

The Dimensional Model of Driver Demand: Extension to Auditory-Vocal and Mixed-Mode Tasks

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

ABSTRACT

The Dimensional Model of Driver Demand is extended to include Auditory-Vocal (i.e., pure “voice” tasks), and *Mixed-Mode* tasks (i.e., a combination of Auditory-Vocal mode with visual-only, or with Visual-Manual modes). The extended model was validated with data from 24 participants using the 2014 Toyota Corolla infotainment system in a video-based surrogate driving venue. Twenty-two driver performance metrics were collected, including total eyes-off-road time (TEORT), mean single glance duration (MSGD), and proportion of long single glances (LGP). Other key metrics included response time (RT) and miss rate to a Tactile Detection Response Task (TDRT). The 22 metrics were simplified using Principal Component Analysis to two dimensions. The major dimension, explaining 60% of total variance, we interpret as the attentional effects of *cognitive demand*. The minor dimension, explaining 20% of total variance, we interpret as *physical demand*. The 22 metrics were well segmented into three groups, using cluster analysis, which independently agreed with the metric groups that were visually evident in the 2-D loadings plot. One cluster was associated with cognitive demand, and the other with physical demand. An additional cluster was discovered which loaded in an opposing fashion to cognitive demand along the cognitive demand dimension, and was called *cognitive facilitation*. The Dimensional Model, by eliminating the correlation between the task scores on the dimensions, also minimizes the biasing effects arising from interactions between metrics. The extended Dimensional Model allows for a simplified understanding of the effects of all modes of secondary tasks on driver performance.

CITATION: Young, R., Hsieh, L., and Seaman, S., "The Dimensional Model of Driver Demand: Extension to Auditory-Vocal and Mixed-Mode Tasks," *SAE Int. J. Trans. Safety* 4(1):2016, doi:10.4271/2016-01-1427.

INTRODUCTION

This study extends the scope of the Dimensional Model to secondary tasks commonly performed while driving, including Visual-Manual, Auditory-Vocal, and Mixed-Mode driver-vehicle interfaces. The current study thus extends the Dimensional Model presented in the companion paper [1], which was limited to Visual-Manual tasks in previously published datasets. The extended Dimensional Model, introduced for the first time in the current paper, describes the effect of all modes of secondary tasks on primary driving subtasks in a single Two-Dimensional Model. The dimensions were solved for using Principal Component Analysis, which simplified 22 driver performance metrics to 2 in an analysis of a new dataset.

The driver metrics included in the Model included event detection and response metrics and their surrogates. These loaded into the major dimension, which we interpreted as representing the attentional effects of *cognitive demand*. We also included surrogate metrics for lateral and longitudinal vehicle variability, including glance metrics, which formed the minor dimension in this study, which we interpreted as *physical demand*. The goal of the study was to determine if all of the metrics for every task in the study could be reduced to two scores for every task, with little loss in explanatory power, as it was possible to do for Visual-Manual tasks [1].

An independent cluster analysis of the same data confirmed the assignment of the 22 metrics to the 2 dimensions. The cluster analysis identified three clusters, of which two formed opposite poles on the cognitive demand dimension. Thus, although there were three clusters of metrics, there were only two underlying dimensions.

It is not straightforward to extend the Visual-Manual Dimension Model in [1] to Auditory-Vocal and Mixed-Mode tasks, because the metrics for such tasks often behave in distinctly different ways than they do for Visual-Manual tasks. For example, the duration of a task (herein, TaskTime) is not as correlated with vehicle metrics such as lane and speed variability for Auditory-Vocal tasks, as it is for Visual-Manual tasks [2]. There are also cross-modality effects that indicate there can be interference between the visual and auditory modes [3]. For example, many studies have shown that Auditory-Vocal tasks can produce slight delays in responses to a visual event; e.g., [4, Figure E.5]. Such cross-modality interference is assumed by some to support a 1-D central attentional resource “bottleneck” model [1]. Indeed, visual, auditory, and tactile stimuli produce a similar pattern of interference with response time in dual-task situations [5], suggesting a single common central attentional resource that is common between modalities. However, only a few studies in the driving safety field explicitly recognize that there is more than one central attentional resource (e.g. [6]). Indeed, modern cognitive

neuroscience has identified three central attentional brain networks (*Orienting*, *Executive*, and *Alerting* attention; see [1] for summary), all of which must be considered for a complete understanding of secondary tasks on driver performance. An Auditory-Vocal task, such as cell phone conversation, causes a slight delay in the processing of visual information, due to its effects on the orienting attention system [7]. However, doing a “moderately” demanding task (which included cell phone conversation in the classifications used in [8]), reduced the relative crash/near-crash risk estimate for drowsiness in naturalistic driving (compared to a baseline matched in time of day, closeness to junction, and other variables) from 55 to 1 to 24 to 1 [9, Table 5]. Although the relative risk of drowsiness still remains, in aggregate, quite high at 24 to 1, the halving of the relative risk by conversation is likely due to the activating effect of the conversation on alerting attention, which by definition countervails drowsiness.

In the current paper, the data from a recent study [10] is examined in an attempt to determine to what degree the Dimensional Model of Driver Demand can be coherently extended from Visual-Manual tasks to pure Auditory-Vocal and Mixed-Mode tasks. The data are represented in a common framework with the same analysis method as used in the companion paper concerning Visual-Manual tasks [1], with similar metric names where they overlap between studies.

The off-road glance metrics in the current dataset are separately analyzed on a single metric basis (that is, metric by metric) in a second companion paper [10]. The current study analyzed these off-road glance metrics in a multivariate analysis with many other driver behavior metrics to adjust for interaction effects. These interaction effects arise from strong correlations between a third metric (say long glance proportion) and two other metrics of interest (say mean single glance duration and total eyes-off-road time). These interaction effects can bias the results in subtle ways when metrics are singly analyzed rather than using a multivariate approach. One advantage of the Dimensional Model is that it is a multivariate method that scores the tasks on orthogonal axes, and therefore it adjusts for biasing effects that may arise from such interactions. The interactions between off-road glance metrics are analyzed in [Appendix C](#), using multivariate and univariate methods to illustrate the differences.

METHOD

Dataset

The dataset analyzed for this study was derived from a video-based “surrogate driving” test using a 2014 Toyota Corolla vehicle in which 24 test participants performed 15 different secondary tasks (see [10] for details of the test protocols). Unlike the Analysis of Variance statistical method, the PCA statistical method requires only the mean across participants for each metric for each task, not the data at the individual participant level. The complete dataset analyzed in this study is given in [Appendix Table A1](#).

Tasks

All 15 secondary tasks in the dataset in [Appendix Table A1](#) were analyzed in the main analysis reported here. See [10] for further description of individual tasks and task steps. Two of the tasks (0-Back and 1-Back) were pure Auditory-Vocal tasks with no manual steps, so they met the SAE J2988 [11] definition of Hands-Free because they did not require more than 1 button press. Five of the tasks were Mixed-Mode tasks, which were Auditory-Vocal interface accompanied by a touch-screen visual display that updates to prompt for further voice or manual inputs and/or to provide confirmation of voice inputs already provided to the voice recognition system by the driver. Four of those five Mixed-Mode tasks were Hands-Free as they required only a single manual button press (of a voice button in this case), and therefore they met the Hands-Free definition in SAE J2988 [11]. The Mixed-Mode Navigation Destination Entry task (NavDest) required a button press to activate voice recognition, and then a touch screen press at the end, so it did not meet the SAE J2988 definition of Hands-Free because it had more than 1 button press. There were also 7 Visual-Manual tasks with no Auditory-Vocal interface modes. Five of the Visual-Manual tasks had a matching counterpart Mixed-Mode task, so it can be determined to what extent using a voice response, instead of a manual response, affected the driver performance metrics. [Appendix C](#) gives an analysis of just these paired tasks.

As mentioned above, five of the tasks in this study had an Auditory-Vocal interface accompanied by a visual display. We are careful not to refer such tasks as “voice” tasks as it would be misleading. The term “voice” task we reserve for a pure Auditory-Vocal task with no visual display, and at most one button press to activate the voice system, thereby meeting the Hands-Free definition in SAE J2988. The “n-Back” tasks in this study meet this definition, but the other tasks do not. The term “Mixed-Mode” (or its synonym “Multi-Modal”) should be used to refer to tasks which have an Auditory-Vocal interface accompanied by a visual display, or accompanied by both a visual display and a manual interface such as button presses, knob turns, or presses on a touch screen. The term “Visual-Manual” should be reserved for tasks with a visual display and with manual input to the system, and no auditory or vocal interface mode.

Task Trials

All metrics were averaged over three task trials while driving, after training to criterion (see [10] for further description of training protocols). The exception to the three trials was for assessment of adherence to the NHTSA Guidelines [12] Visual-Manual glance criteria in a second companion paper [10], for which only the first trial was used, in accordance with the NHTSA Guidelines [12] protocol. In the current paper, only the averages across all 3 trials were used, in accordance with the ISO DRT protocol [4] and Alliance Guidelines protocol [13], to improve the accuracy of the estimates.

Metrics

A number of driver performance metrics or their surrogates were collected, with some key metrics of interest listed below:

- *Task metrics*: task completion time (TaskTime); number of errors made during task performance (Errors).
- *Glance metrics* away from the forward view: Total Eyes-off-Road Time (TEORT_NR), where NR is an abbreviation for “Not Road”; number of glances (Glances_NR); mean single glance duration (MSGD_NR).
- *Glance metrics* to displays or controls of device under test: total glance time (TGT_TR), where “TR” is an abbreviation for “Task-Related”; number of glances to the in-vehicle system (Glances_TR); mean single glance time (MSGD_TR).
- *Event detection and response metrics*: percent of tactile detection response task (TDRT [4]) events missed (Miss%_tdrt); response time to tactile events (RT_tdr).

The abbreviations and definitions for the complete set of 22 metrics are given in [Appendix A, Table A2](#). The main body of this paper analyzes the metrics for glances off the road as well as glances to the in-vehicle displays and controls of the device under assessment.

[Appendix C](#) analyzes just the glance metrics off the road, to investigate the effects of interaction between these glance metrics.

[Appendix C](#) gives an analysis of just the off-road glance metrics.

Statistical Methods

Principal Component Analysis

Primary and secondary task performances were measured and the data were analyzed using Principal Component Analysis (PCA). Citations for the PCA method are given in the companion paper [1]. 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, as in the companion paper [1], all of the driver demand metrics for all tasks were reducible to only two component scores, with little loss in explanatory power for the entire dataset of all metrics.

The PCA in the statistical package in Stata 13.1 [14] was used for the analysis, with results checked and plotted in Minitab® 17.2.1 [15]. Minitab changed the sign of the y-axis, but otherwise the results were identical to the Stata results. The sign of an axis is purely arbitrary in PCA; sometimes an axis will point in a positive direction to indicate an increased effect, and sometimes in a negative direction. Regardless of the direction of an axis, all points in the 2-D space will have the same positions with respect to one another regardless of the sign of the axis. To make it easier for the reader, in the results given here and in the companion paper [1], all axes (whether physical or cognitive) in the positive direction indicate increased demand relative to the average of the tasks in the dataset, while all axes in the negative direction indicate decreased demand relative to the average of all tasks in the dataset.

The PCA used in this study operated on the correlation matrix. In a correlation, the scales of the metrics are converted to a mean of zero by subtracting the mean, and dividing by the standard deviation so the length of the axis is one (i.e., a Z-score). The length of the metric loadings on a given component are then of length one (that is, the square root of the sum of the squared loadings of each metric on each component is one). The metric loadings, although of length one, are not themselves necessarily orthogonal (that is, the correlation between them is not necessarily zero). However, the scores of each secondary task on each component are guaranteed by the PCA method to be orthogonal (i.e., have correlation zero) to the scores of the secondary tasks on every other component. In the most commonly used PCA method, as employed here, the ratio of the variance of each set of scores for each component is set equal by the PCA method to the ratio of the variances explained by each component. For example, assume 60% of the variance in the whole dataset is explained by the first principal component, and 20% by the second component, for a ratio of 3 to 1. Then the variance of the set of task scores for the first set of task scores will be in a ratio of 3 to 1 relative to the variance of the set of task scores for the second component. In this way, the distances in the score plot in the *x* and *y* directions are equated for the relative variance explained by each score vector.

To say that the original dataset with 22 metrics can be reproduced to a high degree of accuracy with only the 2 score metrics is not a trivial statement. Literally, the 2 scores can be used to predict the original scores by back-transforming the scores to the original metric scales if desired, with the removal of all the statistical “noise” represented by the higher-order components. When the transformed metrics are plotted alongside the original metrics, the total sum of the squared distances between them will explain 80% of the variance in the original data set, and provide potentially improved estimates of the original task values (by removing statistical noise). The new estimates of the “true” values of the original metrics are based on the information contained in all 22 metrics. Therefore, the transformed scores should theoretically be closer to the “ground truth” for every original metric, because it effectively has the statistical noise removed, compared to the original metric which includes statistical noise. Basically, the PCA method allows the extraction of the underlying structural components in the dataset. Transformation of the PCA scores back to the scales of the original dataset to predict the original 22 metrics was not done in this study, but could be done in future research by anyone, solely based on the data given in [Appendix A, Table A1](#). In addition, new tasks that were not in the original dataset can be scored using the same set of principal components for validity checking purposes, if all 22 metrics were also available. However, if the metrics set could be reduced to only 2 or 3 original metrics, collecting data and scoring new tasks with regard to previous components would be easier. Again, these additional analyses are left for future research.

There are variations of the PCA method which analyze the covariance matrix rather than correlation matrix, or rotate the scores after extraction to “simplify” the alignment of the scores with the axes, but such alternative analyses were not systematically explored in this study. Any rotation of the score plots destroys the orthogonality in the score sets, and there is no unique mathematical solution as there is with using the PCA scores themselves, as was done here.

Transforming the length of the score axes to unity rather than according to the variance explained would create a uniform score space, such that the confidence interval around each score in the final 2-D plot would be a circle rather than an ellipse, but the issue of confidence intervals around individual points in the score plot, and alternative scales for the final score axes, are left for future research.

The PCA method typically requires the number of tasks to exceed the number of metrics, or a “singular matrix” error often results in a PCA statistical program. In the current study, the number of metrics (22) did exceed the number of tasks (15), but the PCA was completed properly without error, so all 22 metrics were retained for the main analysis.

This study was designed and the data collection was completed before the publication of the SAE Recommended Practice J2944 [16] providing common definitions of driving performance metrics and statistics. However, the definitions of driver performance metrics used in this study are in given [Appendix Table A2](#). We note that in the companion paper [1], the same two dimensions were found after analysis of the datasets in four different studies, despite wide differences in the metric definitions and venues (simulator, closed-road, or open-road). The PCA method is robust to such differences in venues and definitions, because it operates on relative metrics (taken with respect to the average of each metric across tasks), and standardized metrics (the scale for each metric is standardized to length 1 by dividing by the standard deviation across tasks). By avoiding absolute metrics, any slight differences in definitions are ameliorated. In addition, SAE J2396 [17] provides definitions for the glance-related metrics used in the four studies, and the glance definitions as used in this study were consistent with that Recommended Practice.

Partial Correlation Analysis

The methods of partial correlation analysis and methods for reducing the effects of interactions between the metrics in [Appendix C](#) are based on equations and causal diagrams in [18, Sections 3.3 and 3.4]. The Stata statistical program [14] partial correlation command “pcorr” was used to check the partial correlation calculations. These methods of adjustment in statistics are conceptually similar to those used in epidemiology [19,20], and the interested reader can consult most textbooks in statistics or epidemiology to learn more about them. The reader unfamiliar with these methods can still however obtain an understanding of the results of this paper without having detailed knowledge of these statistical methods.

RESULTS

The correlation table between the metrics is given in [Appendix Table B1](#). A high degree of correlation was seen between the metrics within particular subsets, and little correlation between metrics in different subsets. This result was consistent with previous studies [21, [Table 2](#)]. PCA takes advantage of these correlations to eliminate redundancy between the metrics. Hence, in this study, all 22 metrics could be reduced to only two using PCA. Dimension 1 accounted for 60% of the total variance in the dataset, while Dimension 2 accounted for 20% of the total variance (that had not been explained by Dimension 1). The “scree” plot showing the diminishing percentage of the total variance explained by the successive components out to 22 is shown in [Figure 1](#).

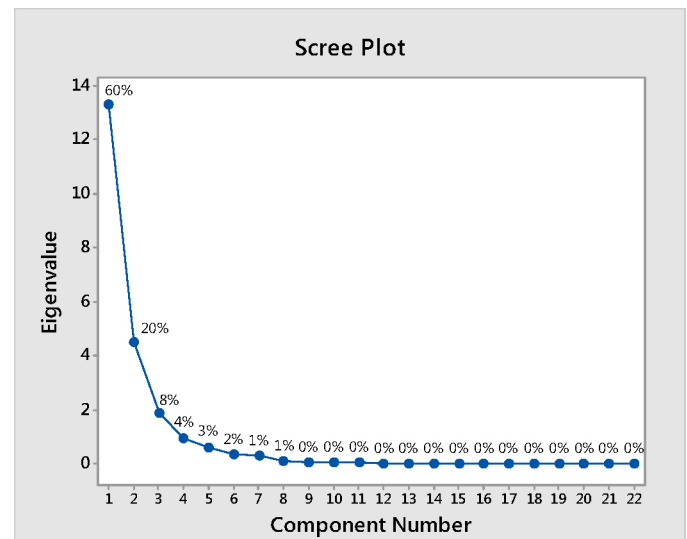


Figure 1. Scree plot from PCA. The percentages indicate the percentage of the total variance explained by each component.

[Figure 1](#) shows that the components beyond the first two contributed little to the total variance. They are essentially statistical noise, and so the higher-order components can be safely discarded, greatly simplifying the analysis. That is, [Figure 1](#) shows that only 2 dimensions can reproduce 80% of the variance on all 22 original driver performance metrics. Hence, the *driver performance* results for all 15 tasks for all 22 metrics can be reproduced to a high degree of accuracy with only two scores for each task on the respective dimension. (Note: We here use the term “*driver performance*” to specify the wider set of event detection and response metrics combined with lane and speed maintenance metrics. The narrower set of only lane and speed maintenance metrics are commonly referred to as “*driving performance*” metrics [13].) Following PCA, the loadings of the metrics on the two main components are plotted in [Figure 2](#).

Using the “Cluster Variables” option in the Minitab program [14] on the original mean data for all 22 metrics (the same data in [Appendix Table A1](#) that was entered into the PCA) objectively determined three major metric clusters ([Figure 3](#)).

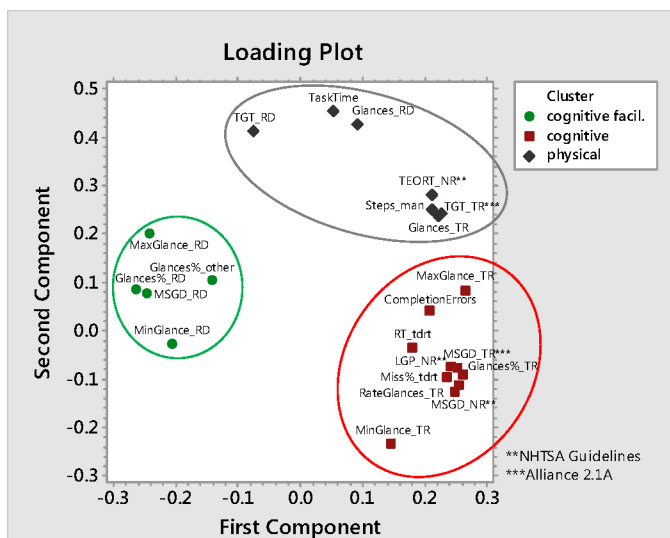


Figure 2. Metric loadings on the two major components (dimensions). The symbols are coded with ellipse according to the three metric clusters found in the cluster analysis in Figure 3.

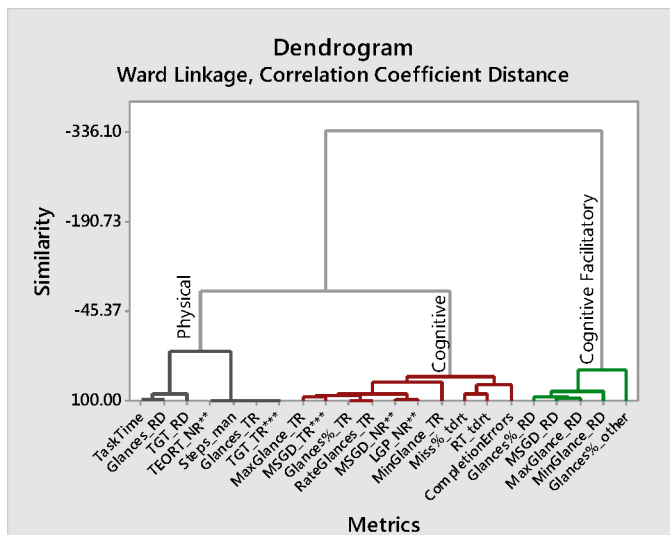


Figure 3. An independent cluster analysis of the same mean data identifies three major metric clusters.

The objective and independent cluster analysis depicted in Figure 3 confirms the visual impression from Figure 2 that there are three major metric clusters, which we here term *physical demand*, *cognitive demand*, and *cognitive facilitation*. (Note that these three clusters of metrics are embedded in the 2-D loading space shown in Figure 2, so there are still only two dimensions in the data despite the fact that there are three clusters of metrics.)

Physical Demand Metrics

The gray lines in the left half of Figure 3 indicate what we here term *physical demand* metrics, which are the metrics in the upper portion of Figure 2 (gray diamonds, surround by gray ellipse). These metrics load positively on Dimension 2 (higher physical demand). (Note that the axes have swapped places from those in study [1] using Visual-Manual tasks only. Physical demand is now on the y-axis, and the attentional effects of cognitive demand on the x-axis. This result

indicates that the inclusion of Mixed-Mode and Auditory-Vocal tasks has made the cognitive demand dimension the major axis, and the physical demand dimension the minor axis.)

The physical demand metrics cluster is composed of surrogates for driving performance (lane and speed variability), as described in the companion paper analyzing Visual-Manual tasks only [1]. Lane and speed maintenance metrics (“Driving Performance”), when measured during actual rather than simulated driving, are regarded as “ground truth” metrics [22]. The surrogate metrics for these ground truth driving performance metrics for physical demand are shown in the “physical demand” metrics cluster in Figures 2 and 3, and include TaskTime, glances to the road (Glances_RD), total glance time to the road (TGT_RD), Total Eyes-Off-Road Time (TEORT_NR**, a NHTSA Guidelines metric), number of manual steps (Steps_man), number of task-related glances (Glances_TR) and total task-related glance time (TGT_TR***, an Alliance Principle 2.1A metric). These metrics are all tightly correlated with each other (see Appendix B1 for correlation matrix), and therefore are essentially redundant, with such high correlations with each other that they are almost collinear. In other words, the metrics that load into relatively high positions on Dimension 2 are nearly-equivalent “surrogate” metrics for the “ground-truth” driver performance metrics concerning lane and speed maintenance, because of the high correlations shown in [1], and they are therefore effective surrogate metrics for physical demand. Hence, they are suitable for representing Dimension 2 in this study, which is the physical demand dimension; it was Dimension 1 in the companion paper [1]. In other words, the metrics that load into relatively high levels of Dimension 2 are equally valid “surrogate” metrics for the “ground-truth” driver performance metrics concerning lane and speed maintenance, because they are highly correlated with them as shown in [1], and are therefore equally suitable for representing the physical demand dimension, which, in this study, is Dimension 2. That is, the PCA method provides the optimal combination of the information in all these physical demand metrics to form Dimension 2. For shorthand, we therefore here label Dimension 2 as a *physical demand* dimension. Note that this dimension also contains the subjective metric of a driver’s perceived workload for Visual-Manual secondary tasks, as described in [1, Example 1]. However, the physical demand dimension mainly concerns metrics that can be objectively measured from direct physical observation (lane exceedance, lane and speed variability, number of glances to a device, task time, etc.).

There is tight coupling of the physical demand metrics with lateral and longitudinal variability driving performance as shown in the Dimensional Model results in [1] for Visual-Manual tasks. This result suggests that it may also be reasonable to regard the physical demand dimension as a “driving performance” dimension rather than a “physical demand” dimension. The non-vehicle metrics (such as TEORT, TaskTime, etc.) that also load onto this dimension might then be regarded as “surrogate” metrics for the “ground truth” driving performance measures of lane crossings and speed variability in the Alliance Guidelines Principle 2.1B.

These physical demand metrics are as tightly correlated with each other, more so than they are correlated with metrics in other cluster. This result is what is effectively shown by the cluster analysis in [Figure 3](#). As an independent verification, consider the matrix of physical demand metric correlations given in [Appendix B, Table B1](#), upper left corner, with the metric names colored in grey, along with the subset of correlations. These correlations have a mean r value of 0.73 within the physical demand cluster. The mean r value of the metrics in the physical demand cluster (gray metric names) with those in the cognitive demand cluster (red metric names) is only 0.30, and with the cognitive facilitation cluster (green metric names) is only -0.26.

Cognitive Demand Metrics

The cluster analysis in [Figure 3](#) also identifies a second major cluster of metrics (center red lines). These metrics are those in the lower right quadrant of [Figure 2](#) (red squares surrounded by red ellipse). This cluster also has tightly-correlated metrics that are largely redundant with each other as shown in [Appendix B1](#) (metric names colored in red), but they are only weakly uncorrelated with metrics loading onto the physical demand dimension (metric names and correlations colored in red). Metrics that are only weakly correlated with one another form different clusters in a cluster analysis ([Figure 3](#)), or load into different dimensions in a PCA ([Figure 2](#)).

As an independent verification, consider the matrix of cognitive demand metric correlations given in [Appendix B, Table B1](#), middle columns and rows, with the metric names colored in red, and the subset of cognitive demand metric correlations colored in red (center of table). These correlations have a mean r value of 0.73 within the cognitive demand cluster red submatrix (center of Table). The mean r value of the metrics in the cognitive demand cluster (red metric names) with those in the physical demand cluster (gray metric names) is only 0.30. The mean r value of the metrics in the cognitive demand cluster (red metric names) with those in the cognitive facilitation cluster (green metric names) is -0.69 (cells colored yellow), indicating that the cognitive facilitation metrics are an “opponent” process to the cognitive demand metrics. The two clusters are at opposite ends of the same cognitive dimension; the cognitive facilitation cluster does not form a third dimension.

The cognitive demand metrics were previously shown for Visual-Manual tasks to load onto the dimension for event detection and response in study [1], for either remote DRT events or real vehicle events such as brake light activation or deceleration of a forward vehicle. This result suggests that it may also be reasonable to regard this dimension as an “event detection and response” dimension rather than a “cognitive demand” dimension, for those who prefer behavioral rather than cognitive science terminology. The metrics besides event detection and response (such as MSGD, LGP, etc.) that also load onto this dimension might then be regarded as “surrogate” metrics for the “ground truth” driver performance metrics of event detection and response to external road events or the detection response task (DRT) [4], as suggested by the CAMP-DWM study findings [39,40] using brake light activations and lead vehicle decelerations.

It is of practical and theoretical interest that mean single glance duration (MSGD) clusters with these cognitive demand metrics, and not with the physical demand metrics. This finding is true for both task-related MSGD (metric 9 in [Appendix Table B1](#), MSGD_TR**, Alliance Principle 2.1A) or off the road MSGD (metric 12 in [Appendix Table B1](#), MSGD_NR**, NHTSA Guidelines).

Likewise, the NHTSA Guidelines LGP metric (metric 13 in [Appendix Table B1](#), LGP_NR**) clusters with the cognitive demand metrics and not the physical demand metrics. Thus, the PCA results ([Figure 2](#)) and the cluster analysis results ([Figure 3](#)) independently suggest that MSGD and LGP are more effective surrogates for the event detection and response (cognitive demand) dimension than they are effective surrogates for metrics indicating physical demand. The loading of MSGD into the cognitive rather than the physical dimension occurs despite the fact that the TGT and TEORT metrics in the physical cluster are formed by the product of MSGD and the number of glances, and thus MSGD adds no information to TGT or TEORT beyond the number of glances. In other words, MSGD and LGP provide independent information about the cognitive dimension of driver demand that is not predicted by TGT, TEORT, or the total number of glances to the device or off the road during a secondary task.

This result may seem strange that multiplying two variables could give rise to a third variable (TEORT or TGT) that is now uncorrelated with one of the original variables (MSGD). However, a multiplication of two variables is a non-linear transformation, so this result is quite possible mathematically and in fact occurs in driving, whether in the surrogate driving venue used in this paper, or as shown in [1], in simulator, closed-road, and open-road venues. These results suggest that any causative effect on lateral and longitudinal vehicle variability) of TEORT or TGT arises solely from the number of glances and not the total duration of those glances, the mean single glance duration, or the long glance proportion (LGP) of those glances.

The converse result is also suggested by the Dimensional Model. The PCA of the data in the current study suggests that MSGD and LGP are independent indicators of event detection and response performance (cognitive demand), and the physical demand metrics have little or no influence on event detection and response. In other orders, event detection and response behavior of a driver is not predictable from number of glances, TGT, or TEORT, or any other physical demand metric. Because the LGP and MSGD metrics in the cognitive cluster are related to the detection and response to events, we believe it is correct to label this dimension a “cognitive” dimension in this study as well as in the companion study [1] of Visual-Manual tasks. However, it is equally valid to label this dimension as an “event detection and response” dimension as previously mentioned, for those who prefer behavioral rather than cognitive science terminology.

The PCA and cluster analysis thus confirm in independent analyses that each metric cluster forms a separate dimension for the set of tasks under the experimental conditions of this study. An objective grouping of the metrics using cluster analysis on the original data

(Figure 3), closely corresponds to the subjective visual impression of the metric groups in the PCA loading space in the Dimensional Model (Figure 2).

Cognitive Facilitation Metrics

The cluster analysis in Figure 3 identified a third group of metrics, not previously well recognized explicitly (green lines to right of Figure 3). These metrics are identified by the green ellipse around the green circles to the left of Figure 2. These metrics are on the negative x-axis, and so form an opponent process to metrics in the cognitive demand cluster on the positive x-axis (red ellipse around the red squares in Figure 2). As such, this cluster of metrics does not represent a third dimension, it represents instead simply an “opposite” or “complement” to the cognitive demand cluster. Metrics in the cognitive facilitation cluster are negatively correlated with the metrics in the cognitive demand cluster. We have termed these metrics “cognitive facilitation,” as they have an opposite effect from the “cognitive demand” metrics.

In other words, the cognitive demand and physical demand metrics were only weakly correlated with one another, and so formed different clusters in the cluster analysis (Figure 3). Or, these two groups load into different dimensions in a PCA (Figure 2). However, metrics that are highly negatively correlated with metrics in another cluster can form a cluster at the opposite pole of one dimension, as did the “cognitive facilitation” metrics (Figure 2, green symbols).

As an independent verification, consider the matrix of cognitive demand metric correlations given in Appendix B, Table B1, with the metric names colored in green, and the subset of cognitive demand metric correlations colored in green (lower right corner of Table B1). These correlations have a mean r value of 0.70 within the cognitive facilitation cluster green submatrix (lower right of Table). The mean r value of the metrics in the cognitive facilitation cluster (green metric names) with those in the physical demand cluster (gray metric names) is only -0.26. The mean r value of the metrics in the cognitive cluster (green metric names) with those in the cognitive demand cluster (red metric names) is -0.69 (cells colored yellow), indicating that the cognitive facilitation metrics are an “opponent” process to the cognitive demand metrics. The two clusters are at opposite ends of the same cognitive dimension; the cognitive facilitation cluster does not form a third dimension.

A relatively high score on the cognitive facilitation metrics indicates the opposite of cognitive demand, or what we here term “cognitive facilitation.” In this study, the cognitive facilitation metrics were all glance metrics that involved glances to the road, or glances to “other” locations than the road or to the displays and controls of the device. Glances to other locations are not a new “dimension,” because if participants are glancing to the device, they cannot simultaneously be glancing to the road or to “other” locations than the device or the road. Likewise, glances to the road are not a new dimension, because if participants are glancing to the road, they cannot simultaneously be glancing to the device or to “other” locations than the device or the road. The cognitive facilitation metrics therefore just represent the opposite or complement of the metrics at the positive end of the cognitive dimension, and do not represent another dimension.

Dimensions Summary

In summary, the overall results for the metrics show that the first two dimensions together captured 80% of the total variance in all 22 metrics assessed. That is, just two metrics from every task can recreate the original data in Appendix A, Table A1 that required all 22 metrics to describe. In a PCA, there are as many potential dimension as metrics (22). As shown in the scree plot (Figure 1), the higher dimensions from 3 to 22 each individually constitute such a small percentage of the remaining 20% of the variance (that is, that which is not explained by the two main dimensions), that the higher dimensions beyond 2 represent nothing more than statistical noise in the data, and so they are not considered further in this study.

An independent cluster analysis of the original mean data in Appendix A, Table A1, confirmed these findings. Cluster analysis successfully segmented the metrics into three groups in the two-dimensional loadings space (Figure 2). We termed the dimensions physical demand, and cognitive demand. The cognitive dimension had one end being “positive” cognitive demand, and the other end “negative” cognitive demand, that we here termed *cognitive facilitation*. The physical dimension has only positive physical demand, because the metrics that load on it cannot physically have negative values (e.g., there cannot be a negative TaskTime, or a negative number of glances, or a negative count of manual task steps).

Task Scores

The tasks can now be scored on these two components by forming the sum of the cross-product of the 22 metrics for each task with the loading coefficients for each component as per standard PCA methods. Figure 4 plots the task scores so calculated.

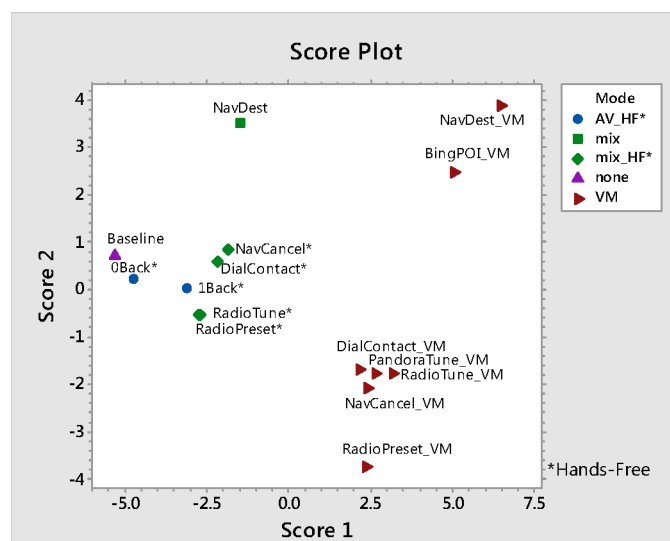


Figure 4. Task scores on the two major components, grouped by the five different interface modes (see legend). AV = Auditory-Vocal; HF = Hands-Free; mix = Mixed-Mode; none = no secondary task; VM = Visual-Manual.

Cognitive Demand Task Scores

Figure 4 shows that all Mixed-Mode tasks (green symbols in center column) scored lower on the first component (cognitive demand) than all Visual-Manual tasks (red arrowhead symbols in the right half of Figure 4). This result confirms previous research showing that the attentional effect of cognitive demand is higher for Visual-Manual tasks than for Mixed-Mode tasks [23]. This point is emphasized by the fact that all the Mixed-Mode interfaces in the current study reduced the attentional effects of cognitive demand as measured by Score 1, even compared to a single press of a radio preset button (RadioPreset_VM at bottom of Figure 4). Despite this reduction in the attentional effects of cognitive demand, the Mixed-Mode tasks still had more attentional effects of cognitive demand than baseline driving with no task (magenta triangle at left), and the pure Auditory-Vocal 0-Back task (leftmost blue circle). The hands-free Mixed-Mode tasks (green diamonds) had about the same cognitive demand as the pure Auditory-Vocal 1-Back task (blue circle near green diamonds).

Physical Demand Task Scores

On the physical demand dimension (y-axis in Figure 4) there was a relatively wide range of task scores within the Mixed-Mode types (green symbols) and Visual-Manual types (red symbols). Figure 4 shows that the hands-free Mixed-Mode types (green diamonds) had higher physical demand (as measured by Score 2) than most Visual-Manual types (red rightward triangles in bottom right of Figure 4). The higher physical demand for Mixed-Mode tasks may seem counter-intuitive given that there was less TEORT for Mixed-Mode tasks compared to Visual-Manual tasks (see second companion paper [10]). This anomaly results from the indirect interaction effects of the high correlation between LGP and MSGD on TEORT (see Appendix C).

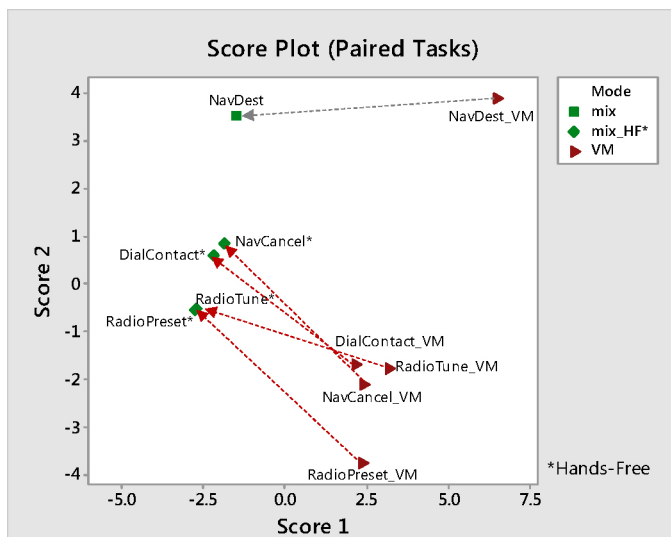


Figure 5. Task scores on the two major components for the five paired tasks.

In other words, the Mixed-Mode tasks consistently showed reduced attentional effects of cognitive demand compared with Visual-Manual tasks, but they also generally showed relative increases in physical

demand according to their PCA scores. This result is illustrated more clearly by considering the *paired tasks*, which have identical task start and goal states, but one member of the pair is performed with a Visual-Manual interface, and the other with a Mixed-Mode interface. Figure 5 replots the data in Figure 4 with just the paired tasks.

In Figure 5, the upward and leftward pointing red dashed arrows indicate a decrease in cognitive demand (Score 1) but an increase in physical demand (Score 2) for the same task with a Mixed-Mode interface vs. a Visual-Manual interface. The arrows in Figure 5 show that four of the five paired tasks had relatively more physical demand and relatively less cognitive demand for the Mixed-Mode condition than they did for the Visual-Manual condition (dashed red arrows pointing in 10 o'clock direction in Figure 5). That is, Mixed-Mode tasks trade off a reduction in the attentional effects of cognitive demand (x-axis), for an increase in physical demand (y-axis), according to this dimensional analysis.

The exception was the NavDest task, with the highest levels of physical demand (top of Figure 5), which the gray dashed arrow shows there was no increase in physical demand for the Mixed-Mode task (green square) compared to the same task executed with Visual-Manual controls (red rightward triangle). That is, the non-hands-free Mixed-Mode NavDest task (green square) scored about the same physical demand as did the Visual-Manual NavDest_VM task, yet had a relative decrease in the attentional effects of cognitive demand.

Task Score Clusters

Figure 6 is a cluster analysis of the task scores in Figure 4.

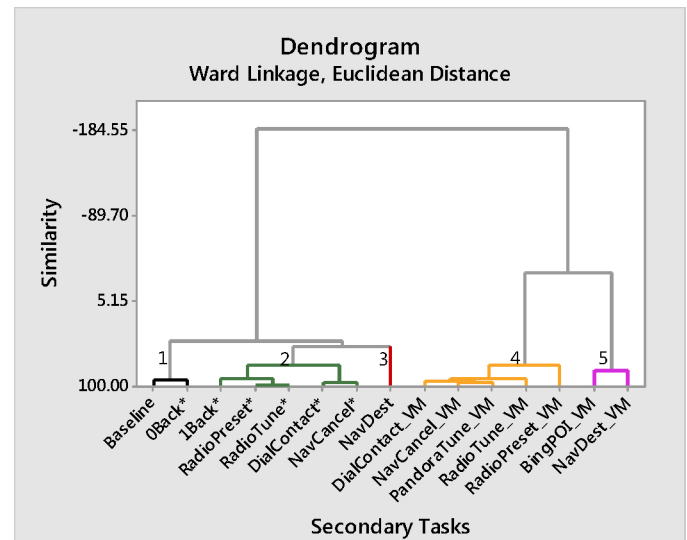


Figure 6. Cluster analysis identifies 5 clusters of the scores in Figure 3.

The five major clusters found in Figure 6 are identified by the elliptical overlays on the points in Figure 4, as shown in Figure 7.

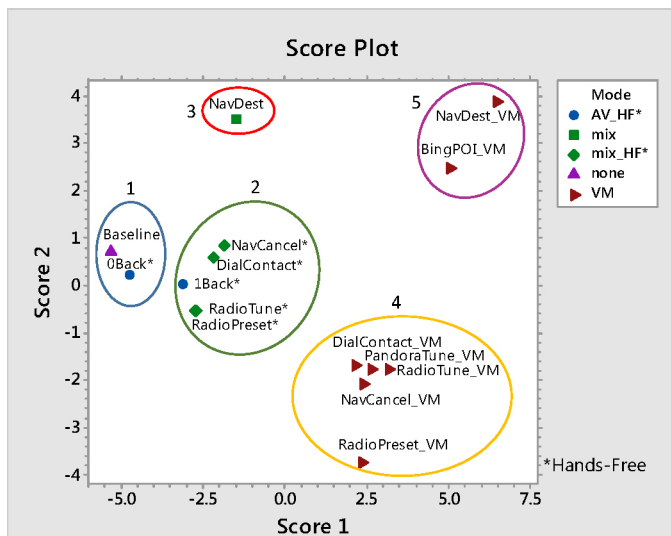


Figure 7. Results of cluster analysis (Figure 6) overlaid on score plot (Figure 3) with ellipses.

Figure 7 confirms the finding shown in Figure 4 that hands-free Mixed-Mode tasks (green ellipse cluster #2) had a higher physical demand (y-axis) than the Visual-Manual tasks in cluster 4 (orange ellipse in lower right of Figure 7), but a lower attentional effect of cognitive demand (x-axis).

Figure 7 also illustrates the 2 groups of outlier clusters found in the cluster analysis in Figure 6. Cluster 5 is composed of Visual-Manual destination entry (NavDest_VM) and searching of a point-of-interest using Bing (BingPOI_VM). The scores of this tasks were exceptionally high in physical demand compared to the other Visual-Manual tasks, and exceptionally high in cognitive demand compared to the Auditory-Vocal and Mixed-Mode tasks. Going to a Mixed-Mode implementation of the navigation destination entry task reduced the cognitive demand (x-axis), but the physical demand stayed just as high as it was for the Visual-Manual mode of the task.

The increased physical demand for Mixed-Mode tasks shown in Figure 7 is a surprising result, given that consideration of single metrics at a time in [10] indicates that there are fewer glances to the screen and smaller TEORT values for Mixed-Mode tasks than Visual-Manual task [10]. The reason for these surprising result has to do with the high correlation between MSGD and LGP, which acts a “suppressor” on TEORT (see Appendix C). When the task data are plotted on orthogonal axes as in Figure 7, the suppressor effects are removed.

Simplification to Two Original Metrics

To the extent that the Dimensional Model has internal validity, then all the metrics within a given metrics cluster are essentially redundant with each other. The Dimensional Model hence 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. There will be some loss of information, because the task scores in the Dimensional

Model combine the information from the all the metrics in a unique and optimal way. However, all the original metrics in the physical demand cluster have a statistically-significant correlation with the task scores (S2) on the second component, and could hypothetically be used as substitutes for the task scores S2. TaskTime had the highest correlation ($r = 0.96$, $df = 13$, $p = 1.5E-08$), with the regression fit shown in Figure 8.

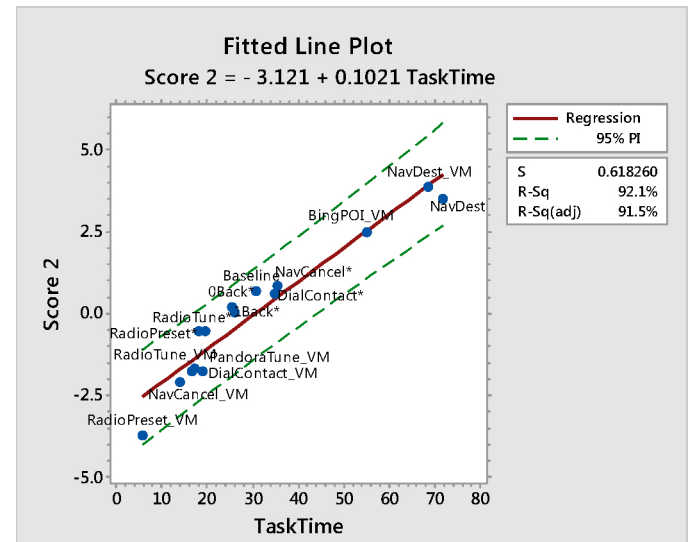


Figure 8. Prediction of Task Score 2 by TaskTime ($r = 0.955$, $df = 76$, $p = 5.5E-42$).

Figure 8 shows that the observed values of the S2 score by TaskTime were all tasks that were well within the 95% prediction interval (dashed lines), confirming the goodness of the fit. That is, TaskTime can be effectively used as a surrogate for the physical demand task score, based on the combined information from all the 22 metrics.

Note that other metrics loading into the physical metrics cluster could have also been used to represent the physical demand task score. For this study, those other metrics would be Glances_RD, TGT_RD, TEORT_NR**, Steps_man, Glances_TR, or TGT_TR***. In addition, the companion study [1] showed that lane crossings, lateral variability, or a longitudinal variability metric, and other physical demand metrics developed for Visual-Manual tasks, could likely have been used, if they had been measured in the current study.

Likewise for Score 1, all the metrics in the cognitive demand cluster had a positive correlation with Score 1, and all the metrics in the cognitive facilitation cluster has a negative correlation with Score 1. MaxGlance_TR (the average duration of the longest task-related glance for every participant) had the highest correlation of all the metrics with the task scores S1 on the first component ($r = 0.97$, $df = 13$, $p = 1.9E-09$). However, the goal is to use non-glance metrics wherever possible, to save the time and expense of eye glance measurements. The highest correlation among the non-glance metrics in Clusters 2 and 3 is MissPct_tdr ($r = 0.86$, $df = 13$, $p = 4.0E-05$), with the regression fit in Figure 9.

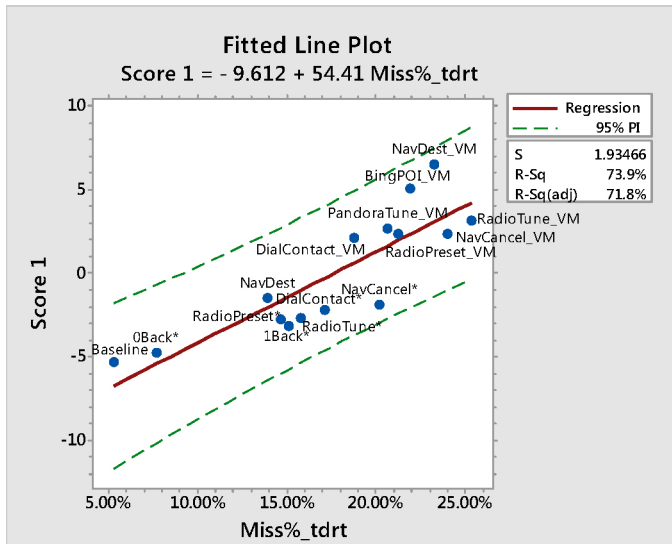


Figure 9. Prediction of Task Score 1 by MissPct_tdr.

The dashed lines show that the 95% prediction interval for Score 2 in Figure 9 is wider than for Score 1 in Figure 8, but the task scores for the attentional effects of cognitive demand in Figure 9 are still all within the 95% prediction interval of Miss%_tdrt.

Note that other metrics loading into the cognitive metrics cluster could have also been used to represent the cognitive demand task score. For this study, these other metrics would be MaxGlance_TR, MSGD_TR***, Glances%_TR, RateGlances_TR, MSGD_NR**, LGP_NR**, MinGlance_TR, Miss%_tdrt, RT_tdr, or CompletionErrors. In addition, any of the cognitive facilitation metrics could have been used with a negative coefficient: MSGD_RD, MaxGlance_RD, MinGlance_RD, or Glances%_other. In addition, the companion study [1] showed that RT and miss rate to the remote DRT, or any other metric loading on the cognitive demand dimension could have been used to represent the attentional effects of cognitive demand, if they had been measured in the current study.

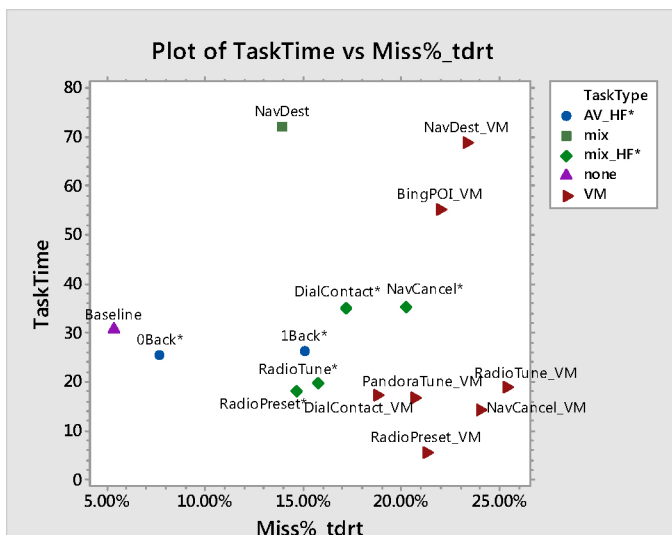


Figure 10. Event response time TaskTime vs. Miss%_tdrt, original means.

Given these encouraging results for the two chosen individual non-glance metrics, the simplification hypothesis was tested by using just TaskTime and MissPct_tdr to span the 2-D score space (Figure 10).

Visual comparison of Figures 4 and 10 shows there are similar relative positions of the task points in the 2-D space.

The original TaskTime and Miss%_tdrt plotted in Figure 10 are nearly virtually uncorrelated or orthogonal (Appendix B, Table B1, $r = -0.03$ $df = 13$, $p = 0.923$). 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 in [1]). Because TaskTime and Miss%_tdrt are already nearly orthogonal, further analysis to find another orthogonal pair of single metrics is not necessary to reasonably represent the physical and cognitive axes for this dataset. For those who wish to use other metrics, any other pairs of metrics could be chosen and examined, as long as one metric is picked from the physical cluster, and one from the cognitive cluster, according to the Dimensional Model. The individual metrics that are most strongly correlated with the task scores, however, should have slightly better prediction results for the “ground truth” task scores based on all the metrics, when the individual metrics are reduced to 2, as in Figure 10.

DISCUSSION

Data was collected from 24 participants during use of the “infotainment” system in a 2014 Toyota Corolla in a video-based surrogate driving venue. There were 15 secondary tasks of varying modes of interaction (various combinations of auditory, vocal, visual, and manual modes), with 6 of them being “paired tasks,” which had the same start and end states, but were performed with different modes of interaction. Data from 22 variables (metrics) were collected, including eye movement data. A Principal Components Analysis found that the 22 metrics could be reduced to 2 principal components (dimensions), with little loss of explanatory power.

The major dimension, explaining 60% of total variance, we interpreted as the attentional effects of *cognitive demand*. The minor dimension, explaining 20% of total variance, we interpret as *physical demand*. The 22 metrics were well segmented into three groups, using cluster analysis, which independently agreed with the metric groups that were visually evident in the 2-D loadings plot. One cluster was associated with cognitive demand, and the other with physical demand. An additional cluster was discovered which loaded in an opposing fashion to cognitive demand along the cognitive demand dimension, and was called *cognitive facilitation*. The Dimensional Model, by eliminating the correlation between the task scores on the dimensions, also minimized the biasing effects arising from interactions between metrics. The extended Dimensional Model allows for a simplified understanding of the effects of all modes of secondary tasks on driver performance.

Simplification to Two Original Metrics

This fact that only two dimensions were need to span the space of all 22 metrics led to a further simplification hypothesis from the Dimensional Model that the two PCA task scores (which combined the information from all 15 metrics) could be reasonably approximated with a reduced set of only 2 of the original metrics. This simplification hypothesis was tested using two non-glance metrics (TaskTime and Miss%_tdrt). These non-glance metrics sufficed to span the two-dimensional PCA space to a reasonable approximation, which had been derived from the complete data set of 22 metrics (including many glance metrics) measured for 15 tasks with different subsets of modes (Visual-Manual, Mixed-Mode, Auditory-Vocal, and baseline just driving). The Dimensional Model predicts that other pairs of metrics besides TaskTime and Miss%_tdrt should also be able to reasonably span this 2-D space (e.g., for example MSGD_TR and TGT_TR, as per the Alliance Guidelines Principle 2.1A [13]). 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 potential options were not further explored for these tasks.

Comparison of Mixed-Mode Tasks to Visual-Manual Tasks

This study found that a hands-free Mixed-Mode interface that uses a display screen has relatively less cognitive demand but relatively greater physical demand than a Visual-Manual task with same start and goal states. It cannot yet be determined with available data whether such a hands-free Mixed-Mode interface will increase or decrease driver performance (and safety-related critical events) compared to a Visual-Manual interface for performing the same task in naturalistic driving data.

However, the results of this study contradict the assertion that the attentional effects of cognitive demand from tasks that employ voice commands is substantial, as claimed in some other studies [24,25,26]. The analysis of the results showed in Figure 7 that all of the six Mixed-Mode tasks investigated here, and both pure Auditory-Vocal tasks, had a reduced attentional effect of cognitive demand compared to pressing a single touch screen preset button one time to tune to a new radio station (the RadioPreset_VM task). Opposite to what is claimed in some studies [24,25,26], it was rather the physical demand from the voice command tasks studied here that showed a relative increase compared to most of the Visual-Manual tasks studied here (except for NavDest_VM and BingPOI_VM). This relative increase in physical demand was only evident in the multivariate methods used here, and was not evident in examining only single glance metrics in a univariate manner [10], which showed a decrease in TEORT for example for Mixed-Mode tasks vs. Visual-Manual paired tasks (see Appendix C for explanation).

Comparison of Mixed-Mode Tasks to Pure Auditory-Vocal Tasks

Adding a display screen to an Auditory-Vocal interface is likely done by vehicle manufacturers and/or their suppliers in an attempt to assist the driver with the voice command interface, such as providing confirmation of voice inputs and enabling users to see the menu

choices instead of, or in addition to, hearing them. Because only voice commands are made in a Mixed-Mode interface, rather than manual button presses (except for pressing a “voice activation button”) such an interface qualifies as “hands-free” according to the SAE J2988 [11]. Indeed, a hands-free Auditory-Vocal interface such as with the OnStar Hands-Free Calling is known to carry no additional crash risk beyond that of baseline driving without using the system [27]. Whether adding a visual display to an Auditory-Vocal hands-free task enhances or detracts from driver performance compared to a pure Auditory-Vocal interface was not tested in this study, nor does such a study exist in the published literature to our knowledge. For example, studies have not been published to our knowledge which test if whether the screen is covered up, that driver performance diminishes or improves when using a pure Auditory-Vocal interface, vs. a Mixed-Mode interface when the screen is visible.

However, given the current findings that participants performed the pure Auditory-Vocal 0-Back task with less cognitive and physical demand than even the Mixed-Mode tasks, it is hypothesized that a pure Auditory-Vocal hands-free task might have reduced the attentional effects of cognitive demand compared to Mixed-Mode tasks. An Auditory-Vocal task such as cell phone conversation may even improve alerting attention and hence reduce drowsiness [9].

Drivers may also reduce speed during performance of Auditory-Vocal tasks because of self-regulation [2,28]). It has not yet been tested to our knowledge whether drivers performing a Mixed-Mode task might also show evidence of self-regulation such as slowed speeds.

A pure Auditory-Vocal system may have relatively poor voice recognition for some dialects, or may have a complex voice tree structure. Such negative attributes might potentially increase driver frustration and reduce customer acceptance to some degree, but the results of this study imply that pure Auditory-Vocal tasks with equivalent cognitive demand to 0-Back will have lower attentional effects from cognitive demand than even Mixed-Mode tasks.

Eye Movements Not Needed to Estimate Demands

A potentially useful contribution of this study is that the methods applied in this dataset (as in the four datasets examined in the companion paper [1]) demonstrate that eye movements are not needed to predict the attentional effects of cognitive demand, nor the physical demand, of secondary tasks (i.e., the location of the secondary tasks in the 2-D score space). The dimensional analysis employed here found that the task locations in the 2-D score space could be reliably approximated by just TaskTime and the percentage of misses to a tactile DRT, at least for the 15 tasks with multiple different types of modalities assessed here. For Visual-Manual tasks, TaskTime and RT to an event were sufficient to approximate the task locations in the 2-D score space.

This result is useful, because the eye glance metrics of TEORT, MSGD, and LGP have complex interactions between them, which causes a task's TEORT value in a univariate analysis method to underestimate the physical demand of the task. In addition, by using eye movement data, a multivariate analysis that considers the