

The Dimensional Model of Driver Demand: Visual-Manual Tasks

Richard Young, Sean Seaman, and Li Hsieh
Wayne State University

ABSTRACT

Many metrics have been used in an attempt to predict the effects of secondary tasks on driving behavior. Such metrics often give rise to seemingly paradoxical results, with one metric suggesting increased demand and another metric suggesting decreased demand for the same task. For example, for some tasks, drivers maintain their lane well yet detect events relatively poorly. For other tasks, drivers maintain their lane relatively poorly yet detect events relatively well. These seeming paradoxes are not time-accuracy trade-offs or experimental artifacts, because for other tasks, drivers do both well. The paradoxes are resolved if driver demand is modeled in two orthogonal dimensions rather than a single "driver workload" dimension. Principal components analysis (PCA) was applied to the published data from four simulator, track, and open road studies of visual-manual secondary task effects on driving. PCA reduced the task metrics to two underlying orthogonal components (hereafter, dimensions) which were consistent across studies, herein designated as *physical* and *cognitive* demand. *Physical demand* is associated with lateral and longitudinal driver performance (lane crossings, standard deviation of lateral position and speed), with correlated surrogate metrics of task time, step count, total glance time, number of glances, and subjective workload. *Cognitive demand* is associated with event detection (RT and miss rate), with correlated surrogate metrics of mean single glance time, long single glances, speed reduction, and task errors. The Dimensional Model of Driver Demand allows for a common simplified understanding of all these measures of visual-manual secondary task effects on driver performance.

CITATION: Young, R., Seaman, S., and Hsieh, L., "The Dimensional Model of Driver Demand: Visual-Manual Tasks," *SAE Int. J. Trans. Safety* 4(1):2016, doi:10.4271/2016-01-1423.

INTRODUCTION

Drivers must manage multiple primary driving subtasks concurrently. These include controlling the vehicle's movement, path and speed via steering, accelerating, and braking; making decisions about speed and lane maneuvers; monitoring and operating vehicle systems; maintaining awareness of the road and traffic situation; detecting and responding to roadway events (such as hazards, traffic lights, brake lights on a forward vehicle, etc.). Drivers also manage various device-related secondary tasks that they may want or need to perform in the vehicle while on the way to a destination, such as radio tuning or song selection, dialing a phone number, entering a destination, etc. To minimize any possible negative effects of device-related secondary tasks on the primary driving subtasks, it is necessary to have a means of determining how, in what ways, and by how much, some secondary tasks may have the potential to interfere with primary driving subtasks. There are non-device-related secondary tasks as well, such as eating, drinking, grooming, interacting with passengers and/or pets, reading signs along the roadway, gawking at stopped vehicles, etc. However, the scope of the current paper is restricted to primary driving subtasks, and the effect of device-related secondary visual-manual tasks on those subtasks.

In psychology, the classical one-dimensional (1-D) attention model [1,2] is that there is an all-or-none, single-channel bottleneck which can deal with data from only one signal at a time, so that data from a signal arriving during the response time (RT) to a previous signal has

to wait until the decision mechanism becomes free. Thus task A must be stopped to perform task B, and task A cannot be continued until task B stops. Other psychology researchers [3,4] extended this classical 1-D model by allowing for a sharing of the assumed single attentional resource rather than all-or-none within the single channel. This extended 1-D model suggests that a task subtracts "attention" from this pool, leaving less "attention" for concurrent tasks, thus task B produces detrimental effects on task A, but not a complete cessation of A. Some form of this extended 1-D model appears to be implicitly held by many in the field of driver attention and performance.

However, other experiments indicated some degree of independent processing of signals. Wickens [5] developed the "Multiple Resource" model, which has four "dimensions" with independent processing across the dimensions, and across the subcategories for each dimension (for review see [6]). The four dimensions and their subcategories were: "1. Codes" (spatial vs. verbal), "2. Modalities" (visual vs. auditory), "3. Stages" (perception, cognition, responding), and "4. Responses" (spatial-manual vs. vocal-verbal).

Later experiments, however, found cross-modality effects that were inconsistent with independent processing for the visual and auditory subcategories within the modality dimension [7]. For example, auditory and visual stimuli (as well as tactile) produce a similar pattern of interference with response time in dual-task situations [8] [9, Figure E.5]. Such cross-modality interference does not necessarily, however, support a 1-D central attentional resource model. Modern

cognitive neuroscience has identified three central attentional resources in the brain that must be considered for a complete understanding of driver attention.

Three Attentional Networks

There is now considerable evidence that any attentional model that involves only one central attentional resource is incomplete. In particular, neuroimaging procedures have identified three brain networks that underlie different types of attention (for reviews see [10,11,12,13]. These attentional networks include:

- *Orienting Attention*, which selects information from sensory input;
- *Executive Attention*, which involves mechanisms for monitoring and resolving conflict among thoughts, feelings and behavioral responses; and
- *Alerting Attention*, which achieves and maintains a state of high sensitivity to incoming stimuli.

The function and processing efficiency of the three networks have been assessed with the Attention Network Test (ANT), a flanker-type task in which differences in response time (RT) between various test conditions are used to evaluate each network [14]. The results of testing with the ANT and neuroimaging procedures indicate that the three networks engage separate brain mechanisms and are functionally independent. However, there are some interactions among them in real life and in certain tasks, and a complete and coherent understanding of the interactions between them is still an active area of research [15,16].

Application of Three Attentional Networks to Driving

Applications of measurements of the three attentional networks to driving is described by [17,18]. Greater understanding of the three attentional processes as applied to driver performance has been further investigated by [19,20,21,22,23].

Our own studies with brain imaging have shown a direct relationship of a specific area in the orienting attention network to event detection and response during simulated driving [24,25]. We measured event detection and response using the Detection-Response Task (DRT) [9], which was recommended in 2015 for publication as an ISO standard. Using the DRT in a simulated driving environment with magnetoencephalography (MEG), our team found that the response time to a red light (with or without a concurrent speech task) is predicted by the level of activation in the right superior parietal lobe [24]. This area is part of the orienting attention network (not the executive attention network, which is commonly assumed to be the only underlying attentional mechanism for secondary task effects on driving). However, the orienting attention network including the right superior parietal lobe can be modulated by top-down influence from the executive attention network in the midline frontal areas, as shown in an functional magnetic resonance imaging (fMRI) study during simulated driving while doing secondary tasks with a DRT [25]. In less technical terms, direct brain imaging during driving-like scenarios shows that the response times to external events are determined first and foremost by bottom-up stimuli to the part of the orienting

attention network whose level of activation almost perfectly predicts response times, but orienting attention can itself be modulated by top-down influences from executive attention networks.

It is intuitively obvious that alerting attention, which is involved in the sleep-wake cycle, “trumps” both orienting attention and executive attention, because these attentional networks cease to process external stimuli when someone falls asleep. Indeed, if the driver has observable drowsiness (such as from falling asleep, head nodding, closed eyelids, or the signs of an observable “behavioral” microsleep [26], the odds of a crash increase to 63 to 1 relative to a matched baseline [27], a far higher odds ratio than any secondary task ever recorded in a naturalistic driving study [97]. The true crash odds ratio for drowsy driving may be even higher than 63 to 1 because “EEG” microsleeps [26], unlike “behavioral” microsleeps, are not outwardly observable, and so would enter into the “not drowsy” category in an odds ratio calculation in the usual naturalistic driving study face video analysis methods, biasing the drowsiness OR downwards. In short, the valid OR for drowsy driving may be even higher than 63 to 1 because it is likely that EEG microsleep episodes were not correctly assigned to the drowsy driving category. Indeed, EEG microsleeps may be what have sometimes been attributed to the attentional effect of cognitive demand in causing crash (e.g., looked-but-did-not-see), when many such crashes may instead be caused by decreased alerting attention, an entirely different attentional effect than cognitive demand. In fact, there is evidence from naturalistic driving studies that cognitive demand from some secondary tasks such as cell phone conversation may reduce crashes, likely by reducing drowsiness somewhat (i.e., improving alerting attention) [27].

In the current paper, two previous publications on the Dimensional Model for visual-manual tasks are briefly reviewed, and the data are represented in a common framework using similar metric names where they overlap. Two additional studies are then analyzed and the results examine to determine if the Dimensional Model can be validated with additional datasets. The companion paper [28] extends the Dimensional Model of Driver Demand to auditory-vocal and *mixed-mode* tasks (i.e., with visual-manual-auditory-vocal modalities in the same task).

GENERAL METHODS

Primary and secondary task performances were measured and the data were analyzed using Principal Component Analysis (PCA). PCA was invented in 1901 by Karl Pearson [29], and it has been widely used in many scientific fields [30]. Its mathematical properties with application examples are available in many textbooks [31,32] or on websites [33]; space does not permit the technical details to be presented here. Simply, PCA reduces the dimensionality of the data set from that of the original number of variables to a few major components; in the current study, all the driver demand metrics were reducible to only two components with little loss in explanatory power in the four examples analyzed.

PCA as employed here operates on the correlation matrix between the metrics. A correlation uses *standardized* metrics (i.e., each metric is converted to a mean of zero by subtracting the original mean, and normalized to a standard deviation of one by dividing by its standard

deviation). The PCA operates on the entire matrix of correlations between all metrics. At times in this report, however, individual correlations between just two metrics are calculated and discussed to clarify the meaning or understanding of the PCA results. The correlation coefficient by itself indicates the strength of the relationship between the two variables. However, the high correlations between many of the metrics in this study is not always appreciated from just the high correlations coefficients. Therefore, the probability values that estimate the statistical significance of the correlation coefficients are given to full precision in this paper using engineering format. The usual publication practice of just using an asterisk to indicate " $p < .001$ " for all statistically significant correlations with a p value less than that does not do justice to the strength of the correlations between the metrics observed in these examples.

The initial stage of PCA finds the "loadings" or weights placed onto each of the metrics that make up each component. Coefficients can then be derived from those loadings using a mathematical process, and the individual tasks can then be scored using those coefficients. For example, if a task originally has 100 metrics measured on it that are highly inter-correlated, then PCA analysis can reduce those 100 metrics to say only 2 scores (these scores are effectively new metrics that are an optimum weighted composite of the original 100 metrics). The original metrics can be reproduced to a high degree of accuracy by just using the 2 scores. That is, the two scores contain almost all the information that was in the original 100 metrics.

The scores of the individual tasks on the components are guaranteed by the PCA to be orthogonal (i.e., to have a correlation of exactly 0). This result guarantees that the two sets of task scores on the two components are accounting for different sources of variance in the original dataset of all the metrics. PCA thus gives rise to a unique solution for the task scores (there is only a single solution which guarantees orthogonality in the task scores between every component with every other). Factor analysis rotation methods can be used to rotate the components after the PCA is completed, to align the final scores slightly better with original individual metric data, or in an attempt to understand better the underlying factors. But rotation will cause the component scores for the tasks to become correlated to varying degrees, depending upon the amount of rotation, so the underlying dimensions after rotation are then more properly referred to as *factors*. To keep it simple here, no rotation of the components was done, except in special circumstances where indicated, with the results placed into appendices.

Note that an intrinsic mathematical property of PCA is that the number of tasks evaluated must be at least one more than the number of metrics evaluated, because otherwise the covariance matrix is singular and no unique PCA solution is possible. In other words, the number of metrics that can be evaluated by a PCA is limited to one less than the number of tasks on which data was collected in the study. It is thus beneficial to have many tasks if possible, because more metrics can then be evaluated at one time. In the current paper, in the four examples given, the number of tasks ranged from 78 (Example 1) to only 5 (Example 4).

The PCA in the statistical package Minitab® 17.2.1 [34] was mainly used for the analyses, with results checked in Stata 13.1 [35]. Note that occasionally one statistical package may present the component

loadings on the negative of an axis compared to the axis loadings in another statistical package (a "sign" flip). This is a well-known property of PCA. A sign flip of an axis in no way affects the orthogonality or the relative positions of the points in the 2-D space. The two solutions can be equated by simply taking the negative of the metric loadings on the axis that is flipped, and the negative of the task scores on that component. Such a sign flip is not a rotation as described above.

For the reader who may be familiar with factor analysis, but not PCA, it may be useful to distinguish the two. Although PCA can be used as the initial stage of a factor analysis, factor analysis attempts to "simplify" the PCA solution even further by rotating the components, or attempting other various linear transformations of the components, in an attempt to "simplify" the factors (i.e., to make each factor have a value of 1 for one metric, and all other metrics have a value of 0, and so forth, for all factors extracted). After such linear transformations, the vectors are no longer components, but factors. There is one and only one principal component solution, but there are an infinite number of factor solutions. The final factors do not necessarily have a simple relationship to the original metrics or components. The unique solution provided by PCA does not attempt to isolate one metric vs. another, but instead attempts to provide the best description of the original data with the least number of components. Therefore, the Kaiser-Meyer-Olkin measure of "factorial simplicity" [36] (available with the "estat" post-estimation command in Stata [35]) is appropriate for factor analysis, but not for PCA. Because PCA provides a unique solution that has a simple relationship to the original data, it is therefore used throughout the main body of this paper rather than factor analysis using rotated components, which are confined to Appendix FS.

This paper provides validation examples for the Dimensional Model of Driver Demand, with closed-road, track, open road, and simulator datasets described in four published studies. A number of glance metrics were evaluated, including glances off the road as well as glances to the device. In the examples provided here, glances to the in-vehicle system under test are presented, rather than glances off the road, for reasons discussed in Example 4 under "metrics."

Since the four studies whose data are analyzed here were completed and published before the recent publication of the SAE Standard J2944 [37] providing common definitions of driving performance measures and statistics, the correspondence of a particular metric in the four studies to the definitions in that Standard is not readily determinable. For example, J2944 (Section 10.2.1.1) provides for 10 different ways to define a lane departure, depending upon which part of the vehicle (front tire, any tire, bounding box of mirror, vehicle body, etc.) encounters which part of the lane marker (middle of lane boundary, inside edge of lane boundary, outside edge of lane boundary, etc.) Although the precise definition of a lane departure likely varied somewhat between the four studies, which could affect the absolute number of lane departures recorded, the PCA method operates on the standardized metrics, which tends to ameliorate any slight differences in definitions of a lane boundary. In fact, as shown, the solutions for the underlying dimensions are similar in the four examples, despite substantial differences in the methods, protocols, or even venues (simulator, track, or open road field study). Similarly, SAE J2396 [38] provides definitions for the Glance-related Metrics

used in the four studies, but again it is difficult to determine after the fact to what extent there were differences, if any, in the definitions of the glance-related metrics in the four studies, which did not relate them to the SAE J2396 standard [38] at the time of the study. Despite these uncertainties, the results of the dimensional analysis of the four studies were consistent, indicating that the PCA method is robustly measuring the underlying demands on the driver, despite differences in the protocols of the studies.

VALIDATION EXAMPLE 1

Dataset 1: Young and Angell (2003)

Dataset 1 was from a closed-road test of 78 visual-manual infotainment and primary driving “anchor” tasks with 82 drivers in 5 vehicles [39] (see also [40]).

Task Set 1

All tasks in the entire dataset were analyzed, but they are grouped for ease of presentation as follows:

- *Address Book* (enter street, name, phone, etc.)
- *Anchor* (primary driving tasks such as adjust mirrors, vent, set cruise control, adjust HVAC, etc.)
- *Car Phone* (call from tag, call from list, 10-digit dial)
- *DialCell* (10-digit dial on portable cell phone)
- *Entertainment* (e.g., tune, change band, find CD track)
- *Internet* (get weather, get stock quote, etc.)
- *Navigation* (find school on map, enter destination)

Metrics 1

All metrics were tabulated over a single task completion while driving, after training to criterion while stationary, with the training-test cycle repeated for each task, one at a time. Fifteen performance metrics were collected:

- *Task metrics*: task completion time (TaskTime); percent unsuccessful task completions (Unsu%).
- *Vehicle metrics*: number of times tire touched or crossed lane marker during task (LaneX); number of times more than 5 mph above or below instructed speed of 40 mph (SpeedX). (Vehicle metrics are often termed “driving performance” metrics (e.g., Alliance Principle 2.1B [47, p. 40]).
- *Glance metrics* to displays or controls of device under test: total glance time (TGT); number of glances to the in-vehicle system (Glances); mean single glance time (MSGD).
- *Driver metrics*: subjective workload (Workload); subjective situation unawareness, (the complement of situation awareness) (Sit_UnAw).

- *Event detection and response metrics*: percent of total visual events missed (Miss%); percent of front visual events missed, where front event is activation of LED light on hood of vehicle (Miss%_front); percent of side visual events missed, where side event is activation of an LED light above the left outside mirror (Miss%_side); response time to all visual events (RT); response time to front visual events (RT_front); response time to side visual events (RT_side).

Results 1

An analysis of the correlation matrix [41, Table 2] shows that there were high correlations within two main subsets of the metrics. Hence, all 15 metrics reduced to only 2 using PCA. Dimension 1 accounted for 61% of the total variance in the dataset, dimension 2 accounted for 17%. As such, these 2 dimensions can reproduce 78% of the variance on all 15 original driver performance metrics. That is, the driver performance results for every task for the 15 metrics can be reproduced with a high degree of accuracy with only 2 scores for that task. (Note: We here use the term “driver performance” to indicate the inclusion of event detection and response metrics with the lane and speed maintenance metrics usually referred as “driving performance” metrics.)

Following PCA, the 2-D metric loadings on the two main components are plotted in Figure 1.

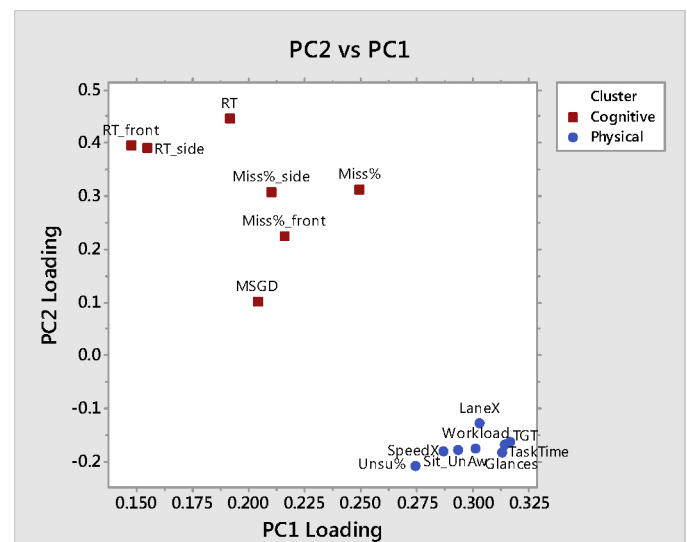


Figure 1. Metric loadings on the two major components (dimensions) (redrawn from [41, Figure 2]).

Cluster analysis [35] on the original mean data for all 15 metrics (the same data used for the PCA) was used to independently solve for the metric clusters, shown in Figure 2.

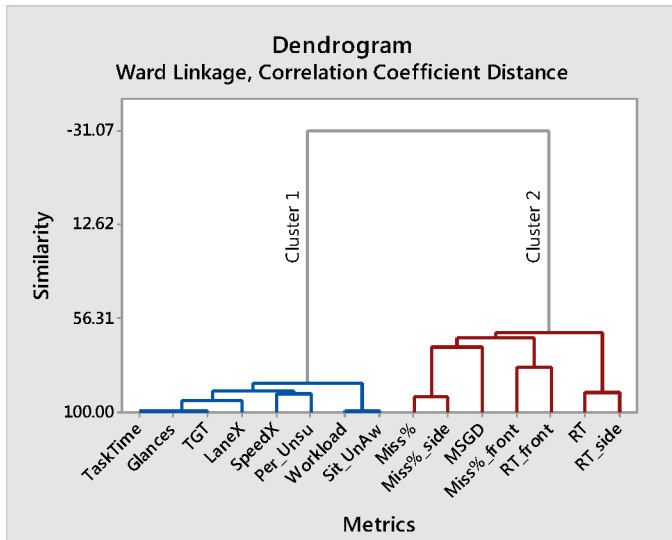


Figure 2. A cluster analysis identifies two major clusters in the metrics.

The cluster analysis in Figure 2 confirms the visual impression from Figure 1 that there are two main clusters of metrics.

Physical Demand Metrics 1

The blue lines in the left half of Figure 2 show that the set of metrics in the lower right quadrant of Figure 1 (blue circles) forms one cluster. These metrics load positively on Dimension 1 and negatively on Dimension 2. This cluster is composed of all the task, vehicle, and driver metrics assessed, as well as the number of glances (Glances), total glance time (TGT), and percent unsuccessful task completions (Unsu%). These metrics are all tightly correlated with each other (see [41, Table 2] for correlation matrix), and therefore are highly redundant. The vehicle metrics concerning lane and speed maintenance (“Driving Performance”), when measured during actual rather than simulated driving, are placed into the category of “ground truth” metrics in [42].

However, the Task, Glance, and Driver metrics that load into Dimension 1 are equally valid “surrogate” metrics for these “ground-truth” Vehicle metrics concerning lane and speed maintenance, because they are highly correlated with them, and are therefore equally suitable for representing Dimension 1. The PCA method provides the optimal combination of the information in all these metrics to form Dimension 1. For shorthand, we here label Dimension 1 as a *physical* demand dimension (even though it also contains the subjective metric of a driver’s perceived workload of a secondary task), because it has metrics that can be measured with direct physical observation (counts of lane and speed deviations, number of glances, task time, etc.). However, the Dimensional Model results indicate that it is equally valid to regard Dimension 1 as a “driving performance” dimension, and treat the non-vehicle metrics loading on this dimension as “surrogate” metrics for the “ground truth” driving performance measures of lane and speed maintenance.

Cognitive Demand Metrics 1

The cluster analysis in Figure 2 (right) also identifies a second main cluster of metrics. These metrics form the cluster in the upper left quadrant of Figure 1 (red squares). This cluster also has tightly-correlated metrics that are redundant with each other (but not with metrics loading onto Dimension 1). These are mainly event detection and response metrics, but it is of interest that mean single glance duration (MSGD) clusters with these metrics, and not the Dimension 1 metrics. Thus, the PCA results (Figure 1) and the cluster analysis results (Figure 2) independently indicate that MSGD is more closely associated with event detection and response than the metrics in cluster 1, despite the fact that the TGT metric in cluster 1 is the product of MSGD and number of glances. That is, this result indicates that MSGD adds no information to TGT beyond that in the number of glances. In other words, MSGD provides independent information about the second dimension of driver demand that is not predictable by the total glance time (TGT) or the number of glances to the device under test (Glances). This result may seem strange to some (that multiplying two variables could give rise to a third variable that uncorrelated with one of the original variables), but a multiplication of two variables is a non-linear transformation. In short, this analysis suggests that any effect on driver performance of Total Eyes-Off-Road Time (TEORT) or TGT to the device arises from the number of glances and not the duration of those glances. The PCA of the data in this example indicated that MSGD, and by implication the proportion of long glances LGP, have a separate effect on event detection performance not determinable from other glance metrics (number of glances, TGT, and TEORT). Because most of the Dimension 2 metrics are related to the detection and response to events, we hence here label this dimension a “cognitive” demand dimension for short-hand, but it is equally valid to label this dimension as an “event detection and response” dimension.

The PCA and cluster analysis thus confirm in independent analyses that each metric cluster forms a separate dimension for this set of tasks under the experimental conditions of this track study [39,41]. In short, an objective grouping of the metrics using cluster analysis on the original data (Figure 2), closely corresponds to the subjective visual impression of the metric groups in the PCA loading space in the Dimensional Model (Figure 1).

The results show that two dimensions capture a high percentage of the total variance in all 15 metrics assessed. The third dimension represents a peripheral/central vision effect for event detection, and is minor, accounting for only 5% of the remaining variance [41], and so is not considered further here. The remaining components constitute so such a small percentage of the little variance individually that are almost certainly statistical noise in the data.

Task Scores 1

The tasks can be scored on these two main components by forming the sum of the cross-product of the 15 metrics for each task with the loading coefficients for the component (for more information on this aspect of PCA, see any multivariate statistical book, for example [31,32]). The task scores are plotted in Figure 3.

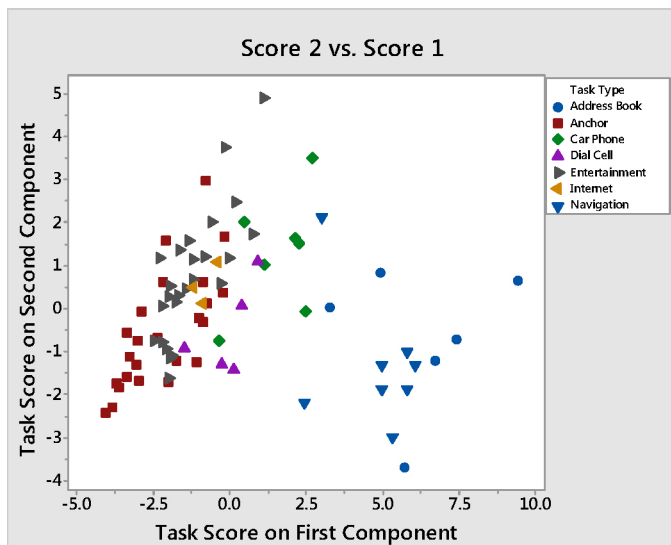


Figure 3. Task scores on the two components.

Tasks Scoring High on Physical Demand

Figure 3 shows that the Address Book and Navigation Tasks (blue circles and blue inverted triangles in the lower right quadrant) have relatively high S1 scores, reflecting the higher physical demand of those tasks (longer task times, more glances, more button presses, etc.). The vehicle metrics (LaneX, SpeedX) are also larger for those tasks. However, this effect may arise because of the confounding factor of TaskTime. With a longer driving time, the lane position varies more, even for baseline driving with no task [43, Figure 166] (a similar confound also happens for the standard deviation of speed). In other words, lateral and longitudinal vehicle variation metrics (e.g., standard deviation of lane position, standard deviation of speed) do not meet the statistical requirement that the samples must be drawn from a time-invariant population. (A similar result holds for the counts of the lane deviations LaneX and speed deviations SpeedX, but not speed itself which tends to load negatively on the cognitive demand dimension because of driver self-regulation [67].)

This confounding with driving time does not mean that the driving performance variability metrics are invalid as metrics for indicating driving safety. It does mean that their properties are not the same as statistical variability measures (whose values do not systematically increase with the number of samples). In other words, the commonly used driving performance variability metrics are inherently confounded by the duration of the temporal window in which they are evaluated. For example, the standard deviation when applied to samples drawn from a fixed population does not systematically increase with the number of samples or the time window over which it is measured, it just produces a more accurate estimate of the true population standard deviation. The fact that the SDLP tends to increase with time (and the number of samples), whether “just driving” with no task [43, Figure 133], or with the same secondary task performed repeatedly, indicates that it and other driving performance measures that are correlated with drive time do not meet the definition of a stationary process in statistics [44], and so they are not valid estimates of a population variance.

This confounding does not mean that driving performance metrics are not useful or valid for addressing “distraction” issues. They are equally valid as any other Dimension 1 metric (TaskTime, Glances, TGT, subjective workload, etc.). If their values on a secondary task are compared to the values of a “just driving” baseline task, or some other benchmark task such as “radio tuning,” it is necessary, however, that the time over which data is collected and the road geometry be the same, or there will be a confounding of driving performance variables by TaskTime. Artificially adjusting all tasks to be equal duration by repeating them over and over again for a fixed time interval is not a satisfactory solution to eliminate such confounding. This method effectively “impeaches” Dimension 1, which is the major underlying component of task demand for visual-manual tasks that was intended to be measured (see further discussion of this topic in Example 4). Artificially forcing all visual-manual tasks to have equal task time effectively equalizes all the Dimension 1 metrics, destroying task differentiation along the Dimension 1 axis, leaving only the differentiation along the Dimension 2 axis.

The fact that the driver performance metrics co-vary with task time (and number of glances, etc.) also does not imply that tasks should simply be made as short as possible. A short task can have poor or good event detection performance, depending upon the design of the task. For example, re-designing a task to be completed as short a time as possible by placing many buttons at a single level rather than in menus can increase mean single glance duration and reduce event detection performance. Conversely, a vehicle manufacturer who “chunks” a short but cognitively-demanding visual-manual task into manageable subtasks that are intentionally strung out over a longer task time (within limits), can arguably make it less cognitively demanding for the driver to perform it, with a plausible net benefit for driving safety.

For example, some tasks, such as fine adjustment of the radio treble (top symbol on y-axis in Figure 3), scored relatively high on the second dimension. These are relatively short tasks with relatively few glances, yet they have relatively high miss rates and long RTs. Conversely, the navigation and address book tasks (blue symbols) with high scores on the first component, generally have low scores on the second component. This is a paradox from a simple one-dimensional driver workload model, but there is no contradiction from the two-dimensional driver demand model presented here. Task scores in the two dimensions are orthogonal, so high scores on one dimension can be associated with either high or low scores on the other dimension; there is a correlation of zero between task scores on the two dimensions, as guaranteed by PCA.

One might also attempt to explain the apparent “paradox” of some short tasks having high cognitive demand as a simple tradeoff between concentrating on lane maintenance or on event detection, in a kind of speed-accuracy trade-off [45,46]. That is, participants attempt to maintain their lane with more accuracy, and so lose speed in their responses to the RT task. However, other tasks have both relatively good lane maintenance and relatively good event detection (tasks in the lower left quadrant of Figure 3 for example), so this paradox cannot be explained as a simple time-accuracy trade-off. It is also not a simple methodological artifact of shorter tasks, because again many shorter tasks have relatively good event detection as well as relatively good lane maintenance.

Task Scores on Cognitive Demand

Figure 3 shows that the task scores are distributed in a band from the lower left corner with “anchor” or primary driving tasks (red squares), up to certain “entertainment” tasks in the upper middle (two rightward pointing grey triangles at the top middle). It is obvious that adding the second dimension causes the tasks to be differentiated from each other much better with both dimensions than using only a single dimension. That is, the Dimensional Model offers improvements over the common practice of investigating single metrics one at a time in a vain attempt to segment the tasks with a single metric, or even a set of single metrics all chosen from a single dimension.

Simplification to Two Original Metrics

If the Dimensional Model is valid, then all the metrics within each set are redundant with each other. Hence, the PCA method gives rise to the “simplification” hypothesis; namely, only two metrics from the original dataset should be needed to characterize each task in terms of the two underlying driver demand dimensions, with little loss of information. Although all 15 original metrics have a statistically significant correlation with the task scores (S1) on the first component, TaskTime had the highest correlation ($r = 0.955$, $df = 76$, $p = 5.5E-42$), with the regression fit shown in Figure 4.

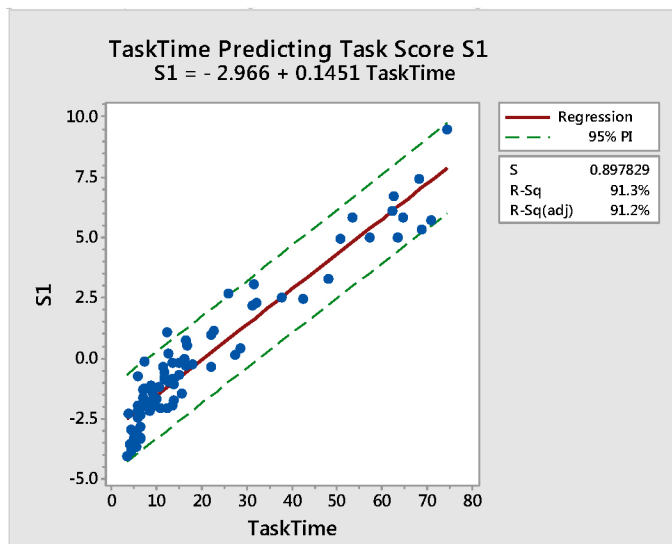


Figure 4. Prediction of Task Score 1 by TaskTime ($r = 0.955$, $df = 76$, $p = 5.5E-42$).

Similarly, RT had the highest correlation of all the metrics with the task scores S2 on the second component ($r = 0.706$, $df = 76$, $p = 5.1E-13$), with the regression fit in Figure 5. The dashed lines that the 95% prediction interval for S2 in Figure 5 is wider than for S1 in Figure 4, but the task scores on S2 in Figure 5 are still within or near the prediction interval.

Given these encouraging results, the simplification hypothesis was tested by using just TaskTime and RT to span the 2-D score space (Figure 6).

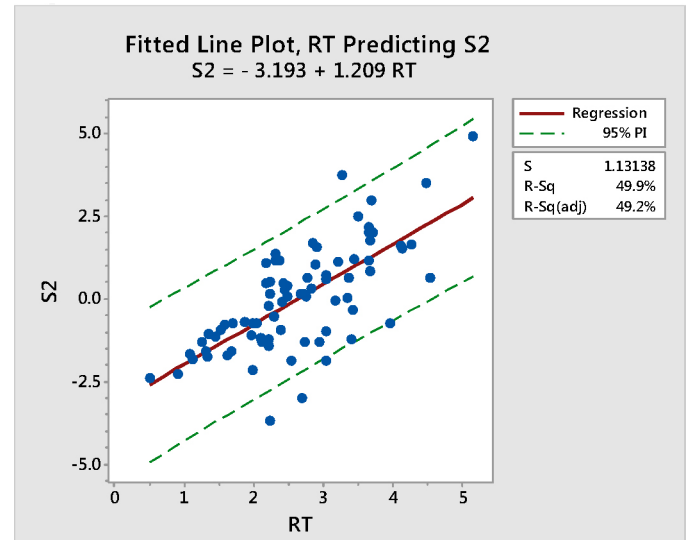


Figure 5. Prediction of Task Score 2 by RT ($r = 0.706$, $df = 76$, $p = 5.1E-13$).

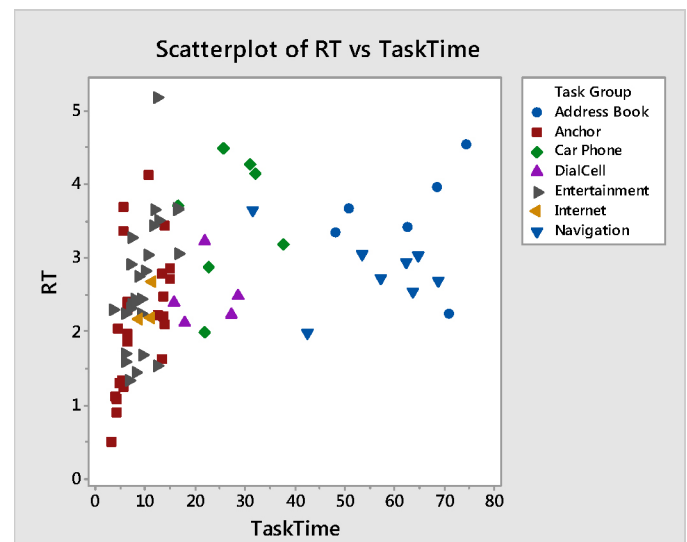


Figure 6. Event response time (RT) vs. TaskTime, original mean metrics.

Visual comparison of Figures 3 and 6 shows similar relative positions of the tasks in the 2-D space, except for a slight counterclockwise rotation of the points in the original RT and TaskTime metrics in Figure 6 vs. Figure 3, which is composed of the scores based on all 15 metrics.

The original RT and TaskTime are weakly correlated [41, Table 2] ($r = 0.281$, $df = 76$, $p = 0.012$). It is preferable to have orthogonal estimates of the underlying task scores, because then the errors across the two metrics are independent, making the error ellipse axes aligned horizontally and vertically rather than tilted (see General Discussion later). Multiple regression analysis can solve for the estimated orthogonal principal component scores $\hat{S}_1$ and $\hat{S}_2$ (see Figure 3 for the original S1 and S2 scores) from the original RT and TaskTime means, as per Equations (1) and (2).

$$\hat{S}_1 = -4.763 + 0.13071 \text{TaskTime} + 0.7920 \text{RT} \quad (1)$$

$$\widehat{S2} = -3.271 - 0.05030 \text{ TaskTime} + 1.6289 \text{ RT} \quad (2)$$

Using Equation 1, S1 is predicted with a slightly higher correlation by TaskTime and RT conjointly ($r = 0.981$, $df = 76$, $p = 7.0E-56$), than by TaskTime alone ($r = 0.955$, $df = 76$, $p = 5.5E-42$, Figure 4).

Using Equation 2, S2 is predicted with a higher correlation by TaskTime and RT conjointly ($r = 0.914$, $df = 76$, $p = 1.7E-31$) than by RT alone ($r = 0.706$, $f = 76$, $p = 5.1E-13$, Figure 5). Figure 7 plots the $\widehat{S1}$ and $\widehat{S2}$ scores as predicted by RT and TaskTime conjointly, using Equations 1 and 2.

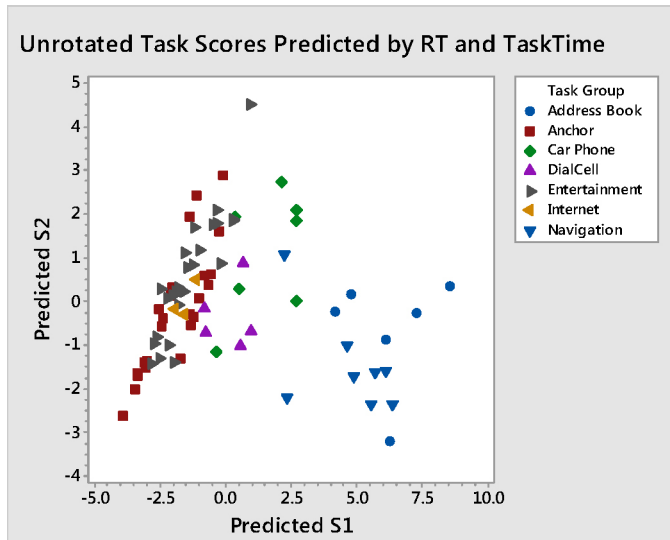


Figure 7. Unrotated task scores on Dimension 2 vs. Dimension 1 predicted by RT and TaskTime conjointly.

By visual inspection, Figure 7 matches even better with Figure 3 (the original task scores solved for using 15 metrics) than did the original individual TaskTime and RT metrics (Figure 6). That is, RT and TaskTime conjointly can accurately predict the orthogonal task scores in Figure 3 that originally required all 15 metrics. The estimated $\widehat{S1}$ and $\widehat{S2}$ from RT and TaskTime conjointly are now virtually orthogonal ($r = -0.06$, $df = 76$, $p = 0.62$), even more so than the already low estimates from RT and TaskTime individually ($r = 0.281$, $df = 76$, $p = 0.012$).

Rotation

The points in the 2-D plots in Figures 6 and Figure 3 line up slightly better after Varimax rotation of the components, which is a standard orthogonal rotation that balances the variances explained by the two components (see detailed rotation results in Appendix F). Because orthogonal estimates of S1 and S2 can be successfully made from TaskTime and RT conjointly (Figure 7 and previous paragraph), rotation was not further pursued in this study.

Control for Correlation between TaskTime and S1, RT and S2 in Example 1

One may argue that the high predictive ability of the single metric TaskTime for S1 (based on 15 metrics) shown in Figure 4, arose simply from the fact that the single metric TaskTime was included in

the calculation of the composite task score S1, so TaskTime is just correlating mainly with itself. The counter-hypothesis to this argument is that the other variables besides TaskTime in Metrics Cluster 1 are sufficiently redundant with TaskTime that a reduced Dimension 1 (with TaskTime left out of the PCA) should give rise to about the same task scores as if TaskTime had been included.

Likewise, other variables than RT in Metrics Cluster 2 should be sufficiently redundant with the RT metric that they give rise to about the same task scores for PCA, whether the three RT variables are included in the PCA or not. In short, the counter-hypothesis predicts that it should make little difference in the relative task score locations in the final 2-D score space, whether TaskTime and the three RT variables are included in the PCA or not. This counter-hypothesis was tested by dropping TaskTime and all 3 RT variables from the original dataset, and performing PCA on the reduced set of metrics.

The results given in Appendix G show that there is little effect whether TaskTime and the three RT metrics are included in the PCA or not, the predictions of RT and TaskTime are still good. This control confirms the redundancy of all those metrics loading within Dimension 1, and all those loading within Dimension 2.

Discussion 1

A Principal Components Analysis of a closed-road study of the effects on driving of 78 secondary tasks of varying types found that 15 driver demand metrics can be reduced to 2 principal components (dimensions), with little loss of explanatory power. This result led to a further simplification hypothesis that the task scores as found by PCA for the 15 metrics may be reasonably approximated with a reduced set of only 2 of the original metrics (instead of 15).

This further simplification hypothesis was tested using two non-glance metrics (TaskTime and RT). These non-glance metrics sufficed to span the two-dimensional space for the complete data set of 15 metrics, which included task, vehicle, glance, driver, and 6 detection and response metrics (3 RTs and 3 miss rates to center and side event lights and their combination). In short, the original dataset for all 15 metrics is reproducible to a high degree of accuracy from only two of the original metrics (TaskTime and RT). A control analysis (Appendix G) was conducted which eliminated TaskTime and the three RT metrics from the PCA. TaskTime and RT could still well predict the task scores in this reduced PCA 2-D score space. That means that TaskTime and RT alone can reproduce the data of the 13 other metrics to a high degree of accuracy, because of the enormous redundancy (i.e., collinearity) within the subsets of the metrics that load onto Dimension 1 and Dimension 2 respectively.

The Dimensional Model predicts that other pairs of metrics besides TaskTime and RT should be able to reasonably span this 2-D space as well (e.g., for example TGT and MSGD, as per the Alliance Guidelines Principle 2.1A [47]). However, the current choice of TaskTime and RT requires fewer resources to test and analyze than glance metrics, allowing more tasks to be tested with a given amount of resources, so these other options were not explored systematically for visual-manual tasks.

The metrics in some current test methods do not span both dimensions of driver demand for visual-manual tasks. For example, the Occlusion Goggles test (e.g. NHTSA Guidelines [48]) offers no additional information beyond that contained in TaskTime, a Dimension 1 metric. For the 13 visual-manual tasks in the CAMPDWM project [51,52], the correlation between static task time [52] and Total Shutter Open Time (TSOT) [52, Table Q-6] for the 13 visual-manual tasks was nearly perfect ($r = 0.993$, $df = 11$, $p = 9.6E-41$; see [Appendix Table E1](#)). TSOT thus provides no information beyond that of static task time, and therefore the Occlusion Goggle test provides no information about how tasks will score on the second dimension of driver demand for visual-manual tasks. In short, the simpler and easier measurement of static task time with just a stop watch will do all that the occlusion goggles test does. NHTSA erroneously rejected the use of task time as a metric (see General Discussion below), and adopted the occlusion goggles test as a final recommended test protocol, despite the fact that the occlusion goggles test is measuring the same underlying physical demand as TaskTime. Perhaps the occlusion goggles test was thought to have better “face validity” because the TSOT gives the appearance of being related to TGT and number of glances metrics, which may be associated to some degree with crash risk according to some studies. But according to the evidence presented here, none of these other Dimension 1 metrics provide any information beyond that contained in TaskTime, or even the number of manual steps in the task [73], which is highly correlated with TaskTime and all the other Dimension 1 metrics [84]. As pointed out, being able to use number of manual steps as an effective surrogate metric is of “great value for phases in product development where no operational prototype may yet exist” [42]. Its relevance to analytically modelling dual-task interactions during actual driving by participants trained on the task under assessment has been well proven by the Example 1 dataset [39, 40].

Likewise, the Alliance Principle 2.1B is based on the driving performance metrics of number of lane exceedances and standard deviation of headway as measured in a simulator. These are considered as closer to “ground truth” metrics [49] in actual driving and may arguably be more closely related to driving safety (as are driver event detection and response metrics) than “surrogate” measures, such as task time or eye glance metrics. These driving performance measures are unfortunately both highly correlated (i.e., confounded) with driving time, even when no task is performed at all (i.e., “just driving”). One way to avoid this confounding is to compare the secondary tasks to a “just driving” baseline, where the time of the baseline driving is the same as the task time of the secondary task. It can then be determined to what extent there is a change from baseline driving, without confounding by task time. It is also possible to statistically control for differences in task time by using partial correlation methods. Even if time matching or partial correlation methods are done, the driver performance metrics as in Alliance Principle 2.1B [47] will still capture only the Dimension 1 demand according to the Dimensional Model. The Model requires that driving performance metrics must be augmented with a Dimension 2 metric (such as an event detection metric) in order to assess both dimensions of driver demand.

As mentioned in the Results 1 section above, there is a trade-off between Dimension 1 and Dimension 2. Thus a long task can be made shorter to reduce its score on Dimension 1, but it has to be carefully designed so as not to increase the mean single glance duration (or long glance proportion), or performance on Dimension 2 will get worse (Event Detection performance will suffer). Likewise, a vehicle manufacturer may deliberately wish to increase the time of a task in order to break into more easily manageable “chunks”, thereby reducing long glances and improving event detection performance on Dimension 2. The overall goal should be to achieve a “sweet spot” as many tasks do, with relatively low scores in the two dimensions, or with relatively lower scores than a benchmark task or tasks measured in the two dimensions.

VALIDATION EXAMPLE 2

Dataset 2: Young (2012)

Young [50] analyzed the driver performance data from 13 visual-manual tasks from the CAMP-DWM study (see [51,52] for task details) on a closed test track with 69 drivers, of which 42 (61%) had eye data scored. Data were also collected in a simulator and on public highways in addition to the track. The simulator eye data were not scored in the CAMP-DWM study because of budget constraints, and so they were not available for validation purposes (that is, they could not be used to determine if the task scores would be the same with and without eye movement metrics).

In addition, the open-road data had fewer tasks than on the track or lab (e.g., some tasks, such as the destination entry task on the 2001 model vehicle tested, was judged to be too unsafe to perform on a public highway with traffic, even in the CAMP-DWM “platoon” scenario [50, Figure 2]).

Therefore, only the track data were analyzed in Example 2, using PCA as previously presented [50]. The original mean data are given in [Appendix B, Table B1](#).

Task Set 2

Complete descriptions of the visual-manual tasks whose data were analyzed in Example 2 are given in [52, [Appendix B](#)]. A description of the HVAC, RadioEasy, and RouteTrace tasks is given in [50, [Appendix](#)].

Metrics 2

The CAMP-DWM track study tabulated over 100 driver and vehicle performance metrics, the mean data for many of which were published in Appendix Q of the CAMP-DWM report [52]. Young [50, Section 2.2.1] selected 10 metrics from the full set that were the closest match to those in the first PCA study [41] of driver demand (see Example 1). (Because of redundancy, many other subsets of metrics from the CAMP-DWM study could have been chosen for the PCA with comparable results.) The selected metrics included objective vehicle measures (number of lane and speed deviations), as well as driver eye movements, task performance variables such as number of steps and

successful task completion, as well as RT and miss rate to real events (forward vehicle brake or decelerate, rear vehicle turn signal as seen through rear-view mirrors). All metrics were tabulated only during a task. The glance metrics evaluated here were based on glances to the display or controls for the device under test, and not all glances off the road, which includes glances to locations other than the road or the device under test, such as mirrors, windows, instrument cluster, etc. The specific metrics selected for the validation test of the Dimensional Model were selected to be close to those in Example 1, with the limitation that only 12 metrics could be picked because there were only 13 visual-manual tasks (as stated previously, the number of metrics must be one less than the number of tasks):

- *Task metrics*: mean task completion time (TaskTime).
- *Vehicle metrics*: standard deviation of lane position (SDLP) - a continuous measure of lane keeping during the task; speed difference (SpeedX) - the difference of maximum and minimum speed during task.
- *Glance metrics* to displays or controls of device under test: total glance time (TGT); mean single glance time (MSGD).
- *Event detection and response metrics*: For the center high mount stop lamp (CHMSL) in the forward vehicle (which was illuminated without an accompanying deceleration) - percent misses (Miss%_CHMSL) and mean response time (RT_CHMSL); for the lead-vehicle deceleration (LVD) event (lead-vehicle “coast-down” deceleration without brake-light illumination) - percent misses (Miss%_LVD) and mean response time (RT_LVD). Note that these events were single surprise events, not the repeated DRT events employed in Validation Example 1.

Other subsets of the 100 or so CAMP-DWM metrics were not systematically explored, but they would be expected to give the similar result that all could be condensed into only two scores per task after PCA. Details of the data collection methods in the CAMPDWM study are given in [51,52], and details of the PCA methods used here are given in [50].

Results 2

Principal Component Analysis (PCA) of the CAMP-DWM track data found that the 10 metrics analyzed reduced to two major components (dimensions) (Figure 8).

Only two components (as in Example 1) captured 78% of the total variance in the original 10 metrics - 53% of the variance on Dimension 1, and 25% on Dimension 2 (Figure 8) [50]. That is, all the data for all 10 metrics for all the tasks tested can be reproduced with only 2 metrics per task instead of 10, with little loss of explanatory power.

The metrics fell into two major metrics clusters according to an independent cluster analysis of the original data (Figure 9). Metrics falling into Cluster 1 were the Task and Vehicle metrics (left half of Figure 14; blue circles in lower right quadrant of Figure 8). Metrics falling to Cluster 2 were the event detection and response metrics and

MSGD (right half of Figure 9; red squares in upper left quadrant of Figure 8). This clustering pattern is consistent with the results for Validation Example 1 (Figure 1).

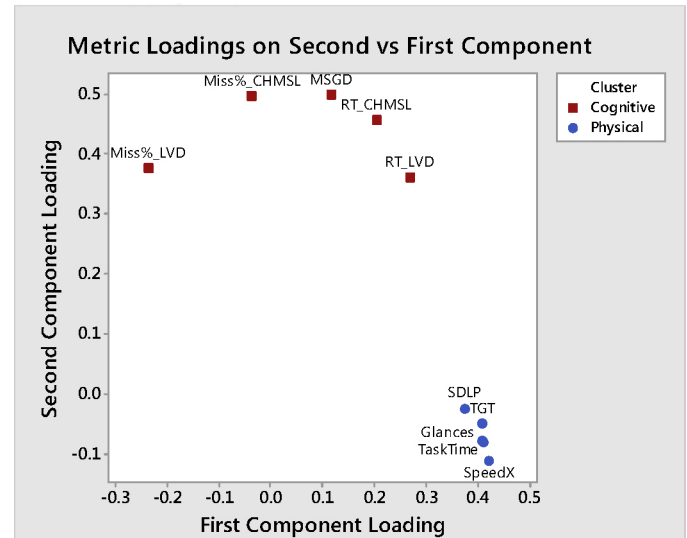


Figure 8. Metric loadings on the two major components (dimensions), and their two major clusters as identified in Figure 9. (Redrawn from [50, Figure 4].)

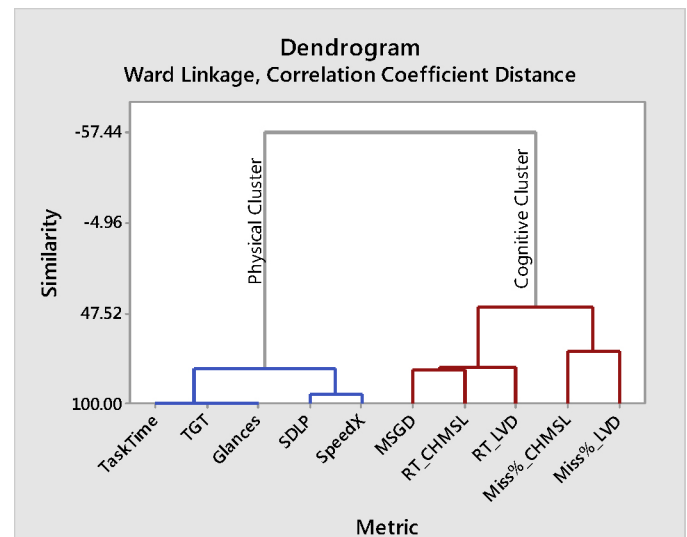


Figure 9. Cluster analysis of the 10 metrics.

These consistencies in the Dimensional Model validation results in Examples 1 and 2, were despite the substantial differences in the testing procedures in Examples 1 and 2, as follows.

1. *Fewer number of tasks*. The CAMP-DWM study in Example 2 had only 13 visual-manual tasks, whereas Example 1 had 78 visual-manual tasks. A larger number of tasks tested in a given venue reduces error in estimating the correlation coefficients and consequently increases the accuracy of the principal component loadings.

2. *Fewer glance measurements.* The CAMP-DWM study measured the glance metrics on only 61% of participants in the CAMPDWM study, due to budget constraints, which increased the uncertainty of the glance metrics (TGT, Glances, and MSGL) in Example 2 relative to the non-glance metrics. Example 1 had all eye glances analyzed.
3. *Surprise event vs. repeated events.* The four event response metrics in the CAMP-DWM study in Example 2 (Miss%_CHMSL, Miss_%LVD, RT_CHMSL, and RT_LVD) were to a single “surprise” event per task trial, either a CHMSL event or lead vehicle deceleration event. Example 1 had many repeated detection-response events per task trial. Hence, Example 2 had larger uncertainty in its detection and response values than the detection-response metrics in Example 1 because of the fewer trials. Also, there is the question whether the data from the repeated events in Example 1 can serve as a surrogate for the single surprise events (many doubt this based on their subjective intuition). This question can be answered empirically by determining if the pattern of responses across tasks to the two types of event is the same or not.
4. *Events in different locations.* Example 2 had vehicle events (CHMSL activation or lead vehicle deceleration) occur in the forward direction only. (The CAMP-DWM study also used side events in the detection of a turn signal in the rear vehicle in the platoon, but the data were noted by the CAMP-DWM investigators as being noisy and so were not analyzed here.) Example 1 used surrogate events randomly presented in the forward direction (DRT light on hood) or side direction (DRT light on top of left outside mirror).
5. *Lead vehicle vs. no lead vehicle.* Example 2 required the participant driver to follow and keep a constant distance and speed to a lead vehicle. Example 1 had only a single vehicle on a closed highway, with instructions to maintain lane position, and control speed at 40 mph (plus or minus 5 mph).

Despite these four substantial methodological differences, the two studies gave comparable results, showing the robustness of the Dimensional Model to major protocol differences. Example 2 may still be considered by some as having more “face validity” for representing an event that has safety relevance in “naturalistic” driving. Nonetheless, the results clearly show that relatively surprising events and repeated events have the same underlying event detection response dimension (which is plausible, since the same orienting attention network in the brain is responsible for determining the detection and response to events, whether expected or not). Therefore, the two major components (dimensions) in both studies were sufficiently similar in the clustering of their metrics that the dimensions are herein labelled as “physical demand” and “cognitive demand,” in both studies, for reasons further explained in Discussion 2 below.

Figure 10 plots the task scores on the two dimensions shown in Figure 8.

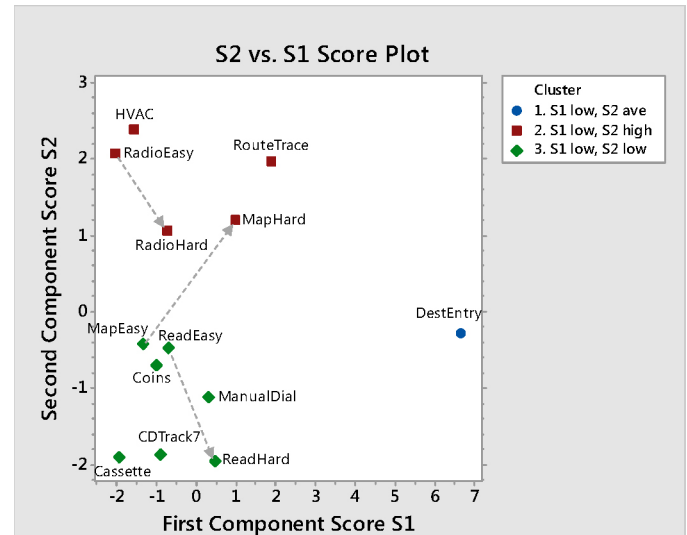


Figure 10. Task scores on the two components. (Redrawn from [50, Figure 6].) The legend indicates the three major clusters as identified by their classifications in Figure 11.

Note in Figure 10 that three tasks (Radio, Read, and MAP), had two “difficulty” levels each, labels which had been created by the CAMPDWM investigators. The methods used by the CAMP-DWM investigators to rate these tasks as “easy” and “hard” were been derived from an analysis that is associated with what we have here termed physical demand (or Dimension 1). Indeed, the three tasks with two difficulty levels each had task scores which were higher on Dimension 1 for the “Hard” type vs. the “Easy” type (i.e., all 3 arrows point to the right, in the direction of higher S1 scores in Figure 15 for Radio, Map, and Read). However, because of the orthogonality between Dimensions 1 and 2, a “Hard” task can be either higher or lower in its Dimension 2 score, regardless of its Dimension 1 score. For example, RadioEasy and ReadEasy each had a higher S2 score than S1 score (the arrow from RadioEasy to RadioHard points downwards in Figure 10). Yet MapHard had a higher S2 score than MapEasy (the arrow from MapEasy to MapHard points upwards in Figure 10). These results shows the independence of the two dimensions. That is, whether a task was deemed “Hard” or “Easy” by the CAMP-DWM investigators appears to have been based on the single dimension that has classically been called “driver workload,” which corresponds to dimension 1 in the Dimensional Model of the data. However, a task that is relatively easy on Dimension 1, can have either relatively higher cognitive demand (e.g. MapEasy vs. MapHard) or relatively lower cognitive demand (i.e., RadioEasy vs. RadioHard or ReadEasy vs. ReadHard), because the two dimensions are orthogonal according to the Dimensional Model.

A cluster analysis of the 13 visual-manual tasks in the original mean data in Appendix Table B1 was done. This analysis is independent of the PCA that gave rise to the task scores in Figure 10. Figure 11 shows the three major task groups that were identified by this objective clustering method.

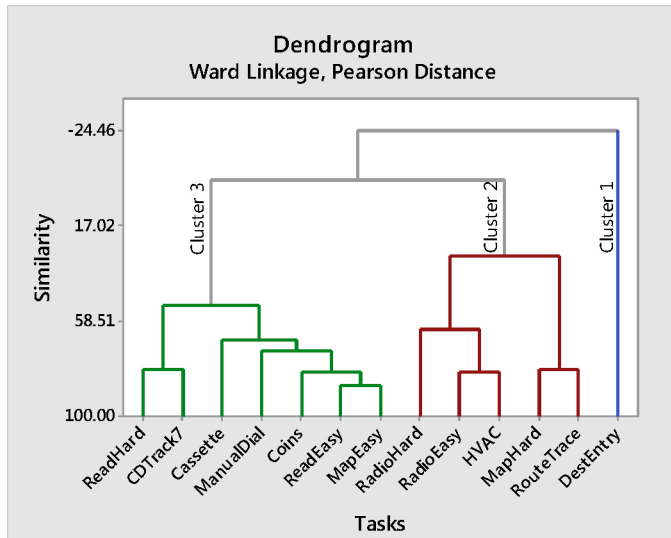


Figure 11. Cluster analysis of the 13 visual-manual tasks, based on all 10 metrics.

Figure 11 shows that DestEntry is by itself in Cluster 1 (blue rightmost line in Figure 11). Figure 10 shows that DestEntry had the highest S1 score of the 13 tasks tested (blue circle to right of Figure 10), but an average S2 score.

Figure 11 identifies the Cluster 2 tasks by the red center lines. Figure 10 shows that the corresponding PCA scores (red squares in upper left quadrant) were relatively high on S2, and relatively low on S1, compared to the other tasks.

Figure 11 identifies the Cluster 3 tasks by the green leftmost lines. Figure 10 shows that the corresponding PCA scores (green diamonds in lower left quadrant) were relatively low on both S2 and S1, compared to the other tasks.

In sum, an objective grouping of the tasks using cluster analysis on the original data (Figure 11), closely corresponds to the subjective visual impression of the task groups in the PCA score space in the Dimensional Model (Figure 10). In other words, the cluster analysis in Figure 11 confirms the visual groupings of the tasks in Figure 10 as scored by the Dimensional Model.

Simplification to Two Original Metrics

As in Example 1, the results of the PCA indicate all the metrics that are highly correlated (i.e., redundant) with each other within each cluster. Hence, there is an opportunity to simplify the set of metrics even further. Theoretically, only two metrics from the original dataset should be enough to characterize each task on the two underlying driver demand dimensions, with little loss of information because of the high redundancy. In particular, TaskTime again had a high correlation with S1 (the task scores on the first component) ($r = 0.945$, $df = 11$, $p = 1.1E-06$), as it did in Example 1. RT_CHMSL had a statistically significant correlation with S2 (the task scores on the second component) ($r = 0.716$, $df = 11$, $p = 0.005$). (See Appendix H, Figures H.1 and H.2 for a visual confirmation of these findings using regression plots.) Given the statistically significant correlations between S1 and TaskTime, and S2 and RT_CHMSL, the Dimensional

Model predicts that the RT_CHMSL vs. TaskTime plot with only two original metrics (Figure 12) should place the tasks in about the same locations and clusters as they were in the S2 vs. S1 score plot (Figure 10). That is, the Dimensional Model predicts that only two original metrics (e.g., as shown in Figure 12) should be able to segment the tasks as well as the task scores based on 15 metrics could in the original dataset, after PCA (as shown in Figure 10).

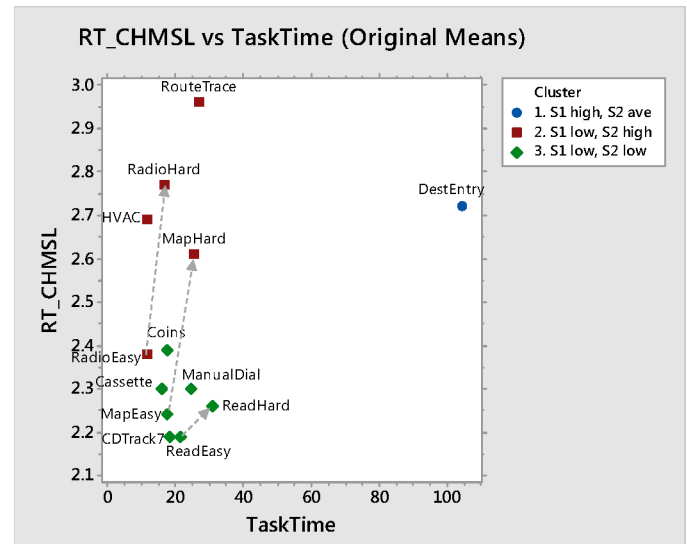


Figure 12. RT to CHMSL event vs. TaskTime (original means across participants). The legend indicates the three major clusters as identified by their classifications in Figure 11, based on all 10 metrics.

Figure 12 plots RT (substituting for S2) vs. TaskTime (substituting for S1). Figure 12 shows that the data points were mostly in similar locations as in Figure 10 with the original S1. (The legend in Figure 12 identifies the task clusters as identified by the analysis in Figure 11, which was based on all 10 metrics.) By visual inspection, the tasks in Figure 12 (based on only two original metrics) generally fall into the same clusters as they did in Figure 10 (based on all 10 original metrics). For example, MapEasy fell into Cluster 3 (green diamonds), and MapHard fell into Cluster 2 (red squares) in both Figure 12 (based on 2 metrics) and Figure 10 (based on 10 metrics).

Of the 13 tasks, the only one misclassified after simplification to 2 original metrics was RadioEasy. It was in Cluster 2 (S1 low, S2 high) in Figure 11, but fell near the tasks in Cluster 3 (S1 low, S2 low) in Figure 12 (note red square oddball RadioEasy task mixed in with the green diamond tasks in the lower left quadrant in Figure 12). RadioEasy was relatively low in its RT_CHMSL metric compared to the other metrics that loaded onto Dimension 2. The S2 metric combines the effect of all the Dimension 2 metrics (LVD and CHMSL event detection, and MSGD) into a single score, and so represents the “gold standard” task score locations along the cognitive demand dimension. The RT_CHMSL metric by itself is generally consistent with this “gold standard” S2 dimensional score, except for the RadioEasy task. An explanation and discussion of this cluster misclassification is given in Appendix I, which cites references [53,54,55].