Educational and Psychological Measurement
Volume 66 Number 6
December 2006 984-1000
© 2006 Sage Publications
10.1177/0013164406288170
http://epm.sagepub.com
hosted at
http://online.sagepub.com

Measuring Attentional Ability in Older Adults

Development and Psychometric Evaluation of DriverScan

Lesa Hoffman
The Pennsylvania State University
Xiangdong Yang
The University of Kansas
James A. Bovaird
University of Nebraska–Lincoln
Susan E. Embretson
Georgia Institute of Technology

Although deficits in visual attention are often postulated as an important component of many declines in cognitive processing and functional outcomes in older adults, surprisingly little emphasis has been placed on evaluating psychometric instruments with which individual differences in visual attention ability can be assessed. This article reports the development and beginning psychometric evaluation of DriverScan, a change detection measure of attentional search for older adults. A constrained graded response model is used to approximate response speed and accuracy with categories of immediate, delayed, or no response. DriverScan items are shown to have excellent reliability over the studied sample, and the distribution of items is shown to adequately cover the difficulty continuum and to be maximally sensitive at distinguishing individuals with lower than average abilities (i.e., individuals with attention deficits). Item design features representing goal-directed and stimulus-driven attentional processing significantly predict item difficulty as hypothesized.

Keywords: *visual attention; aging; item response theory; change detection*

Attention is the mechanism by which certain aspects of the environment are selected for further processing while others are inhibited and appears to be an integral component of most cognitive tasks. Although generally paired with saccadic eye movements, it has repeatedly been shown that visual attention also operates independently (for general reviews, see Chun & Wolfe, 2001; Wolfe, 2000). Extensive research has demonstrated that attention has a limited capacity—that is, we cannot simultaneously attend to everything perceived by the optical system. This limited capacity restricts the speed and accuracy of performance in a variety of situations.

With respect to aging, it has become increasingly necessary to examine how the capacity or speed of attentional processing can change with age and to understand the extent to which these changes can impact the visual functioning of a rapidly growing older adult population. Given that attention is a necessary component in many everyday skills that require search and prioritization of information, such as when driving or interacting with technology (e.g., Web sites, ATMs), measuring individual differences in attentional abilities among older adults would allow prediction of continued success with tasks such as these. In addition, age-related deficits in attentional processing are often postulated as one of the mediating factors in other types of cognitive decline (Craik, 1977; Hasher & Zacks, 1988), and thus reliable measurement of attentional abilities may be necessary in evaluating theories of cognitive aging.

Age-related changes in attention are often investigated by examining the extent to which experimental manipulations create quantitative or qualitative differences in the speed and accuracy of performance of older adults as compared to younger adults. In general, this research has suggested that only some aspects of attention appear compromised in older age (see McDowd & Shaw, 2000, for an excellent review). Older adults appear similar to younger adults in their ability to ignore distracting information when the target has a known location or is easily distinguished from distractors (D'Aloisio & Klein, 1990; Plude & Doussard-Roosevelt, 1989) and show similar responses to external events that prompt shifts of attention, such as a new object (Gottlob & Madden, 1998). Older adults have also been shown to be similar to younger adults in their use of task contingencies and expectancies to improve performance (Humphrey & Kramer, 1997; Madden, Gottlob, & Allen, 1999). Age-related deficits have been found when visual search is necessary, however, and become more pronounced with increasing numbers of distractors or target-distractor similarity (Folk & Lincourt, 1996; Scialfa & Joffe, 1997) or when insufficient processing time is allotted to encode and respond to cues that direct endogenous attention shifts (Broduer & Enns, 1997; Folk & Hoyer, 1992).

Age deficits are most often evaluated at the group level by comparing convenience samples of younger and older adults on their performance of carefully controlled and necessarily contrived experimental tasks. Accordingly, individual differences observed within this paradigm are regarded as error, a nuisance to be controlled as much as possible. Yet such individual differences in attentional abilities are likely to be relevant to many cognitive and functional outcomes. The challenge then, is how experimental research findings can be integrated into psychometric instruments with which individual differences in attention can be assessed.

Authors' Note: This work was supported by a dissertation award to the first author from the State Farm Companies Foundation. We thank Scott Hofer and Joan McDowd for their helpful comments in preparing this article and Devon Land, Alicia Mackay, Christian Stopp, and Imad Uddin for their assistance in collecting data. Please direct any inquiries to Lesa Hoffman, Gerontology Center, The Pennsylvania State University, 135 E. Nittany Avenue, Suite 405, State College, PA 16801-5363, e-mail: lesa2@psu.edu.

One such alternative is the Useful Field of View© (UFOV), a three-subtest experimental task presumed to measure the spatial extent of the attentional window (Ball & Owsley, 1993; Owsley, Ball, Sloane, Roenker, & Bruni, 1991). A central discrimination task is performed, with or without a peripheral localization task, and the presentation time needed for 75% accuracy in each subtest is the primary outcome. Psychometric evaluation of the UFOV has mostly been limited to examination of its prediction of accident risk in older drivers; when participants have been sampled heavily for visual impairments and prior history of accidents, the UFOV has shown good sensitivity and specificity in predicting accidents (Owsley et al., 1998). Yet given considerable evidence suggesting that older adults do not appear differentially affected relative to younger adults by the narrowing of their attentional window (e.g., Seiple, Szlyk, Yang, & Holopigian, 1996; Sekuler, Bennett, & Mamelak, 2000), it is clear that other dimensions should be included when measuring attentional ability in older adults.

As reviewed above, older adults appear relatively more impaired in situations in which one must actively search in the presence of competing information for an object needed for further cognitive processing and when contextual or environmental assistance is limited. This ability to orient attention efficiently throughout the visual field in response to internal goals or external stimuli has been described as *attentional search* or *shifting* (as opposed to *attentional scaling*, the ability to adjust the size and focus of attention, for which the UFOV was developed; see Greenwood & Parasuraman, 2004). To date, we are aware of no psychometric instruments with which individual differences in attentional search ability in older adults can be assessed.

The focus of the current studies was the development and initial psychometric evaluation of DriverScan, a measure of attentional search ability in older adults. DriverScan was derived from a change detection task developed by Rensink, O'Regan, and Clark (1997). Original (A) and modified (A′) digital photographs are presented for 280 milliseconds (ms), and blank screens are interspersed for 80 ms. In this presentation (*A, blank, A, blank, A′, blank, A′, blank* . . .), search for a change between repeated presentations of an otherwise identical scene must be conducted through controlled processing, because local luminance cues at the change location are unable to direct attention in the presence of a global luminance change. It often takes considerably longer to notice even large, salient changes than when such changes are presented without interruption, a phenomenon known as *change blindness*. Both goal-directed and stimulus-driven orienting have been found to facilitate change detection (Scholl, 2000; Werner & Thies, 2000), and deficits in change detection speed have been differentiated from deficits in spatial attention (Pringle, Irwin, Kramer, & Atchley, 2001). The change detection measure of attentional search described in this article incorporates the context of driving. Besides the obvious practical application, driving scenarios provide a natural visual environment in which certain objects and locations are inherently prioritized and in which other salient events can also be important, which permits use of goal-directed and stimulus-driven processing, respectively.

The current work expands on previous research by interfacing an experimental task with latent trait methodology. This approach to instrument development is novel within the study of cognitive aging in two regards. First, a cognitive design systems approach can be used to incorporate cognitive theory into item generation to enhance construct validity (Embretson, 1998). As a result, *construct representation*, or the meaning of a construct, can be differentiated from *nomothetic span*, or the instrument's utility in measuring individual differences, as evaluated through correlational studies (i.e., whether expected relationships are observed with measures of theoretically related constructs). These two dimensions are often confounded in test development, such that the validity of a test is implied by the strength of its relationship with other tests (see Embretson, 1983; Smith, 2001).

Construct representation is addressed by designing items with features that reflect processes or knowledge thought essential to the ability being measured and evaluating the role of these features in predicting item difficulty. In DriverScan, three features based on findings from previous research will be used to predict item difficulty. The first, *visual clutter*, or the congestion of the scene, reflects the greater difficulty of detecting changes in the presence of many competing sources of information. The second, *change relevance*, or the extent to which the change would be meaningful to the driver in the photograph, reflects the contribution of goal-directed orienting in improving search speed. The third, *change brightness*, or the physical salience or conspicuity of the change, reflects the greater difficulty of detecting changes when they are smaller in size or contrast, given the visual impairments that often occur in older adults, and reflects the contribution of stimulus-driven orienting in improving performance.

The second way in which the current approach to instrument development is novel is in its incorporation of *latent trait* or *item response* modeling (Embretson & Reise, 2000). In item response theory (IRT) modeling, because latent abilities and item difficulties are placed along the same metric, scores on different test versions can be equated through linking procedures, and change scores can also be meaningfully compared. Differences among the items in terms of their stimulus features, content, or discrimination can be included in ability and item estimation. Once item properties are calibrated, IRT tests can be customized for specific needs (i.e., general versus point-specific measurement) or even specific persons through adaptive testing, and precision of measurement (i.e., test information) can be evaluated at each level of the latent trait.

In summary, the purpose of the current studies is to develop a psychometric instrument to assess individual differences in attentional search in older adults using an experimentally derived change detection task within the context of driving. In Study 1, DriverScan is developed and refined through pilot testing. In Study 2, the revised instrument is administered to 155 older adults to examine its psychometric properties, including dimensionality, test information and reliability of item responses in the sample, distributions of item difficulty and person ability, and the relation between the item design features and item difficulty.

Study 1: Instrument Development

The goal of Study 1 is to develop the DriverScan instrument and obtain values for the instrument design features from younger and older adults. The pilot version of DriverScan is then administered to younger and older adults to examine the range in item performance.

Method

DriverScan development. Sixty-four items were created using Adobe Photoshop 6.0 from pictures of several midwestern metropolitan areas. Changes were introduced as object deletions, color changes, or lettering changes on signs. Items were first categorized into conditions of low or high visual clutter by an experimenter. Items were then assigned into conditions of central or peripheral (greater or less than 7° visual angle) change location and low or high change relevance pseudo-randomly (i.e., eight pictures could not be modified as assigned and were reassigned into different conditions) to prevent any confounding between change location and the other factors. A change of high relevance was operationalized as involving an alteration to an object usually important for safe driving, or as an alteration that changed the importance of the object relative to the safety of the driver. For example, removal of a car near the driver's vehicle would be a change of high relevance, because the presence of a nearby car would likely impact a driver's response, whereas removal of a logo on the same car would be considered a change of low relevance, because the actions of the driver would not be affected. High-relevance changes included changes to stoplights, pedestrians, construction markers, turn signals, road or street signs, and removal of near cars. Low-relevance changes included changes to logos on pedestrians or cars, billboards, light poles, buildings, trees, and relatively distant cars.

Change brightness was then measured for each item via differences in the changed area of the picture in the luminance and red/green/blue color channels. Six measures were created: absolute mean difference in luminance, summed color, and luminance and summed color combined, as well as each of these weighted by the number of pixels altered. Because these measures were strongly correlated, a linear composite variable was generated through principal component analysis, *physical change brightness*, which was uncorrelated with assigned visual clutter, change location, and change relevance (r = –.02, .06, –.08, respectively).

To summarize, 64 items were created independently varying across the dimensions of visual clutter, physical brightness of the change, relevance of the change to driving, and balanced for central or peripheral change location. Correspondence between the design features of clutter and relevance as assigned and as rated subjectively by groups of younger and older adults was examined next. Furthermore, these observers also provided subjective ratings for change brightness to account for the effects of perceptual grouping and other top-down mechanisms.

Participants. Ratings for the design features were provided by 7 men and 13 women aged 68 to 89 years ($M = 76.8$, $SD = 5.7$) and 7 men and 13 women aged 18 to 24 years ($M = 19.9$, $SD = 1.2$). The pilot instrument was then administered to 6 men and 5 women aged 73 to 87 years ($M = 77.6$, $SD = 4.1$) and 6 men and 14 women aged 18 to 23 years ($M = 20.3$, $SD = 1.5$). Older adults were recruited by phone and received $10 each. Younger adults from a large midwestern university participated in partial fulfillment of a psychology course requirement.

Apparatus. DriverScan was presented for the older adults (OA) on a 17-inch (in.) CRT monitor at a distance of about 30 in., and for the younger adults (YA) on a 21-in. CRT monitor at 38 in. to approximate 24° visual angle for each group. Each participant was tested individually.

Ratings task procedure. Participants were told they would be viewing photographs of real-world driving scenes and asked to answer questions about each photograph. For each design feature, participants were first familiarized with the construct and shown low and high exemplars. Four pictures not included in the current study were also presented to illustrate independent variation in change relevance and change brightness. Items were then presented in a random order. Participants were instructed to rate between 1 and 100 their perception of the magnitude of each design feature, relative to the exemplars. Visual clutter was always rated first. Half of the participants then rated change brightness second and change relevance third, with this order reversed for the other half. Participants could view the change as many times as needed.

Change detection task procedure. Participants were told that they would be viewing digital photographs of real-world driving scenes with one change made between successive presentations. They were instructed to find the change as quickly as possible and to respond with the mouse and verbal report of the change. Eprime software was used to present the items and record response times, and accuracy was recorded by the experimenter. The original (A) and modified (A′) photographs were presented for 280 ms along with blank screens presented for 80 ms in the sequence *A, blank, A, blank, A′, blank, A′, blank* for 45 seconds (sec.) or until a response. This duration was chosen on the basis of previous research in which the change was unlikely to be detected after 45 sec. of viewing. After reviewing the instructions and several practice trials (three for younger adults, eight for older adults), items were presented in a random order.

Results

Change detection task. Not detecting the change within 45 sec., or *time-outs*, composed 97% and 99% of incorrect responses for OA and YA, respectively. Ten items with < 75% accuracy for the YA and 6 additional items with < 50% accuracy for the

OA were removed because their difficulty was likely to be too high for the target population of OAs.

Ratings task. Ratings were within-person centered by calculating individual *Z*-scores across the responses for each participant to remove bias from differential use of response ranges, and the arithmetic mean of the transformed ratings across observers within each age group was used as the design feature value for each item within each age group. Although both the item distributions for visual clutter and change brightness were normal, the distribution for change relevance was bimodal, and as such, items were divided into groups of low and high change relevance. Two items that were outliers in their design features were removed: one item with a physical brightness score 5 *SD* above the mean, and another item with Cook's leverage values in the multivariate relations among the design features above the recommended cutoff of .13, as calculated from $2p/N$ (Kutner, Nachtsheim, Neter, & Li, 2004).

In the 46 final items, there was strong correspondence between the experimenter-assigned groupings (low or high) and sample-obtained ratings (unit-normal) as indicated by tetrachoric correlations for visual clutter (OA $r = .82$, YA $r = .83$) and for change relevance (OA $r = .82$, YA $r = .86$). However, Pearson correlations between physical change brightness and obtained change brightness ratings (unit-normal) were close to zero (OA $r = .05$, YA $r = .13$). Although it was anticipated that objective and subjective measures of change brightness would differ due to top-down influences on perception, the complete lack of correspondence is surprising. Furthermore, although the obtained ratings for visual clutter were uncorrelated with those for change brightness (Pearson $r < .15$) and change relevance (tetrachoric $r < .02$), a strong relationship was obtained between ratings of change relevance and change brightness, as indicated by tetrachoric correlations (OA $r = .85$, YA $r = .69$). This suggests that these constructs may be inherently related subjectively, despite our efforts to disentangle them experimentally. Finally, similarity between OA and YA in the rank ordering of the items per dimension was excellent, as indicated by Pearson correlations for visual clutter ($r = .89$) and change brightness ($r = .73$) and by tetrachoric correlations for change relevance ($r = .99$). Given the target population, only ratings from the OA were averaged to create values for each of the three design features for the 46 items. A description of each item and its design features is available from the first author, and item examples are available at http://www.personal.psu.edu/lrh15/research/aging/DriverScan.

Study 2: DriverScan Administration

In Study 1, the DriverScan instrument was developed and refined through pilot testing, and values for the design features were obtained. We next evaluate its psychometric properties through latent trait modeling on a larger sample. IRT models are primarily used for dichotomous or polytomous responses, such as accuracy or partial accuracy, whereas confirmatory factor analysis (CFA) models are primarily used for continuous responses. In this study, however, persons of higher ability should detect changes more quickly, whereas persons of lower ability should detect changes less quickly or not at all (i.e., a time-out). Modeling only response accuracy would ignore potentially informative variability in response time. Yet modeling response time as a continuous outcome would be misleading, given that response time is right-censored (i.e., a 45-sec. limit was imposed in administering the items), and thus the slowest responses for each person may not be fully observed. To address this issue, a censored response model (CRM; Maddala, 1983) will be estimated as shown in Equation 1:

$$X_{is}^{*} = b_i + a\theta_s + e_{is} \quad \text{where} \quad \begin{array}{ll} X_{is} = X_{is}^{*} & \text{if } X_{\min} \leq X_{is}^{*} \leq X_{\max} \text{ and} \\ X_{is} = X_{\max} & \text{if } X_{is}^{*} > X_{\max} \end{array}, \tag{1}$$

where θ_s is the ability of person *s*, b_i is the intercept of item *i*, *a* is the common slope, and *e* is the residual. The model links X_{is}^{*}, the unobserved potential response time, with X_{is}, the observed response time for item *i* for person *s*, as shown, where $X_{\min} = 0$ and $X_{\max} = 45$.

However, no direct correspondence to IRT parameters has been developed for the CRM. To approximate censored responses within an IRT model, we opted to create an ordinal (polytomous) variable that captures both time and accuracy in empirically defined categories (as will be explained in the Results section) of *immediate*, *delayed*, and *no response* (scored 2, 1, and 0, respectively). We then examined to the extent possible the congruence between parameters from the CRM and those from a constrained version of the graded response model (GRM; Samejima, 1969). For a polytomous item with $j + 1$ categories, *j* between-category threshold (difficulty) parameters will be modeled within the GRM. Let $X_{ijs} = 1$ if person *s* falls in category *j* or above for item *i*. The *j*th threshold can then be modeled as shown in Equation 2:

$$P(X_{ijs} = 1 | \theta_s, \beta_{ij}, \alpha_i) = \frac{\exp[\alpha_i(\theta_s - \beta_{ij})]}{1 + \exp[\alpha_i(\theta_s - \beta_{ij})]}, \tag{2}$$

where θ_s is the ability of person *s*, β_{ij} is the *j*th threshold parameter for item *i*, and α_i is the discrimination parameter for item *i*. β_{ij} is the point on the latent trait where person *s* has a 50% probability of responding in the category *j* or above for item *i* (i.e., where $\theta_s = \beta_{ij}$).

The GRM was chosen instead of other available models for polytomous data (e.g., partial credit, nominal) because response time has a clear underlying continuum. The GRM is consistent with the categorization of the continuous responses in that the property of additivity of the model holds (i.e., finer recategorization or combining of two or more categories together are possible), and a continuous response IRT model is a logical extension of the GRM (see Samejima, 1995).

Discrimination parameters (α_i) were constrained to be equal across items for three reasons. First, the interpretation of abilities in relation to item features is more complicated within a two-parameter model, in which items are weighted by the relative strength of each item component (Embretson, 1998). Second, examination of the item-total correlations suggests that discrimination is relatively homogeneous across items (i.e., correlations between .2 and .4). Finally, the sample (able to be obtained) in Study 2 was not of recommended size to estimate both difficulty and discrimination parameters (Truskosky, 2000).

To examine whether the three item features incorporated into the design of DriverScan are related to item difficulty as hypothesized, item difficulty was modeled in the GRM as a linear combination of the item features by substituting for β_{ij} in Equation 2 as follows:

$$\text{Difficulty} = c + (\tau_{\text{clutter}} \times q_{i,\text{clutter}}) + (\tau_{\text{relevance}} \times q_{i,\text{relevance}}) + (\tau_{\text{brightness}} \times q_{i,\text{brightness}}), \quad (3)$$

where the τs represent estimated regression weights for each feature, the q_is represent values for each feature, and c is a scaling constant (Fischer, 1973). For each item, the first threshold was specified as a linear combination of the item features, and an additive term was estimated with which to obtain the second threshold (see Rijmen, Tuerlinckx, De Boeck, & Kuppens, 2003).

To summarize, latent trait modeling was used in Study 2 to investigate three issues: the psychometric properties of DriverScan, the convergence of parameters from the GRM with those from the CRM, and the extent to which design features could predict item difficulty.

Method

Participants. A sample of 155 community-dwelling, currently licensed drivers were recruited by phone, consisting of 68 men (44%) and 87 women (56%) between 63 and 87 years of age ($M = 75.2$, $SD = 4.7$). The majority of participants were White ($n = 149$, 96%), and the rest were African American ($n = 6$, 4%). Participants each received $30 as compensation.

Apparatus and procedure. DriverScan was presented on a 17-in. LCD monitor at a distance of about 30 in., subtending 24° visual angle. DriverScan was presented as described in Study 1, except that participants completed eight practice trials before the 46 items. Participants also completed measures of visual acuity, contrast sensitivity, spatial attention, and simulated driving as part of a larger study (Hoffman, McDowd, Atchley, & Dubinsky, 2005).

Model estimation. The CRM was estimated in Mplus 3.0 (Muthén & Muthén, 1998-2004) using robust maximum likelihood. The GRM was estimated in PARSCALE 4.1 (Scientific Software International; Du Toit, 2003). Person parameters were estimated via *expected a posteriori* scoring ($M = 0$, $SD = 1$). Item parameters were estimated via marginal maximum likelihood. The item location and category parameters were used to calculate the j item category thresholds, as defined in Equation 2. Item feature weights (τ) were estimated in SAS Proc Nlmixed (De Boeck & Wilson, 2004). To empirically define *immediate* and *delayed* responses for the GRM, overall response frequency and the number of items with sufficient responses in each category were examined across several cut-points. A cut-point of 8 sec. was chosen, classifying 30% of responses as *immediate* (within 8 sec.) and 48% as *delayed* (between 8 and 45 sec.). Time-outs (no response within 45 sec.) occurred for 19% of responses; less than 3% of responses were wrong or inadvertent and were treated as missing at random. Of the 46 items given, the 38 in which > 3 responses were observed within each category were analyzed.

Results

Unidimensionality in the ability underlying the DriverScan items was examined by fitting a one-factor model with categorical indicators using the WLSMV estimator in Mplus 3.0. A one-factor model fit acceptably when slopes were constrained to equality (i.e., a one-parameter IRT model), root mean square error of approximation (RMSEA) = .05; and when slopes were estimated per item, RMSEA = .04. The constrained GRM was then estimated and fit acceptably, $\chi^2(429) = 422$, $p = .59$. Only one item exhibited statistically significant misfit, $p = .02$.

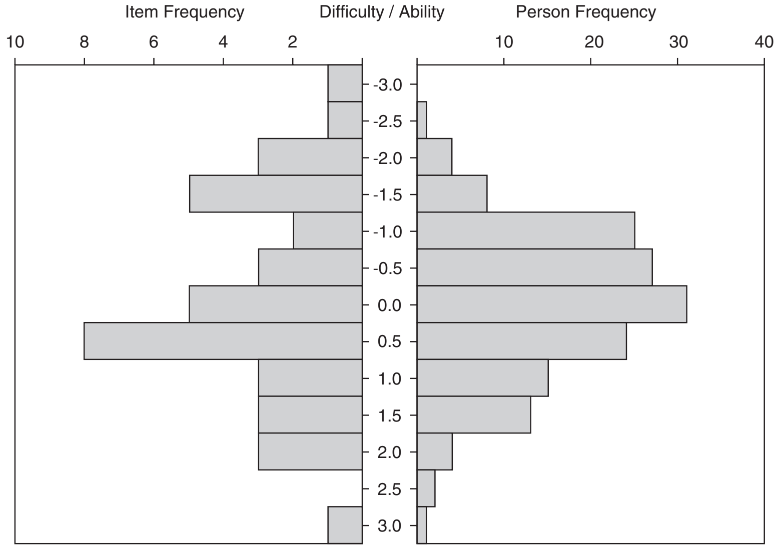
Figure 1 displays the distributions for the overall item locations on the left and abilities on the right, where the *y*-axis is the latent metric of difficulty/ability, and the *x*-axis on each side represents frequency. Although persons and items were relatively well matched, there was a slightly greater concentration of less-difficult items. Item location estimates ranged from –3.01 to 2.99 (*M* = –0.07, *SD* = 1.46); standard errors ranged from .28 to .34 (*M* = .31, *SD* = .02). Ability estimates ranged from –2.31 to 2.95 (*M* = 0, *SD* = 1); standard errors ranged from .52 to .61 (*M* = .55, *SD* = .14) and were largest at the extremes where responses were less frequent.

Figure 2 displays the distribution of item category thresholds under the GRM. First category thresholds (i.e., delayed or immediate vs. no response) ranged from –5.32 to 0.71 (*M* = –2.36, *SD* = 1.54); standard errors ranged from .26 to .65 (*M* = .36, *SD* = .11). Second category thresholds (i.e., immediate vs. delayed or no response) ranged from –1.09 to 5.39 (*M* = 2.36, *SD* = 1.58); standard errors ranged from .26 to .74 (*M* = .36, *SD* = .13).

The top of Figure 3 displays the item category characteristic curves for an item with an overall location parameter of .40. As shown, no response (i.e., a time-out) is most likely for this item until ability reaches –1.0, above which point a response between 8 and 45 sec. is most likely until ability reaches 1.8, above which point a response under 8 sec. becomes most likely. The bottom of Figure 3 displays the overall test information curve from the GRM, which shows that measurement precision is highest among lower ability levels. Overall test reliability for the sample was calculated as .88, as derived for polytomous models (Lord, 1980, p. 52).

The extent to which the polytomous treatment of time and accuracy could approximate the censored response time was examined via correlations of GRM and CRM parameters. The item locations obtained in the GRM were strongly related to the item

Figure 1
Distribution of DriverScan Item Difficulty (Left) and Ability (Right) Parameters



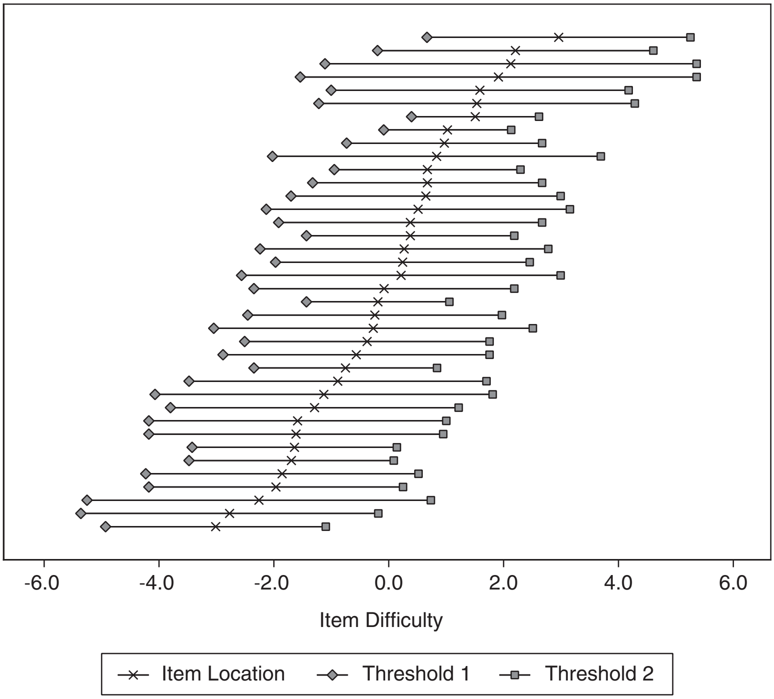
Note: The *y*-axis is the common metric of difficulty/ability, and the *x*-axis on each side is frequency.

intercepts obtained in the CRM, $r(36) = .93$, $p < .001$, as were the ability parameters obtained in the GRM and the factor scores obtained in the CRM, $r(153) = .95$, $p < .001$. There was a slight trend for items with locations of large values and persons of higher abilities to be better differentiated in the CRM.

The extent to which visual clutter, change relevance to driving, and change brightness could predict item location was then examined. Item locations from the weighted combinations of the item features were significantly correlated with item locations from a saturated model (in this case, the constrained GRM, in which separate locations were estimated per item), $r(36) = .49$, $p = .002$. The weight (standard error) for each of the item features was statistically significant at $p < .05$: visual clutter $\tau = .18$ (.04), change relevance $\tau = –.45$ (.07), and change brightness, $\tau = –.49$ (.08). As expected, greater levels of visual clutter, lower levels of change relevance, and lower levels of change brightness were related to greater item difficulty.

Given the moderate prediction of item locations from the hypothesized item features, it is likely that other item features may play a role in determining item difficulty. One such feature was chosen for further examination: whether the change was made to a legible sign, which occurred on 13 items. Incorporating this feature resulted into a

Figure 2
Distribution of Item Locations and Category Thresholds



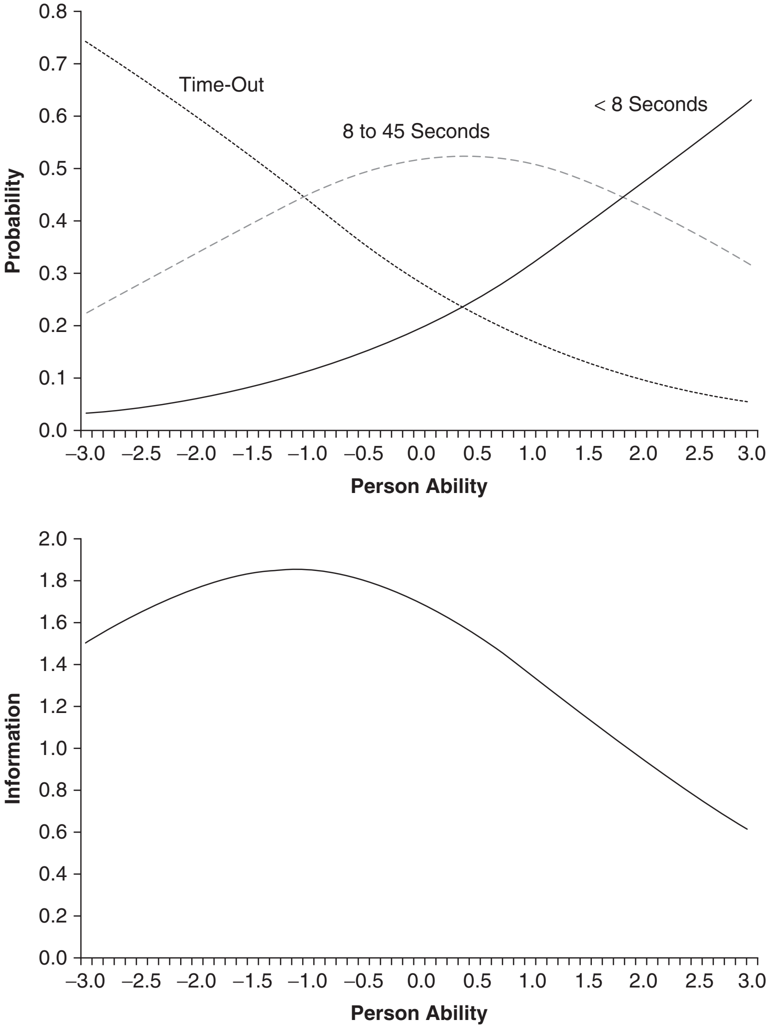
Note: Each line represents one item.

significant feature weight as well as a statistically significant correlation between feature-predicted and obtained item locations, $r(36) = .62, p < .001$; items with changes to legible signs were less difficult.

General Discussion

Although deficits in visual attention are often postulated as an important component of many declines in cognitive processing and functional outcomes in older adults, surprisingly little emphasis has been placed on evaluating psychometric instruments with which individual differences in attentional ability can be assessed. The current work reports the development and beginning psychometric evaluation of an instrument to measure attentional search, or the ability to orient attention in response to internal goals or external stimuli. In DriverScan, observers are asked to detect changes between successive presentations of an otherwise identical scene,

Figure 3
Top: Example Item Characteristic Curve for an Item With a Location of .40. Bottom: Test Information Curve Across All Items.



Note: Each line represents one item.

with blank screens interspersed between each scene to minimize the extent to which local luminance cues can direct attention. Efficient allocation of attention and eye movements in response to goal-directed, endogenous factors (e.g., scene context) as well as stimulus-driven, exogenous factors (e.g., change size or relative salience) is required to detect the change.

Latent trait models were used to estimate attentional abilities and to examine the psychometric properties of the item responses. A constrained version of the GRM was estimated to appropriately model both speed and accuracy of response. Ability estimates and item locations from a GRM were almost perfectly correlated with factor scores and item intercepts from a censored response model, suggesting that the categories of *immediate*, *delayed*, or *no response* successfully approximated right-censored response time. Given the positive skew in the response times (i.e., a preponderance of relatively fast response times), we did not feel there were sufficient responses in the right tail of the distribution to create multiple categories; thus, a cut-point of 8 sec. was chosen to create categories of *immediate* and *delayed* response only. Although this is an admittedly arbitrary criterion, similar analyses conducted using alternative cut-points (e.g., 10, 12, and 15 sec.) suggested that higher abilities were not as well differentiated under those criteria. Replication in additional samples is clearly needed to best assess where such a cut-point might be, as well as what an optimal time limit might be for distinguishing *delayed* from *no response* (i.e., greater or less than 45 sec.).

The psychometric properties of DriverScan were then examined. The item responses in the sample were found to be sufficiently unidimensional and reliable, and a broad range of item difficulty was represented. DriverScan had maximum test information at lower ability levels, such that individuals with attentional deficits were measured most precisely. However, because DriverScan was built on the principles of IRT, the test content and, accordingly, the range of ability it can measure, need not be seen as fixed or finite. New items could be added to extend the range of difficulty or for specific assessment purposes (i.e., discrimination around a criterion score). Because abilities are estimated in reference to items and not to group norms, ability estimates will remain comparable as the instrument undergoes revision. Further study with larger samples is also needed to evaluate potential differences among the items in discrimination, or the strength of their relation to the latent trait; the efficiency of DriverScan may be improved by removing or replacing less discriminating items in the future.

Construct representation was addressed by examining the relation between DriverScan item locations and item features representing goal-directed and stimulus-driven attentional processes. Difficulty in attentional search was associated with the amount of visual clutter in the picture, the brightness of the change, and the relevance of the change to driving. The effect of visual clutter replicates previous work in which performance of older adults was hampered by increasing amounts of distracting information (e.g., D'Aloisio & Klein, 1990; Folk & Lincourt, 1996; Madden et al., 1999; Scialfa & Joffe, 1997). The effect of change brightness replicates previous work within the change detection paradigm of the relation between change salience and performance (e.g., Pringle et al., 2001; Williams & Simons, 2000). In addition to these stimulus-driven factors, the effect of change relevance replicates previous work within the change detection paradigm demonstrating the impact of context on performance (e.g., Hollingworth & Henderson, 2000; Rensink, O'Regan, & Clark, 1997; Werner &

Thies, 2000) and suggests that it is likely that participants were using goal-directed expectations about scene content to guide their eye movements and attention. An additional factor identified post hoc, whether the change was made to a legible sign, further reflects goal-directed processing in measuring attentional search.

Although statistically significant, the relationships between the item locations from the item design features and those from the saturated model were not overly strong (*r*s ≈ .6). Thus, other item features may also be relevant in predicting item difficulty, especially given that these items are natural scenes in which many other properties could differ than those investigated. Future work should examine additional features that may be related to attentional search ability or, alternatively, strive to control for extraneous factors that may be related to item difficulty but that are unrelated to ability. Reliable prediction of difficulty from item features could ultimately enable the creation of items via computer-generated graphics, which would allow for greater flexibility in test configuration and reduce retest effects arising from repeated item exposure.

Although evidence for the validity of the inferences made from an instrument can be demonstrated through the underlying relations between item features and performance, another crucial aspect of validity is the extent to which scores on the instrument demonstrate expected or predictive relationships with other constructs (i.e., nomothetic span). As presented in Hoffman et al. (2005), DriverScan has been shown to be related to other measures of attention, such as the subtests of the UFOV and has predicted simulated driving performance in older adults.

In sum, the current work describes the development and psychometric examination of DriverScan, a change-detection measure of attentional search ability in older adults. Latent ability and item threshold parameters were estimated from a constrained version of a graded response model of speed and accuracy. The distribution of test information in DriverScan was shown to be maximally sensitive at lower abilities (i.e., deficits), and item responses within the sample were shown to be sufficiently unidimensional and reliable. Item locations were significantly predicted by the item features of visual clutter, change brightness, change relevance, and whether the change was made to a legible sign. Although validation is always an ongoing process, this preliminary research suggests DriverScan to be a useful measure of individual differences in attention search ability in older adults.

References

Ball, K. K., & Owsley, C. (1993). The Useful Field of View test: A new technique for evaluating age-related declines in visual function. *Journal of the American Optomological Association, 63,* 71-79.

Brodeur, D. A., & Enns, J. T. (1997). Covert visual orienting across the lifespan. *Canadian Journal of Experimental Psychology, 51,* 20-35.

Chun, M. M., & Wolfe, J. M. (2001). Visual attention. In E. B. Goldstein (Ed.), *Blackwell's handbook of perception* (pp. 272-310). Oxford, UK: Blackwell.

Craik, F. I. M. (1977). Age differences in human memory. In J. E. Birren & K. W. Schaie (Eds.), *Handbook of the psychology of aging* (pp. 384-420). New York: Van Nostrand Reinhold.

D'Aloisio, A., & Klein, R. M. (1990). Aging and the deployment of visual attention. In J. Enns (Ed.), *Advances in psychology* (Vol. 69, pp. 447-466). North Holland: Elsevier.

De Boeck, P., & Wilson, M. (2004). *Explanatory item response models*. New York: Springer.

Du Toit, M. (2003). *IRT from SSI*. Lincolnwood, IL: Scientific Software International.

Embretson, S. E. (1983). Construct validity: Construct representation versus nomothetic span. *Psychological Bulletin, 93*, 179-197.

Embretson, S. E. (1998). A cognitive design system approach to generating valid tests: Application to abstract reasoning. *Psychological Methods, 3,* 380-396.

Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists*. Mahwah, NJ: Lawrence Erlbaum.

Fischer, G. H. (1973). The linear logistic test model as an instrument in educational research. *Acta Psychologica, 37,* 359-374.

Folk, C. L., & Hoyer, W. J. (1992). Aging and shifts of visual spatial attention. *Psychology and Aging, 7,* 453-465.

Folk, C. L., & Lincourt, A. E. (1996). The effects of age on guided conjunction search. *Experimental Aging Research, 22,* 99-118.

Gottlob, L. R., & Madden, D. J. (1998). Time course allocation of visual attention after equating for sensory differences: An age-related perspective. *Psychology and Aging, 13,* 138-149.

Greenwood, P. M., & Parasuraman, R. (2004). The scaling of spatial attention in visual search and its modification in healthy aging. *Perception & Psychophysics, 66*(1), 3-22.

Hasher, L., & Zacks, R. T. (1988). Working memory, comprehension, and aging: A review and a new view. In G. K. Bower (Ed.), *The psychology of learning and motivation* (Vol. 22, pp. 193-225). San Diego, CA: Academic Press.

Hoffman, L., McDowd, J. M., Atchley, P., & Dubinsky, R. (2005). The role of visual attention in predicting driving impairment in older adults. *Psychology and Aging*, *20*(4), 610-622.

Hollingworth, A., & Henderson, J. M. (2000). Semantic informativeness indicates the detection of changes in natural scenes. *Visual Cognition, 7,* 213-235.

Humphrey, D. G., & Kramer, A. F. (1997). Age differences in visual search for feature, conjunction, and triple-conjunction targets. *Psychology and Aging, 12,* 704-717.

Kutner, M. H., Nachtsheim, C. J., Neter, J., & Li, W. (2004). *Applied linear statistical models* (5th ed.). New York: McGraw-Hill.

Lord, F. M. (1980). *Applications of item response theory to practical test problems.* Hillsdale, NJ: Lawrence Erlbaum.

Maddala, G. S. (1983). *Limited-dependent and qualitative variables in econometrics*. New York: Cambridge University Press.

Madden, D. J., Gottlob, L. R., & Allen, P. A. (1999). Adult age differences in visual search accuracy: Attentional guidance and target detectability. *Psychology and Aging, 14,* 683-694.

McDowd, J. M., & Shaw, R. J. (2000). Attention and aging: A functional perspective. In F. I. M. Craik & T. A. Salthouse (Eds.*)*, *The handbook of aging and cognition* (2nd ed., pp. 221-292). Mahwah, NJ: Lawrence Erlbaum.

Muthén, B. O., & Muthén, L. K. (1998-2004). *Mplus user's guide* (3rd ed.). Los Angeles, CA: Muthén & Muthén.

Owsley, C., Ball, K., McGwin, G., Jr., Sloane, M. E., Roenker, D. L., White, M. F., et al. (1998). Visual processing impairment and risk of motor vehicle crash among older adults. *Journal of the American Medical Association, 279,* 1083-1088.

Owsley, C., Ball, K. K., Sloane, M. E., Roenker, D. L., & Bruni, J. R. (1991). Visual/cognitive correlates of vehicle accidents in older drivers. *Psychology and Aging, 6,* 403-415.

Plude, D. J., & Doussard-Roosevelt, J. A. (1989). Aging, selective attention, and feature integration. *Psychology and Aging, 4,* 98-105.

Pringle, H. L., Irwin, D. E., Kramer, A. F., & Atchley, P. (2001). The role of attentional breadth in perceptual change detection. *Psychonomic Bulletin and Review, 8,* 89-95.

Rensink, R. A., O'Regan, J. K., & Clark, J. J. (1997). To see or not to see: The need for attention to perceive changes in scenes. *Psychological Science, 8,* 368-373.

Rijmen, F., Tuerlinckx, F., De Boeck, P., & Kuppens, P. (2003). A nonlinear mixed model framework for item response theory. *Psychological Methods, 8,* 185-205.

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika Monograph*, No. 17.

Samejima, F. (1995). Acceleration model in the heterogeneous case of the general graded response model. *Psychometrika, 60,* 549-572.

Scholl, B. J. (2000). Attenuated change blindness for exogenously attended items in a flicker paradigm. *Visual Cognition, 7,* 377-396.

Scialfa, C. T., & Joffe, K. M. (1997). Age differences in feature and conjunction search: Implications for theories of visual search and generalized slowing. *Aging, Neuropsychology, and Cognition, 4,* 227-246.

Seiple, W., Szlyk, J. P., Yang, S., & Holopigian, K. (1996). Age-related functional field losses are not eccentricity dependent. *Vision Research, 36,* 1859-1866.

Sekuler, A. B., Bennett, P. J., & Mamelak, M. (2000). Effects of aging on the useful field of view. *Experimental Aging Research, 26,* 103-120.

Smith, E. V., Jr. (2001). Evidence for the reliability of measures and validity of measure interpretation: A Rasch measurement perspective. *Journal of Applied Measurement, 2,* 281-311.

Truskosky, D. M. (2000). *An empirical examination of classical test theory and item response theory parameters: Implications for research and practice in small- and large-sample assessments*. Unpublished doctoral dissertation, Southern Illinois University, Carbondale.

Werner, S., & Thies, B. (2000). Is "change blindness" attenuated by domain-specific expertise? An expert-novice comparison of change detection in football images. *Visual Cognition, 7,* 163-173.

Williams, P., & Simons, D. J. (2000). Detecting changes in novel, complex three-dimensional objects. *Visual Cognition, 7,* 297-322.

Wolfe, J. (2000). *Visual attention*. In K. K. De Valois (Ed.), *Seeing* (2nd ed., pp. 335-386). San Diego, CA: Academic Press.