Optimal Experimental Design of Human Appraisals for Modeling Consumer Preferences in Engineering Design

Christopher Hoyle
Ph.D. Candidate
e-mail: cj-hoyle@u.northwestern.edu

Wei Chen
Professor
e-mail: weichen@northwestern.edu

Bruce Ankenman
Associate Professor
e-mail: ankenman@iems.northwestern.edu

Northwestern University,
2145 Sheridan Rd.,
Evanston, IL 60208–3111

Nanxin Wang
Technical Leader
Ford Research and Advanced Engineering,
2101 Village Road,
Dearborn, MI 48124
e-mail: ankenman@iems.northwestern.edu

Human appraisals are becoming increasingly important in the design of engineering systems to link engineering design attributes to customer preferences. Human appraisals are used to assess consumers' opinions of a given product design, and are unique in that the experiment response is a function of both the product attributes and the respondents' human attributes. The design of a human appraisal is characterized as a split-plot design, in which the respondents' human attributes form the whole-plot factors while the product attributes form the split-plot factors. The experiments are also characterized by random block effects, in which the design configurations evaluated by a single respondent form a block. An experimental design algorithm is needed for human appraisal experiments because standard experimental designs often do not meet the needs of these experiments. In this work, an algorithmic approach to identify the optimal design for a human appraisal experiment is developed, which considers the effects of respondent fatigue and the blocked and split-plot structures of such a design. The developed algorithm seeks to identify the experimental design, which maximizes the determinant of the Fisher information matrix. The algorithm is derived assuming an ordered logit model will be used to model the rating responses. The advantages of this approach over competing approaches for minimizing the number of appraisal experiments and model-building efficiency are demonstrated using an automotive interior package human appraisal as an example. [DOI: 10.1115/1.3149845]

1 Introduction

Human appraisal experiments are used in a variety of contexts in product design to elicit consumer feedback on current or future product designs. The link between consumer preferences and engineering design has received much attention in literature recently [1–7]. Such design approaches have created the need for methods to assess human preferences for hypothetical or actual product designs to enable the desired linkage between consumer preferences and engineering design. In our previous work [7], a hierarchical choice modeling approach was developed in which a hierarchy of customer preference models is used to estimate consumer preferences for a given system design. Such an approach requires the collection of customer opinion for given system and subsystem designs. These designs are generally represented by prototype hardware for human appraisals, more recently by a highly flexible computer-controlled programmable hardware [8], which can assume a wide array of unique configurations for human evaluation. Complimentary developments in experimental design are needed to fully exploit such prototype hardware to estimate useful predictive models of customer preferences. The previous approaches to human appraisals in the design literature have generally assumed that the customer preference data are readily available, generally from a marketing source [7], or that a standard experiment design [3,4] (e.g., full factorial or fractional factorial) can be given to each respondent for the purpose of collecting the desired preference data. As will be presented in this work, the large number of factors and the experimental structure of a human appraisal for a complex system, such as an automobile, generally preclude the use of standard designs in such experiments. It will be shown in this work that in such cases, it is more efficient as well as necessary to provide each survey respondent with a different set of configurations.

A human appraisal is characterized by an interaction between the human respondent and the product design; therefore, the sets of factors, which influence the response from a given respondent for a given product *configuration*, are both *product* attributes, denoted by **A**, and respondent *human* attributes, denoted by **S**. Product attributes are characteristics of the product, such as its performance, appearance, features, and cost. Human attributes are defined as characteristics, primarily anthropomorphic characteristics such as stature or body mass index (BMI), of a respondent, which influence how the respondent experiences the system. In human appraisal experiments, the response for a given experiment could be the identification of a preferred configuration, or *choice*, from the configuration set, a *rank-ordering* of the configurations evaluated, or a *rating* for each configuration [9]. In this work, the response considered is in the form of a discrete *rating*, on a scale selected by the survey administrator. The number of rating categories should be limited to between 4 and 11 categories [10,11] (scales of 0–10, 1–5, and 1–7 are popular in application) to balance the competing desires of maximizing information recovery (i.e., maximize number of categories) versus minimizing scale usage heterogeneity (i.e., minimize number of categories). Rating responses represent an *ordinal* scale, in which higher ratings represent stronger positive preference for a given product configuration. The most popular models for estimating ratings as a function of independent variables are the ordered *probit* [12] and *ordered logit* [13] models. These models assume a respondent rating is a discrete realization of a continuous underlying opinion, or *utility*, for a given product configuration. In this work, the ordered logit model is used; however, the approach presented can easily be

Contributed by the Design Theory and Methodology Committee of ASME for publication in the JOURNAL OF MECHANICAL DESIGN. Manuscript received April 11, 2008; final manuscript received March 26, 2009; published online June 24, 2009. Review conducted by Janet K. Allen. Paper presented at the ASME 2008 Design Engineering Technical Conferences and Computers and Information in Engineering Conferences (DETC2008), Brooklyn, NY, August 3–July 6, 2008.

Downloaded From: http://mechanicaldesign.asmedigitalcollection.asme.org/ on 08/06/2014 Terms of Use: http://asme.org/terms

adapted to the ordered probit model (or other related models).

1.1 Issues Unique to Human Appraisal Experiments. In human appraisal experiments, a single respondent often evaluates several product configurations in sequence due to time and cost constraints. This implies that human appraisal experiments will naturally have a *random block effect*, as each person's ratings will have some level of correlation depending on the rating style of the respondent. A block is a set of experiments conducted under homogeneous but uncontrolled external conditions. Also, human appraisals are naturally *split-plot* designs [14], because it is unrealistic to completely randomize human attributes since a single respondent represents a set of fixed human attributes, and it is the most efficient to have a single respondent evaluate an entire block of experiments, or configurations, at a single time. Split-plot experiments are characterized by one or more factors remaining unchanged for a given set of experiments. In general, the goal of a human appraisal experiment is to create a response surface model, thus requiring a minimum of three levels of each product attribute (three levels cannot always be achieved for human attributes, which are categorical, such as gender). The desire to create a response surface is based on findings in psychometrics, in which it has been found that the human sensation magnitude to a given stimuli intensity follows a power law relationship [15]. A three level experiment enables approximation of the power law relationship using linear and quadratic terms in the prediction model (e.g., the ordered logit model).

A key issue to consider in human appraisal experiments is user fatigue [16]. Unlike computer or industrial experiments, fatigue will create additional error in the response in a human appraisal experiment. The number of trials or configurations, B, given to each respondent must be managed to ensure that the effects of fatigue are minimized. Another important issue in human appraisal experiments is the inclusion or exclusion of certain (experimental) design points of interest. The reason for specific inclusion or exclusion of design points is due to the interaction effects of certain factors, which may be theorized to be highly significant and important. If the interaction effect is achievable in the product, it would be of particular interest to study the impact of the interaction, whereas if the interaction is unachievable in the real product, it may be of interest to exclude such a combination. The design of experiments (DOE) with excluded combinations has been studied previously, e.g., Ref. [17], but the general case of inclusion or exclusion of certain design points has not been examined.

1.2 Comparison of Human Appraisal Experiments to Other Classes of Experiments. Human appraisal experiments can be differentiated from other types of experiments in literature. Industrial and scientific designs of experiments have been well documented [14,18] and utilized in practice. This class of experiments is characterized by random error in the response due to uncontrolled nuisance factors. While blocked and split-plot designs are used in this class of experiments, the reasons are typically due to nuisances or compromises in the experimental design, which introduce additional error or prevent full randomization, as opposed to being an integral feature of the design. Computer experiments have been studied extensively [19,20] for the purpose of metamodeling, and are characterized by a lack of random error, and thus methods of blocked or split-plot designs are not used. Conjoint experiments have been used for product or service evaluations in the marketing field [9,21], and are characterized by random error in the response and blocks corresponding to each respondent; however, they have not considered human attributes **S** in the design of the experiments but rather have treated the **S** as covariates (i.e., quantities recorded during the experiment but not used in the design of the experiment). Garneau and Parkinson [22] demonstrated that both systematic and random anthropomorphic heterogeneities are significant predictors of preferences for product designs in which the design interacts with the human (e.g., an exercise bicycle seat); however, a general approach for designing experiments for such human appraisals and methods to separate respondent level variation from random variation was not presented.

The human appraisal experiment is presented as a separate class of experiment in this work, specific to product evaluations in which the human attributes of the respondent have an *observable, systematic* influence on the response, in addition to the random effect captured by the random block effect as in a general conjoint analysis. Standard experimental designs and other experimental design approaches for human appraisals are generally not suitable for these experiments, which are conducted with the goal of creating a response surface model to understand respondent preferences as a function of product and human attributes. Standard split-plot designs based on standard full factorial or fractional factor designs for response surface creation, considering significant respondent blocking, do not exist [14]. Orthogonal array designs [23], such as the L_{18} design, are small enough such that each person can complete the entire experiment and blocking is not required; however, while such designs allow estimation of linear and quadratic terms, interactions cannot generally be estimated. Experiments specifically for human appraisals, with the goal of minimizing the number of configurations for each respondent to evaluate, have been developed for certain situations. Adaptive conjoint analysis [24] uses a prescreening of preferences for factor levels to optimize the configurations presented; however, this approach requires gaining access to resources for the prescreening tests and ignores the importance of factor interactions. One-factor-at-a-time experiments [25] have been developed to reduce the number of configurations needed when the goal of the experiment is to identify an optimal configuration. While this approach is effective for optimization, the goal of the human appraisal experiment in this work is to create a response surface model over a design space to understand response behavior. Based on the limitations of existing approaches, an approach using the D-optimality criterion is implemented as the method for selecting a human appraisal experiment.

An example of a human appraisal used throughout this work is the design of an automotive occupant package. A respondent's rating of a particular package configuration is dependent not only on the product attributes (**A**), such as the amount of head room, knee room, etc., in the package, but also the human attributes of the respondent (**S**), such as his/her stature, weight, gender, etc. Also, these experiments are characterized by a block effect because, after controlling for the respondents' human attributes, each respondent will retain a certain correlation among their ratings, which must be accounted for in the resulting model.

To summarize, the focus of this paper is the development of a design of experiments methodology for human appraisal experiments, considering the split-plot and block structures of these experiments, and the use of ordered logit (or probit) to estimate the subsequent response model. The developed methodology enables the number of configurations, B, provided to each respondent to be controlled and minimized, and will also allow certain factor combinations to be included or excluded. The remainder of the paper is organized as follows: Section 2 provides background for the DOE and modeling methodology, Sec. 3 presents the experimental design methodology for human appraisals, Sec. 4 discusses implementation of the methodology, Sec. 5 provides a case study, and Sec. 6 provides conclusions.

2 Design of Experiments and Modeling Methodologies

2.1 Blocked and Split-Plot Experimental Designs. Blocked and split-plot designs have been used extensively in physical experimentation. The difference between blocked and split-plot designs is illustrated in Fig. 1. The larger experimental unit (composed of many individual experimental design points or configurations) in a blocked experiment is called a *block*, whereas

Downloaded From: http://mechanicaldesign.asmedigitalcollection.asme.org/ on 08/06/2014 Terms of Use: http://asme.org/terms

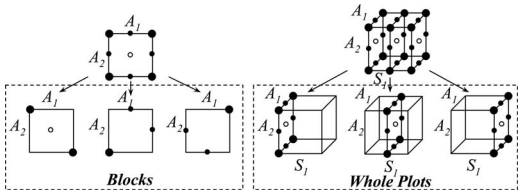

Fig. 1 The structure of blocked and split-plot experiments [28]

the larger experimental unit in a split-plot experiment is called a *whole plot*. Each block or whole plot consists of a number of experimental design factors, $\mathbf{x}=(A_1,A_2,\ldots,A_j)$, the values of which are determined by a design criterion, such as the *D*-optimality to be discussed in Sec. 2.2. The primary difference between a blocked versus split-plot design is that in a split-plot design, *whole-plot factors*, such as a human factor S_1, remain unchanged for a given experimental run. In blocked experiments, there are no corresponding larger experimental unit, or block level, factors such as the whole-plot factors. Therefore, the goal of a split-plot design is the selection of the design points *under* each whole-plot factor, whereas in a blocked design the goal is the allocation of design points to each block.

A demonstration experiment with two **A** and one **S**, which is presumably used to estimate a linear regression model, quadratic in **A** and linear in **S**, is used to demonstrate the terminology used in optimal DOE. In the proposed experimental design approach, both **A** and **S** comprise the experimental factor set **x** as follows:

$$\mathbf{x}=\begin{bmatrix}A_1 & A_2 & S_1\end{bmatrix} \tag{1}$$

The complete set of terms, **A** and **S**, which appear in the resulting prediction model (e.g., the ordered logit model), such as an intercept and linear, quadratic, and interaction terms, forms the extended design point, denoted by $\mathbf{f}(\mathbf{x})$ as follows:

$$\mathbf{f}(\mathbf{x})=\begin{bmatrix}1 & A_1 & A_2 & S_1 & A_1^2 & A_2^2 & A_1A_2 & S_1A_1 & S_1A_2\end{bmatrix} \tag{2}$$

The matrices of all extended design points in the complete experiment form the extended design matrix, denoted as **X**,

$$\mathbf{X}=\begin{bmatrix}1 & -1 & -1 & -1 & +1 & +1 & +1 & +1 & +1\\ 1 & +1 & 0 & -1 & +1 & 0 & 0 & -1 & 0\\ 1 & 0 & +1 & -1 & 0 & +1 & 0 & 0 & -1\\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots\end{bmatrix} \tag{3}$$

The motivation for split-plot design methodology is the inclusion of "hard-to-change" factors, e.g., a respondent's human attributes, in the experimental design. These hard-to-change factors are the whole-plot factors, which are not completely randomized as with the other design factors, and remain at a fixed level during the completion of a given whole-plot experiment. Alternatively, blocked experiments are motivated by the need to minimize the effects of known or theorized uncontrollable factors, such as the rating style of each respondent, not included as a design factor (i.e., **A** or **S**), but believed to have an influence on the experiment response. Therefore, the goal in blocked experiments is to distribute the experimental design points among homogeneous blocks, or respondents, to minimize the effects of uncontrollable factors.

2.2 Optimal Design of Experiments Methodology. To select experimental designs for human appraisals, given a constraint on the number of configurations rated by a single respondent (due to fatigue) and multiple product and human attributes, optimal design of experiment methods is adapted to the specific needs of this class of experiments. Optimal DOE has been studied for a variety of applications, such as industrial, agricultural, or scientific experiments (see Ref. [27] for a comprehensive treatment of the topic), and conjoint experiments [16,28]. The methodology has been extensively developed for ordinary least-squares (OLS) modeling [27] and has been extended recently to generalized least-squares (GLS) to account for the error variance structure in blocked or split-plot experiments [26]. Optimal DOE methodology has also been applied to multinomial logit (MNL) discrete choice analysis models [16,29,30], as well as general logistic regression, including ordered logit and ordered probit [31,32]; however, a general approach to account for the combined split-plot and block structures of the human appraisal experiments has not been presented and is therefore a focus of this work.

In optimal DOE, a *candidate set* of design points *G*, typically the design points of a full factorial experiment in the desired number of factors, is provided to an algorithm, which uses a defined *criterion* to select the optimal design points from the set to achieve a design of any arbitrary size, *M*. Various criteria for selecting the optimal experimental design are available, the most widely used being *D*-optimality. The *D*-optimality criterion selects the design, which minimizes the generalized variance of the model parameters, $\boldsymbol{\beta}$. Other design selection criteria can be utilized, but as will be discussed in Sec. 3, *D*-optimality is the most appropriate for the human appraisal experiments considered in this work. A key concept in optimal DOE is that the form of the model to be estimated, i.e., the form of the extended design point $\mathbf{f}(\mathbf{x})$, must be specified a priori to determine the optimal design, which supports the specified model.

2.3 Modeling of Rating Responses Using Ordered Logit. While experimental design methods to fit linear regression models are prevalent, to fit a predictive model to survey ratings, or *ordinal data* (e.g., 1=poor, 2=fair, and 3=good; rating from 1 to 10), alternative methods to linear regression are required. A key assumption of linear regression is violated when used to fit ordinal data because the expected model error cannot be assumed to be of zero mean with constant variance: The dependent variable (i.e., predicted rating) is not a linear function of the explanatory variables $\mathbf{f}(\mathbf{x})$ (defined in Eq. (2)). For this reason, *ordered logit* [13] is employed in this work to estimate models for ordinal customer ratings. The ordered logit approach assumes that the *P* ordered ratings, **R**, are discrete representations of a continuous, underlying *utility*, u_{ni}, associated with each design configuration, *i*, rated by each survey respondent, *n*. This underlying utility measure, u_{ni}, is the sum of a parametrized observable component, $\boldsymbol{\beta}\cdot\mathbf{f}(\mathbf{x})'$, and an unobserved error component ε_{ni} [7].

3 Optimal Experimental Design Method for Human Appraisals Using Rating Responses

In our proposed experimental design method, the human appraisal experiment is considered as both a split-plot and a blocked experiment. The human attributes **S** form the whole-plot factors because they represent hard-to-change factors. As discussed in the Introduction, a single respondent, characterized by a fixed human profile **S**, rates several configurations in succession due to the expense and inconvenience of requiring people to evaluate configurations randomly over time. Also, each whole-plot experiment may be too large for a single respondent to complete due to the fatigue issues discussed in Sec. 1.1. Each whole-plot may therefore be distributed among multiple survey respondents, each with the same **S**, in the form of blocks. The blocked split-plot design is illustrated in Fig. 2. In this diagram, the respondent human factors, **S**, are the whole-plot factors, and the product factors, **A**, are the split-plot factors.

In general, several criteria exist for selecting a preferred experimental design. The popular criteria in literature are *D*, *A*, *G*, and *V* (also known as *I*, *Q*, or *IV*) optimality, which are all functions of the Fisher information matrix, **M**, of the extended design matrix, **X**. The *D* and *A* criteria are related to making precise estimates of the model parameters ($\boldsymbol{\beta}$), whereas the *G* and *V* criteria are concerned with minimizing the overall prediction variance of the resulting model. While any optimality criterion can be used with the

Downloaded From: http://mechanicaldesign.asmedigitalcollection.asme.org/ on 08/06/2014 Terms of Use: http://asme.org/terms

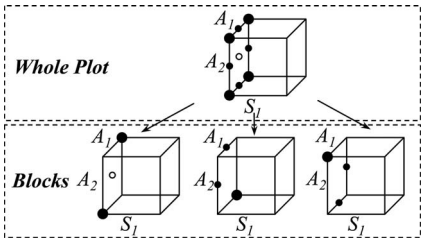

Fig. 2 The structure of the human appraisal blocked split-plot experiment [28]

approach presented in this work, the approach is presented using the *D*-optimality criterion for several reasons. First, *D*-optimality is widely used as an optimality criterion and is computationally inexpensive for experiment selection compared with some of the other criteria, such as *V*-optimality. Additionally, *D*-optimal designs have been shown to be highly efficient (i.e., provide efficient model building) with respect to the other optimality criteria (i.e., *G*, *A*, and *V*), whereas *G*-, *A*-, and *V*-optimal designs generally are not efficient with respect to *D*-optimality [26]. Also, because the models estimated must be validated in some manner, *D*-optimal designs provide precise estimates of the resulting model parameters ($\boldsymbol{\beta}$), which can be interpreted for expected sign and magnitude as part of the model validation process. *D*-optimality is achieved algorithmically through maximization of the determinant of the Fisher information matrix, $\mathbf{M}$, or the *D*-criterion, of a given experiment design

$$\max \ \det(\mathbf{M}) \tag{4}$$

The Fisher information matrix for the OLS fixed-effect model parameters, $\boldsymbol{\beta}$, can be expressed as [27]

$$\mathbf{M} = \sigma_\varepsilon^{-2}\mathbf{X}'\mathbf{X} \tag{5}$$

As seen in Eq. (5), $\mathbf{M}$ for an OLS model is a function of the extended design matrix, $\mathbf{X}$, and the random error variance, σ_ε (which, without loss of generality, can be assumed to be 1 for experiment optimization purposes), both of which are independent of the model parameters $\boldsymbol{\beta}$. In the case of GLS, Goos [26] derived the information matrix for the random block effect model, in which each experiment respondent forms a block. The variance-covariance matrix of the rating observations, $\mathbf{R}$, for a single respondent n, $\text{cov}(\mathbf{R}_n)$, is of the form

$$\mathbf{V}_n = \begin{bmatrix} (\sigma_\varepsilon^2 + \sigma_u^2) & \sigma_u^2 & \cdots & \sigma_u^2 \\ \sigma_u^2 & (\sigma_\varepsilon^2 + \sigma_u^2) & \cdots & \sigma_u^2 \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_u^2 & \sigma_u^2 & \cdots & (\sigma_\varepsilon^2 + \sigma_u^2) \end{bmatrix} \tag{6}$$

where σ_u is the variance at the *respondent* level, and σ_ε is the variance at the *observation* level. The information matrix for all observations can then be written as

$$\mathbf{M} = \mathbf{X}'\mathbf{V}^{-1}\mathbf{X} = \sigma_\varepsilon^{-2}\left\{\mathbf{X}'\mathbf{X} - \sum_{n=1}^{N} \frac{\rho}{1+\rho(B_n-1)}(\mathbf{X}_n' 1_{B_n})(\mathbf{X}_n' 1_{B_n})'\right\}$$

where

$$\rho = \frac{\sigma_u^2}{(\sigma_\varepsilon^2 + \sigma_u^2)} \tag{7}$$

B_n is the number of configurations in block n (of N blocks), and 1_{B_n} is a square matrix of ones of size B_n. In this case, an estimate of ρ, which is a measure of the ratio of *across*-respondent to *within*-respondent variance, is needed to select the optimal design. For this reason, such experimental designs are referred to as *semi-Bayesian* designs, since they require a prior estimation of ρ. The expression for $\mathbf{M}$ given in Eq. (7) is only valid if the model to be estimated is a (least-squares) linear regression model. It is therefore not valid for the human appraisal experiments in this work, which are to be modeled using ordered logit.

A complementary derivation is proposed in this work to support estimation of the ordered logit model. The ordered logit model can be written as

$$\Pr(R_{ni} = R_{nip}) = \pi_{nip}(\boldsymbol{\beta}) = F(k_p - \mathbf{x}_{ni}'\boldsymbol{\beta}) - F(k_{p-1} - \mathbf{x}_{ni}'\boldsymbol{\beta}) \tag{8}$$

where R_{ni} is the discrete rating for respondent, or block, n (of N blocks) and configuration i (of B configurations), k is an ordered logit cutpoint, p is a rating category (of P categories, such as 1–10), and F is the cumulative distribution function (CDF) of the logistic distribution (this CDF can be replaced with the standard normal CDF, $\boldsymbol{\Phi}$, if the ordered probit model is to be used).

To enable selection of a *D*-optimal design to support the ordered logit model, an expression for the information matrix (needed to calculate the *D*-criterion) that can be estimated without prior knowledge of the resulting model parameters, i.e., $\boldsymbol{\beta}$, is needed. In general, the information matrix for the ordered logit model can be expressed as [33]

$$\mathbf{M} = \sum_{n=1}^{N} \mathbf{D}_n'\mathbf{V}_n^{-1}\mathbf{D}_n \tag{9}$$

where $\mathbf{V}_n$ is the asymptotic variance-covariance matrix for block n. $\mathbf{D}_n$ is the derivative of $\boldsymbol{\pi}_n$ with respect to $\boldsymbol{\beta}$ as follows:

$$\mathbf{D}_n = \mathbf{D}_n(\boldsymbol{\beta}) = d\boldsymbol{\pi}_n(\boldsymbol{\beta})/d\boldsymbol{\beta} \tag{10}$$

where the $(P-1)$ vector of rating probabilities for a single individual n for configuration i is given as $\boldsymbol{\pi}_{ni} = (\pi_{ni1}, \pi_{ni2}, \ldots, \pi_{ni,P-1})$ and $\boldsymbol{\pi}_n = (\boldsymbol{\pi}_{n1}, \boldsymbol{\pi}_{n2}, \ldots, \boldsymbol{\pi}_{nB})'$. The asymptotic variance-covariance matrix, $\mathbf{V}_n$, for the ordinal model, such as ordered logit, can be written in block-matrix form as [34]

$$\mathbf{V}_n = \begin{bmatrix} \mathbf{V}_{11} & \mathbf{V}_{12} & \cdots & \mathbf{V}_{1T} \\ \mathbf{V}_{21} & \mathbf{V}_{22} & \cdots & \mathbf{V}_{2T} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{V}_{T1} & \mathbf{V}_{T2} & \cdots & \mathbf{V}_{TT} \end{bmatrix} \tag{11}$$

where the on-diagonal matrices are *multinomial* covariance matrices, $\mathbf{V}_{tt} = \text{diag}(\boldsymbol{\pi}_{ni}) - \boldsymbol{\pi}_{ni}\boldsymbol{\pi}_{ni}'$, and the off-diagonal matrices, $\mathbf{V}_{ts}(t \neq s)$, are the *within-block* covariance matrices between any two responses in a block. These matrices are generally calculated as part of the model estimation process using collected data; therefore, a method for estimating them for experimental design purposes must be devised.

The on-diagonal multinomial covariance matrices ($\mathbf{V}_{tt}$) can be calculated from knowledge of the rating probabilities; however, the within-block covariance matrix ($\mathbf{V}_{ts}$) requires additional derivation. In general, the $\mathbf{V}_{ts}$ matrix follows the form [33]

$$\mathbf{V}_{ts} = (\mathbf{B}^{1/2})'\mathbf{P}_{tn}(\mathbf{B}^{1/2}) \tag{12}$$

where $\mathbf{P}$ is the "working" correlation matrix, and $\mathbf{B}$ is a matrix determined by the correlation structure. The selection of $\mathbf{B}$ and $\mathbf{P}$ depends on the form of the model to be estimated with the experimental response data [35,36]. The proper specification for $\mathbf{P}_{tn}$ for the random-effect ordered logit model has been found to be the "exchangeable" structure. In the exchangeable structure, $\mathbf{P}_{tn}$ is a diagonal matrix with all diagonal elements of $\mathbf{P}_{tn} = \alpha$, implying equal correlation among all observations in a given block. In this formulation, α is the pairwise correlation coefficient between elements in the $\mathbf{V}_{tt}$ matrices, similar to the correlation coefficient ρ applicable for the scalar variance-covariance matrix of Eq. (7). The recommended specification for $\mathbf{B}$ for the random-effect model is $\mathbf{V}_{tt}$ [35]. Therefore $\mathbf{V}_{ts}$ can be written as

$$\mathbf{V}_{ts} = (\mathbf{V}_{tt}^{1/2})'\text{diag}(\alpha)(\mathbf{V}_{tt}^{1/2}), \quad t \neq s \tag{13}$$

Downloaded From: http://mechanicaldesign.asmedigitalcollection.asme.org/ on 08/06/2014 Terms of Use: http://asme.org/terms

In viewing Eqs. (10), (11), and (13), it can be seen that in order to calculate $\mathbf{M}$, estimates for $\boldsymbol{\pi}_n$ and α are required. The pairwise correlation coefficient α is not reported in the random-effect ordered logit modeling process, which provides a challenge to determining a reasonable estimate for α from previous experiments or literature. However, the coefficient ρ is reported in the modeling process, and it has been found that α can be estimated using ρ by the relation $\alpha \approx \rho / P$ to enable calculation of $\mathbf{V}_{ts}$. This estimate is based on the assumption that ρ should be "distributed" over the P rating categories in the working correlation matrix, such that the influence of α and ρ is equivalent in the respective information matrix calculations of Eqs. (7) and (9). Because a rating prediction model is not available before the experiment is conducted, the rating category (e.g., 1–10) probabilities, $\boldsymbol{\pi}_n$, must be estimated directly. They can be estimated from prior knowledge from a previous experiment, or if no prior knowledge is available, an equal probability of each rating category can be assumed. Because estimates of the entire response probability vectors, $\boldsymbol{\pi}_{ni}$, are needed to calculate $\mathbf{V}_n$ and $\mathbf{D}_n$ to compute $\mathbf{M}$, such experimental designs are referred to as *Bayesian* designs [27].

To verify the formulation of $\mathbf{M}$ for the ordered logit model and the estimates for $\boldsymbol{\pi}_n$ and α, two test data sets with equal probability of each rating (i.e., ratings 1–10) are created. In one data set, the average correlation ρ of ratings from a single respondent is set to 0 (data set 1) and in the second data set, the ratings were distributed such that the average rating correlation, ρ, is 0.40 (data set 2). The purpose of this verification is to ensure that the proposed calculation of the information matrix (Eq. (9)), in which the rating probabilities are estimated a priori and the correlation of responses is estimated using α, is consistent with the information matrix calculated from actual data. Ordered logit models are estimated using both data sets in the statistical modeling software STATA™ [37]. The information matrices calculated by STATA (labeled *stat*) are compared with the information matrices calculated using the proposed derivation using estimates for $\boldsymbol{\pi}_n$ and α (labeled *der*). For data set 1, the information matrices calculated by STATA and Eq. (9) are identical, and the determinants of $\mathbf{M}$ identical ($\det_{stat} = \det_{der} = 1.16 \times 10^{20}$). For data set 2, the difference in the determinants is 7.62% ($\det_{stat} = 5.15 \times 10^{17}$, $\det_{der} = 5.54 \times 10^{17}$), most likely because only the average correlation could be controlled in the created data set and α is approximated as described previously. A study of the sensitivity of the algorithm to misspecification of ρ has been investigated for the GLS algorithm by Goos [26]. He found that a misspecification of $\pm 50\%$ results in only a 4–8% error in the information matrix. In a further study [28], it was found that the actual experiment design selection was robust to larger misspecifications of ρ (range 0.1–0.9), indicating that an exact estimate of ρ is not needed for design selection purposes.

The challenges of optimal experimental design for a random-effect ordered logit model can be understood through a comparison to the generalized least-squares approach presented at the beginning of this section. In the least-squares approach, the on-diagonal terms of the $\mathbf{V}_n$ matrix in Eq. (6) are scalars of estimated within-block and across-block variances, whereas in the ordered logit approach, the on-diagonal terms of the $\mathbf{V}_n$ matrix of Eq. (11) are matrices that are a function of estimated response probabilities. Comparing the off-diagonal terms of Eqs. (6) and (11) indicates that the least-squares method requires only a scalar estimate of the across-block variance whereas the ordered logit requires estimation of a matrix (i.e., $\mathbf{V}_{ts}$). This comparison indicates the difficulties in design optimization for ordinal data models in that the computation is more expensive due to the replacement of scalar quantities with matrices, and that estimates of both $\boldsymbol{\pi}_n$ and α are required.

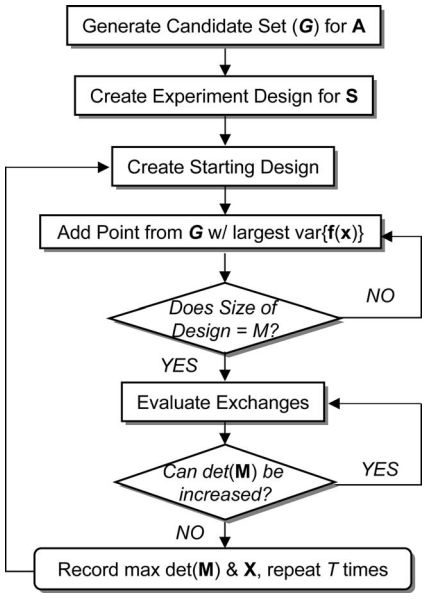

Fig. 3 Algorithmic implementation of the optimal experimental design method

4 Algorithmic Implementation

The algorithmic implementation for selecting the optimal blocked split-plot design follows the approach provided in Ref. [38], with the least-squares information matrix of Eq. (7) used in their approach replaced with that of Eq. (9) for the new approach. In general, the experimental design is built sequentially, with points from the candidate set (G) having the highest prediction variance added to the experiment to maximize the D-criterion. An overview of the algorithm is shown in Fig. 3 and described as follows:

1. Generate a set of candidate points, G, for the product attributes, $\mathbf{A}$, from which to select the optimal set. G is typically the points of a full factorial experiment in the number of factors desired. Specific factor combinations to be specifically excluded from the candidate set, or specifically included in the final experiment design, are also specified.
2. Create an experimental design for the desired human whole-plot factors, $\mathbf{S}$. This design can be a full or fractional factorial in human attributes, depending on the size of $\mathbf{S}$ and the number of respondents. Randomly assign the whole-plot factors to each block, n.
3. Create a starting design. To begin building the experimental design, a starting design is composed of a randomly selected small number of points from the candidate set and randomly assigned to the blocks. Compute the initial information matrix, $\mathbf{M}$, and the determinant, $\det(\mathbf{M})$.
4. Determine the point in the candidate set G with the largest prediction variance, $\text{var}\{\mathbf{f}(\mathbf{x})\}$, given by [39]

$$\text{var}\{f(\mathbf{x})\} \approx \sum_{p=1}^{P} \left(\frac{d\pi_p(\boldsymbol{\beta})}{d\boldsymbol{\beta}} \right)' \mathbf{M}^{-1} \left(\frac{d\pi_p(\boldsymbol{\beta})}{d\boldsymbol{\beta}} \right)$$

   Randomly assign this point to a block, and update $\mathbf{M}$ and $\det(\mathbf{M})$. Repeat this process until each block n receives B configurations, forming an experiment design of size M.
5. Evaluate exchanges. Since the design was started with a random selection of points, there may be points in the candidate set G, which will increase the D-criterion. Each point in the current design is evaluated to determine if its replacement

Downloaded From: http://mechanicaldesign.asmedigitalcollection.asme.org/ on 08/06/2014 Terms of Use: http://asme.org/terms

Fig. 4 PVM used for conducting human appraisal experiments [8]

with a point in the candidate set will increase the *D*-criterion. This is continued until no further increases can be established.

6. Record the *D*-criterion and repeat steps 3–5. Steps 3–5 constitute a single *try*, each with a local maximum for *D*-optimality based on the starting design of step 3. *T* tries (e.g., 100 tries) can be conducted to search for the global maximum.

5 Automotive Occupant Packaging Case Study

A case study using an automotive occupant packaging human appraisal is used to demonstrate the methodology, as well as the advantages of using the blocked split-plot experimental design methodology for human appraisal. The occupant packaging appraisal is performed on a programmable vehicle model (PVM) as shown in Fig. 4, which is capable of creating a wide range of parametric representations of an occupant package through a computer-controlled interface.

5.1 Design of Experiments. A human appraisal experiment has been previously conducted by Ford Motor Co., Dearborn, MI using the PVM to evaluate occupant package design specifically for headroom. In the experiment conducted, headroom design is characterized by three dimensions as defined by the Society of Automotive Engineers (SAE) J1100 [40]: L38 (frontal), W35 (lateral), and H61 (vertical). These three product factors ($\mathbf{x}^* = [A_1, A_2, A_3]$) were used to create a full $3^2 4^1$ factorial experiment (i.e., 36 trials), which was given to each of 100 human appraisal respondents, for a total of 3600 rating responses. The responses were given on a (discrete) scale of 2–10, with 10 representing highest satisfaction with the headroom, and 2 representing the least satisfaction, leading to $P=9$. Human profile ($\mathbf{S}$) factors were not used in the design of the experiment; however, the $\mathbf{S}$ were treated as *covariates* in that the human profile of each person was recorded, but no attempts were made to control the profiles of the respondents in the experimental design process. The data set with rating responses was used to create a full quadratic response surface model, used to predict a customer headroom rating for a given occupant package design and a given target market human. This data set is referred to as data set Full in the case study. Conducting an experiment of this size was very time consuming and costly for Ford, and methods to conduct more efficient experiments are needed. Using this example in which data have already been collected, we will demonstrate that the experimental design methodology presented in this paper allows selection of an experimental design, which can be used to estimate a comparable model with significantly fewer experimental design points than used in the Full data set. In the new methodology, the $3^2 4^1$ factorial experiment forms the candidate set for the optimization algorithm. Additionally, a set of potentially significant human attributes, $\mathbf{S}$, is included in the design of the experiment as whole-plot factors. The human profile attributes included are respondent gender (Gen) and stature (Stat). An issue to address in the experimental design of $\mathbf{S}$ is that exact levels cannot be practically achieved for all $\mathbf{S}$ (e.g., stature) in a real human appraisal design. In this case, human attribute *ranges* are assigned to a level in the design of an experiment, for example, statures between 54 in. and 57 in. are coded as the −1 level and those between 73 in. and 76 in. are coded as the +1 level. These human attribute "bins" are needed to ensure that the proper respondents are selected for the experiment; however, the actual human measurements (e.g., stature, weight, and age) are to be used in the model estimation process. A criterion for selecting the bins is to ensure that 5% and 95% human-measurement respondents of the target population are included in the bins. If more levels (i.e., bins) can be afforded, respondents closer to the human mean (e.g., 50%) should be included; however, it is most important from a *D*-optimality perspective to include 5% and 95% respondents. At the time of the experiment, additional human and socio-economic attributes of a respondent can be recorded and treated as covariates in the modeling process.

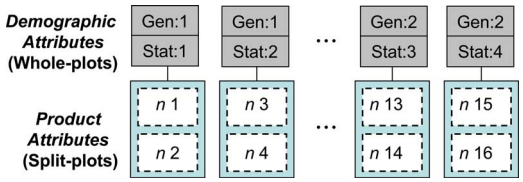

Fig. 5 Occupant package blocked split-plot human appraisal experiment

To demonstrate the ability of the new method to manage the size of an experiment, the number of configurations given to each respondent is reduced from 36 to a block size of 18. The whole-plot experiment design is composed of two levels of gender (i.e., male and female) and four levels of stature (using stature ranges), leading to a $2^1 4^1$ whole-plot experiment design. Two respondents (i.e., blocks) will be assigned to each whole plot for a total of 16 respondents (or blocks, *n*), leading to a total of $M=288$ total trials, versus 3600 in the Full experiment described above. A summary of the experimental design is shown in Fig. 5.

The exact form of the model to be estimated is known for this case study from previous work, enabling specification of model form $\mathbf{f}(\mathbf{x})$ as defined in Eq. (2). The model form contains full quadratic terms for $\mathbf{A}$ (linear, squared, interaction) and linear terms for $\mathbf{S}$ (no $\mathbf{S}\cdot\mathbf{A}$ interactions). With $\mathbf{f}(\mathbf{x})$ specified, the algorithm can be used to select the best 18 configurations to give to each of the 16 respondents. As discussed in Sec. 3 a *prior* rating probability estimate is needed to calculate $\mathbf{M}$. For this study, it was assumed that the probability, π_{nip}, of each rating R_p for each respondent *n* and each configuration *i* is equally probable, i.e., $\pi_{nip}=1/9=0.11$. Also, it was known from a previous experiment that the correlation among ratings of a single respondent is $\rho=0.3$. The use of equal rating probabilities assumes that there is *no* prior information about the rating responses. If prior information is available (e.g., middle ratings are more likely than extreme ratings) such information can be incorporated to improve the experiment design. In this experiment, the best experiment as selected by the algorithm presents each respondent with a *different* set of configurations, demonstrating that the use of the same 18 point fractional factorial experiment (of the original $3^2 4^1=36$ experiment) for each respondent would not be optimal for a human appraisal experiment. The data set with observations based on this design is labeled *D*-Opt. For comparison, an additional set of experimental designs is created. In these designs, 16 respondents are randomly selected from the original 100 respondents and 18 observations are randomly selected from the 36 total observations for each respondent. A total of 100 such random experiments are created, such that experimental design comparisons are made to the *mean* random experimental design, to ensure that any com-

Downloaded From: http://mechanicaldesign.asmedigitalcollection.asme.org/ on 08/06/2014 Terms of Use: http://asme.org/terms

Table 1 Summary of experiment and model statistics

	No. of experiments	*D*-efficiency	Model fit, ρ^2	Prediction error (%)
Full	3600	—	0.373	2.80
D -Opt	288	79.7%	0.485	6.90
Rand	288	35.8% ± 1.6% [a]	0.375	14.60

[a]The mean and ±1 standard deviation are shown.

parisons are made based on a typical random experiment and not on an outlying design. This set of experimental designs is labeled Rand.

5.2 Results of Model Estimation. With the three experimental designs established, a random-effect ordered logit model was estimated using each of the three data sets. A summary of the experimental efficiency as measured by *D*-efficiency, model fit as measured by ρ^2 [41], and average rating prediction error [42] are shown in Table 1. In the case of the Rand experiment designs, the model was fitted using an experiment design with a mean *D*-efficiency.

D-efficiency is a measure of the relative efficiency of an experiment versus a base experiment, for example, the Full experiment in this work. As seen in the table, the *D*-efficiency of the *D*-Opt experiment is high, ensuring low variance estimates of the model parameters, whereas the mean *D*-efficiency of the Rand experiment is quite low and will result in poor model parameter estimates. The ρ^2 statistic varies between 0 and 1 and is a function of the log-likelihood of the estimated model, with higher ρ^2 indicating a better "model fit." The ρ^2 for the *D*-Opt model is significantly higher than that of the Full model. The explanation for this can be provided by reviewing the assumptions of ordered logit modeling and the nature of ratings. Ratings tend to have higher variance in the middle ratings versus those at the extremes [12]. *D*-optimality tends to bias toward including those configurations with the most extreme settings. Thus by selecting the *D*-optimal configurations from the full PVM data set, a more efficient estimation of the model $\boldsymbol{\beta}$ parameters, and hence utility, is accomplished for the assumed model. The fit of this mean Rand model is similar to the Full model, which is consistent with the fact that the points were randomly selected, so similar model fits would be expected. The prediction error is the rating misclassification error when using the three models to estimate ratings in the full 3600 observation data set. The effects of the prediction error on the resulting rating predictions can be seen graphically in Fig. 6. As shown, the prediction error of the mean Rand model is significantly higher than the other two models.

The estimated model parameters, $\boldsymbol{\beta}$, for the utility function are shown in Table 2, along with the standard errors of the parameters (the cutpoints, **k**, are not shown since these estimates are similar for all three models). We can compare model attributes, such as the relative magnitudes and signs of parameters and the general

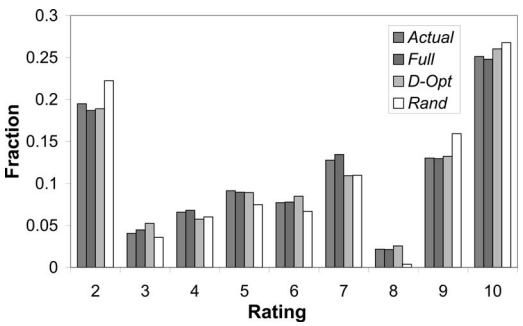

Fig. 6 Comparison of rating predictions to actual ratings

Table 2 Summary of headroom rating model parameters

	Full model		*D*-Opt model		Rand model	
	Coef	Std. err.	Coef.	Std. err.	Coef.	Std. err.
L38	2.61	0.368	4.50	0.731*	2.49	1.449
W35	2.03	0.359	2.11	0.970	2.83	1.376*
H61	12.09	0.491	13.01	2.165**	10.61	1.838*
$L38^2$	−0.74	0.292	−0.76	0.852	−0.23	1.111**
$W35^2$	−1.23	0.291	−1.08	1.562	−2.14	1.104*
$H61^2$	−2.55	0.354	−2.40	1.693	−0.89	1.325*
L38*W35	0.19	0.211	−0.16	0.820	0.13	0.826
L38*H61	−0.32	0.270	−0.16	0.949	−1.15	1.093*
W35*H61	0.49	0.261	0.20	0.857	0.85	1.010
Gender	−0.78	0.494	−0.56	0.726	0.14	1.115**
Stature	−2.24	1.008	−1.81	1.425	−0.94	2.763
Resp. σ^2_u	2.95	0.452	1.73	0.780	2.57	1.071

interpretation of the models, in addition to the model statistics. Considering the model estimated on the Full data set to be the baseline, it is seen that the model estimated using the *D*-Opt data set is close in interpretation. The signs of the parameters agree (except for the insignificant L38*W35 interaction). The ranking of parameter importance as measured by the parameter magnitudes is the same in both models. Vertical headroom clearance (H61) is found to be the most important dimension influencing a respondent's perception of headroom. The next most important dimension is frontal headroom clearance (L38), followed by lateral headroom clearance (W35). The human attributes indicate that taller respondents and female respondents (gender is a dummy variable: 0=male, 1=female) systematically respond with lower headroom ratings (on average) than shorter and male respondents, respectively. The ratio of parameters (e.g., W35/H61) is similar in both models, with the exception of L38, which is more important in the *D*-Opt model. The reason for this could be explained by the improved model fit statistic, ρ^2, of the *D*-Opt model as described previously.

The model parameters in the *D*-Opt and Rand models are compared with those in the Full model using a *t*-test, in which the null hypothesis is that the model parameters are not different. The model parameters in which the null hypothesis can be rejected with 95% confidence are marked with *, whereas those rejected with 90% confidence are marked with ** in Table 2. As seen in the table, the Rand model contains significantly more parameters, which differ from the Full model than the *D*-Opt model. Such results are expected due to the lower *D*-efficiency of the Rand experiment, which results in less precise estimates of the model parameters than the higher efficiency *D*-Opt model.

While the *D*-optimization algorithm has been shown to be effective for this example, its true utility is in experiments with large numbers of product attribute factors (e.g., 6–9) and several human attributes. In such a case, the candidate set will be several hundred to several thousands of potential points, and the task of choosing the appropriate set of points for each respondent is not as straightforward as in the previous example. To demonstrate, an experiment designed for the PVM to elicit preferences for the roominess and ingress/egress of the vehicle occupant package is used. In this simplified experiment, eight product factors are examined by eight respondents, and it is desired to estimate all linear, quadratic, and all 2-factor $\mathbf{A}\cdot\mathbf{A}$ and $\mathbf{A}\cdot\mathbf{S}$ interactions. Respondents are selected based on three human factors at two levels (a 2^3 full factorial human experiment). The experiment design for the product attributes is conducted by selecting 18 points from a 3^8 full factorial (i.e., C^{18}_{6561}) for each respondent. In this example, the

Downloaded From: http://mechanicaldesign.asmedigitalcollection.asme.org/ on 08/06/2014 Terms of Use: http://asme.org/terms

Table 3 Comparison of three-factor to eight-factor human appraisal experiment

Product factors	3 factors, 3 levels		8 factors, 3 levels	
Human factors	$2^1 4^1$		2^3	
	Mean	St. dev	Mean	St. dev
D-optimal Exp.	2.24×10^{59}		8.10×10^{140}	
Random Exp.	6.17×10^{52}	5.3×10^{52}	9.77×10^{105}	1.1×10^{107}
D-efficiency random	45.0%	2.1%	29.4%	4.2%

D-optimal experimental design is found with the algorithm, and 100 randomly selected experimental designs are also generated for comparison as in the previous example. In this comparison, the *D*-optimal experiment is the baseline for the efficiency comparison, since comparison to an experiment in which each respondent receives the 3^8 full factorial, product factor experiment (i.e., 6561 configurations) is not a realistic baseline. In Table 3, the mean *D*-efficiency of the 8-factor random experiments in this example is compared with the mean *D*-efficiency of the random 3-factor experiments of the previous example. As shown, the efficiency of the random 3-factor experiment has a mean *D*-efficiency of 45.0%, whereas the random 8-factor experiment has a mean *D*-efficiency of 29.4%. The variance of the random 8-factor experiment is higher than the 3-factor experiment as would be expected in selecting 18 points from 6561 (C_{6561}^{18}) versus 36 (C_{36}^{18}) points for each respondent. As shown previously in Table 2, reduced *D*-efficiency results in reduced precision in estimating model parameters.

6 Conclusion

An algorithmic approach for selection of the optimal design of experiments for human appraisal experiments has been developed, demonstrated, and validated in this paper. An algorithmic approach is necessary for human appraisals due to the large number of potential design and human attributes, coupled with issues of respondent fatigue in such experiments. Human appraisal experiments have been shown to be unique in that the experiment response is a function of both the product attributes and the human attributes of the respondent. They are characterized as split-plot designs, in which the respondent human attributes form the hard-to-change whole-plot factors while the product attributes form the split-plot factors. The experiments are also characterized by random block effects, in which the configurations evaluated by a single respondent form a block. The experimental design algorithm presented seeks to identify the experimental design, which maximizes the determinant of the Fisher information matrix, or *D*-criterion, of a given design, assuming that the model to be estimated is an ordered logit model.

The case study and subsequent discussion demonstrate many of the key features of the optimization algorithm. Most importantly, it was shown that the algorithm allows efficient model estimation with a minimal number of experiment points. For the vehicle headroom appraisal, previous methods had used 3600 experiment points, while a comparable model was estimated using 288 experiment points selected using the proposed algorithm. Also, it was shown that randomly selecting 288 points from the full 3600 point experiment produces an inferior model, and the utility of the algorithm increases as the number of experiment factors increases. The optimization algorithm distributes a different set of experiment points to each respondent, demonstrating that using a standard fractional factorial to reduce the number of trials per person is not the best alterative for human appraisals. The algorithm can also be applied for experiments, which do not use a factorial experiment as the candidate set, such as a vehicle survey, which uses current automobiles in the market. The algorithm can select the vehicles to use in the survey, e.g., the set of midsize sedans, which will result in the best subsequent model estimation from the survey data. In summary, this generic methodology can be used in the design of many types of human appraisal experiments. Improvements to the efficiency of the search algorithm will be investigated using memetic or stochastic evolutionary algorithms.

Acknowledgment

Grant support from the National Science Foundation (Grant No. CMMI—0700585) and the Ford URP (University Research Program) are greatly appreciated.

References

[1] Li, H., and Azarm, S., 2000, "Product Design Selection Under Uncertainty and With Competitive Advantage," ASME J. Mech. Des., **122**(4), pp. 411–418.
[2] Petiot, J. F., and Yannou, B., 2004, "Measuring Consumer Perceptions for a Better Comprehension, Specification and Assessment of Product Semantics," Int. J. Ind. Ergonom., **33**(6), pp. 507–525.
[3] Michalek, J. J., Feinberg, F. M., and Papalambros, P. Y., 2005, "Linking Marketing and Engineering Product Design Decisions Via Analytical Target Cascading," J. Prod. Innovation Manage., **22**(1), pp. 42–62.
[4] MacDonald, E., Gonzalez, R., and Papalambros, P. Y., 2007, "Preference Inconsistency in Multidisciplinary Design Decision Making," Proceedings of the 2007 International Design Engineering Technical Conferences, Las Vegas, NV, Sept. 4–8.
[5] Besharati, B., Azarm, S., and Farhang-Mehr, A., 2002, "A Customer-Based Expected Utility Metric for Product Design Selection," Proceedings of ASME 2002 IDETC Conference, Montreal, Canada, Sept. 29–Oct. 2.
[6] Wassenaar, H. J., and Chen, W., 2003, "An Approach to Decision-Based Design With Discrete Choice Analysis for Demand Modeling," ASME J. Mech. Des., **125**(3), pp. 490–497.
[7] Kumar, D., Hoyle, C., Chen, W., Wang, N., Gomez-Levi, G., and Koppelman, F., 2009, "Incorporating Customer Preferences and Market Trends in Vehicle Package Design," Int. J. Prod. Dev., **8**(3), pp. 228–251.
[8] Wang, N., Kiridena, V., Gomez-Levi, G., and Wan, J., 2006, "Design and Verification of a New Computer Controlled Seating Buck," *Proceedings of the 2006 ASME IDETC/CIE*, Philadelphia, PA, Sept. 10–13.
[9] Louviere, J. J., Hensher, D. A., and Swait, J. D., 2000, *Stated Choice Methods: Analysis and Application*, Cambridge University Press, New York.
[10] Cox, E. P., III, 1980, "The Optimal Number of Response Alternatives for a Scale: A Review," J. Mark. Res., **17**(4), pp. 407–422.
[11] Green, P. E., and Rao, V. R., 1970, "Rating Scales and Information Recovery. How Many Scales and Response Categories to Use?," J. Marketing, **34**(3), pp. 33–39.
[12] McKelvey, R. D., and Zavoina, W., 1975, "A Statistical Model for the Analysis of Ordinal Level Dependent Variables," J. Math. Sociol., **4**(1), pp. 103–120.
[13] McCullagh, P., 1980, "Regression Models for Ordinal Data," J. R. Stat. Soc. Ser. B (Methodol.), **42**(2), pp. 109–142.
[14] Box, G. E. P., Hunter, J. S., and Hunter, W. G., 2005, *Statistics for Experimenters: Design, Innovation, and Discovery*, Wiley, New York.
[15] Stevens, S. S., 1986, *Psychophysics: Introduction to Its Perceptual, Neural, and Social Prospects*, Transaction Publishers, New Brunswick, NJ.
[16] Kuhfeld, W. F., Tobias, R. D., and Garratt, M., 1994, "Efficient Experimental Design With Marketing Research Applications," J. Mark. Res., **31**(4), pp. 545–557.
[17] Steckel, J. H., DeSarbo, W. S., and Mahajan, V., 1991, "On the Creation of Acceptable Conjoint Analysis Experimental Designs," Decision Sci., **22**(2), pp. 435–442.
[18] Montgomery, D. C., 2005, *Design and Analysis of Experiments*, Wiley, New York.
[19] Simpson, T. W., Poplinski, J. D., Koch, P. N., and Allen, J. K., 2001, "Metamodels for Computer-Based Engineering Design: Survey and Recommendations," Eng. Comput., **17**(2), pp. 129–150.
[20] Jin, R., Chen, W., and Simpson, T. W., 2001, "Comparative Studies of Metamodeling Techniques Under Multiple Modeling Criteria," Struct. Multidiscip. Optim., **23**(1), pp. 1–13.
[21] Green, P. E., and Srinivasan, V., 1978, "Conjoint Analysis in Consumer Research: Issues and Outlook," J. Consum. Res., **5**(2), pp. 103–123.
[22] Garneau, C. J., and Parkinson, M. B., 2007, "Including Preference in Anthropometry-Driven Models for Design," 2007 ASME Design Engineering Technical Conference (DETC), Las Vegas, NV.
[23] Phadke, M. S., 1995, *Quality Engineering Using Robust Design*, Prentice-Hall, Upper Saddle River, NJ.
[24] Green, P. E., Krieger, A. M., and Agarwal, M. K., 1991, "Adaptive Conjoint Analysis: Some Caveats and Suggestions," J. Mark. Res., **28**(2), pp. 215–222.
[25] Frey, D. D., Engelhardt, F., and Greitzer, E. M., 2003, ""A Role for One-Factor-at-a-Time" Experimentation in Parameter Design," Res. Eng. Des., **14**(2), pp. 65–74.
[26] Atkinson, A. C., and Donev, A. N., 1992, *Optimum Experimental Designs*, Oxford University Press, Oxford.
[27] Kessels, R., Goos, P., and Vandebroek, M., 2008, "Optimal Designs for Con-

Downloaded From: http://mechanicaldesign.asmedigitalcollection.asme.org/ on 08/06/2014 Terms of Use: http://asme.org/terms

joint Experiments," Computational Statistics & Data Analysis, **52**(5), pp. 2369–2387.
[28] Goos, P., 2002, *The Optimal Design of Blocked and Split-Plot Experiments*, Springer-Verlag, New York.
[29] Sandor, Z., and Wedel, M., 2001, "Designing Conjoint Choice Experiments Using Managers' Prior Beliefs," J. Mark. Res., **38**(4), pp. 430–444.
[30] Kessels, R., Goos, P., and Vandebroek, M., 2006, "A Comparison of Criteria to Design Efficient Choice Experiments," J. Mark. Res., **43**(3), pp. 409–419.
[31] Zocchi, S. S., and Atkinson, A. C., 1999, "Optimum Experimental Designs for Multinomial Logistic Models," Biometrics, **55**(2), pp. 437–444.
[32] Chipman, H. A., and Welch, W. J., 1996, "D-Optimal Design for Generalized Linear Models," unpublished.
[33] Liang, K., and Zeger, S. L., 1986, "Longitudinal Data Analysis Using Generalized Linear Models," Biometrika, **73**(1), pp. 13–22.
[34] Williamson, J. M., Kim, K., and Lipsitz, S. R., 1995, "Analyzing Bivariate Ordinal Data Using a Global Odds Ratio," J. Am. Stat. Assoc., **90**(432), pp. 1432–1437.
[35] Hines, R., 1998, "Comparison of Two Covariance Structures in the Analysis of Clustered Polytomous Data Using Generalized Estimating Equations," Biometrics, **54**(1), pp. 312–316.
[36] Zorn, C. J. W., 2001, "Generalized Estimating Equation Models for Correlated Data: A Review With Applications," Am. J. Pol. Sci., **45**(2), pp. 470–490.
[37] Stata Corporation, 2007, STATASE 9.2, College Station, TX.
[38] Goos, P., and Vandebroek, M., 2004, "Outperforming Completely Randomized Designs," J. Quality Technol., **36**(1), pp. 12–26.
[39] Tamhane, A. C., and Dunlop, D. D., 2000, *Statistics and Data Analysis: From Elementary to Intermediate*, Prentice-Hall, Upper Saddle River, NJ.
[40] Society of Automotive Engineers, 2002, "Surface Vehicle Recommended Practice-Motor Vehicle Dimensions," Human Accommodation and Design Devices Standards Commission, SAE Paper No. J1100.
[41] Train, K. E., 2003, *Discrete Choice Methods With Simulation*, Cambridge University Press, Cambridge.
[42] Johnson, V. E., and Albert, J. H., 1999, *Ordinal Data Modeling*, Springer, New York.

Downloaded From: http://mechanicaldesign.asmedigitalcollection.asme.org/ on 08/06/2014 Terms of Use: http://asme.org/terms