Safety, Washington, DC.

*Cuřín, J., Labský, M., Macek, T., Kleindienst, J., Young, H., Thyme-Gobbel, A., … König, L., 2011. Dictating and editing short texts while driving. Proceedings of the 3rd International Conference on Automotive User Interfaces and Interactive Vehicular Applications - AutomotiveUI '11. ACM Press, New York, New York, USA, pp. 13. http://dx.doi.org/10.1145/2381416.2381418.

Dingus, T.A., Guo, F., Lee, S., Antin, J.F., Perez, M., Buchman-King, M., Hankey, J., 2016. Driver crash risk factors and prevalence evaluation using naturalistic driving data. Proc. Natl. Acad. Sci. 113 (10), 60–68.

Fitch, G.M., Soccolich, S.A., Guo, F., McClafferty, J., Fang, Y., Olson, R.L., Perez, M.A., Hanowski, R.J., Hankey, J.M., Dingus, T.A., 2013. The impact of hand-held and hands-free cell phone use on driving performance and safety-critical event risk (report no. DOT HS 811 757). National Highway Traffic Safety Administration, Washington, DC.

Governors Highway Safety Association (GHSA), 2015. Cell phone and texting laws. Retrieved July 9, 2015, from http://www.ghsa.org/html/stateinfo/laws/cellphone_laws.html.

*Graham, R., Carter, C., 2001. Voice dialling can reduce the interference between concurrent tasks of driving and phoning. Int. J. Vehicle Design 26 (1), 30. http://dx.doi.org/10.1504/IJVD.2001.001925.

Green, P., 2016. Where do drivers look while driving (and for how long)? In: Smiley, A. (Ed.), Human Factors in Traffic Safety, 3rd ed. Lawyers and Judges Publishing, Tucson, AZ, pp. 57–86.

*Greenberg, J., Tijerina, L., Curry, R., Artz, B., Cathey, L., Kochhar, D., … Grant, P., 2003. Driver distraction: evaluation with event detection paradigm. Transport. Res. Rec. 1843, 1–9. http://dx.doi.org/10.3141/1843-01.

Hamilton, B.C., Arnold, L.S., Tefft, B.C., 2013. Distracted driving and perceptions of hands-free technologies: Findings from the 2013 Traffic Safety Culture Index. AAA Foundation for Traffic Safety, Washington, DC.

*Harbluk, J.L., Burns, P.C., Hernandez, S., Tam, J., Glazduri, V., 2013. Detection response tasks: Using remote, headmounted and tactile signals to assess cognitive demand while driving. Proceedings of the Seventh International Driving Symposium on Human Factors in Driver Assessment, Training, and Vehicle Design 78–84.

*Harbluk, J.L., Burns, P.C., Lochner, M., Trbovich, P.L., 2007. Using the Lane Change Test (LCT) to assess distraction: Tests of visual-manual and speech-based operation of navigation system interfaces. Proceedings of the 4th International Driving Symposium on Human Factors in Driver Assessment, Training and Vehicle Design 16–22.

*He, J., Chaparro, A., Nguyen, B., Burge, R.J., Crandall, J., Chaparro, B., … Cao, S., 2014. Texting while driving: is speech-based text entry less risky than handheld text entry? Accident Anal. Prevent. 72, 287–295. http://dx.doi.org/10.1016/j.aap.2014.07.014.

*He, J., Choi, W., McCarley, J.S., Chaparro, B.S., Wang, C., 2015. Texting while driving using Google Glass^{TM}: promising but not distraction-free. Accident Anal. Prevent. 81, 218–229. http://dx.doi.org/10.1016/j.aap.2015.03.033.

Horrey, W.J., Wickens, C.D., 2006. The impact of cell phone conversation on driving using meta-analytic techniques. Human Factors 48 (1), 196–205.

*Itoh, K., Miki, Y., Yoshitsugu, N., Kubo, N., Mashimo, S., 2004. Evaluation of a Voice-Activated System Using a Driving Simulator (SAE Technical Paper 2004-01-0232). SAE Technical Paper 2004-01-0232http://dx.doi.org/10.4271/2004-01-0232.

Klein, M., 2015. 26 Actually Useful Things You Can Do with Siri. Retrieved December 9, 2015, from http://www.howtogeek.com/229308/26-actually-useful-things-you-can-do-with-siri/.

Klauer, S.G., Guo, F., Simons-Morton, B.G., Ouimet, M.C., Lee, S.E., Dingus, T.A., 2014. Distracted driving and risk of road crashes among novice and experienced drivers. N. Engl. J. Med. 370, 54–59.

Lakens, D., 2013. Calculating and reporting effect sizes to facilitate cumulative science: a practical primer for t-tests and ANOVAs. Front. Psychol. 4, 1–12. http://dx.doi.org/10.3389/fpsyg.2013.00863.

Land, M.F., Lee, D.N., 1994. Where do we look when we steer? Nature 369, 742–744.

*Lee, J.D., Caven, B., Haake, S., Brown, T.L., 2001. Speech-based interaction with in-vehicle computers: the effect of speech-based e-mail on drivers' attention to the roadway. Human Factors 43 (4), 631–640. http://dx.doi.org/10.1518/001872001775870340.

Loftus, G.R., Masson, M.E.J., 1994. Using confidence intervals in within-subject designs. Psychonomic Bull. Rev. 1 (4), 476–490. http://dx.doi.org/10.3758/BF03210951.

*Maciej, J., Vollrath, M., 2009. Comparison of manual vs. speech-based interaction with in-vehicle information systems. Accident Anal. Prevent. 41 (5), 924–930. http://dx.doi.org/10.1016/j.aap.2009.05.007.

*McCallum, M.C., Campbell, J.L., Richman, J.B., Brown, J.L., Wiese, E., 2004. Speech recognition and in-vehicle telematics devices: potential reductions in driver distraction. Int. J. Speech Technol. 7 (1), 25–33. http://dx.doi.org/10.1023/B:IJST.0000004804.85334.35.

McCartt, A.T., Kidd, D.G., Teoh, E.R., 2014. Driver cellphone and texting bans in the United States: evidence of effectiveness. Ann. Adv. Automotive Med. 58, 99–114.

*McWilliams, T., Reimer, B., Mehler, B., Dobres, J., Coughlin, J.F., 2015. Effects of age and smartphone experience on driver behavior during address entry. Proceedings of the 7th International Conference on Automotive User Interfaces and Interactive Vehicular Applications - AutomotiveUI '15. ACM Press, New York, New York, USA, pp. 150–153. http://dx.doi.org/10.1145/2799250.2799275.

*Mehler, B., Reimer, B., Dobres, J., Coughlin, J.F., 2015a. Assessing the Demands of Voice Based In-Vehicle Interfaces - Phase II Experiment 3-2015 Toyota Corolla (2015b) (MIT AgeLab Technical Report 2015-14). Massachusetts Institute of Technology, Cambridge, MA.

*Mehler, B., Reimer, B., Dobres, J., McAnulty, H., Coughlin, J.F., 2015b. Assessing the Demands of Voice-Based In-Vehicle Interfaces - Phase II Experiment 1-2014 Chevrolet Impala (2014b) (MIT AgeLab Technical Report 2015-6A). Massachusetts Institute of Technology, Cambridge, MA.

*Mehler, B., Reimer, B., Dobres, J., McAnulty, H., Mehler, A., Munger, D., Coughlin, J.F., 2014. Further Evaluation of the Effects of a Production Level "Voice-Command" Interface on Driver Behavior: Replication and a Consideration of the Significance of Training Method. Massachusetts Institute of Technology, Cambridge, MA.

*Mehler, B., Reimer, B., McAnulty, H., Dobres, J., Lee, J., Coughlin, J.F., 2015c. Assessing the Demands of Voice Based In-Vehicle Interfaces - Phase II Experiment 2-2014 Mercedes CLA (2014t) (MIT AgeLab Technical Report 2015-8). Massachusetts Institute of Technology, Cambridge, MA.

Moher, D., Liberati, A., Tetzlaff, J., Altman, D.G., 2009. Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. PLoS Med. 6 (7), e1000097. http://dx.doi.org/10.1371/journal.pmed.1000097.

*Munger, D., Mehler, B., Reimer, B., Dobres, J., Pettinato, A., Pugh, B., Coughlin, J.F., 2014. A simulation study examining smartphone destination entry while driving. Proceedings of the 6th International Conference on Automotive User Interfaces and Interactive Vehicular Applications - AutomotiveUI '14. ACM Press, New York, New York, USA, pp. 1–5. http://dx.doi.org/10.1145/2667317.2667349.

National Highway Traffic Safety Association, 2012. Visual-manual NHTSA driver distraction guidelines for in-vehicle electronic devices. Docket No. NHTSA-2010-0053. Department of Transportation, Washington, D.C.

National Highway Traffic Safety Association, 2014. Visual-manual NHTSA driver distraction guidelines for portable and aftermarket electronic devices. Docket No. NHTSA-2013-0137. Department of Transportation, Washington, D.C.

National Highway Traffic Safety Association (NHTSA), 2015. Traffic Safety Facts Research Note: Distracted Driving 2013. NHTSA, Washington, DC.

*Neurauter, M., Hankey, J., Schalk, T., Wallace, G., 2012. Outbound texting. Transport. Res. Rec. 2321 (1), 23–30. http://dx.doi.org/10.3141/2321-04.

Pickrell, T., KC, S., 2015. Driver electronic device use in 2014 (Traffic Safety Factos: Research Note DOT HS 812 197). Washington, DC.

*Ranney, T.A., Harbluk, J.L., Ian Noy, Y., 2005. Effects of voice technology on test track driving performance: implications for driver distraction. Human Factors 47 (2), 439–454. http://dx.doi.org/10.1518/0018720054679515.

*Reimer, B., Mehler, B., Coughlin, J.F., Roy, N., Dusek, J.A., 2011. The impact of a naturalistic hands-free cellular phone task on heart rate and simulated driving performance in two age groups. Transport. Res. Part F: Traffic Psychol. Behav. 14 (1), 13–25. http://dx.doi.org/10.1016/j.trf.2010.09.002.

*Reimer, B., Mehler, B., D'Ambrosio, L.A., Fried, R., 2010. The impact of distractions on young adult drivers with attention deficit hyperactivity disorder (ADHD). Accident Anal. Prevent. 42 (3), 842–851. http://dx.doi.org/10.1016/j.aap.2009.06.021.

*Reimer, B., Mehler, B., Dobres, J., Coughlin, J.F., 2013. The Effects of a Production Level "Voice-Command" Interface on Driver Behavior: Reported Workload, Physiology, Visual Attention, and Driving Performance (Technical Report 2013-17A). Massachusetts Institute of Technology, Cambridge, MA https://doi.org/2013-18A.

*Reimer, B., Mehler, B., Dobres, J., Coughlin, J.F., 2015. Assessing the Demands of Voice Based In-Vehicle Interfaces - Phase II Experiment 4 - An Exploratory Study of Driver Behavior With and Without Assistive Cruise Control (ACC) (2015a) (MIT AgeLab Technical Report 2015-15). Massachusetts Institute of Technology, Cambridge, MA.

*Reimer, B., Mehler, B., Reagan, I., Kidd, D., Dobres, J., 2016. Multi-modal demands of a smartphone used to place calls and enter addresses during highway driving relative to two embedded systems. Ergonomics 1–21. http://dx.doi.org/10.1080/00140139.2016.1154189.

Rosenthal, R., DiMatteo, M.R., 2001. Meta-analysis: recent developments in quantitative methods for literature reviews. Annu. Rev. Psychol. 52, 59–82.

International, S.A.E., 2015. Surface Vehicle Recommended Practice: Operational Definitions of Driving Performance Measures and Statistics. Warrendale, PA.

*Salvucci, D., 2001. Predicting the effects of in-car interface use on driver performance: an integrated model approach. Int. J. Human-Comput. Stud. 55 (1), 85–107. http://dx.doi.org/10.1006/ijhc.2001.0472.

*Salvucci, D.D., Macuga, K.L., 2002. Predicting the effects of cellular-phone dialing on driver performance. Cognit. Syst. Res. 3 (1), 95–102. http://dx.doi.org/10.1016/S1389-0417(01)00048-1.

Schmidt, F., Hunter, J. (Eds.), 2014. Methods of Meta-Analysis: Correcting Error and Bias in Research Findings, 3rd ed. Sage Publications, Inc, Thousand Oaks, CA.

*Schreiner, C., Blanco, M., Hankey, J.M., 2004. Investigating the effect of performing voice recognition tasks on the detection of forward and peripheral Events. Proceedings of the Human Factors and Ergonomics Society Annual Meeting, Vol. 48 2354–2358. http://dx.doi.org/10.1177/154193120404801932.

*Schreiner, C.S., 2006. The effect of phone interface and dialing method on simulated driving performance and user preference. Proceedings of the Human Factors and Ergonomics Society Annual Meeting, Vol. 50 2359–2363. http://dx.doi.org/10.1177/154193120605002202.

*Serafin, C., Wen, C., Paelke, G., Green, P., 1993. Car phone usability: a human factors laboratory test. Proceedings of the Human Factors and Ergonomics Society 37th Annual Meeting 220–224.

Shutko, J., Mayer, K., Lansoo, E., Tijerina, L., 2009. Driver workload effects of cell phone, music player, and text messaging tasks with the Ford SYNC voice interface versus hand-held visual-manual interfaces (Paper No. 2009-01-0786). Society of Automotive Engineering.

Simmons, S., Hicks, A., Caird, J.K., 2016. Safety critical events associated with cell phone distraction tasks as measured through naturalistic studies: a systematic review and meta-analysis. Accident Anal. Prevent. 87, 161–169.

Steel, P.D., Kammeyer-Mueller, J.D., 2002. Comparing meta-analytic moderator estimation techniques under realistic conditions. J. Appl. Psychol. 87 (1), 96.

*Strayer, D., Cooper, J.M., Turrill, J., Coleman, J.R., Hopman, R.J., 2015a. The Smartphone and the Driver's Cognitive Workload: A Comparison of Apple, Google, and Microsoft's Intelligent Personal Assistants. AAA Foundation for Traffic Safety, Washington, DC.

Strayer, D.L., Cooper, J.M., Turrill, J., Coleman, J., Hopman, R.J., 2016. Measuring cognitive distraction in the automobile. Cognit. Res.: Principles Implications 1 (16), 1–17. http://dx.doi.org/10.1186/s41235-016-0018-3.

*Strayer, D.L., Cooper, J.M., Turrill, J., Coleman, J., Medeiros-Ward, N., Biondi, F., 2013. Measuring Cognitive Distraction in the Automobile. AAA Foundation for Traffic Safety, Washington, DC.

*Strayer, D.L., Cooper, J.M., Turrill, J., Coleman, J., Medeiros-Ward, N., Biondi, F., 2015b. Measuring Cognitive Distraction in the Automobile III: A Comparison of Ten 2015 In-Vehicle Information Systems. AAA Foundation for Traffic Safety, Washington, DC.

*Strayer, D.L., Turrill, J., Coleman, J.R., Ortiz, E.V., Cooper, J.M., 2014. Measuring Cognitive Distraction in the Automobile II: Assessing In-Vehicle Voice-Based Interactive Technologies. AAA Foundation for Traffic Safety, Washington, DC.

Sutton, A.J., 2009. Publication bias. In: Cooper, H., Hedges, L.V., Valentine, J.C. (Eds.), The Handbook of Research Synthesis and Meta-analysis, 2nd ed. Russell Sage Foundation, New York, pp. 435–452.

*Terken, J., Visser, H.-J., Tokmakoff, A., 2011. Effects of speech-based vs handheld e-mailing and texting on driving performance and experience. Proceedings of the 3rd International Conference on Automotive User Interfaces and Interactive Vehicular Applications - AutomotiveUI '11. ACM Press, New York, New York, USA, pp. 21. http://dx.doi.org/10.1145/2381416.2381419.

Tijerina, L., 2016. Driver distraction and road safety. In: Smiley, A. (Ed.), Human Factors in Traffic Safety, 3rd ed. Lawyers and Judges Publishing, Tuscon, AZ, pp. 219–276.

*Törnros, J.E.B., Bolling, A.K., 2005. Mobile phone use—effects of handheld and handsfree phones on driving performance. Accident Anal. Prevent. 37 (5), 902–909. http://dx.doi.org/10.1016/j.aap.2005.04.007.

*Truschin, S., Schermann, M., Goswami, S., Krcmar, H., 2014. Designing interfaces for multiple-goal environments. ACM Trans. Comput.-Human Interact. 21 (1), 1–24. http://dx.doi.org/10.1145/2544066.

*Tsimhoni, O., Smith, D., Green, P., 2004. Address entry while driving: speech recognition versus a touch-screen keyboard. Human Factors 46 (4), 600–610. http://dx.doi.org/10.1518/hfes.46.4.600.56813.

White, C., Caird, J.K., 2010. The blind date: the effects of change blindness, passenger conversation and gender on looked-but-failed-to-see (LBFTS) errors. Accident Anal. Prevent. 42, 1822–1830. http://dx.doi.org/10.1016/j.aap.2010.05.003.

Whitener, E.M., 1990. Confusion of confidence intervals and credibility intervals in meta-analysis. J. Appl. Psychol. 75, 315–321.

Wickens, C.D., 2002. Multiple resources and performance prediction. Theoretical Issues Ergon. Sci. 3 (2), 159–177.

Wickens, 2008. Multiple resources and mental workload. Human Factors 50 (3), 449–455.

Wickens, C.D., Hollands, J.G., Bansbury, S., Parasuraman, R., 2013. Engineering Psychology and Human Performance, 4th ed. Pearson, Montréal.

Wilson, F.A., Stimpson, J.P., 2010. Trends in fatalities from distracted driving in the United States, 1999 to 2008. Am. J. Public Health 100 (11), 2213–2219. http://dx.doi.org/10.2105/AJPH.2009.187179.

World Health Organization (WHO), 2015. Global status report on road safety 2015. Geneva, Switzerland.

*Yager, C., 2013. An Evaluation of the Effectiveness of Voice-To-Text Programs at Reducing Incidences of Distracted Driving (Report 600451-00011-1). Texas A & M Transportation Institute, College Station, TX. Retrieved from http://swutc.tamu.edu/publications/technicalreports/600451-00011-1.pdf.