Rater Reliability and Rater Effects of the Safe Driving Behavior Measure

KEY WORDS
- aged
- automobile driving
- dangerous behavior
- reproducibility of results
- safety
- self report

Sherrilene Classen, Pey-Shan Wen, Craig A. Velozo, Michel Bédard, Sandra M. Winter, Babette A. Brumback, Desiree N. Lanford

We used the Safe Driving Behavior Measure (SDBM) to determine rater reliability and rater effects (erratic responses, severity, leniency) in three rater groups: 80 older drivers (mean age = 73.26, standard deviation = 5.30), 80 family members or caregivers (age range = 20–85 yr), and two driving evaluators. Rater agreement was significant only between the evaluators and the family members or caregivers. Participants rated driving ability without erratic effects. We observed an overall rater effect only between the evaluator and family members or caregivers, with the evaluators being the more severe rater group. Training family members or caregivers to rate driving behaviors more consistently with the evaluator's ratings may enhance the SDBM's usability and provide a role for occupational therapists to interpret proxy reports as an entry point for logical and efficient driving safety interventions.

Classen, S., Wen, P.-S., Velozo, C. A., Bédard, M., Winter, S. M., Brumback, B. A., et al. (2012). Rater reliability and rater effects of the Safe Driving Behavior Measure. *American Journal of Occupational Therapy, 66,* 69–77. doi: 10.5014/ajot.2012.002261

Sherrilene Classen, PhD, MPH, OTR/L, is Graduate Faculty Member, Rehabilitation Science Doctoral Program; Associate Professor, Department of Occupational Therapy; and Director, Institute for Mobility, Activity, and Participation, College of Public Health and Health Professions, University of Florida, PO Box 100164, Gainesville, FL 32610; sclassen@phhp.ufl.edu

Pey-Shan Wen, PhD, OTR/L, is Postdoctoral Fellow, Department of Occupational Therapy, College of Public Health and Health Professions, University of Florida, Gainesville.

Craig A. Velozo, PhD, OTR/L, is Graduate Faculty Member, Rehabilitation Science Doctoral Program, Department of Occupational Therapy; Institute for Mobility, Activity, and Participation, College of Public Health and Health Professions, University of Florida, Gainesville; and Research Health Scientist, Rehabilitation Outcomes Research Center and Brain Rehabilitation Research Center, North Florida/South Georgia Veterans Health System, Gainesville.

Michel Bédard, PhD, is Professor, Centre for Research on Safe Driving, Lakehead University, Thunder Bay, Ontario.

Sandra M. Winter, PhD, OTR/L, is Postdoctoral Associate, Institute for Mobility, Activity, and Participation, Department of Occupational Therapy, College of Public Health and Health Professions, University of Florida, Gainesville.

Babette A. Brumback, PhD, is Associate Professor, Department of Biostatistics, College of Public Health and Health Professions, College of Medicine, University of Florida, Gainesville.

Desiree N. Lanford, MOT, CDRS, OTR/L, is Certified Driving Rehabilitation Specialist, Department of Occupational Therapy and Institute for Mobility, Activity, and Participation, College of Public Health and Health Professions, University of Florida, Gainesville.

Statistics have shown that older drivers' motor vehicle crash, injury, and fatality rates continue to be a concern for occupational therapists because of this population's current and future growth. Accurate and precise measurement of older adults' unsafe driving behaviors is an essential first step in curtailing crashes and preventing adverse effects such as injuries and fatalities (Classen et al., 2010). The comprehensive driving evaluation (clinical tests and an on-road test conducted by a driving rehabilitation specialist), the gold standard measure for driving evaluation, is highly valid and reliable but has limitations such as being time consuming, providing limited access, and holding an element of threat (mandatory or ethical reporting with driver failure). Self-report can be used, complementary to other forms of assessment, to identify older adults' unsafe driving behaviors, increase driving safety awareness and knowledge, and promote behavior change and safer driving outcomes (Eby, Molnar, Shope, Vivoda, & Fordyce, 2003; McGee & Tuokko, 2003).

Advances have been made in developing self-report measures for older drivers. Such tools include *Drivers 55 Plus: Check Your Own Performance* (AAA Foundation for Traffic Safety, 1994) and the computer-based *Roadwise Review* (American Automobile Association, 2004). These measures have strengths and limitations. For example, the computer-based *Roadwise Review* has great face validity, but it takes approximately 40 minutes to complete and may be challenging to use with older adults with low computer fluency. Although these two self-report measures provide meaningful descriptions of a driver's ability level, they emphasize person and environment factors, with a gap in items addressing the vehicle. Understanding a participant's level of safe driving behaviors is a critical step toward providing an entry point for logical and efficient occupational therapy interventions, identifying optimal training parameters, and predicting future safe driving ability.

In ongoing work, we have developed a self-report Safe Driving Behavior Measure (SDBM; Classen et al., 2010; Winter et al., 2011). We have tested it among 80 older drivers, 80 family members and caregivers (F–C), and two driving evaluators, and we have conducted psychometric analyses (Classen et al., in press). Findings from the pilot work encouraged us to further refine the SDBM as a precise and accurate measure for detecting safe driving behaviors among older adults. As such, the objective of this study was to quantify the rater reliability and rater effects, using Item Response Theory (IRT), among three rater groups (older drivers, F–C, driving evaluators).

Interrater Reliability

Interrater reliability is defined as the extent to which different raters agree on the same people or characteristics. The terms *interrater reliability, rater agreement,* and *rater correlation* are often used interchangeably. Two studies investigated the relationship of driving performance rated by evaluators and older drivers (Marottoli & Richardson, 1998; Wild & Cotrell, 2003). Marottoli and Richardson (1998) investigated the relationship between on-road driving performance rated by evaluators and self-reported driving ability rated by older drivers using the Pearson correlation coefficient. They did not find a significant association between the ratings of these two groups. Wild and Cotrell (2003) investigated the differences between evaluators' ratings and older drivers' ratings on the Driving Safety Evaluation scale using *t* tests. They found that only 2 of 10 items showed significant differences between evaluators' ratings and older drivers' ratings.

Neither the Pearson correlation coefficient nor the *t*-test statistic can accurately determine the potential rater effects. Although the Pearson correlation coefficient can indicate the strength of the relationship between two sets of data (the concordance of the data), it cannot detect whether the value of one set of data is consistently greater than the value of the other, which might indicate that one rater is more severe or lenient than the other. The *t*-test statistic detects the significant difference of the means of two sets of data; however, using the mean may partial out the individual differences that exist within the rater group. Moreover, the Pearson correlation coefficient and *t*-test statistic cannot provide information regarding the response pattern, that is, whether someone responds to the items erratically (i.e., rating inconsistently). Thus, although the Pearson correlation coefficient and *t*-test statistic provide necessary methods for assessing rater agreement, they are not sufficient to make an accurate determination of rater effects. Examining rater effects is critical, especially when people will be reporting on safety aspects of driving.

Rater Effects

Rater effects are a function of *severity* or *leniency,* defined as the tendency for a rater to assign ratings consistently higher or lower than those of other raters (Myford & Wolfe, 2004). In addition to having tendency to assign higher or lower ratings, raters may also assign ratings erratically (erratic response pattern); that is, the raters inappropriately assign low scores (*cannot do*) to drivers with a higher ability level or high scores (*no difficulty*) to drivers with a lower ability level. The Many Facets Rasch Model (MFRM; Linacre, 2004), an extension of the Rasch model, is useful in investigating rater severity and response patterns. The Rasch model, a one-parameter IRT model, converts ordinal scales into interval measures (using logit as its unit) and provides a useful, efficient, and objective framework for developing, evaluating, and revising measures.

Five published driving studies applied Rasch analysis to develop or evaluate driving scales (Kay, Bundy, & Clemson, 2008, 2009; Myers, Paradis, & Blanchard, 2008; Patomella, Kottorp, & Tham, 2008; Patomella, Tham, & Kottorp, 2006). Patomella and colleagues (2006) first applied Rasch analysis to examine the Performance Analysis of Driving Ability (P-Drive) using a driving simulator with 31 people with brain injury; they later (Patomella et al., 2008) used Rasch analysis to evaluate the P-Drive with 101 people with stroke. Kay et al. (2008) applied Rasch analysis to a standard on-road test to transform the on-road test into a linear interval measure with hierarchical ordered tasks. Myers and colleagues (2008) also used a Rasch model to examine the structure of a scale assessing driving confidence. Most recently, Kay and colleagues (2009) applied Rasch analysis to a simulated test rated by trained professionals and an awareness test to investigate the construct validity and internal reliability of the simulated test and awareness test. Although we have seen an increased application of Rasch analysis in developing and evaluating assessments, no driving-related published study has yet applied the Rasch model to assess rater effects.

Beyond estimating item difficulties and person abilities, the MFRM includes an additional parameter, the rater, to detect whether the response differences are caused by systematic rater severity or leniency. Moreover, by fitting data to the Rasch model, the MFRM can detect the erratic raters.

Rater effect is particularly important in our field of study as we develop an older driver and proxy self-report tool, the SDBM. When comparing older drivers' self-reports with F–C reports or driver evaluator reports, we anticipate a discrepancy. That is, we expect that older drivers may be the least severe in their self-ratings (e.g., do not want to lose their license), and evaluators may be the most severe in their ratings (e.g., are trained to focus on deficits). The F–C may be somewhat in the middle with their ratings of their loved one's driving safety; some may be overly severe (i.e., really want the driver to stop driving), and some may be less severe (i.e., do not want them to lose their means of transportation).

Purpose

In this study, we addressed interrater reliability among three groups of raters (older driver, F–C, and driving evaluators) and investigated the rater effects between two of the groups (F–C and driving evaluators) on the SDBM's 41 items. We expected our findings to provide the evidence to use the self- or proxy-report SDBM as a reliable measure of safe driving among older adults, their F–C, and occupational therapists conducting such evaluations.

Method

This study received approval from the institutional review boards of the University of Florida and Lakehead University.

Design

A convenience sample of older drivers and F–C was recruited from north Florida and Thunder Bay, Ontario. All participants completed an SDBM. The older drivers underwent an on-road driving evaluation conducted by trained driving evaluators, who also completed an SDBM after the on-road test.

Participants

We recruited participants in north Florida, United States, and Ontario, Canada, by means of advertisements in newspapers, word-of-mouth referrals, and flyers distributed to local community facilities and through an aging registry. Participants were included if they were older drivers (ages 65–85 yr); had a valid driver's license; were driving at the time of recruitment; had the cognitive ability to complete the SDBM, as evidenced during the telephone interview; and had the ability to participate in an on-road driving test (behind-the-wheel test in a dual-brake vehicle with the driving rehabilitation specialist using a standardized scoring sheet to evaluate driving performance (Justiss, Mann, Stav, & Velozo, 2006), as evidenced in not having missing limbs or a severe psychiatric diagnosis. The participants generally met the inclusion criteria. Participants were excluded if they had been medically advised not to drive, had experienced uncontrolled seizures in the past year, or took medications that caused central nervous system impairment. F–C between ages 18 and 85 yr were included if they were able to report (on the basis of observation) on the older adult's driving behavior and excluded if they showed the presence of a physical or mental condition that impaired their ability to make an active contribution. At the primary site, the certified driving rehabilitation specialist (Desiree Lanford) with 7 years of clinical practice experience collected the data. At the Canadian site, the driving evaluator, who was a driving instructor accredited by the Province of Ontario with >10 yr of experience, completed the on-road test and evaluator SDBM. Thus, the rater groups were older drivers, F–C, and driving evaluators.

Procedure

We standardized the SDBM and clinical test administration, as well as the on-road driving evaluation test, across sites by (1) using a set testing protocol for the two sites, (2) having the driving evaluator at the primary site (Florida) conduct a 3-day training session with the evaluator at the Canadian site, and (3) ensuring 100% congruence between the two on-road driving evaluators by using a 4-point scale to rate the driving of three healthy volunteers. All older drivers and their F–C provided consent in a private research office before completing demographic information and the SDBM, which was part of a larger battery of clinical and on-road assessments (Classen et al., 2008; Stav, Justiss, McCarthy, Mann, & Lanford, 2008). The two evaluators, who were blinded to the participants' SDBM self-ratings or proxy ratings, also completed an SDBM on each driver after the on-road test. All participants received $50 for their study participation.

Instrument

The SDBM is a 68-item self-report or proxy measure to assess safe driving behaviors (Classen et al., 2010; Winter et al., 2011). The measure score (derived from Rasch analysis) represents the reported level of difficulty for the items given the participant's ability level. Difficulty with the driving task is rated via a 5-point adjectival scale (ranging from 1 = *cannot do* to 5 = *not difficult*). The SDBM items are displayed in the Appendix.

Data Collection and Management

All the data (SDBM, demographic information, scores on the clinical tests and the on-road tests) of the older drivers and F–C were collected and recorded by research assistants in a central, secure, and password-protected data repository, which was located at the primary site, the University of Florida. Data entry was monitored by Sherrilene Classen, the principal investigator, and quality control spot checks and corrections were made intermittently to ensure data completion and accuracy.

Data Analysis

Item Inclusion and Exclusion. We excluded 27 items from the analysis, 22 items that were not observable by the driving evaluator at the time of testing (e.g., driving in snow) and 5 items that added little or no variance to the responses. For example, >95% of respondents used the same rating category (i.e., *not difficult*) for 5 items.

Interrater Reliability. We conducted an interclass correlation (ICC) to examine rater reliability on the 41 remaining items. We used SPSS version 17.0 (SPSS, Inc., Chicago) for the analyses and considered $p \leq .05$ significant for the correlations.

Rater Effects. We used the MFRM to analyze rater effects using Facets software version 3.57 (Linacre, 2004). The MFRM extends the rating-scale Rasch model by adding one component or facet (Cj) to calibrate rater severity:

$$\mathrm{Log}\left[P_{nik}/\left(P_{ni(k-1)}\right)\right] = Bn - Dg_i - F_{gk} - C_j,$$

where P_{nik} is the probability of observing category k for person n who answers item i; $P_{ni(k-1)}$ is the probability of observing category $k-1$; Bn is person ability; Dg_i is item difficulty for item i in group g; F_{gk} is the difficulty of being observed in category k relative to category $k-1$ for an item in group g; and C_j is the severity of judge j, who gives rating k to person n on item i.

We used Facet ruler, fit statistics, fixed χ^2, and paired comparisons to investigate the rater effects. Facet ruler, displaying three facets (rater, item difficulty, person ability) in the same linear continuum, provides a visual map to compare the relative hierarchy within and between facets. To illustrate the relative distribution of the drivers' abilities and item difficulties simultaneously, we anchored the mean of the rater severity to 0.

We used fit statistics (infit mean squares [*MnSq*s] and outfit *MnSq*s) to detect erratic raters, that is, raters who assign high scores to drivers with a low ability level and low scores to drivers with a high ability level. Infit statistics are more responsive to the variance of those well-targeted observations, and outfit statistics are sensitive to the variance of outliers or extreme observations. *Ideal fit* occurs when the observed response patterns exactly match the predicted pattern (*MnSq* = 1) of the model. We considered infit *MnSq* and outfit *MnSq* ranging from 0.6 to 1.4 an adequate fit for survey data (Bond & Fox, 2001). The measure represents the average ratings of the rater in logits, with higher scores indicating greater severity in rating.

We used fixed χ^2 to examine whether at least one rater group, on the overall scale level, consistently used the ratings differently from other rater groups. If the fixed χ^2 test was significant, then we performed paired comparisons to identify item-level rater effects. For example, if three rater groups are tested, a significant fixed χ^2 statistic means that at least one of these three rater groups is more severe or lenient in their ratings on the overall scale.

We then performed paired comparisons to identify which rater group was significantly more severe or lenient in its ratings or to show which items the raters rated significantly more severely or leniently. We used an α level of .05 to determine a significant rater effect.

Results

Demographics

Table 1 presents the driver and F–C demographics. We tested 80 licensed drivers with a mean age of 73.26 yr (standard deviation [*SD*] = 5.30), even gender distribution, and mainly White. They were an educated and healthy older community-dwelling group. We also tested 80 F–C (age range = 20–85 yr), all of whom were community dwelling, who were mainly female and White; most lived with a partner or spouse.

Interrater Reliability

The ICCs among the ratings of the three rater groups was significant but weak (ICC = .256, $p < .001$, 95% confidence interval [CI] = .118–.403). Of the 41 items, we found a significant correlation only between the ratings of the evaluator and F–C groups (ICC = .462, $p < .001$, 95% CI = .271–.618). We observed no significant correlations between the ratings of the older driver and F–C groups (ICC = .127, $p = .129$) or between the older driver and evaluator groups (ICC = .088, $p = .217$).

Rater Effects

Facet Ruler of the SDBM. Figure 1 depicts three facets (raters, drivers, items) on the linear interval scale for the SDBM. The first column, Measure, is the interval scale expressed as a logit unit. The second column displays the

Table 1. Demographics and Driving Characteristics of Older Drivers ($N = 80$) and Their Caregivers ($N = 80$)

Characteristic	Older Drivers		Caregivers	
Age, yr, mean (*SD*)	73.26	(5.30)	64.85	(14.03)
Age range, yr	65–85		20–85	
Gender, *n* (%)				
Male	41	(51.25)	18	(22.50)
Female	39	(48.75)	62	(77.50)
Race, *n* (%)				
White	77	(96.25)	79	(98.75)
Other	2	(2.50)	1	(1.25)
Missing data	1	(1.25)	0	
Education, *n* (%)				
College or university	51	(63.75)	39	(48.75)
Vocational training or some college	10	(12.5)	23	(28.75)
Associate's degree	3	(3.75)	3	(3.75)
≤ High school	16	(20.00)	15	(18.75)
Drive 7 days/wk	34	(42.50)	40	(50.00)
Drive <7 days/wk	46	(57.50)	40	(50.00)
Licensed driver	80	(100.00)	80	(100.00)
Living alone	25	(31.25)	9	(11.25)
Living with partner or spouse	49	(61.25)	55	(68.75)
Other	6	(7.50)	16	(20.00)
MMSE, mean (*SD*)	27.67	(1.85)	NA	
No. of self-reported medications, mean (*SD*)	6.57	(3.87)	NA	

Note. MMSE = Mini-Mental State Examination; NA = not applicable; *SD* = standard deviation.

severity of raters, representing, from bottom to top, lenient to severe raters. The third column shows the distribution of the drivers' safe driving ability, from bottom to top, representing the drivers with safe driving abilities ranging from poor to good. The fourth column displays item difficulties representing, from bottom to top, the essentially easy items and then progressing to levels of increasing difficulty. The fifth column shows the likelihood of applying the rating scale in relation to the raters' abilities; that is, when a driver's estimated ability is between 1 and 2 logits, he or she will likely receive a rating of 4 on this measure. In the second column, the driving evaluator is located above the caregiver, indicating that the driving evaluator is the more severe rater. The distribution of the drivers' abilities was on the upper part of the ruler, as displayed in the third column, and the distribution of the item difficulties was on the lower part of the ruler, as displayed in the fourth column. This distribution indicated that the drivers had, generally speaking, high safe driving abilities as measured on this 41-item scale.

Fit Statistics of the Rater Groups. The infit *MnSq*s and outfit *MnSq*s for both rater groups were between 0.97 and 1.05, well within the defined criteria of 0.6 and 1.4 (Bond & Fox, 2001; Linacre, 2002; Wright & Linacre, 1994).

Fixed χ^2. The fixed χ^2 value, 166.9 with one degree of freedom, was statistically significant ($p < .001$). The ratings between the F–C and evaluator groups showed significant rater effects: The evaluator group was more severe, considering that the measure of the evaluator group is higher (mean = −3.32, *SD* = 0.03) than that of the F–C group (mean = −3.98, *SD* = 0.04).

Paired Comparisons. The results of the paired comparisons showed significant rater effects on 17 items (Table 2 and Figure 2). Although the evaluators' ratings were more severe on the overall scale, the F–C group rated 10 of 17 items more severely than did the evaluator group.

Discussion

In this study, we addressed interrater reliability among three groups of raters (older drivers, F–C, and driving evaluators) and investigated the rater effects between the evaluators and the F–C to identify erratic responses and to determine the severity or leniency of the groups' ratings on the 41 items of the SBDM.

Interrater Reliability

We found no statistically significant correlation between the ratings of the driver and evaluator groups or between the ratings of the driver and F–C groups. However, we found a significantly moderate agreement (.53) between the evaluator and F–C groups. Two studies have previously investigated the relationship of driving performance rated by evaluators and older drivers (Marottoli & Richardson, 1998; Wild & Cotrell, 2003) and found no significant correlation between the evaluators' rating and the drivers' rating (Marottoli & Richardson, 1998) and no significance for 8 of 10 items rating the drivers' driving performance (Wild & Cotrell, 2003). Our study's findings are therefore somewhat consistent with these two studies in that the evaluators' ratings were not associated with the drivers' ratings, but they are correlated with the F–C's ratings.

Rater Effects

Facet Ruler of the SDBM. The distribution of the drivers' ability relative to the distribution of the items' difficulty indicates that the participants in this study performed well on the instrument. As can be seen in Figure 1, many of our items are on the same logit level. Taking into account that only the means of the items are represented, we have more overlap among the items because each item consists of five difficulty levels corresponding to a 5-point adjectival scale. Having different items at the same difficulty level in the item pool may be redundant for paper-and-pencil tests; however, it will

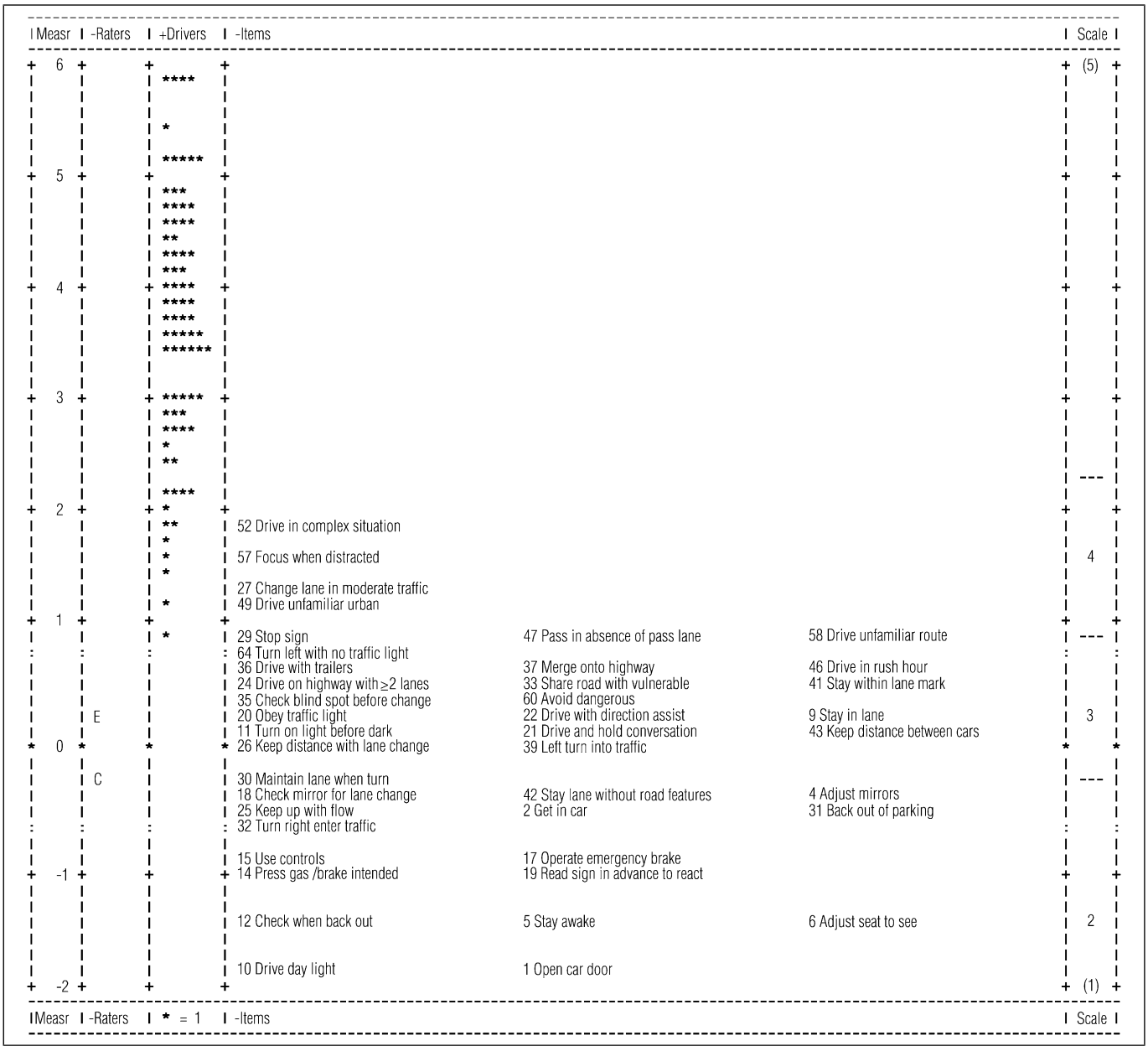

Figure 1. Facet ruler of 41 items of the Safe Driving Behavior Measure.

Note. Each number represents an item; Appendix A contains each item by item number. C = family member or caregiver; E = driving evaluator; Measr = measure.

increase the item pool, which will in turn provide more choices for future applications, such as using computer adaptive testing (the next step in the development of our instrument).

Fit Statistics of the Three Rater Groups. The fit statistics across the rater groups (evaluators and F–C) showed that no rater group was erratic and that overall the evaluators were more severe raters (Facets) than the F–C.

Fixed χ^2 and Paired Comparisons. Although the evaluator group rated more severely than the F–C group on the overall scale, the F–C group rated 10 items more severely than did the evaluator group. However, the evaluator group rated 7 items more severely than did the F–C group. Evaluators have the formal training to rate driving behaviors according to the standards of regulatory bodies such as the Department of Motor Vehicle and Highway Safety licensure guidelines, and one can therefore expect that they will be more technical and more stringent in their ratings. The F–C group did not have such formal training and were rating the drivers on their perceptions of their loved one's driving safety. The tendency for evaluators to rate more severely (than the F–C) may be influenced by their training to focus on identifying deficits. The F–C, however, may be influenced by concern for their loved one's safety, thus rating more stringently, or may be concerned with maintaining their own independence in transportation and rating leniently, especially given that 31.3% of F–C stated that their

Table 2. Items With Significant Rater Effects

Item	Measure of Caregiver	Measure of Evaluator	Contrast	Joint *SE*	*T*	*p*
F–C group rated more severely than evaluator group on 10 items						
2. Get in his or her car.	−0.12	−1.13	1.25	0.42	2.99	.003
5. Stay awake while driving.	−0.7	−2.38	1.68	0.69	2.44	.016
17. Operate the emergency brake.	−0.18	−1.35	1.19	0.47	2.55	.012
18. Check car mirrors when changing lanes.	−0.25	−1.02	1.27	0.40	3.19	.002
30. Maintain lane when turning.	0.25	−0.51	0.86	0.36	2.37	.019
42. Stay within proper lane in the absence of road features.	0.05	−0.84	0.89	0.39	2.27	.024
49. Drive in an unfamiliar urban area.	1.56	0.85	0.71	0.25	2.87	.005
58. Drive in an unfamiliar area.	1.31	0.44	0.86	0.26	3.34	.001
60. Avoid dangerous situations.	0.88	0.09	0.79	0.29	2.74	.007
64. Turn left across multiple lanes when no traffic light.	1.11	0.58	0.53	0.26	2.03	.044
Evaluator group rated more severely than F–C group on 7 items						
9. Stay in the proper lane.	−0.46	0.58	−1.04	0.37	−2.78	.006
24. Drive on a highway with ≥2 lanes in each direction.	0.05	0.82	−0.77	0.32	−2.44	.016
27. Change lanes in moderate traffic.	−0.46	1.97	−2.43	0.36	−6.81	<.001
29. Brake at a stop sign so car stops completely before the marked line.	0.42	1.02	−0.6	0.28	−2.11	.037
36. Drive with surrounding tractor trailers.	0.31	1.04	−0.74	0.29	−2.53	.012
43. Keep distance between his or her car and others.	−0.84	0.51	−1.35	0.43	−3.15	.002
57. Stay focused on driving when there are distractions.	1.03	1.87	−0.83	0.24	−3.49	.001

Note. F–C = family member–caregiver; *SE* = standard error.

independence would be affected if the older driver stopped driving. In future studies, we may want to control for this variable by stratifying F–C on the basis of whether their independence will or will not be affected if the older driver stops driving.

The generalizability of our findings is limited because we used only two evaluators, had a convenience sample, and had a sample size of 80 F–C and 80 older drivers. Our driver sample was skewed to include mainly White (97.5%) and educated participants (63.8% had some college education or a university degree). The F–C sample was mainly female (77.5%) and White (98.8%), and 48.8% had completed a college or university degree.

Despite the study's limitations, our findings suggest that because of the significant relationship between F–C and evaluator findings, we may train caregivers to more accurately recognize older adults' unsafe driving behaviors. As such, after a short training program for F–C, we

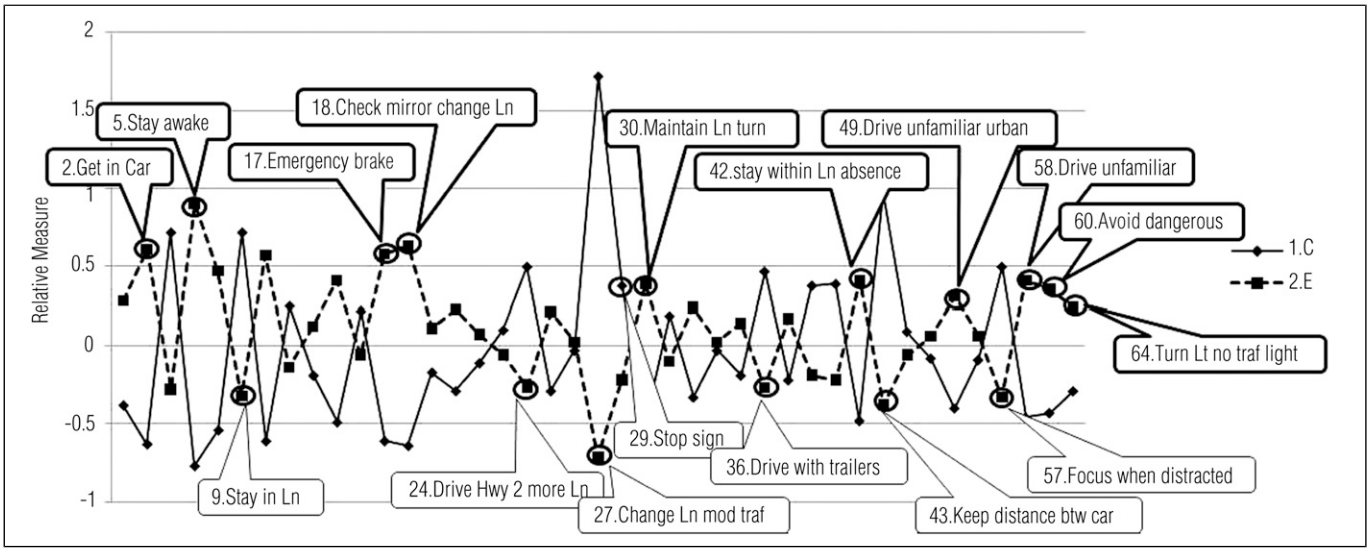

Figure 2. Bias analysis map for the evaluator and family member–caregiver rater groups.

Note. The map shows significant rater effects on 17 items: 2, 5, 9, 17, 18, 24, 27, 29, 30, 36, 42, 43, 49, 57, 58, 60, and 64. The ratings of the F–C group were significantly more severe than the ratings of the evaluator group on 10 items—2, 5, 17, 18, 30, 42, 49, 58, 60, and 64—and the ratings of the evaluator group were significantly more severe than those of the F–C group on 7 items—9, 24, 27, 29, 36, 43, and 57. See the Appendix for a detailed description of the items by item number. btw = between; C = family member or caregiver; E = driving evaluator; Hwy = highway; Ln = lane; Lt = left; mod = moderate; traf = traffic.

expect that the paired comparisons of the identified items (displayed in Table 2) may show improved congruence between F–C and evaluators. We are developing a caregiver training protocol to test this hypothesis.

Implications for Occupational Therapy Practice

Because occupational therapists play an important role in driving rehabilitation, understanding the differences in ratings among raters can help occupational therapists accurately identify the problems and efficiently develop driving safety interventions. This article presents several implications for occupational therapy practice:

- The SDBM can be used as a reliable tool to assess older drivers' safe driving behavior in occupational therapy practice.
- Driving evaluators may tend to rate more severely than F–C.
- When reviewing the ratings of the SDBM, occupational therapists must take rater effects into account.
- With adequate training, F–C may be the most available, accurate, and reliable resource for following up on the safe driving behaviors of older people.

Conclusion

Our findings address an understudied area in the older driver safety literature: the reliability, leniency, and severity of F–C and evaluator ratings of older drivers through the SDBM. This study makes it clear that a correlation exists between the evaluator and the F–C ratings, that neither of these groups is erratic in their rating responses, that the driving evaluator is the most severe rater, and that the F–C show potential to be trained to increase the accuracy of their ratings. A future implication is to devise, implement, and test an F–C training protocol to enhance the accuracy and reliability of their ratings. If this proves successful, then the SDBM will have the potential to be used by F–C as a proxy self-report tool for identifying safe and unsafe driving behaviors. Occupational therapists may play a critical role in interpreting the findings of such proxy reports and identifying entry points for logical and efficient driver safety interventions. ▲

Acknowledgments

This project was funded by National Institute on Aging Grant (R21) PAR-06-247 (Principal Investigator, Sherrilene Classen) and the University of Florida's Center for Multimodal Studies on Congestion Mitigation (CMS) 00063055 (Principal Investigator, Sherrilene Classen). We acknowledge the support of the Institute for Mobility, Activity, and Participation at the University of Florida and the Centre for Research on Safe Driving at Lakehead University.

References

AAA Foundation for Traffic Safety. (1994). *Drivers 55 plus: Check your own performance: A self-rating form of questions, facts and suggestions for safe driving.* Washington, DC: AAA Foundation for Traffic Safety.

American Automobile Association. (2004). *Roadwise review.* Heathrow, FL: AAA Public Affairs.

Bond, T. G., & Fox, C. M. (2001). *Applying the Rasch Model: Fundamental measurement in the human sciences.* Mahwah, NJ: Erlbaum.

Classen, S., Horgas, A., Awadzi, K., Messinger-Rapport, B., Shechtman, O., & Joo, Y. (2008). Clinical predictors of older driver performance on a standardized road test. *Traffic Injury Prevention, 9,* 456–462. doi: 10.1080/15389580802260026

Classen, S., Wen, P., Velozo, C., Bédard, M., Winter, S. M., Brumback, B., et al. (in press). Psychometrics of the self-report Safe Driving Behavior Measure for older adults. *American Journal of Occupational Therapy.*

Classen, S., Winter, S. M., Velozo, C. A., Bédard, M., Lanford, D. N., Brumback, B., et al. (2010). Item development and validity testing for a self- and proxy report: The Safe Driving Behavior Measure. *American Journal of Occupational Therapy, 64,* 296–305. doi: 10.5014/ajot.64.2.296

Eby, D. W., Molnar, L. J., Shope, J. T., Vivoda, J. M., & Fordyce, T. A. (2003). Improving older driver knowledge and self-awareness through self-assessment: The driving decisions workbook. *Journal of Safety Research, 34,* 371–381. doi: 10.1016/j.jsr.2003.09.006

Justiss, M. D., Mann, W. C., Stav, W., & Velozo, C. (2006). Development of a behind-the-wheel driving performance assessment for older adults. *Topics in Geriatric Rehabilitation, 22,* 121–128.

Kay, L. G., Bundy, A. C., & Clemson, L. M. (2008). Predicting fitness to drive using the visual recognition slide test (USyd). *American Journal of Occupational Therapy, 62,* 187–197. doi: 10.5014/ajot.62.2.187

Kay, L. G., Bundy, A. C., & Clemson, L. M. (2009). Predicting fitness to drive in people with cognitive impairments by using DriveSafe and DriveAware. *Archives of Physical Medicine and Rehabilitation, 90,* 1514–1522. doi: 10.1016/j.apmr.2009.03.011

Linacre, J. M. (2002). What do infit and outfit, mean-square and standardized mean? *Rasch Measurement Transactions, 16,* 878

Linacre, J. M. (2004). *Facets: Rasch measurement computer program.* Available at www.winsteps.com/index.htm

Marottoli, R. A., & Richardson, E. D. (1998). Confidence in, and self-rating of, driving ability among older drivers. *Accident Analysis and Prevention, 30,* 331–336. doi: 10.1016/S0001-4575(97)00100-0

McGee, P., & Tuokko, H. (2003). *The older and wiser driver: A self-assessment program.* Victoria, BC: Centre on Aging, University of Victoria.

Myers, A. M., Paradis, J. A., & Blanchard, R. A. (2008). Conceptualizing and measuring confidence in older drivers:

Development of the Day and Night Driving Comfort Scales. *Archives of Physical Medicine and Rehabilitation, 89,* 630–640. doi: 10.1016/j.apmr.2007.09.037

Myford, C. M., & Wolfe, E. W. (2004). Detecting and measuring rater effects using many-facet Rasch measurement: Part II. *Journal of Applied Measurement, 5,* 189–227.

Patomella, A. H., Kottorp, A., & Tham, K. (2008). Awareness of driving disability in people with stroke tested in a simulator. *Scandinavian Journal of Occupational Therapy, 15,* 184–192. doi: 10.1080/11038120802087600

Patomella, A. H., Tham, K., & Kottorp, A. (2006). P-drive: Assessment of driving performance after stroke. *Journal of Rehabilitation Medicine, 38,* 273–279. doi: 10.1080/16501970600632594

Stav, W. B., Justiss, M. D., McCarthy, D. P., Mann, W. C., & Lanford, D. N. (2008). Predictability of clinical assessments for driving performance. *Journal of Safety Research, 39,* 1–7.

Wild, K., & Cotrell, V. (2003). Identifying driving impairment in Alzheimer disease: A comparison of self and observer reports versus driving evaluation. *Alzheimer Disease and Associated Disorders, 17,* 27–34. doi: 10.1097/00002093-200301000-00004

Winter, S. M., Classen, S., Bédard, M., Lutz, B., Velozo, C. A., Lanford, D. N., et al. (2011). Focus group findings for a self-report Safe Driving Behavior Measure. *Canadian Journal of Occupational Therapy, 78*(2), 72–79. doi: 10.2182/cjot.2011.78.2.2

Wright, B. D., & Linacre, J. M. (1994). Reasonable mean-square fit values. *Rasch Measurement Transactions, 8,* 370.

Appendix.
Items on the Safe Driving Behavior Measure

Response options: *cannot do, very difficult, somewhat difficult, a little difficult, not difficult.*

How difficult is it for him or her to . . .

1. Open the car door?
2. Get in his or her car?
3. Turn the steering wheel?
4. Adjust the car mirrors?
5. Stay awake while driving?
6. Adjust the driver's seat so he or she can see above the steering wheel?
7. Stop for pedestrians crossing the roadway?
8. Drive in good weather?
9. Stay in the proper lane?
10. Drive during daylight hours?
11. Remember to turn on the headlights before driving in the dark?
12. Check for a clear path when backing out from a driveway or parking space?
13. Reach the gas pedal (accelerator) and brake pedal?
14. Press the gas or the brake when intended?
15. Use the car controls (such as the turn signals, windshield wipers, or headlights)?
16. Place the car in the correct gear (such as drive or reverse)?
17. Operate the emergency brake?
18. Check car mirrors when changing lanes?
19. Read road signs far enough in advance to react (such as make a turn)?
20. Obey varied forms of traffic lights (such as green arrow for turn lane or flashing lights)?
21. Drive and hold a conversation with one or more passengers?
22. Drive with a passenger who is providing driving directions or assistance?
23. Drive in light rain?
24. Drive on a highway with two or more lanes in each direction?
25. Keep up with the flow of traffic?
26. Keep distance from other vehicles when changing lanes?
27. Change lanes in moderate traffic?
28. Drive cautiously (to avoid collisions) in situations when others are driving erratically (such as speeding, road rage, crossing lane lines, or driving distracted)?
29. Brake at a stop sign so car stops completely before the marked line?
30. Maintain lane when turning (not cut corner or go wide)?
31. Back out of parking spots?
32. Enter the flow of traffic when turning right?
33. Share the road with vulnerable road users such as bicyclists, scooter drivers, motorcyclists?
34. Drive on graded (unpaved) road?
35. Check blind spots before changing lanes?
36. Drive with surrounding tractor trailers (transport trucks)?
37. Merge onto a highway?
38. Use a map while driving?
39. Make a left-hand turn crossing multiple lanes and entering traffic (with no lights or stop signs)?
40. Parallel park?
41. Stay within the lane markings unless making a lane change?
42. Stay within the proper lane in the absence of road features such as clearly marked lane lines, reflectors, or rumble strips?
43. Keep distance between his or her car and others (allow time to react to hazards)?
44. Look left and right before crossing an intersection?
45. Drive in a construction zone?
46. Drive in dense traffic (such as rush hour)?
47. Pass (overtake) a car in the absence of a passing lane?
48. Pass (overtake) a larger vehicle such as an RV, tractor trailer (transport truck), or dump truck in the absence of a passing lane?
49. Drive in an unfamiliar urban area?
50. Control his or her car when going down a steep hill?
51. Exit an expressway or interstate from a left-hand lane?
52. Drive in a highly complex situation (such as a large city with high-speed traffic, multiple highway interchanges, and several signs)?
53. Control the car (brake hard or swerve) to avoid collisions?
54. Drive a different car (such as another person's car or a rental car)?
55. Alter his or her driving in response to changes in health (such as vision, reaction time, fatigue, thinking, joint stiffness, medications)?
56. Drive when upset (anxious, worried, sad, or angry)?
57. Stay focused on driving when there are distractions (such as radio, eating, drinking, pet in the car)?
58. Drive in an unfamiliar area?
59. Drive at night?
60. Avoid dangerous situations (such as car door opening, car pulling out, road debris, or animal darting in front of car)?
61. Drive when there is fog?
62. Drive at night on a dark road with faded or absent lane lines?
63. Drive when there is glare or the sun is in his or her eyes?
64. Turn left across multiple lanes when there is no traffic light?
65. Drive in a thunderstorm with heavy rains and wind?
66. Control his or her car on a wet road?
67. Drive on a snow-covered road?
68. Drive on an icy road?

Note. The complete Safe Driving Behavior Measure is available on request from Sherrilene Classen.