Reliability and validity of on-road driving tests in vulnerable adults: a systematic review
Tatsunori Sawada, Kounosuke Tomori, Haruka Hamana, Kanta Ohno, Yosuke Seike, Yo Igaric and Yoshio Fujita

The on-road driving test is considered a ‘gold standard’ evaluation; however, its validity and reliability have not been sufficiently reviewed. This systematic review aimed to map out and synthesize literature regarding on-road driving tests using the Consensus-based Standards for the Selection of Health Measurement Instruments checklist. Cochrane Library, PubMed, CINAHL, and Web of Science databases were searched from initiation through February 2018. All articles addressing reliability or validity of on-road driving tests involving adult rehabilitation patients were included. The search output identified 513 studies and 36 articles, which were included in the review. The Washington University Road Test/Rhode Island Road Test, performance analysis of driving ability, test ride for investigating practical fitness-to-drive, and K-score demonstrated high reliability and validity in regard to the Consensus-based Standards for the Selection of Health Measurement Instruments checklist. The Washington University Road Test/Rhode Island Road Test and test ride for investigating practical fitness-to-drive were analyzed based on Classical Test Theory techniques, and performance analysis of driving ability and K-score were analyzed based on Item Response Theory techniques. The frequency of studies were Washington University Road Test/Rhode Island Road Test (n=9), Test Ride for Investigating Practical fitness-to-drive (n=8), performance analysis of driving ability (n=4), and K-score (n=1).

From the viewpoint of accuracy and generalization, the Washington University Road Test/Rhode Island Road Test, test ride for investigating practical fitness-to-drive, and performance analysis of driving ability were identified as highly qualified concerning on-road driving tests. However, the ability to assess real-world driving depends on various environmental conditions. International Journal of Rehabilitation Research 42: 289–299 Copyright © 2019 The Author(s). Published by Wolters Kluwer Health, Inc.

Keywords: automobile driver examination, neuropsychological tests, patient outcome assessment, systematic review, validation studies

Introduction
With technological advances, safe driving has become more possible, and the number of traffic accidents is decreasing. Simultaneously, however, the driver population has increased and changed with becoming an inclusive society. The number of elderly drivers has almost doubled since the 1990s, and people with disabilities, such as those post-stroke, hope to return to driving (Yu et al., 2016). The evidence shows that driving cessation in these people contributes to a variety of health problems, particularly depression or functional limitation (Chihuri et al., 2016; Shimada et al., 2016). However, Azami-Aghdash et al. (2018) reported that the traffic-related mortality rate in elderly people is almost twice that of the non-elderly [odd = 2.57 (1.2–5.4 95% CI)]. Actually, 50% of countries recorded a rise in the number of road deaths among elderly people, and in 30% of countries, the elderly have the highest mortality rate in traffic of all age groups (Forum, 2018). Elderly people have become more mobile and more exposed to traffic risks. Therefore, it is of great importance to accurately evaluate the driving ability of these individuals.

There are two types of evaluations for testing driving skills: off-road and on-road tests. Off-road tests assess driving skills related to cognitive ability by paper-based or computer-based testing. Bliokas et al. (2011) demonstrated that some neuropsychological measures could predict the pass/fail classification of the on-road test with 73% sensitivity and 76% specificity. The Stroke Drivers Screening Test was developed to predict stroke patient’s driving ability. It was determined to be useful for not only assessing cognitive ability but also for predicting on-road driving ability (Nouri and Lincoln, 1992; Edwards et al., 2005). Neuropsychological measures are an important component of a multidisciplinary approach for evaluation of driving capacity (Wolfe and Lehockey, 2016).
In the on-road study, Brooke et al. (1992) clarified that for closed head injury patients, off-road tests did not correspond to a pass/fail rating of an on-road test. Fox et al. (1998) showed that on-road driving assessments examined proficiency in operating a motor vehicle, but not the ability to drive in traffic, and, thus, were not an accurate prediction of safe driving. Marshall et al. (2007) reported on-road testing as the ‘gold standard’ of driving ability in his review because off-road tests are not always appropriate for understanding one’s actual driving capabilities. These results show that it is important to evaluate vulnerable adults’ driving ability using multiple perspectives (on-road and off-road). Several off-road tests have demonstrated their reliability and validity, and it is useful for predicting patient’s driving ability by systematic review and meta-analysis (Reger et al., 2004; Devos et al., 2011; Hird et al., 2016). However, the reliability and validity of on-road tests have not been adequately researched. To the best of our knowledge, no prior systematic review has solely verified on-road test reliability and validity. Although some studies have assessed on-road test reliability and validity, it is still unclear which on-road tests are most reliable and valid. These problems of on-road tests cause ambiguity in testing the real driving ability. Therefore, the aim of the current systematic review was follows:

1. to map out and synthesize literature on on-road driving
2. to clarify which on-road tests are most reliable and valid

**Method**

**Research design**

This systematic review focuses on the reliability and validity of on-road tests in consideration of various health conditions and aging in relation to driving. A systematic review was conducted, focusing on reliability and validity of on-road tests in reference to the Preferred Reporting Items for Systematic Reviews and Meta-Analysis statement (Liberati et al., 2009; Moher et al., 2009). The Consensus-based Standards for the Selection of Health Measurement Instruments (COSMIN) guidelines and recommendations for evaluating methodological quality were followed.

**Inclusion and exclusion criteria**

Peer-reviewed in current study, only academic papers describing on-road driving tests for individuals with various health conditions were assessed. Systematic literature reviews, study protocols, conference proceedings, commentary papers, studies with simulated on-road tests (excluded for the reliability and validity of on-road test), off-road tests developed by the authors for this study, and studies without reliability or validity assessments were excluded (e.g. simply comparing pass or fail groups by using the result of on-road tests). Criterion validity usually means comparing on-road with on-road tests. However, many studies compared on-road with off-road tests. We included the off-road test studies that assessed criterion validity for comparison with on-road tests (e.g. combining several neuropsychological assessments, Useful Field of View Test), even if the aim of the study was not to investigate on-road tests. Therefore, we changed the check item of reliability to ‘on-road to on-road’ and ‘on-road to off-road.’ Only studies published in English were included.

**Literature search**

We searched the Cochrane Library, PubMed, CINAHL, and Web of Science databases using keywords (Table 1) for searching the relevant articles on 21 February 2018. Search words included driving, road, route, way, motor vehicles, automobile, measurement, outcome, test, and assessment. Disease terms included stroke, traumatic brain injury, mild cognitive impairment, dementia, Alzheimer’s, cognitive dysfunction, physical dysfunction, spinal cord injury, and elderly (Table 1). We did not adopt the term ‘reliability’ or ‘validity’ because we tried to find as many on-road tests as possible. If eligible articles had studies on reliability or validity in their references, those references were checked in a manual search. In this case, only published research articles were adopted (e.g. excluding PhD theses).

**Eligibility criteria**

The titles and abstracts were first reviewed by two authors independently, following the removal of duplicates. If they were unsure whether the article met the criteria, the two authors independently screened the full-text papers and confirmed the articles that met include/exclude criteria. Differences of opinion were resolved through discussion with other reviewers. A structured abstract was then created considering the article objective, subject, method, and results.

**Methodological quality evaluation**

The COSMIN is one valid methodology for determining study quality (Prinsen et al., 2018). The COSMIN consists of 4 steps (Fig. 1). At first, measurement properties (Box A to I) are evaluated in each article (Step 1). Although the responsiveness and interpretability are neither reliability nor validity items, the COSMIN checklist adopted these items for standardization. There are three

| Table 1 | Search strategy |
|---------|----------------|
| Driving AND (‘road’ or ‘route’ or ‘way’) AND (‘motor vehicles’ or ‘automobile’) AND (‘stroke’ or ‘traumatic brain injury’ or ‘mild cognitive impairment’ or ‘dementia’ or ‘Alzheimer’s’ or ‘cognitive dysfunction’ or ‘physical dysfunction’ or ‘spinal cord injury’ or ‘elderly’) AND (‘measurement’ or ‘outcome’ or ‘test’ or ‘assessment’) |
Reliability and validity of on-road driving tests

Sawada et al. 291

types of reliability (test-retest, interrater, and intrarater) in Box B. If the statistical method used in the article was based on item response theory (IRT), it was evaluated by the IRT box (4 items) in Step 2 (Table 2). We performed the systematic review protocol in reference to a previous study methodology using a COSMIN 4-point modular checklist (Mokkink et al., 2010; Wales et al., 2016). Each checklist item in COSMIN is scored on a four-point ordinal rating scale to evaluate the methodological quality of each measurement item’s property: excellent, good, fair, or poor. A score for a given box was obtained by using the lowest score for any item (‘worst score counts method’).

If one item is scored as ‘poor,’ the overall score for the study on that box will be ‘poor’ (Mokkink et al., 2010). More than good was regarded as high quality. This scoring method is used in Steps 2 and 3. In Step 3, the evaluator completes the corresponding boxes marked in Step 1. Each corresponding box should be completed for each measurement property that was detected in Step 1. The researcher determines if the measurement properties were assessed according to the standards for methodological quality in Step 3. We showed the measurement property Box B as an example of method in Table 3. Finally, the generalizability box must be checked. Since there is

---

**Table 2** Step 2. Determining if the statistical method used in the article are based on classical test theory or item response theory

| Step | Condition | Excellent | Good | Fair | Poor |
|------|-----------|-----------|------|------|------|
| 1.   | Was the IRT model used adequately described? e.g. OPLM, Partial Credit Model, Graded Response Model | IRT model adequately described | IRT model not adequately described |
| 2.   | Was the computer software package used adequately described? e.g. RUMM2020, WINSTEPS, OPLM, MULTLOG, PARSSCALE, BILOG, NLMIXED | Software package adequately described | Software package not adequately described |
| 3.   | Was the method of estimation used adequately described? e.g. conditional maximum likelihood, marginal maximum likelihood | Method of estimation adequately described | Method of estimation not adequately described |
| 4.   | Were the assumptions for estimating parameters of the IRT model checked? e.g. unidimensionality, local independence, and item fit (e.g. differential item functioning) | Assumptions of the IRT model checked | Assumptions of the IRT model partly checked | Assumptions of the IRT model not checked or unknown |

IRT, item response theory; OPLM, one-parameter logistic model.

---

COSMIN checklist process. COSMIN is standardized checklist methodology of the reliability and validity. It consists of four steps. COSMIN, Consensus-based Standards for the Selection of Health Measurement Instruments; IRT, item response theory.
no scoring system in this box, it was recommended by the previous study to use this box for extracting data on the characteristics of the study (Terwee et al., 2012). We adopted this methodology.

Although COSMIN was developed for Patient-Reported Outcomes, a COSMIN checklist can be used for functional assessments, as has been done by occupational therapists for determining severe criteria within a systematic review (Wales et al., 2016). As the present study assessed reliability and validity from a standard point of view, even if the assessment does not focus on patient-reported outcomes, COSMIN criteria are useful for on-road test investigations.

The authors also independently assessed specific properties of each on-road assessment. When results of this assessment were difficult to decipher, additional authors were included to arrive at a consensus.

**Results**

**Study selection**

A total of 513 papers were initially screened. A flow diagram depicting study selection is shown in Fig. 2. Initially, we selected 342 individual studies. We filtered this number down to 64 studies, excluding 282 studies that failed to meet criteria based on the titles and abstracts, as well as 171 duplicates. We then read full texts of these remaining papers. An additional 9 studies were included after a manual search. Next, we filtered down to studies that included reliability or validity assessments. This left us with 37 total studies (three duplicates).

**Measurement properties for identified on-road assessments (Step 1)**

Twenty-nine types of on-road tests were identified (Table 4). Some had various versions. For instance, the Test Ride for Investigating Practical fitness-to-drive (TRIP) Belgian Version 3 (named for purposes of the present review) was the most frequently assessed (Table 4). The Washington University Road Test (WURT) and the TRIP had the most versions (different number of items). There were 8 TRIP studies and 4 Rhode Island Road Test (RIRT) studies. However, the RIRT included the same items as the WURT to be applicable to Rhode Island (Brown et al., 2005; Ott et al., 2008). Therefore, we regarded these two on-road tests as the same one. When combining the WURT and RIRT, 9 studies were included.

---

**Table 3** Step 3. Determining if a study meets the standards for good methodological quality (Box A. Internal consistency)

| Item 1 is used to determine whether internal consistency is relevant for the instrument under study. It is not used to rate the quality of the study. |
|---|
| 1. Does the scale consist of effect indicators, i.e. is it based on a reflective model? |
| Design requirements |
| 2. Was the percentage of missing items given? |
| 3. Was there a description of how missing items were handled? |
| 4. Was the sample size included in the internal consistency analysis adequate? |
| 5. Was the unidimensionality of the scale checked? i.e. was factor analysis or IRT model applied? |
| 6. Was the sample size included in the unidimensionality analysis adequate? |
| 7. Was an internal consistency statistic calculated for each (unidimensional) (sub)scale separately? |
| 8. Were there any important flaws in the design or methods of the study? |
| Statistical methods |
| 9. For CTT, continuous scores: Was Cronbach’s alpha calculated? |
| 10. For CTT, dichotomous scores: Was Cronbach’s alpha or KR-20 calculated? |
| 11. For IRT: Was a goodness of fit statistic at a global level calculated? |

| Percentage of missing items described |
|---|
| Percentage of missing items NOT described |
| Not clear how missing items were handled |
| Adequate sample size (≥ 100) |
| Authors refer to another study in which factor analysis was performed, but not in a similar study population |
| Authors refer to another study in which factor analysis was performed, |
| No other important methodological flaws in the design or execution of the study |
| Only item-total correlations calculated |
| Goodness of fit statistic at a global level calculated |

Cronbach’s alpha calculated |

Only item-total correlations calculated |

No Cronbach’s alpha and no item-total correlations calculated |

Other minor methodological flaws in the design or execution of the study |

Internal consistency statistic NOT calculated for each subscale separately |

Other important methodological flaws in the design or execution of the study |

Internal consistency statistic calculated for each subscale separately |

No other important methodological flaws in the design or execution of the study |

No Cronbach’s alpha or KR-20 calculated |

Goodness of fit statistic at a global level NOT calculated |

Cronbach’s alpha or KR-20 calculated |

Only item-total correlations calculated |

No Cronbach’s alpha or KR-20 and no item-total correlations calculated |

Small sample size (< 30) |

< 5 number of items |

< 100 |

≥ 100 OR 6–7 number of items but < 100 |

≥ 100 |

7 number of items and ≥ 100 |

5 number of items and ≥ 100 OR 6–7 number of items but < 100 |

< 5 number of items |

< 100 OR 6–7 number of items but < 50 |

< 5 number of items |

< 100 OR 6–7 number of items but < 50 |

< 5 number of items |

< 100 OR 6–7 number of items but < 50 |

< 5 number of items |

< 100 OR 6–7 number of items but < 50 |

< 5 number of items |
Reliability and validity of on-road driving tests
Sawada et al. 293

Item response theory analysis of each on-road test (Step 2)
Performance Analysis of Driving Ability (P-drive) and K-score were analyzed by IRT method. These two on-road tests indicated good to excellent quality in IRT items (Table 5) (Patomella et al., 2004; Kay et al., 2008). Most studies used classical test theory.
| On-road test | Author Year | Object n | Items | Total score | Age (range) | Female % | Setting | Country |
|--------------|-------------|----------|-------|-------------|------------|----------|---------|---------|
| WURT Pilot   | Hunt et al 1993 | Alzheimer’s disease 38a | 38 | | 73.4 ± 6.2 | 55.3% | Designed route urban | United States |
|             | Original Hunt et al 1997 | Alzheimer’s disease 63–121 | 54 | | b 108 | b 74.2 ± 9.0 (52–90) | 37.0% | 9.6 km urban course | United States |
|             | Modified Carr et al 2011 | Dementia 41 | b | | b | | b | 12 miles closed and open course | United States |
|             | Modified Barco et al 2014 | Stroke 33 | b | | b | | b | 59.3 ± 13.0 (31–88) | 46.0% | 13 miles route including parking lot | United States |
| RRT Original | Brown et al 2005 | Alzheimer’s disease 20 | 54 | 108 | 73.2 ± 8.3 | 41.2% | Daylight hours under good | United States |
|             | Original Brown et al 2005 | Very mild dementia 55a | 54 | 108 | 72.0–76.9 | 45.5% | Daylight hours under good | United States |
|             | Original Ott et al 2008 | Alzheimer’s disease 20–121 | 54 | 108 | 73.6–75.8 | 45.1% | Daylight hours under good | United States |
|             | Version 2 Ott et al 2012 | Cognitive impairment 80a | 28 | 960 | 73.1 ± 7.3 | 46.2% | 6.5 miles of urban terrain | United States |
|             | Version 2 Davis et al 2012 | Cognitive impairment 103a | 28 | 960 | 73.9 ± 7.2 (80–90) | 49.2% | 6.5 miles of urban terrain | United States |
| TRIP Original | Tant et al 2002 | Homonymous hemianopia 28 | 55 | 220 | 53.0 (24–76) | 21.4% | b | Netherlands |
|             | Belgian version 1 De Raedt and Ponjaert-Kristoffersen, et al 2001 | Old drivers 11–32 84 | 11 | 209 | 78.6 ± 6.8 (65–96) | 28.6% | b | Standardized 35 km route including highway | Belgium |
|             | Belgian version 1 Stapleton et al 2012 | Stroke 11–32 11 | 11 | 209 | 63.5 ± 13.4 (29–83) | 19.6% | Participant’s own home town area | Ireland |
|             | Belgian version 2 Akinwuntan et al 2003 | Stroke 27 | 55 | 220 | 60.0 ± 2.6 | 18.5% | 20 km including closed courses and highway | Belgium |
|             | Belgian version 3 Akinwuntan et al 2005 | Stroke 38 | 49 | 196 | 53.9 ± 12.8 (24–73) | 18.4% | 17 km including premise and highway | Belgium |
|             | Belgian version 3 Akinwuntan et al 2006 | Stroke 68 | 49 | 196 | 53.0 ± 13.0 | 16.2% | b | Standardized 20 km road | Belgium |
|             | Belgian version 3 Devos et al 2014 | Huntington’s disease 30 | 49 | 196 | 50.2 ± 12.4 | 26.7% | b | Standardized 20 km road | Belgium |
|             | Belgian version 3 Devos et al 2017 | Multiple sclerosis 102 | 49 | 196 | 47.8 ± 2.7 (26–65) | 86.0% | b | Country and urban route | United States |
|             | Original Patomella et al 2004 | Brain damage 31 | 21 | 84 | 87.0 ± 12.2 (22–77) | 29.0% | b | Simulator fixed route | Sweden |
|             | Version 2 Patomella et al 2010 | Stroke, MCI, dementia 205 | 27 | 104 | 69.0 ± 11.0 (33–86) | 16.0% | b | set route | Sweden |
|             | Version 2 Seilandera et al 2011 | Old drivers 85 | 27 | 104 | 72.0 ± 5.3 (65–85) | 47.0% | b | 39.7 km fixed route | Sweden |
|             | Version 3 Vaucher et al 2015 | Old drivers 24 | 26 | 104 | 77 (75–85) | 4.0% | b | 21 km including highway | Switzerland |
|             | NNDA Radford et al 2004 | Parkinson’s disease 49 | 25 | b | 64.4 ± 9.1 (44–85) | 20.0% | b | Standard route | United Kingdom |
|             | Lincol et al 2012 | Dementia 6 | 25 | b | (70–85) | 16.7% | b | Standard route | United Kingdom |
|             | UWO Classen et al 2016 | Multiple sclerosis 34a | 41 | b | 48.3 ± 9.8 | 60.0% | Including parking lot, highway | Canada |
|             | Classen et al 2017 | Multiple sclerosis 34a | 41 | b | b | 58.8% | Including parking lot, highway | Canada |
|             | PBDE Odenheimer et al 1994 | Dementia 26–30 | 75 | b | 72.2 (61–89) | 13.0% | 23 miles fixed route including parking lot, Saturday morning | United States |
|             | SRT Fitten et al 1995 | Alzheimer’s disease 8–43 | 41 | b | b | 10 miles, week day | 27% | Unstructural route United States |
|             | HES Dobbs et al 1998 | Cognitive impairment 253a | 37 | b | 72.7 ± 9.1 | 27.7% | Closed and open-road course | Canada |
|             | MDPE Janke et al 1998 | Old drivers 106 | 6 | b | 75.7 (60–91) | 36.0% | Fixed route | United States |
|             | ADPE Janke et al 1998 | Old drivers 106 | 6 | b | b | 20 miles including parking lot, highway | United States |
|             | NHS Richardson et al 2003 | Old drivers 26–357 | 36 | 72 | b | b | 20 miles including parking lot, highway | United States |
|             | OTADP Mallon et al 2004 | Old drivers 137a | 106 | 100 (%) | 70.6 ± 6.2 | b | 15 km predetermined course | United States |
Confirming for each property of box (Step 3)

Reliability

In the number of studies, interrater reliability was tested most frequently (13 on-road tests), followed by internal consistency (seven on-road tests), and test-retest reliability (three on-road tests). There were no studies regarding intrarater reliability and measurement error. In total, WURT/RIRT and Behind-the-Wheel Driving Performance Assessment had the most COSMIN checklist reliability items (3 items).

In the quality of item, five on-road tests (Ott et al., 2012), TRIP (De Raedt and Ponjaert-Kristoffersen, 2001), P-drive (Patomella et al., 2010), Behind-the-Wheel Driving Performance Assessment (Justiss et al., 2006), and K-score (Kay et al., 2008) indicated high quality of internal consistency items (good or excellent). Only WURT indicated high quality in the interrater and test-retest items (Table 5).

WURT/RIRT had the highest quality items (two good and one excellent) in COSMIN reliability.

Validity

In a number of studies, criterion validity was examined most frequently. There were 9 studies comparing on-road tests to on-roads tests in existence (Selander et al., 2011; Vaucher et al., 2015) and there were 14 studies comparing on-road tests to off-road tests (Hunt et al., 1993). Structural validity was studied in P-drive (two studies) and K-score (one study) using IRT. Hypotheses testing was studied in TRIP (De Raedt and Ponjaert-Kristoffersen, 2001) and Sepulveda Road Test (Fitten et al., 1995) (each one study). Content validity was examined only in one study (University of Western Ontario’s on-road assessment) (Classen et al., 2017). There were no studies that verified cross-cultural validity, responsiveness, and interpretability.

In the quality of items, WURT/RIRT (Ott et al., 2008; Hunt et al., 1997), Hazardous Error Score (Dobbs et al., 1998), Occupational Therapy Assessment of Open-Road Driving Performance (Mallon and Wood, 2004), and Ryd On-road Assessment had high quality (good) in criterion validity items. In the structural validity items, P-drive and K-score showed high quality. TRIP (De Raedt and Ponjaert-Kristoffersen, 2001) had good quality in the hypotheses testing items.

P-drive had the highest quality items (one good and one excellent) in COSMIN validity.

Generalizability (Step 4)

Some on-road tests had more than two studies: WURT/RIRT (9), TRIP (8), P-drive (4), Nottingham Neurological Driving Assessment (2), University of Western Ontario’s on-road assessment (2); others had only one. Although the subjects’ age in some studies was unclear because of extracting a part of the article, the average age of the
| On-road test | Number of articles | IRT | Internal consistency | Inter-rater reliability | Intrarater reliability | Retest reliability | Measurement error | Hypotheses testing | Content validity | Structural validity | Cross-cultural validity | On-road | Off-road | Responsiveness | Interpretability |
|--------------|-------------------|-----|----------------------|-------------------------|------------------------|-------------------|------------------|-------------------|----------------|----------------|----------------------|----------|-----------|---------------|-----------------|
| Pilot        | 1                 |     | +                    |                         |                        |                   |                  |                   |                |                |                      |         |           |               |                 |
| Original     | 1                 |     | ++                   |                         |                        |                   |                  |                   |                |                |                      |         |           |               |                 |
| Modified     | 2                 |     | +, +                 |                         |                        |                   |                  |                   |                |                |                      |         |           |               |                 |
| RIRT         |                   |     |                      |                         |                        |                   |                  |                   |                |                |                      |         |           |               |                 |
| Original     | 3                 |     | ±, ±                 |                         |                        |                   |                  |                   |                |                |                      |         |           |               |                 |
| Version 2    | 2                 |     | ++                   |                         | ±                       |                   |                  |                   |                |                |                      |         |           |               |                 |
| TRIP         |                   |     |                      |                         |                        |                   |                  |                   |                |                |                      |         |           |               |                 |
| Original     | 1                 |     | ±                     |                         |                        |                   |                  |                   |                |                |                      |         |           |               |                 |
| Belgian      | 2                 |     | ++                   |                         | ±                       |                   |                  |                   |                |                |                      |         |           |               |                 |
| Version 1    |                   |     |                      |                         |                        |                   |                  |                   |                |                |                      |         |           |               |                 |
| Belgian      | 1                 |     | ±                     |                         |                        |                   |                  |                   |                |                |                      |         |           |               |                 |
| Version 2    |                   |     |                      |                         |                        |                   |                  |                   |                |                |                      |         |           |               |                 |
| Belgian      | 4                 |     | +                    |                         |                        |                   |                  |                   |                |                |                      |         |           |               |                 |
| Version 3    |                   |     |                      |                         |                        |                   |                  |                   |                |                |                      |         |           |               |                 |
| P-Drive      |                   |     |                      |                         |                        |                   |                  |                   |                |                |                      |         |           |               |                 |
| Original     | 1                 |     | +++                  | ±                       |                        |                   |                  |                   |                |                |                      |         |           |               |                 |
| Version 2    | 2                 |     | +++                   | +++                     | +                      |                   | ±                |                   |                |                |                      |         |           |               |                 |
| Version 3    | 1                 |     | ++                   |                        | ±                       |                   |                  |                   |                |                |                      |         |           |               |                 |
| NNDA         | 2                 |     | ±                     |                         |                        |                   |                  |                   |                |                |                      |         |           |               |                 |
| UWO          | 2                 |     | +                     |                         |                        |                   | ±                |                   |                |                |                      |         |           |               |                 |
| PBDE         | 1                 |     | ±                     |                         | ±                       |                   | ±                |                   |                |                |                      |         |           |               |                 |
| SRT          | 1                 |     | ±                     |                         |                        |                   | ±                |                   |                |                |                      |         |           |               |                 |
| HES          | 1                 |     | +                     |                         |                        |                   | ±                |                   |                |                |                      |         |           |               |                 |
| MDPE         | 1                 |     | +                     |                         |                        |                   | ±                |                   |                |                |                      |         |           |               |                 |
| ADPE         | 1                 |     | +                     |                         |                        |                   |                    |                   |                |                |                      |         |           |               |                 |
| NHS          | 1                 |     | +                     |                         |                        |                   | ±                |                   |                |                |                      |         |           |               |                 |
| OTADP        | 1                 |     | +                     |                         |                        |                   | ±                |                   |                |                |                      |         |           |               |                 |
| RIDE         | 1                 |     | +                     |                         |                        |                   | ±                |                   |                |                |                      |         |           |               |                 |
| BTW          | 1                 |     | +                     | +                       | ±                       |                   | ±                |                   |                |                |                      |         |           |               |                 |
| Kascoe       | 1                 |     | +++                   | +                       | +                       |                   | ±                |                   |                |                |                      |         |           |               |                 |
| BOST         | 1                 |     | +                     |                         |                        |                   | ±                |                   |                |                |                      |         |           |               |                 |
| ROA          | 1                 |     | +                     |                         |                        |                   | ±                |                   |                |                |                      |         |           |               |                 |
| JRHREF       | 1                 |     | +                     |                         |                        |                   | ±                |                   |                |                |                      |         |           |               |                 |
| DBOG         | 1                 |     | ±                     |                         |                        |                   | ±                |                   |                |                |                      |         |           |               |                 |
| RODE         | 1                 |     | ±                     |                         |                        |                   | ±                |                   |                |                |                      |         |           |               |                 |

Criterion validity regarding on-road means comparing on-road test with on-road test and off-road means comparing on-road tests with off-road tests. ADPE, Area Driving Performance Evaluation; BOST, Basic Operator Skills Test; BTW, Behind-the-Wheel Driving Performance Assessment; DBOG, Driving Behaviors Observation Grid; HES, Hazardous Error Score; IRT, item response theory; JRHREF, Jewish Rehabilitation Hospital Road Evaluation Form; MDPE, Modified Driving Performance Evaluation; NHS, New Haven Score; NNDA, Nottingham Neurological Driving Assessment; OTADP, Occupational Therapy Assessment of Open-Road Driving Performance; P-drive, Performance Analysis of Driving Ability; PBDE, Performance-Based Driving Evaluation; RIDE, Rhode Island Driving Evaluation; RIRT, Rhode Island Road Test; ROA, Ryd On-road Assessment; RODE, Record of Driving Errors; SRT, Sepulveda Road Test; TRIP, Test Ride for Investigation Practical fitness-to-drive; UWO, University of Western Ontario’s on-road assessment; WURT, Washington University Road Test; ++++, excellent; ++, good; +, fair; ±, poor.
subject in available studies was 65.2 (47.9–78.6; lowest to highest average age).

WURT/RIRT has mostly studies from dementia subjects (one of stroke patients), while TRIP has four studies of stroke patients and four studies with other patient groups (homonymous hemianopia, elderly drivers, Huntington’s disease, and Multiple Sclerosis). Also, other tests included fewer female subjects (less than 40%—Record of Driving Errors, Jewish Rehabilitation Hospital Road Evaluation Form, K-score, Area Driving Performance Evaluation, Modified Driving Performance Evaluation, Hazardous Error Score, PDBE, and Nottingham Neurological Driving Assessment). In the setting of on-road tests, standardized route was 60.0%, fixed distance was 55.0% and both of them was 40% (Table 4). Nottingham Neurological Driving Assessment, Hazardous Error Score, Occupational Therapy Assessment of Open-Road Driving Performance, Behind-the-Wheel Driving Performance Assessment, and Ryd On-road Assessment had high-quality items, but they also consisted of only one item.

**Detail of high-quality on-road tests**

WURT/RIRT and P-drive had four high-quality COSMIN items, followed by K-score and TRIP (three items). However, K-score was conducted in only one study.

In the WURT study, interrater reliability was high (k = 0.85 to 0.96), and test-retest reliability correlations were 0.53 to 0.76. In the criterion validity, the quantitative score from the investigator and the global rating from the driving instructor were highly positively correlated (Kendall T-b = 0.60; P < 0.001) (Hunt et al., 1997). One RIRT study examined internal consistency reliability. Results revealed a homogeneous cluster of 21 RIRT items with a strong intraclass correlation (ICC = 0.40) and high internal consistency (Cronbach’s α = 0.93). Furthermore, Spearman rank correlation between the RIRT and the Composite Driving Assessment Scale (P = 0.62, P < 0.001) indicated criterion validity (Ott et al., 2012).

A Rasch analysis (IRT analysis) has been previously employed to develop the P-drive test. The first study to use this test observed adequate internal consistency and structural validity for the 21 items analyzed (χ² mean square = 0.6–1.3; z = −1–1), while also demonstrating unidimensionality (structural validity) (Patomella et al., 2004). The P-drive Version 2 included additional items 27. This version demonstrated adequate structural validity and internal consistency, as well as unidimensionality (Patomella et al., 2010). Combined sensitivity/specificity curves crossed at 85, providing an optimal cutoff value for the P-drive protocol. In terms of criterion validity, P-drive scores were related to driving instructors’ subjective evaluations (R² = 0.44) (Vaucher et al., 2015).

The TRIP showed Cronbach’s α reliabilities were high (range = 0.86 to 0.97) in internal consistency (De Raedt and Ponjaert-Kristoffersen, 2001). TRIP demonstrated hypotheses testing items, using accidents ratio (De Raedt and Ponjaert-Kristoffersen, 2001). Criterion validity is based on correlations between the tremendous driving assessment (e.g. Useful Field of View Test, Trail Making Test) and TRIP scores (Devos et al., 2017). Another study revealed very similar results in that TRIP produced significant correlations (test for attentional performance, Stroke Drivers’ Screening Assessment; 0.36 to 0.39) (Akinwuntan et al., 2006).

**Discussion**

To our knowledge, this is the first systematic review of studies regarding on-road driving tests. From the 513 studies, 37 were extracted and evaluated for quality by the COSMIN checklist. Most studies could not meet COSMIN checklist criteria. However, WURT/RIRT, P-drive, and TRIP have met on many items of the COSMIN checklist. These on-road tests have good reliability, validity, and generalizability of the on-road tests.

In recent years, a shift has occurred from the use of traditional statistical methods of Classical Test Theory to the recommended use of newer statistical methods of IRT or Rasch Measurement Theory analyses for developing and evaluating outcome measurement instruments. It is difficult to decide which is superior between IRT and Classical Test Theory (Kohli et al., 2015; Jabrayilov et al., 2016), however, the test using IRT method tends to have high quality according to the COSMIN checklist (Prinsen et al., 2018).

Although it is difficult to gain high quality by the IRT method (e.g. sample size issues), the studies using the IRT method solved this problem and had relatively high quality in this review. In fact, our review showed that P-drive and K-score using IRT have high reliability and validity. Our finding indicates that the use of IRT is still low. Therefore, we suggest the use of IRT methodology in future studies. Although K-score also had high-quality items, there was only one study about its reliability and validity. Thus, we did not include it in the recommendation list.

On the other hand, more classical test theory method studies were reviewed than IRT studies. The WURT/RIRT and TRIP assessments included several verified items, and many reliability and validity studies were conducted. One could argue that these tests are of adequate quality given the repeated assessments undertaken for determining reliability and validity. Together, our findings suggest that WURT/RIRT, P-drive, and TRIP are useful on-road tests for drivers.

Next, we describe focusing on each item of reliability and validity. Interrater reliability was the most frequent item
assessed based on the checklist. Medical assessments, such as off-road tests (e.g., Trail Making Test, Stroke Drivers Screening Assessment), were usually conducted by medical staff, while on-road tests were typically administrated by driving instructors (Hunt et al., 1997; De Raedt and Ponjaert-Kristoffersen, 2001; Akinwuntan et al., 2006; Ott et al., 2012). Therefore, it might be easy for the researchers to check the interrater reliability.

Another examiner such as an occupational therapist might be added to further simplify the research process. Interrater reliability means that regardless of who evaluates the test, it is possible to generalize these reliabilities. However, no prior studies verified measurement error of on-road driving tests. Measurement error is important for a decision whether the changing score is due to interventional effect or error (bias). Our results suggest that further study is needed to confirm the measurement error item. In the items of validity, we included the criterion validity study that was conducted to test the ‘off-road’ comparing to ‘on-road.’ Because these studies used a cross-sectional study design, the gold standard test of on-road driving tests had been unclear. After this study, WURT/RIRT, P-drive, and TRIP would be expected to be used as a gold standard test.

Moreover, some studies (De Raedt and Ponjaert-Kristoffersen, 2001; Fitten et al., 1995) demonstrated validity items based on drivers’ accident ratios. It is important to consider for what purpose the on-road test is to be conducted, and what it intends to predict in the end. It is recommended that future studies clarify actual on-road behavior, especially in cases where there is no gold-standard on-road test. The current review also observed that few studies verified content and structural validity. These forms of validity assume that a theoretical rationale underlies the test instrument. Such verifications are related to construct validity and are the highest form of empirical evidence for an instrument’s utility (Rice and Cutler, 2012). Therefore, it is better for future studies to develop road tests to verify content and structural validity. No prior studies verified responsiveness and interpretability. Since patients with some neurological diseases (e.g., traumatic brain injury, stroke) may improve their driving performance in the future, these components are important for determining effectiveness of any training, exercise, or treatment of driving performance.

Many researchers are more focused on predicting driving skills of subjects than training effectiveness. In fact, there is no strong evidence for training effectiveness when engaging in driving interventions within these populations. For instance, randomized control trial studies show that there was some effectiveness in the subgroup (Mazet et al., 2003; Mazet et al., 2015). Mazet et al. (2003) clarified that participants with moderate impairment who received simulator training were more likely to pass the driving test compared with those in the control group (86% versus 17%). However, they were unable to demonstrate the effectiveness of a driving training program in the main group/outcome. Furthermore, only one Cochran review report has assessed driving rehabilitation among stroke patients, and results indicated insufficient evidence for reaching conclusions regarding improved on-road driving skills post-stroke (George et al., 2014). Our result showed that the COSMIN checklist items for the effect of interventions have not been demonstrated. These items will be demonstrated in the near future because more researchers are becoming interested in the effect of driving rehabilitation.

Limitations
The study has several limitations. First, we reviewed only studies published in the English literature. Therefore, we could not learn about the on-road tests performed in non-English-speaking countries. Second, we adopted the COSMIN methodology in the current study, but there are some limitations in confirming the validity. For example, some studies reported on construct validity (Classen et al., 2017) and ecological validity (Vlahodimitrakou et al., 2013). Although these methods were used to assess on-road test reliability and validity, their analysis methods did not fit for the COSMIN. In the different items, a cross-cultural validity item is present on the COSMIN checklist. This item deals predominantly with language (e.g., back translation). The ability to assess real-world driving depends on various environmental conditions. For instance, some countries require motorists to drive on the left, others on the right. Therefore, cross-cultural validity of on-road tests should not only depend on language but also on various contexts. Along these lines, the RIRT was created by adapting the WURT, but it did not match with the cross-cultural validity items of COSMIN methodology. From these reasons, without verification from the COSMIN checklist, some of the on-road tests reviewed may be of limited quality. Future studies assessing on-road test validity within various environments are recommended.

Conclusion
The WURT/RIRT, P-drive, and TRIP were identified as highly qualified on-road driving tests. Future studies should confirm measurement error, content validity, structural validity, responsiveness, and interpretability of these tools.

Acknowledgements
The author disclosed receipt of the following support for the research, authorship, or publication of this article: This work was supported by the Japan Society for the Promotion of Science (grant number 70434529).

Conflicts of interest
There are no conflicts of interest.
De Raedt R, Ponjaert-Kristoffersen I (2001). Prediction of at-fault car accidents of drivers. *Neuropsychologia* 39:203–213.

Azami-Aghdash S, Aghaei MH, Sadeghi-Bazarghani H (2018). Epidemiology of road traffic injuries among elderly people: A systematic review and meta-analysis. *Bull Emerg Trauma* 6:279–291.

Bliokas V V, Taylor JE, Leung J, Deane FP (2011). Neuropsychological assessment of fitness to drive following acquired cognitive impairment. *Brain Inj* 25:471–487.

Brooke MM, Questad KA, Patterson DR, Valois TA (1992). Driving evaluation after traumatic brain injury. *Am J Phys Med Rehabil* 71:177–182.

Brown LB, Ott BR, Papandontos GD, Sui Y, Ready RE, Morris JC (2005). Prediction of on-road driving performance in patients with early Alzheimer’s disease. *J Am Geriatr Soc* 53:94–98.

Chihuri S, Mielenz TJ, DiMaggio CJ, Betz ME, DiGuiseppi C, Jones VC, Li G, Bliokas V V, Taylor JE, Leung J, Deane FP (2011). Neuropsychological assessment of fitness to drive following acquired cognitive impairment. *Brain Inj* 25:471–487.

Kohli N, Koran J, Henn L (2015). Relationships among classical test theory and meta-analysis of studies that evaluate health care interventions: prediction and elaboration. *Plos Med* 11:e1000100.

Marshall SC, Molnar F, Man-Son-Hing M, Blair R, Brosseau L, Finestone HM, et al. (2007). Predictors of driving ability following stroke: a systematic review. *Top Stroke Rehabil* 14:98–114.

Mazer B, Gélinas I, Duquette J, Vanier M, Rainville C, et al. (2015). A randomized clinical trial to determine effectiveness of driving simulator retraining on the driving performance of clients with neurological impairment. *Br J Occup Ther* 78:369–376.

Mazer BL, Sofer S, Korner-Bilensky N, Gélinas I, Hanley J, Wood-Dauphinee S (2003). Effectiveness of a visual attention retraining program on the driving performance of clients with stroke. *Arch Phys Med Rehabil* 84:541–550.

Moher D, Liberati A, Tetzlaff J, Altman DG, PRISMA Group (2009). Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. *J Clin Epidemiol* 62:1006–1012.

Mokkink LB, Terwee CB, Patrick DL, Alonso J, Stratford PW, Knol DL, et al. (2010). The COSMIN checklist for assessing the methodological quality of studies on measurement properties of health status measurement instruments: an international delphi study. *Qual Life Res* 19:539–549.

Nouri FM, Lincoln NB (1992). Validation of a cognitive assessment: predicting driving performance after stroke. *Clin Rehabil* 6:275–281.

Ott BR, Festa EA, Amick MM, Grace J, Davis JD, Heindel WC (2008). Computerized maze navigation and on-road performance by drivers with dementia. *J Geriatr Psychiatry Neurol* 21:118–25.

Otto BR, Papandontos GD, Davis JD, Barco PP (2012). Naturalistic validation of an on-road driving test of older drivers. *Hum Factors* 54:663–674.

Patomella AH, Caneman G, Kottorp A, Tham K (2004). Identifying scale and person response validity of a new assessment of driving ability. *Scand J Occup Ther* 11:77–80.

Patonella AH, Tham K, Johansson K, Kottorp A (2010). P-drive on-road: internal scale validity and reliability of an assessment of on-road driving performance in people with neurological disorders. *Scand J Occup Ther* 17:86–93.

Prinson CAC, Mokkink LB, Bouter LM, Alonso J, Patrick DL, de Vet HCW, Terwee CB (2018). COSMIN guideline for systematic reviews of patient-reported outcome measures. *Qual Life Res* 27:1147–1157.

Rogers MA, Welsh RK, Watson GS, Cholerton B, Baker LD, Craft S (2004). The relationship between neuropsychological functioning and driving ability in dementia: a meta-analysis. *Neuropsychology* 18:85–93.

Rice M, Cutler SK (2012). Clinical research in occupational therapy. 5th ed. NY: Delmar.

Selanders L, Lee HC, Johansson K, Falkmer T (2011). Older drivers: on-road and off-road test results. *Arch Phys Med Rehabil* 92:363–40.

Hirad MA, Egote P, Fischer CE, Naglie G, Schweizer TA (2016). A systematic review and meta-analysis of on-road simulator and cognitive driving assessment in Alzheimer’s disease and mild cognitive impairment. *J Alzheimers Dis* 53:713–729.

Hunt L, Morris JC, Edwards D, Wilson BS (1993). Driving performance in persons with mild senile dementia of the Alzheimer type. *J Am Geriatr Soc* 41:747–752.

Hunt LA, Murphy CF, Carr D, Duchek JM, Buckley V, Morris JC (1997). Reliability of the walthington university road test. A performance-based assessment for drivers with dementia of the Alzheimer type. *Arch Neuro* 54:707–712.

Jabrayilov R, Emmons WHM, Sijtsma K (2016). Comparison of classical test theory and item response theory in individual change assessment. *Appl Psychol Meas* 40:559–572.

Justiss MD, Man WC, Staw W, Velozo C (2006). Development of a behind-the-wheel driving performance assessment for older adults. *Top Geriatr Rehabil* 22:121–128.

Kay L, Bundy A, Clemson L, Jolly N (2008). Validity and reliability of the on-road driving assessment with senior drivers. *Accid Anal Prev* 40:751–759.

Kohi N, Koran J, Henn L (2015). Relationships among classical test theory and item response theory frameworks via factor analytic models. *Educ Psychol Meas* 75:389–405.

Liberati A, Altman DG, Tetzlaff J, Mulrow C, Gotzsche PC, Ioannidis JP, et al. (2009). The PRISMA statement for reporting systematic reviews and meta-analyses of studies that evaluate health care interventions: prediction and elaboration. *Plos Med* 6:e1000100.