INTERNATIONAL CONFERENCE ON ENGINEERING DESIGN, ICED'09
24 - 27 AUGUST 2009, STANFORD UNIVERSITY, STANFORD, CA, USA

UNDERSTANDING HETEROGENEITY OF HUMAN PREFERENCES FOR ENGINEERING DESIGN

Christopher Hoyle[1], Wei Chen[1], Nanxin Wang[2], and Gianna Gomez-Levi[2]
(1) Northwestern University (2) Ford Research and Advanced Engineering

ABSTRACT
In today's competitive market, it is essential for producers to provide products which not only achieve high performance, but also appeal to the tastes of consumer. Therefore, a key element of design is an understanding of human preferences for products and features. In this work, a human appraisal experiment is conducted to understand preferences for automobile occupant package design. The experiment is conducted to build predictive parametric models of consumer preferences. An issue with this class of experiment is that the heterogeneity of the experimental respondents contributes to the response, and this heterogeneity must be understood to separate the influence of design factors from that of human factors. Latent class analysis is used to combine multiple responses of the human appraisal respondents to an appropriate set of measures. Cluster analysis and smoothing spline regression are used to gain an understanding of respondent rating styles and preference heterogeneity. These analyses allow estimation of ordered logit models for prediction of consumer occupant package preferences. Methods from machine learning are also investigated as an alternative to parametric modeling.

Keywords: Human appraisal, heterogeneity, ordered logit, latent class analysis, smoothing spline regression, machine learning

1 INTRODUCTION
In today's highly competitive market, it is essential for producers to provide products which not only achieve a high level of performance, but also appeal to the tastes of a broad range of consumers. Therefore, a key element of the design process is forecasting potential consumer opinion for the features to be incorporated in a new product. In our previous work [1] we developed an experimental design methodology for human appraisal experiments which provides optimal sampling over both *product* attributes (**A**), or attributes of the design, and *human* attributes (**S**), or attributes of the consumer. It is important to account for consumer heterogeneity in both experiments and modeling because each consumer has a unique preference: an "average" or group preference does not exist [2]. Therefore, an understanding of consumer preferences requires an investigation into preference heterogeneity. An experiment was conducted using the Ford Programmable Vehicle Model (PVM) [3] to determine preferences for automobile occupant package design, specifically regarding the roominess and ingress-egress quality of the package. The design of an automobile occupant package is used in this paper as a motivating example because of the interaction between the product design and the consumers' human attributes, such as their height or weight, in determining preferences. In the experiment, each respondent is presented with several package *configurations*, for which they evaluate and express their opinion in the form of a *rating* (e.g., 1-5, 0-10), a standard method for quantifying preferences for subjective attributes [4]. The intent is to use the data collected in the experiments to build *ordered logit* models to predict consumer preferences (i.e. ratings) for a given set of consumers and for a given occupant package design.
Analyzing and creating models from data collected from a human appraisal experiment presents unique issues not encountered with data collected from the typical industrial and scientific experiments usually considered in design of experiments methodology [5, 6]. The key issues in human appraisals are that the responses are more difficult to elicit, respondents may utilize different rating styles, the shape of the response-factor curve may not be approximately linear, and interactions may be highly significant. To address these issues, several analysis and modeling methodologies are employed in this work and examined for their applicability to human appraisal experiments. The data

collected in the PVM experiments are analyzed to combine multiple consumer responses into a set of combined measures, to understand the influence of respondent heterogeneity on rating responses, and to gain further insight into the experiment using alternate data analysis methods. In the human appraisals, multiple responses are often collected from the respondent for a single sub-system design. The reason multiple responses are collected for certain sub-systems is because it can be challenging to devise a single survey question to capture the respondents' true opinion of the subsystem design as a whole, and multiple questions are used to assess opinion for different aspects of the design. To determine a measure to use in the modeling process, *Latent Class Analysis* (LCA) [7] is used to create a combined subsystem measure for each respondent to fully describe his/her overall opinion of the subsystem design. Heterogeneity of the survey respondents has much influence on the rating responses given. The effect of *systematic* heterogeneity, which is heterogeneity that can be captured with a human variable in the model, is investigated using *Smoothing Spline Regression* (SSR) [8]; *random* heterogeneity, which is heterogeneity not directly observed but rather captured in a distribution of respondent-specific intercepts, is investigated using *Cluster Analysis* (CA) [9]. The previous analyses allow estimation of parametric *Random-Effects Ordered Logit* (RE-OL) models for the prediction of ratings for a given population and given package design. In addition to the parametric ordered logit models, methods from machine learning are also explored. Decision trees and Bayesian networks [10] are used to gain insights into the data not easily seen in the previous analyses or parametric modeling methods. The methods developed in this work for the analysis of data collected from human appraisal experiments compliment our previous work in human appraisal experimental design [1]. These methods provide a clear understanding of the heterogeneous preferences within a consumer population, applicable for understanding preferences for system, subsystem, or component design.

2 PVM ROOMINESS/INGRESS/EGRESS EXPERIMENTS

To understand preferences for the occupant package design, a comprehensive set of human appraisal experiments to access vehicle package overall roominess, ingress and egress preferences is conducted. Conducting such experiments is motivated by the fact that occupant package design is determined by the occupant package dimensions, as well as by the exterior vehicle dimensions and the structural design dimensions, illustrated in Figure 1.Therefore, understanding human preferences for occupant package dimensions is a key element in vehicle design.

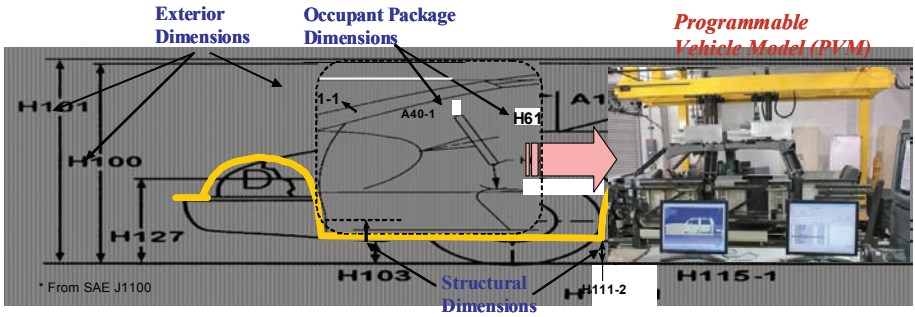

Figure 1. Example of Key Dimensions which Influence Occupant Package Design

The full design of the Programmable Vehicle Model (PVM) roominess/ingress/egress experiment is created using the optimal design of human appraisal experiments (DOE) methodology presented in [1]. The combined experiment consists of eight product factors that have previously been found to influence roominess, ingress and egress preference. The eight factors used in the human appraisal experiment correspond to dimensions defined for control of the Ford PVM are:

1. a_1: Hinge position in X (**HNG_X**)
2. a_2: Rocker position in Y (**ROK_Y**)
3. a_3: Heel position in Z (**HEL_Z**)
4. a_4: Ground position in Z (**GRD_Z**)
5. a_5: Sill position in Z (**StoH**)
6. a_6: Roof position in Z (**HR_Z**)

7. a_7: Front Header position in X (**HR_X**)
8. a_8: Side Rail position in Y (**HR_Y**)

All product factors, a_1-a_8, assume three levels to create a response surface ordered logit model. Three human attributes have been hypothesized to influence roominess/ingress/egress opinions:

1. s_1: Gender (**gend**)
2. s_2: Body Mass Index (**BMI**)
3. s_3: Stature (**stat**)

In this experiment, gender assumes two levels, BMI three levels, and stature (or height) four levels. In addition to the human attributes used for the *design* of the experiment, each respondent's age and seated height (seat) was collected at the time of the experiment. Because the total experiment is quite large, it is divided into a series of *blocks*, each of which represents a fraction of the larger experiment. For the appraisal, 30 respondents are used, each of whom evaluates 18 occupant package configurations: the 18 configurations represent a single block. Respondents provide ratings on a 1-5 scale for 10 questions regarding the occupant package subsystems as follows:

1. *Ingress*: Three questions related to ease/difficulty
2. *Roominess*: Headroom, Leftroom, Kneeroom, and Overall Roominess
3. *Egress*: Three questions related to ease/difficulty

The goal of the experiments is to create probabilistic rating models to predict a rating, R_p, as a function of product and human attributes, $\mathbf{Z}$ ($\mathbf{Z} = \{\mathbf{A}, \mathbf{S}\}$), using the ordered logit (OL) model [11]

$$\Pr\left[R = R_p \mid Z_1, \ldots, Z_J\right] = F\left(k_p - \boldsymbol{\beta}'\mathbf{Z}\right) - F\left(k_{p-1} - \boldsymbol{\beta}'\mathbf{Z}\right), \tag{1}$$

where k_p are the OL cutpoints and F is the cumulative logistic distribution function. In the RE-OL model, a random-effect parameter, β_0, is used to capture random heterogeneity, σ_u [12].

$$\boldsymbol{\beta}'\mathbf{Z} = \beta_0 + \boldsymbol{\beta}'\mathbf{A} + \boldsymbol{\beta}'\mathbf{S} \text{ where } \beta_0 \sim N\left(0, \sigma_u\right). \tag{2}$$

Note that in this work, all factor values are normalized on the scale [0, 1].

3 LATENT CLASS ANALYSIS FOR COMBINING MULTIPLE RESPONSES

A primary issue with the collected data is that it is desired to create predictive preference models for each major sub-system attribute, for example ingress, egress and interior roominess. However, in the survey three responses were collected each for ingress and egress (i.e., acceptability, effort, and space) and it is not clear how a single measure of ingress or egress preference can be inferred from the multiple responses. A correlation matrix is estimated which shows significant correlation among the three ingress and three egress responses, respectively. Based on this observation, a formal analysis of the responses is conducted using Latent Class Analysis (LCA). LCA is a general method for data reduction for discrete categorical or ordinal data, analogous to factor analysis used for continuous variables [7]. LCA assumes that several discrete variables, such as the three ratings given by each person for ingress or egress, are *indicators* of an overall discrete *latent class* (LC), such as an overall opinion of ingress or egress. LCA provides a single latent class response for each subsystem response (e.g. ingress), based upon the value of the indicators. This predicted LC can be used as the ingress or egress response in a parametric model, analogous to the use of factor scores resulting from factor analysis for continuous variables.

LCA analysis assumes that the several response indicators are correlated, and seeks to divide the subsystem responses into a number of latent classes such that the indicators are *conditionally independent* within each class. Conditional independence implies that the correlation between the indicators is no higher than "chance" correlation in any class. In order to determine the division of subsystem responses to LCs, the number of LCs must be defined *a priori* for model estimation. The division of subsystem responses is achieved using maximum likelihood estimation to estimate the conditional probabilities of each subsystem response given the LC, and the probability of each LC. Among different models (i.e. different assumptions on the *a priori* number of LCs), the Akaike Information Criterion (AIC), which is a function of likelihood and the number of classes, is used for model selection. The model with the lowest AIC is the preferred model, i.e. the model which balances goodness of fit with the complexity of the model.

LCA is conducted for ingress, assuming the three ingress questions (i.e., acceptability, effort, and space) are indicators of each persons overall opinion of the ingress quality. Different numbers of latent

classes, between 1 and 10, were tested in the modeling process. These ten models are compared based on the AIC criteria, indicating that the 7 LC model is the preferred model. A comparison using the latent class ingress measure versus the original 3 ingress measures is shown in Table 1 using ordered logit models for comparison (numbers in table are ordered logit β coefficients). As seen in the models, the coefficients in the latent class model are within the max. and min. range of the coefficients in the models using the three indicators as responses, indicating the latent class is capturing the effect of all three of the ingress indicators.

Table 1. Ordered Logit Coefficient Comparison of Ingress Measures

	acceptability	effort	space	*Range*	*latent class*
HEL_Z	2.017	2.272	1.344	1.34 — 2.27	**1.912**
GRD_Z	-2.026	-2.261	-1.162	-2.26 — -1.16	**-1.875**
StoH	-1.124	-1.268	-0.731	-1.27 — -0.73	**-0.985**
HR_Z	2.270	1.745	2.703	1.75 — 2.70	**2.196**
HR_X	0.550	0.512	0.476	0.48 — 0.55	**0.527**
Stat	-2.814	-2.790	-4.954	-4.95 — -2.79	**-3.150**
Age	3.402	2.878	2.493	2.49 — 3.40	**2.956**
BMI	-2.382	-1.686	-2.003	-2.38 — -1.69	**-2.158**
ρ_0^2	0.1886	0.1873	0.1873		0.139

LCA is also conducted for the 3 egress responses, with a similar result to ingress: the preferred number of classes is found to be 7. LCA was used to create a model for all six ingress/egress responses, assuming ingress and egress responses are indicators of an overall opinion of the vehicle opening; however, an acceptable model was not identified using any number of assumed latent classes. Therefore, it can be concluded that the combination of the three ingress responses are indicators of a respondent's opinion of ingress, whereas the combination of the three egress responses are indicators of egress opinion, respectively.

4 UNDERSTANDING FACTOR AND RESPONDENT IMPORTANCE

In the previous section, latent class analysis was used to understand the relationship among responses, in situations in which multiple responses are assumed to be related to a single unobserved latent class. In this section, methods will be used to understand the relationship among the factors (product and human attributes), human respondents, and the responses. In order to understand how the overall variance in the responses is partitioned among the explanatory variables, an Analysis of Variation (ANOVA) analysis is conducted. ANOVA analysis is an investigation of how the total sum of squares SS_T is decomposed into the sum of squares contributions from the model, SS_M, and the error, SS_E. The SS_M can be further decomposed to understand the influence of the individual product factors, SS_{TR}, and the individual human factors, SS_R, including the *block effect* attributable to individual respondents. The block effect is the portion of the respondent response not explained by the human factors, with the effect of different configurations and human attributes removed. It is realized in a model as a respondent-specific intercept (i.e. 24 unique intercepts). The magnitude of the sum of squares is a measure of the contribution of each factor and respondent, as well as the error, in explaining the variation in the responses (i.e. the ratings).

An ANOVA analysis is conducted for each response, and each of the product and human attributes. Several insights into the collected data are provided by the analysis, particularly the importance of the respondent block effect. The magnitude of the SS block effect versus the magnitude of the SS human factors is approximately equal, indicating that there is much heterogeneity in responses not captured by the human factors. This unexplained heterogeneity can be attributed to human or socio-economic attributes not recorded and therefore not included in the analysis (e.g. income, usage), or individual *rating styles*. It has been found in previous research that respondents often display distinct rating styles, such as rating systematically high or low. Attempts have been made previously to identify these behaviors and control for them in the modeling process [13, 14] by using the mean and variance of each person's set of ratings for normalization; however, in this experiment, respondents were not given the same set of configurations to evaluate and it is expected that respondents with different human attributes rate differently. Thus, comparison of the mean rating of each person is meaningless.

For this reason, a general method to control for rating style must be developed which does not assume respondents have evaluated the same set of configurations or that each respondent should have the same mean rating. In this work, the block effect is used as a means of comparison among different respondents. With the respondent block effect available for each respondent for each of the 10 responses, cluster analysis is conducted to determine unique clusters of respondent ratings styles, for example, a high block effect indicates a systematically high rater, whereas a low block effect indicates a systematically low rater. Cluster analysis is conducted using complete linkage hierarchical clustering [9], with results shown in Figure 2. The three cluster model separates the respondents into groups in which each respondent's block effect is close to zero (Neutral Raters), positive (High Raters), or negative (Low Raters).

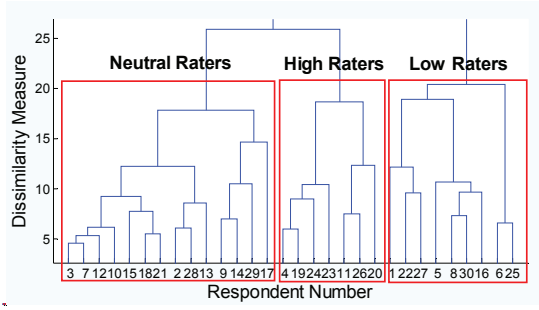

Figure 2. Cluster Analysis to Understand Respondent Rating Style

To capture style in the modeling process, rating style is represented using two dummy variables: one for high rating style, $styl_H$, and one for neutral rating style, $styl_N$, to represent the three clusters of rating styles. To demonstrate the effect of the rating style variables, random-effects ordered logit models are estimated with and without inclusion of the style variables, illustrated using the LC ingress response in Table 2.

Table 2. Comparison of Random-Effects Ordered Logit Models for Ingress w/ and w/o Style

	Without Style		With Style	
	coef.	*t-value*	*coef.*	*t-value*
ROK_Y	0.324	*1.91*	0.320	*1.89*
HEL_Z	2.371	*11.04*	2.371	*11.06*
GRD_Z	-2.207	*-7.40*	-2.217	*-7.42*
StoH	-0.863	*-4.00*	-0.866	*-4.02*
HR_Z	2.822	*13.33*	2.818	*13.32*
HR_X	0.745	*4.22*	0.746	*4.23*
gend	-0.380	*-0.64*	0.619	*1.35*
stat	-0.313	*-1.03*	0.341	*1.39*
BMI	0.229	*0.68*	0.100	*0.40*
age	-1.164	*-1.44*	0.408	*0.64*
$styl_H$			2.307	*5.38*
$styl_N$			1.611	*2.04*
σ_u	1.48		0.64	
ρ_0^2	0.184		0.194	

Random respondent variation is reduced significantly with the inclusion of an explanatory variable for ratings style: the fraction of unexplained variance at the respondent level, σ_u, reduces from 1.48 to 0.64 with the inclusion of this style variable. In addition, the goodness-of-fit of the model, ρ_0^2 (a measure between 0 and 1), improves from 0.184 to 0.194. This indicates there is less unexplained rating heterogeneity among respondents with inclusion of the style term, i.e. rating style accounts for a significant portion of unexplained heterogeneity. The benefit of including the style term in the predictive model is an improved understanding of the heterogeneity in rating responses. Assuming the population sampled in the experiment is representative of the population as whole, controlling for the

rating style explicitly in the model will provide better predictions than those obtained by integrating over the respondent variance. Also, by knowing people have certain ratings styles, a pre-experiment calibration technique could be implemented to determine a respondent's rating style before the appraisal is conducted to ensure better consistency in rating style in future experiments [13].

5 SPLINE REGRESSION TO UNDERSTAND RESPONSE BEHAVIOR

With a set of responses determined in Section 3 and an understanding of heterogeneity in Section 4, the modeling process can begin. A remaining issue is an understanding of the functional relationship between the factors and responses. It has been found in the study of psychophysics that a human response to stimuli follows a power law relationship [15], which provides guidance for determining the form of the product factors in the model. However, in the case of human or socio-economic attributes, such a general theory does not exist. In addition, the actual human attributes of each person were collected during the experiment and will be used in model estimation, such that higher ordered terms (e.g., quadratic, cubic) can be estimated for these terms. A general method to understand the relationship between the response and a factor is the use of *smoothing spline regression*. Smoothing spline regression is similar to piecewise linear regression; however, the breakpoints are connected with polynomials as opposed to lines.

Using spline regression, we can better understand the relationship between response and factor, and decide upon the factor forms (e.g., linear, quadratic, cubic) to include in the subsequent RE-OL models. In this work, smoothing spline linear regression models are fit to the PVM human appraisal data and the results used to provide guidance in determining factor forms for the ordered logit modeling, in which the utility function is linear additive. Plots of representative smoothing spline regression relations are shown in Figure 3 a), b), and c) (dashed lines represent 95% confidence intervals).

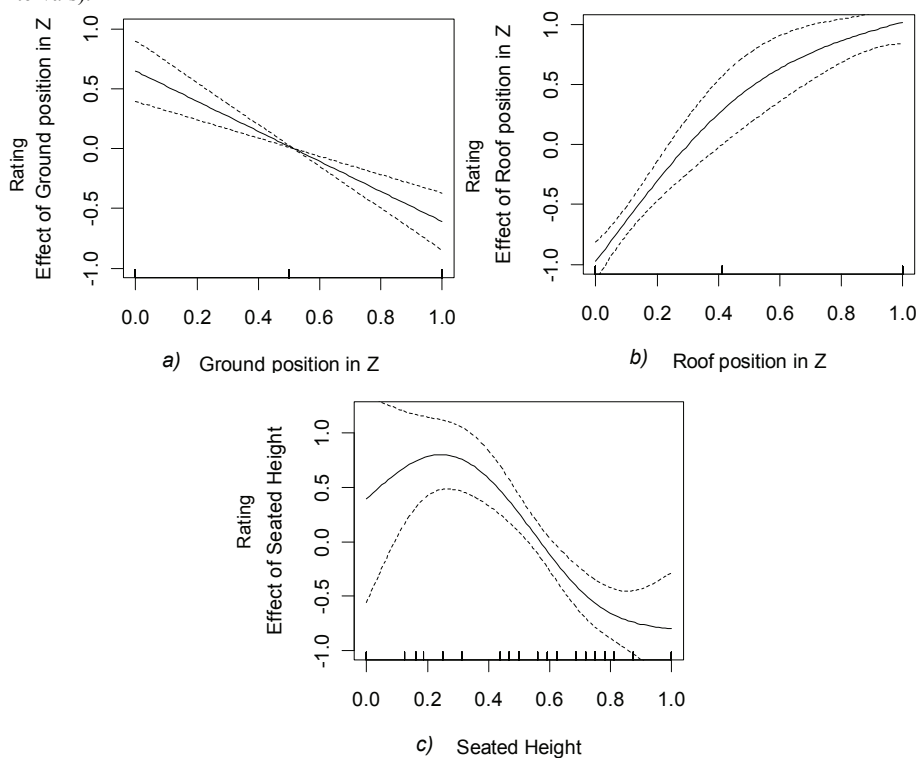

Figure 3. Examples of Linear, Power Law, and Critical Level Attributes

These three plots represent the three dominant types of relationships found in the modeling process.

1. *Linear Relationship*: As illustrated in Figure 3 a) using the SgRP to Ground factor (GRD_Z) as an

example, many of the factors, both product and demographic, have a linear relationship with the rating response (e.g. the LC ingress measure).

2. *Power Law Relationship*: As illustrated in Figure 3 b) using the SgRP to Roof Z factor (HR_Z) as an example, several of the product factors exhibit a power law relationship. In such a relationship the rate of increase of the rating response decreases as the magnitude of the stimuli increases. This is important to capture in the modeling process for use in design because improving the magnitude of these dimensions, such as HR_Z, result in a diminishing rate of increase in the expected rating. Therefore, it may be more advantageous to improve the vehicle dimensions of other factors appearing in the ordered logit model, such as those exhibiting a linear response.
3. *Critical Level Relationship*: As illustrated in Figure 3 c) using Seated Height as an example, several of the human attributes display a critical level relationship. In such a relationship, the rating response is constant over certain factor levels, such as very small (0.0−0.2) or very large (0.8−1.0) seated heights, but displays a linear (or higher) relationship over other levels of the factor, such as medium statures.

With an understanding of the various relationships created using smoothing splines, a straightforward method is required to approximate these relationships in the random-effects ordered logit models. The three relationships identified can be approximated closely through combinations of linear, quadratic, and cubic terms in the model. The linear relationship only requires a linear term, the power relationship a linear and quadratic term, and the critical level relationship a linear, quadratic, and cubic term. This method is utilized and demonstrated in a random-effects ordered logit model for the latent class ingress rating response. The results of the model are shown in Table 3.

Table 3. Random-Effects Ordered Logit for Ingress Response with Factor Shapes

Product Attributes

	ROK_Y	$\mathbf{HEL_Z}$	$\mathbf{HEL_Z^2}$	GRD_Z	StoH	$\mathbf{HR_Z}$	$\mathbf{HR_Z^2}$	$\mathbf{HR_X}$	$\mathbf{HR_X^2}$
coef.	0.29	6.95	-4.75	-1.87	-0.79	35.95	-33.15	5.43	-4.73
t-value	*1.71*	*2.83*	*-1.92*	*-6.03*	*-3.58*	*3.72*	*-3.45*	*1.81*	*-1.57*

Human Attributes

	gender	**age**	$\mathbf{age^2}$	BMI	**seat**	$\mathbf{seat^2}$	$\mathbf{seat^3}$	$styl_H$	$styl_N$
coef.	1.25	-5.02	5.12	1.37	832.33	-1637	805.13	3.05	1.41
t-value	*1.61*	*-1.8*	*1.61*	*1.88*	*1.68*	*-1.7*	*1.65*	*4.72*	*2.04*

Additional higher-ordered terms were tested in the RE-OL model, but the relationships identified in the smoothing spline regression were found to be applicable for the RE-OL, and thus no other higher ordered terms were found to be significant. Similar findings were made for the other collected PVM responses, i.e. headroom, leftroom, kneeroom, roominess, and egress. Based upon this study, we conclude that smoothing spline regression is an effective method for better understanding the response-factor relationship, and guiding the selection of terms to be included in the prediction model.

6 RANDOM-EFFECTS ORDERED LOGIT MODELS

With the set of responses determined using latent class analysis in Section 3, an understanding of the rating style in Section 4, and an understanding of the shape of the factor-response relationship in Section 5, random-effects ordered logit models are fit to the data. The previous methods did not study the effect of the interactions, which will be investigated in the modeling process. As an example model, the RE-OL model for the LC ingress response with significant terms, including interactions, is shown in Table 4. Models for the other responses are estimated similarly.

In comparing among the RE-OL models for ingress-egress and roominess, factors thought to be primarily associated with roominess, such as HR_Z, HR_X, and HR_Y, appear in the ingress-egress models, and factors thought to be associated with ingress-egress, such as HNG_X, appear in the roominess models. The reason for this could be two-fold: respondents' opinions of ingress-egress also influence their opinions of roominess, or the factors actually contribute to the ingress-egress or roominess experience directly. Different human attributes and human attribute interactions appear in the models. For example, gender, seated height, and age appear in the ingress model, whereas only age appears in the egress model. This could be explained by the fact that it is generally easier for

respondents to exit the vehicle than enter the vehicle, and thus factors such as seated height and anthropomorphic gender differences do not influence the rating for egress as they do for ingress.

The effect of including both systematic (**S**) and random heterogeneity (σ_u) (Eq. (2)) on the rating predictions can be seen using a simple example in which the headroom model is re-estimated without **S**, without σ_u, and without both **S** and σ_u. The models estimated with different representations of heterogeneity are compared in terms of their ability to match the first four moments of the actual ratings distribution, shown in Table 5.

Table 4. Random-Effects Ordered Logit for Ingress Response with Interactions

Product Attributes

	ROK_Y	HEL_Z	HEL_Z^2	GRD_Z	StoH	HR_Z	HR_Z^2	HR_X	HR_X^2
coef	0.28	-16.75	-5.43	-1.75	-4.09	45.78	-47.01	7.71	-10.49
t-value	*2.18*	*-3.8*	*-2.01*	*-1.33*	*-5.53*	*4.2*	*-4.33*	*2.12*	*-2.99*

Product Interactions

	ROK_Y·HEL_Z	**ROK_Y·GRD_Z**	**ROK_Y·HR_Y**	**HEL_Z·GRD_Z**	**HEL_Z·StoH**	**HEL_Z·HR_Z**
coef	2.78	-5.48	4.84	6.21	4.66	22.94
t-value	*1.39*	*-2.52*	*2.97*	*2.25*	*5.09*	*7.69*

Human Attributes

	gend	age	seat	$seat^2$	$seat^3$	**seat·gend**	**seat·age**	$styl_H$	$styl_N$
coef	-33.53	-51.67	-786.25	1399.74	-632.79	35.65	55.64	2.16	0.86
t-value	*-2.35*	*-3.33*	*-1.43*	*1.29*	*-1.18*	*2.37*	*3.37*	*3.65*	*1.88*

Table 5. Comparison of Inclusion of Different Forms of Heterogeneity in the Ratings Model

	OL		***OL with S***		***OL with σ_u***		***OL with S & σ_u***	
	Sample	*Error*	*Sample*	*Error*	*Sample*	*Error*	*Sample*	*Error*
Mean	3.321	*-0.15%*	3.322	*-0.14%*	3.315	*-0.35%*	3.318	*-0.25%*
Variance	2.049	*-26.41%*	2.315	*-16.86%*	2.203	*-20.85%*	2.389	*-14.20%*
Skewness	-0.109	*-68.19%*	-0.249	*-27.35%*	-0.168	*-50.93%*	-0.264	*-22.94%*
Kurtosis	1.178	*-18.92%*	1.349	*-7.11%*	1.276	*-12.14%*	1.360	*-6.36%*
ρ_0^2	0.380		0.483		0.518		0.536	

A primary difference among the models can be seen in the goodness-of-fit, ρ_0^2, which increases as either systematic, random, or both, types of heterogeneity are included in the model. The effect of the improved model goodness-of-fit results in improved moment matching, as can be seen in the decreasing error in each moment as heterogeneity is more explicitly represented. An exception to this is the ability of any of the models to match the mean, since all models are unbiased estimates of the mean. The improved model fit can be seen graphically using a comparison of histograms of the OL model without **S** and σ_u versus the OL with **S** and $\boldsymbol{\sigma_u}$ model in Figure 4.

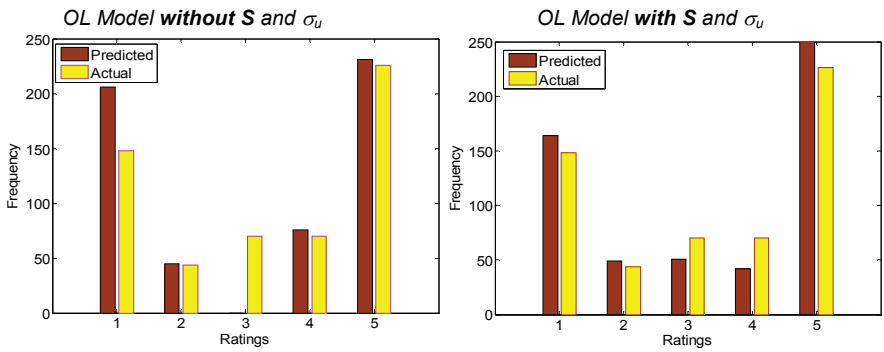

Figure 4. Comparison of Lowest to Highest Goodness of Fit OL Ratings Models

It can be seen that the OL model without **S** and σ_u does a poor job of matching the actual ratings distribution, whereas the OL model with **S** and σ_u is much better at matching the actual ratings distribution.

7 ALTERNATE DATA ANALYSIS METHODS

While the focus of this work is to estimate parametric models to make predictions of human preferences, data mining machine learning methods are also investigated to gain further insight into the data, and to confirm the RE-OL modeling approach. The data mining methods investigated in this work are classification methods, i.e. methods to predict the ratings class (i.e. 1-5 rating) based upon the attribute values. Two applicable approaches to classification data mining are investigated: a Decision Tree and Supervised and Unsupervised Bayesian networks. The five classes to be estimated are the five ratings: 1, 2, 3, 4, 5. An issue with these approaches is that the mainstream implementation of the Decision Tree and Bayesian network is based upon the assumption that attribute values, **Z**, including both product attributes **A** and human attributes **S**, are discrete categorical variables. This is not a significant issue for the PVM product factors, which only assume three levels and therefore can be considered discrete; however, they will be treated as nominal as opposed to interval (or ratio) level variables. The demographic attributes are generally continuous interval level variables (except gender), and thus will be divided into discrete categories based upon their continuous values.

7.1 Decision Tree Analysis of the PVM Dataset

A Decision Tree was created using the PVM dataset. A decision tree is created through a process in which a number of observations or cases, *c*, within a training data set, *Tn*, are *classified* into a number subsets with respect to a class variable (i.e. a response), R_p, based upon a rule concerning a "splitting" attribute value, **Z** (i.e. a product or demographic attribute). The tree building process continues to add branches until no further information can be gained. The decision tree is then pruned using a cost criterion to maximize the classification accuracy relative to the complexity of the tree [10]. The goal is to create a non-parametric model capable of predicting the class (rating) based on the value of the attributes. In this respect, a decision tree is similar to the ordered logit model, in that the goal is to predict a rating (i.e. rating is the class) based upon attribute values (e.g. HR_Z, Stature). Therefore, a decision tree can be viewed as a non-parametric alternative to the ordered logit model. As an example, a simplified decision tree is built for the *headroom* response as shown in Figure 5 (variables un-normalized for clarity). All units in the figure are in *cm*, except for BMI in standard units of kg/m^2; the number in the box is the rating class, and the number below the rating class is the number of predicted observations belonging to each rating class based on the classification rule.

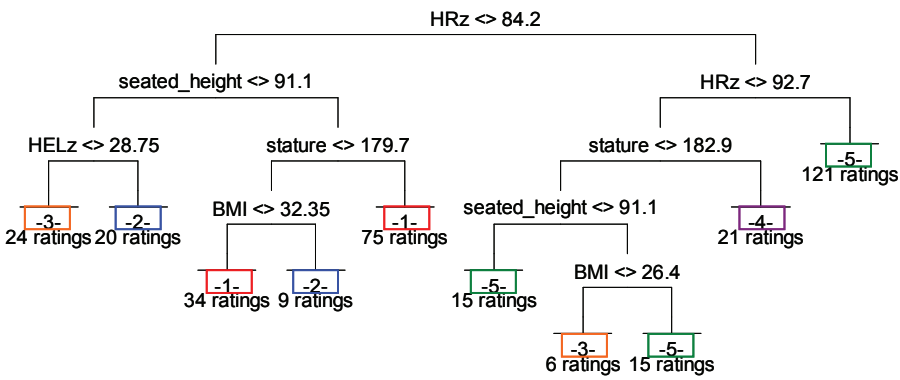

Figure 5. Simplified Decision Tree for the Headroom Response

The decision tree can provide insights not gained readily in traditional parametric modeling methods, such as ordered logit modeling. One such observation is that 85% of configurations receiving a rating of 5 occur when HR_Z (a_6) is at its maximum value, regardless of other product or demographic attribute values. This indicates that increasing HR_Z is a straightforward method for achieving a high headroom rating. While the HR_Z attribute is dominant in the ANOVA analysis, the decision tree

provides information regarding how specific attribute values influence specific rating frequencies. Another interesting observation is that the combination of low values of HR_Z coupled with respondents of large seated height and overall height account for the majority of the low ratings (69%). A more enlightening finding is that HR_Z at its minimum value coupled with high seated height, low height, and low BMI account for 31% ratings of 1. This could possibly be explained by the seating position of low BMI respondents versus high BMI respondents, because low BMI respondents may position their seat differently in terms of lateral position and tilt angle, leading to a different experience of headroom for a given configuration for respondents of the same height. This can be captured in a model through the inclusion of a BMI-seated height interaction term, which should be positive in sign. In conclusion, these findings indicate that HR_Z, BMI, seated height, height and a BMI-seated height interaction are important variables in the parametric modeling process.

7.2 Bayesian Networks for the PVM Dataset

The use of Bayesian networks in analyzing and modeling the PVM data is investigated in this subsection. The Bayesian network can be used in two distinct implementations: supervised and unsupervised. In the supervised implementation, the Bayesian network is used as a classifier in which attribute values are used to predict a class, e.g. a rating. In the unsupervised implementation, no assumption is made regarding responses, or dependent variables, and factors, or independent variables, but rather the network identifies dependent and independent variables. The two implementations of the Bayesian network will be investigated.

7.2.1 Supervised Bayesian Network

The supervised Bayesian network is a classifier in which a class R_p, such as a rating, is predicted based upon the conditional probability of the attribute **Z** values. In the supervised network, the class to be predicted is defined *a priori*. Therefore, the Bayesian network is used as a method to determine the probability of being in each class R_p, (i.e. each rating category) for each observation (i.e. each respondent). The probability of the class assuming a certain value R_p (i.e. rating of 1,2,3,4 or 5), given a set of attribute values **Z**, is determined using Bayes law:

$$\Pr\left[R = R_p \mid Z_1,\ldots,Z_J\right] = \frac{\Pr\left\{Z_1,\ldots,Z_J \mid R = R_p\right\}\cdot \Pr\left\{R_p\right\}}{\Pr\left\{Z_1,\ldots,Z_J\right\}}. \tag{3}$$

The Bayesian network uses the assumption of conditional independence. Conditional independence requires that each attribute, Z_j, is conditional only on the immediate, or parent, attributes and not upon the distant relative attributes (i.e. grandparents, great-grandparents, etc.). Using this assumption, the conditionally independent probabilities can be multiplied to find the joint probability of **Z**:

$$\Pr\left\{Z_1,\ldots,Z_J \mid R = R_p\right\} = \prod_{j=1}^{J} \Pr\left\{Z_j \mid R, \text{parents}\left(Z_j\right)\right\}. \tag{4}$$

Eq (3) demonstrates that the supervised Bayesian network is a form of non-parametric regression. An advantage of the Bayesian network is that no assumptions are made on the error distribution (i.e. logistic or normal distribution) because it is non-parametric. The Bayesian network ratings predictions can therefore be directly compared to the ordered logit regression predictions given by Eq. (1).

In the comparison, it is found that the Bayesian network results in similar ratings classification to the ordered logit model. The superior performance of the ordered logit model can be attributed to the enforcement of the ordinal constraint (i.e. adjacent ratings are correlated), as opposed to the nominal assumption of the Bayesian network, and the discretizing of attributes to nominal categories in the Bayesian network. The Bayesian network also identifies the conditional relationships; in this study, the effect of a_6 (i.e. HR_Z) is conditional on the value of seated height, indicating that an interaction term of HR_Z·seated should be investigated in a parametric modeling process.

7.2.2 Unsupervised Bayesian Network

As opposed to the supervised Bayesian network which can be viewed as an alternative to the ordered logit model, the unsupervised Bayesian network is used to understand relationships in the data. In the unsupervised Bayesian network, no distinction is made between responses and factors. For this reason,

the focus is upon identifying the joint distribution of attributes **Z** (in this case the rating response is considered another attribute) in terms of the conditional distributions:

$$\Pr[Z_1,\ldots,Z_J] = \prod_{j=1}^{J} \Pr\{Z_j \mid \text{parents}(Z_j)\} \tag{5}$$

An unsupervised Bayesian network for the PVM dataset was conducted. The primary finding is that not all human attributes collected, such as height and seated height, are independent, i.e. there are significant correlations among the human attributes. For example, age is conditional upon the values of gender, height, seated height, and BMI, which can be confirmed using a standard linear regression analysis. The issues created by this correlation among human attributes in the ordered logit modeling process are *redundancy* and *suppression* [16]. Redundancy and suppression occur when certain correlation patterns are present among multiple independent (**Z**) and dependent variables (**Y**).These patterns cause the magnitudes of the correlated attributes to be over- or under-predicted in the presence of these phenomena. Redundancy and suppression do not necessarily diminish the predictive ability of the RE-OL model, but they make interpretation of the model parameters difficult, which complicates the model validation process. A straightforward solution to address these issues is to use either height or seated height, but not both, in a model together with BMI and age, which are not highly correlated with height or seated height. Gender can be used in the modeling since it is only moderately correlated with height and seated height; however, it may cause some level of suppression or redundancy in the model and may not be easily interpreted.

8 CONCLUSION

Methods for the analysis of human appraisal experiments to understand and predict consumer preferences for new or existing product designs were developed in this work. The methods developed are for the purpose of preprocessing data, reduction of data, capturing respondent heterogeneity, and creating random effects ordered logit models for understanding consumer preferences. Latent Class Analysis is shown to be effective for combining several responses given by a consumer during an appraisal into a smaller number of latent classes related to their overall opinion of key product features. ANOVA analysis is used to understand the relative importance of the product and human attributes on the different rating responses provided in the survey. In these analyses, the respondent block effect, or unexplained respondent heterogeneity is found to be large. Cluster analysis of the block effect is used to indentify systematic ratings styles of the respondents, which explain a significant portion of the unexplained heterogeneity. Adding a new variable to control for rating style in the modeling process significantly reduces the unexplained heterogeneity. The use of smoothing spline regression is demonstrated to be an effective tool to understand the shape of the response-factor curve and guide the form of factors (i.e. linear, quadratic, cubic) to be introduced in the subsequent ordered logit modeling.

With data preprocessing, response reduction, and an understanding of respondent heterogeneity, random effects ordered logit models are estimated for each response. The importance of interactions and the benefits of explicitly modeling systematic heterogeneity and random heterogeneity are demonstrated in the ability of the distribution of the predicted ratings to match the actual distribution of ratings, an important feature of a model to be used to predict preferences for different populations and different designs. Machine learning methods from data mining are also applied to the PVM data. The decision tree provides additional insights into the relationship among the product factors, human factors, and rating responses not easily identified in the parametric ordered logit. The unsupervised Bayesian network provided insights into the relationships among the human factors not easily seen in methods such as correlation analysis. In summary, the methods employed in this work are important for analyzing data collected in human appraisal experiments and should be implemented as standard practice for analyzing and creating effective models of human preference. The models estimated to predict consumer preferences for occupant package attributes can be used in trade-off studies with structural models and exterior design appraisals to optimize the vehicle design.

ACKNOWLEDGEMENT

Grant support from National Science Foundation (CMMI—0700585) and the Ford URP (University Research Program) are greatly appreciated.

REFERENCES

[1] Hoyle, C., Chen, W., Ankenman, B. and Wang, N. Optimal Experimental Design of Human Appraisals for Modeling Consumer Preferences in Engineering Design. *Proceedings of the 2008 ASME IDETC/CIE*, New York, NY, August 3-5 2008.

[2] Hazelrigg, G.A. The Implications of Arrow's Impossibility Theorem on Approaches to Optimal Engineering Design. *Journal of Mechanical Design*, 1996, 118(2), 161-164.

[3] Wang, N., Kiridena, V., Gomez-Levi, G. and Wan, J. Design and Verification of a New Computer Controlled Seating Buck. *Proceedings of the 2006 ASME IDETC/CIE*, Philadelphia, Pennsylvania, September 10-13 2006.

[4] Keeney, R.L. and Raiffa, H. *Decisions with Multiple Objectives: Preferences and Value Tradeoffs*, 1993 (Cambridge University Press, New York).

[5] Box, G.E.P., Hunter, J.S. and Hunter, W.G. *Statistics for Experimenters: Design, Innovation, and Discovery*, 2005 (John Wiley & Son, New York).

[6] Montgomery, D.C. *Design and Analysis of Experiments*, 2005 (John Wiley and Sons, Inc., New York).

[7] McCutcheon, A.L. *Latent Class Analysis*, 1987 (Sage Publications, Beverly Hills, CA).

[8] Wood, S.N. mgcv: GAMs with GCV smoothness estimation and GAMMs by REML/PQL. (R Foundation for Statistical Computing, Vienna, Austria, 2004).

[9] Johnson, R.A. and Wichern, D.W. *Applied Multivariate Statistical Analysis*, 2002 (Prentice Hall, Upper Saddle River, NJ).

[10] Witten, I.H. and Frank, E. *Data Mining: Practical Machine Learning Tools and Techniques*, 2005 (Morgan Kaufmann, San Francisco, CA).

[11] McCullagh, P. Regression Models for Ordinal Data. *Journal of the Royal Statistical Society. Series B (Methodological)*, 1980, 42(2), 109-142.

[12] Hedeker, D. and Gibbons, R.D. A Random-Effects Ordinal Regression Model for Multilevel Analysis. *Biometrics*, 1994, 50(4), 933-944.

[13] Greenleaf, E.A. Improving Rating Scale Measures by Detecting and Correcting Bias Components in Some Response Styles. *Journal of Marketing Research*, 1992, 29(2), 176-188.

[14] Rossi, P.E., Gilula, Z. and Allenby, G.M. Overcoming Scale Usage Heterogeneity: A Bayesian Hierarchical Approach. *Journal of the American Statistical Association*, 2001, 96(453).

[15] Stevens, S.S. *Psychophysics: Introduction to Its Perceptual, Neural, and Social Prospects*, 1986 (Transaction Publishers.

[16] McClendon, M.K.J. *Multiple Regression and Causal Analysis*, 1994 (FE Peacock Publishers, Itasca).

Contact: Wei Chen
Northwestern University
Department of Mechanical Engineering
Evanston, IL 60208-3111
USA
Tel: +1 (847) 491-7019
Fax: +1 (847) 491-3915
E-mail: weichen@northwestern.edu

Wei Chen is Professor and Director of the Integrated Design Automation Laboratory in Mechanical Engineering at Northwestern University, Evanston, IL, USA. Her research focus is computational design methodology; specifically, design optimization, design under uncertainty, multidisciplinary design optimization, simulation-based design, and design theory & methodology.