Submitted to the Annals of Applied Statistics
arXiv: math.PR/0000000

A NEW MULTIVARIATE MEASUREMENT ERROR MODEL WITH ZERO-INFLATED DIETARY DATA, AND ITS APPLICATION TO DIETARY ASSESSMENT

By Saijuan Zhang [*], Douglas Midthune, Patricia M.Guenther, Susan M. Krebs-Smith , Victor Kipnis , Kevin W. Dodd , Dennis W. Buckman , Janet A. Tooze , Laurence Freedman and Raymond J. Carroll[*,†]

Texas A&M University, National Cancer Institute, U.S. Department of Agriculture, National Cancer Institute, National Cancer Institute, National Cancer Institute, Information Management Services, Inc., Wake Forest University, Sheba Medical Center and Texas A&M University

In the United States the preferred method of obtaining dietary intake data is the 24-hour dietary recall, yet the measure of most interest is usual or long-term average daily intake, which is impossible to measure. Thus, usual dietary intake is assessed with considerable measurement error. Also, diet represents numerous foods, nutrients and other components, each of which have distinctive attributes. Sometimes, it is useful to examine intake of these components separately, but increasingly nutritionists are interested in exploring them collectively to capture overall dietary patterns. Consumption of these components varies widely: some are consumed daily by almost everyone on every day, while others are episodically consumed so that 24-hour recall data are zero-inflated. In addition, they are often correlated with each other. Finally, it is often preferable to analyze the amount of a dietary component relative to the amount of energy (calories) in a diet because dietary recommendations often vary with energy level. The quest to understand overall dietary patterns of usual intake has to this point reached a standstill. There are no statistical methods or models available to model such complex multivariate data with its measurement error and zero inflation. This paper proposes the first such model, and it proposes the first workable solution to fit such a model. After describing the model, we use survey-weighted MCMC computations to fit the model, with uncertainty estimation coming from balanced repeated replication.

[*]This paper forms part of Zhang's Ph.D. dissertation at Texas A&M University. Zhang and Carroll's research was supported by a grant from the National Cancer Institute (CA57030). This work was also supported by National Science Foundation Instrumentation grant number 0922866.

[†]Corresponding Author.

Keywords and phrases: Bayesian methods, Dietary assessment, Latent variables, Measurement error, Mixed models, Nutritional epidemiology, Nutritional surveillance, Zero-Inflated Data

The methodology is illustrated through an application to estimating the population distribution of the Healthy Eating Index-2005 (HEI-2005), a multi-component dietary quality index involving ratios of interrelated dietary components to energy, among children aged 2-8 in the United States. We pose a number of interesting questions about the HEI-2005 and provide answers that were not previously within the realm of possibility, and we indicate ways that our approach can be used to answer other questions of importance to nutritional science and public health.

1. INTRODUCTION. This paper presents statistical models and methodology to overcome a major stumbling block in the field of dietary assessment. More nutritional background is provided in Section 2: a summary of the key conceptual issues follows.

- Nutritional surveys conducted in the United States typically use 24-hour (24hr) dietary recalls to obtain intake data, i.e., an assessment of what was consumed in the past 24 hours.
- Because dietary recommendations are intended to be met over time, nutritionists are interested in "usual" or long-term average daily intake.
- Dietary intake is thus assessed with considerable measurement error.
- Consumption patterns of dietary components vary widely; some are consumed daily by almost everyone, while others are episodically consumed so that 24-hour recall data are zero-inflated. Further, these components are correlated with one another.
- Nutritionists are interested in dietary components collectively to capture patterns of usual dietary intake, and thus need multivariate models for usual intake.
- These multivariate models for usual intakes, taking into account episodically consumed foods, do not exist, nor do methods exist for fitting them.

One way to capture dietary patterns is by scores, although our work is not limited to scores. The Healthy Eating Index-2005 (HEI-2005), described in detail in Section 2, is a scoring system based on a priori knowledge of dietary recommendations, and is on a scale of 0 to 100. Ideally, it consists of the usual intake of 6 episodically consumed and thus 24hr-zero inflated foods, 6 daily-consumed dietary components, adjusts these for energy (caloric) intake, and gives a score to each component. The total score is the sum of the individual component scores. Higher scores indicate greater compliance with dietary guidelines and, therefore, a healthier diet. Here are a few questions

that nutritionists have not been able to answer, and that our approach can address.

- What is the distribution of the HEI-2005 total score, and what % of Americans are eating a healthier diet defined for example, by a total score exceeding 80?
- What is the correlation between the individual score on each dietary component and the scores of all other dietary components?
- Among those whose total HEI-2005 score is > 50 or ≤ 50, what is the distribution of usual intake of whole grains, whole fruits, dark green and orange vegetables and legumes (DOL) and calories from solid fats, alcoholic beverages and added sugars (SoFAAS)?
- What % of Americans exceed the median score on all 12 HEI-2005 components?

In this paper, to answer public health questions such as these that can have policy implications, we build a novel multivariate measurement error model for estimating the distributions of usual intakes, one that accounts for measurement error and zero-inflation, and has a special structure associated with the zero-inflation. Previous attempts to fit even simple versions of this model, using nonlinear mixed effects software, failed because of the complexity and dimensionality of the model. We use survey-weighted Monte Carlo computations to fit the model with uncertainty estimation coming from balanced repeated replication. The methodology is illustrated using the HEI-2005 to assess the diets of children aged 2-8 in the United States. This work represents the first analysis of joint distributions of usual intakes for multiple food groups and nutrients.

The paper is outlined as follows. In Section 2 we give the background for the data we observe. In particular, we provide more information about the HEI-2005. Section 3 describes our model which is a highly nonlinear, zero-inflated, repeated measures model with multiple latent variables. The model also has a patterned covariance matrix with structural zeros and ones. We derive a parameterization that allows estimated covariance matrices to be actual covariance matrices. We also define technically what we mean by usual intake, and illustrate the use of simulation methods used to answer the questions posed above, as well as many others.

Section 4 describes our estimation procedure. Previous attempts using nonlinear mixed effects models to estimate the distribution of episodically consumed food groups (Tooze, et al., 2006; Kipnis, et al., 2009) do not work here because of the high dimensionality of the problem. We instead develop a Monte Carlo strategy based on the idea of Gibbs sampling; although because

of sampling weights, we treat the method as a frequentist (non-Bayesian) one. This section describes some of the basics of the methodology; the full technical details of implementation are given in an appendix.

Section 5 describes the analysis of the HEI-2005 components using the 2001-2004 National Health and Nutrition Examination Survey (NHANES) for children ages 2-8. Important contextual points arise because of the nature of the data. For example, if whole grains are consumed, then necessarily total grains are consumed with probability one, a restriction that a naive use of our model cannot handle. We develop a simple novel device to uncouple consumption variables that are tightly linked in this way. Finally in this section, we provide the first answers to the four questions we have posed. In Section 6, we discuss various additional aspects of the problem and the data analysis. Concluding remarks and a policy application are given in Section 7.

There are a number of general reviews of the measurement error field (Fuller, 1987; Gustafson, 2003; Carroll, et al., 2006; Buonaccorsi, 2010). Recent papers that focus on estimating the density function of a univariate continuous random variable subject to measurement error include Delaigle (2008), Delaigle and Hall (2008, 2010), Delaigle and Meister (2008), Delaigle, et al. (2008), Staudenmayer, et al. (2008) and Wand (1998). The field of measurement error in regression continues to expand rapidly, with some recent contributions including Küchenhoff, et al (2006), Guolo (2008), Liang, et al. (2008), Messer and Natarajan (2008) and Natarajan (2009). There is also a large statistical literature on measurement error as it relates to public health nutrition: some recent papers relevant to our work include Carriquiry (1999, 2003), Ferrari, et al. (2009), Fraser and Shavlik (2004), Kott, et al. (2009), Nusser, et al. (1996, 1997), Prentice (1996, 2003), and Tooze, et al. (2003, 2006).

2. Data and the HEI-2005 Scores. Here we give more detail about the nutrition context that motivates this work.

In surveys conducted in the United States, the preferred method of obtaining intake data is the 24-hour dietary recall because it limits respondent burden and facilitates accurate reporting; yet the measure of greatest interest is "usual" or long-term average daily intake. Thus dietary intake is assessed with considerable measurement error. Also, diets are comprised of numerous foods, nutrients, and other components, each of which may have distinctive attributes and effects on nutritional health. Sometimes, it is useful to examine intake of these components separately, but increasingly nutritionists are interested in exploring them collectively to capture patterns of dietary

intake. Consumption patterns of these components vary widely; some are consumed daily by almost everyone while others are episodically consumed so that 24-hour recall data are zero-inflated. In addition, these various components are often correlated with one other. Finally, it is often preferable to analyze the amount of a dietary component relative to the amount of energy (calories) in a diet because dietary recommendations often vary with energy level, and this approach provides a way of standardizing dietary assessments.

One of the US Department of Agriculture's (USDA's) strategic objectives is "to promote healthy diets" and it has developed an associated performance measure, the Healthy Eating Index-2005 (HEI-2005, http://www.cnpp.usda.gov/HealthyEatingIndex.htm). The HEI-2005 is based on the key recommendations of the 2005 Dietary Guidelines for Americans (http://www.health.gov/dietaryguidelines/dga2005/document/default.htm). The index includes ratios of interrelated dietary components to energy. The HEI-2005 comprises 12 distinct component scores and a total summary score. See Table 1 for a list of these components and the standards for scoring, and see Guenther et al. (2008) for details. Intakes of each food or nutrient, represented by one of the 12 components, are expressed as a ratio to energy intake, assessed, and ascribed a score.

The HEI-2005 is used to evaluate the diets of Americans to assess compliance with the 2005 Dietary Guidelines, yet use of the HEI-2005 is limited by the challenges described above. Until recently, there have been no solutions to these challenges, so published evaluations have been limited to analyses of mean scores for the population and various subgroups. Freedman, et al. (2010) have described a method of estimating the population distribution of a single component of HEI-2005, and the prevalence of high or low scores on that component; but there has been to date no satisfactory way to determine the prevalence of high or low total HEI-2005 scores, considering all of its interrelated components simultaneously. In addition, answers to the complex questions posed in the Introduction remain unavailable. This paper aims to provide a means to do these crucial evaluations.

The 12 HEI-2005 components represent 6 episodically consumed food groups (total fruit, whole fruit, total vegetables, dark green and orange vegetables and legumes or DOL, whole grains and milk), 3 daily-consumed food groups (total grains, meat and beans and oils), and 3 other daily-consumed dietary components (saturated fat; sodium; and calories from solid fats, alcoholic beverages and added sugars, or SoFAAS). The classification of food groups as "episodically" and "daily" consumed is based on the number of individuals who report them on 24hr recalls. If there are only a few zeros for a component, we treat that as a daily-consumed food, and replace all

TABLE 1

Description of the HEI-2005 scoring system. Except for saturated fat and SoFAAS, density is obtained by multiplying usual intake by 1000 and dividing by usual intake of kilo-calories. For saturated fat, density is 9×100 *usual saturated fat (grams) divided by usual calories, i.e., the percentage of usual calories coming from usual saturated fat intake. For SoFAAS, the density is the percentage of usual intake that comes from usual intake of calories, i.e., the division of usual intake of SoFAAS by usual intake of calories. Here, "DOL" is dark green and orange vegetables and legumes. Also, "SoFAAS" is calories from solid fats, alcoholic beverages and added sugars. The total HEI-2005 score is the sum of the individual component scores.*

Component	Units	HEI-2005 score calculation
Total Fruit	cups	$\min(5, 5 \times (\text{density}/.8))$
Whole Fruit	cups	$\min(5, 5 \times (\text{density}/.4))$
Total Vegetables	cups	$\min(5, 5 \times (\text{density}/1.1))$
DOL	cups	$\min(5, 5 \times (\text{density}/.4))$
Total Grains	ounces	$\min(5, 5 \times (\text{density}/3))$
Whole Grains	ounces	$\min(5, 5 \times (\text{density}/1.5))$
Milk	cups	$\min(10, 10 \times (\text{density}/1.3))$
Meat and Beans	ounces	$\min(10, 10 \times (\text{density}/2.5))$
Oil	grams	$\min(10, 10 \times (\text{density}/12))$
Saturated Fat	% of energy	if density ≥ 15 score $= 0$ else if density ≤ 7 score $= 10$ else if density > 10 score $= 8 - (8 \times (\text{density} - 10)/5)$ else, score $= 10 - (2 \times (\text{density} - 7)/3)$
Sodium	milligrams	if density ≥ 2000 score=0 else if density ≤ 700 score=10 else if density ≥ 1100 score $= 8 - \{8 \times (\text{density} - 1100)/(2000 - 1100)\}$ else score $= 10 - \{2 \times (\text{density} - 700)/(1100 - 700)\}$
SoFAAS	% of energy	if density ≥ 50 score $= 0$ else if density ≤ 20 score=20 else score $= 20 - \{20 \times (\text{density} - 20)/(50 - 20)\}$

zeros with 1/2 the minimum value of the non-zeros for that food. However, the crucial statistical aspect of the data is that six of the food groups are zero-inflated. The percentages of reported non-consumption of total fruit, whole fruit, whole grains, total vegetables, DOL, and milk on any single day are 17%, 40%, 42%, 3%, 50% and 12%, respectively.

We are interested in the usual intake of foods for children aged 2-8. The data available to us, described in more detail in Section 5, came from the National Health and Nutrition Examination Survey, 2001-2004 (NHANES). The data used here consisted of $n = 2,638$ children, each of whom had a survey weight w_i for $i = 1, ..., n$. In addition, one or two 24hr dietary recalls were available for each individual. Along with the dietary variables, there are covariates such as age, gender, ethnicity, family income and dummy

variables that indicate a weekday or a weekend day, and whether the recall was the first or second reported for that individual.

Using the 24hr recall data reported, for each of the episodically consumed food groups, two variables are defined: (a) whether a food from that group was consumed; and (b) the amount of the food that was reported on the 24hr recall. For the 6 daily-consumed food groups and nutrients, only one variable indicating the consumption amount is defined. In addition, the amount of energy that is calculated from the 24hr recall is of interest. The number of dietary variables for each 24hr recall is thus 12+6+1 = 19. The observed data are Y_{ijk} for the i^{th} person, the j^{th} variable and the k^{th} replicate, $j = 1, \ldots, 19$ and $k = 1, \ldots, m_i$. In the data set, at most two 24hr recalls were observed, so that $m_i \leq 2$. Set $\widetilde{Y}_{ik} = (Y_{i1k}, ..., Y_{i,19,k})^{\mathrm{T}}$, where

- $Y_{i,2\ell-1,k}$ = Indicator of whether dietary component # ℓ is consumed, with $\ell = 1, 2, 3, 4, 5, 6$.
- $Y_{i,2\ell,k}$ = Amount of food # ℓ consumed. This equals zero, of course, if none of food #ℓ is consumed, with $\ell = 1, 2, 3, 4, 5, 6$.
- $Y_{i,\ell+6,k}$ = Amount of non-episodically consumed food or nutrient #ℓ, with $\ell = 7, 8, 9, 10, 11, 12$.
- $Y_{i,19,k}$ = Amount of energy consumed as reported by the 24hr recall.

3. Model and Methods.

3.1. *Basic Model Description.* Our model is a generalization of work by Tooze et al. (2006) and Kipnis, et al. (2009) for a single food and Kipnis, et al. (2010) and Zhang, et al. (2010) for a single food and nutrient. Observed data will be denoted as Y, and covariates in the model will be denoted as X. As is usual in measurement error problems, there will also be latent variables, which will be denoted by W.

We use a probit threshold model. Each of the 6 episodically consumed foods will have 2 sets of latent variables, one for consumption and one for amount, while the 6 daily-consumed foods and nutrients as well as energy will have 1 set of latent variables, for a total of 19. The latent random variables are ϵ_{ijk} and U_{ij}, where $(U_{i1}, \ldots, U_{i,19}) = \text{Normal}(0, \Sigma_u)$ and $(\epsilon_{i1k}, \ldots, \epsilon_{i,19,k}) = \text{Normal}(0, \Sigma_\epsilon)$ are mutually independent. In this model, food $\ell = 1, ..., 6$ being consumed on day k is equivalent to observing the binary $Y_{i,2\ell-1,k}$, where

$$\begin{aligned} Y_{i,2\ell-1,k} = 1 \quad &\Longleftrightarrow \quad W_{i,2\ell-1,k} \\ &= \quad X_{i,2\ell-1,k}^{\mathrm{T}} \beta_{2\ell-1} + U_{i,2\ell-1} + \epsilon_{i,2\ell-1,k} > 0. \end{aligned} \tag{3.1}$$

If the food is consumed we model the amount reported $Y_{i,2\ell,k}$ as

$$[g_{\mathrm{tr}}(Y_{i,2\ell,k},\lambda_\ell)|Y_{i,2\ell-1,k}=1] = W_{i,2\ell,k} = X_{i,2\ell,k}^{\mathrm{T}}\beta_{2\ell}+U_{i,2\ell}+\epsilon_{i,2\ell,k}, \tag{3.2}$$

where $g_{\mathrm{tr}}(y,\lambda)=\sqrt{2}\{g(y,\lambda)-\mu(\lambda)\}/\sigma(\lambda)$, $g(y,\lambda)$ is the usual Box-Cox transformation with transformation parameter λ, and $\{\mu(\lambda),\sigma(\lambda)\}$ are the sample mean and standard deviation of $g(y,\lambda)$, computed from the non-zero food data. This standardization is simply a convenient device to improve the numerical performance of our algorithm without affecting the conclusions of our analysis.

The reported consumption of daily consumed foods or nutrients $\ell = 7,\ldots,12$ are modeled as

$$g_{\mathrm{tr}}(Y_{i,\ell+6,k},\lambda_\ell) = W_{i,\ell+6,k}=X_{i,\ell+6,k}^{\mathrm{T}}\beta_{\ell+6}+U_{i,\ell+6}+\epsilon_{i,\ell+6,k}. \tag{3.3}$$

Finally, energy is modeled as

$$g_{\mathrm{tr}}(Y_{i,19,k},\lambda_{13}) = W_{i,19,k}=X_{i,19,k}^{\mathrm{T}}\beta_{19}+U_{i,19}+\epsilon_{i,19,k}. \tag{3.4}$$

As seen in (3.3)-(3.4), different transformations $(\lambda_1,...,\lambda_{13})$ are allowed to be used for the different types of dietary components, see Section A.12.

In summary, there are latent variables $\widetilde{W}_{ik}=(W_{i1k},...,W_{i,19,k})^{\mathrm{T}}$, latent random effects $\widetilde{U}_i=(U_{i1},...,U_{i,19})^{\mathrm{T}}$, fixed effects $(\beta_1,...,\beta_{19})$, and design matrices $(X_{i1k},...,X_{i,19,k})$. Define $\widetilde{\epsilon}_{ik}=(\epsilon_{i1k},...,\epsilon_{i,19,k})^{\mathrm{T}}$. The latent variable model is

$$W_{ijk} = X_{ijk}^{\mathrm{T}}\beta_j+U_{ij}+\epsilon_{ijk}, \tag{3.5}$$

where $\widetilde{U}_i=\mathrm{Normal}(0,\Sigma_u)$ and $\widetilde{\epsilon}_{ik}=\mathrm{Normal}(0,\Sigma_\epsilon)$ are mutually independent.

3.2. *Restriction on the Covariance Matrix.* Two necessary restrictions are set on Σ_ϵ. First, following Kipnis, et al. (2009, 2010), $\epsilon_{i,2\ell-1,k}$ and $\epsilon_{i,2\ell,k}$, $(\ell=1,\ldots,6)$ are set to be independent. Second, in order to technically identify $\beta_{2\ell-1}$ and the distribution of $U_{i,2\ell-1}$ $(\ell=1,\ldots,6)$, we require that $\mathrm{var}(\epsilon_{i,2\ell-1,k})=1$, because otherwise the marginal probability of consumption of dietary component $\#\ell$ would be $\Phi\{(X_{i,2\ell-1,k}^{\mathrm{T}}\beta_{2\ell-1}+U_{i,2\ell-1})/\mathrm{var}^{1/2}(\epsilon_{i,2\ell-1,k})\}$, and thus components of β and Σ_u would be identified only up to the scale $\mathrm{var}^{1/2}(\epsilon_{i,2\ell-1,k})$.

So that we can handle any number of episodically consumed dietary components and any number of daily consumed components, suppose that there

are J episodically consumed dietary components, and K daily consumed dietary components, and in addition there is energy. Then the restrictions defined above lead to the covariance matrix

$$\Sigma_\epsilon = \begin{pmatrix} 1 & 0 & s_{13} & s_{14} & \dots \\ 0 & s_{22} & s_{23} & s_{24} & \dots \\ s_{13} & s_{23} & 1 & 0 & \dots \\ s_{14} & s_{24} & 0 & s_{44} & \dots \\ \vdots & \vdots & \vdots & \vdots & \ddots \\ s_{1,2J+1} & s_{2,2J+1} & s_{3,2J+1} & s_{4,2J+1} & \dots \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ s_{1,2J+K+1} & s_{2,2J+K+1} & s_{3,2J+K+1} & s_{4,2J+K+1} & \dots \end{pmatrix} \tag{3.6}$$

$$\begin{pmatrix} \dots & s_{1,2J+1} & \dots & s_{1,2J+K+1} \\ \dots & s_{2,2J+1} & \dots & s_{2,2J+K+1} \\ \dots & s_{3,2J+1} & \dots & s_{3,2J+K+1} \\ \dots & s_{4,2J+1} & \dots & s_{4,2J+K+1} \\ \ddots & \vdots & \dots & \vdots \\ \dots & s_{2J+1,2J+1} & \dots & s_{2J+1,2J+K+1} \\ \vdots & \vdots & \ddots & \vdots \\ \dots & s_{2J+1,2J+K+1} & \dots & s_{2J+K+1,2J+K+1} \end{pmatrix}.$$

The difficulty with parameterizations of (3.2) is that the cells that are not constrained to be 0 or 1 cannot be left unconstrained, otherwise (3.2) need not be a covariance matrix, i.e., positive semidefinite.

We have developed an unconstrained parameterization that results in the structure (3.2). Consider an unconstrained lower triangular matrix V and define $\Sigma_\epsilon = VV^{\mathrm{T}}$. This is positive semidefinite and therefore qualifies Σ_ϵ as a proper covariance matrix. The form of V is

$$V = \begin{pmatrix} v_{11} & 0 & \dots & 0 \\ v_{21} & v_{22} & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ v_{2J+K+1,1} & v_{2J+K+1,2} & \dots & v_{2J+K+1,2J+K+1} \end{pmatrix}.$$

To achieve the desired pattern (3.2), we derive the following four restrictions:

$$\begin{aligned} v_{11} &= 1; \\ v_{21} &= 0; \\ \textstyle\sum_{p=1}^{q} v_{qp}^2 &= 1; \quad q = 3, 5, \dots, 2J-1; \\ \textstyle\sum_{p=1}^{q} v_{qp} v_{q+1,p} &= 0; \quad q = 3, 5, \dots, 2J-1. \end{aligned}$$

The third restriction can be ensured by the further parameterization

$$\begin{aligned}
v_{31} &= r_1 \sin(\theta_1);\\
v_{32} &= r_1 \cos(\theta_1);\\
v_{33} &= \sqrt{1-r_1^2};\\
v_{2q+1,1} &= r_q \sin(\theta_{1+(q-1)^2});\\
v_{2q+1,p} &= r_q \cos(\theta_{1+(q-1)^2}) \times \cdots \times \cos(\theta_{p-1+(q-1)^2}) \sin(\theta_{p+(q-1)^2}),\\
&\qquad p = 2, \ldots, 2q-1;\\
v_{2q+1,2q} &= r_q \cos(\theta_{1+(q-1)^2}) \times \cdots \times \cos(\theta_{q^2});\\
v_{2q+1,2q+1} &= \sqrt{1-r_q^2},
\end{aligned}$$

where $q = 2, 3, \ldots, J-1$; $|r_t| \le 1$, $t = 1, \ldots, J-1$, and $|\theta_s| \le \pi$, $s = 1, \ldots, (J-1)^2$.

Similarly, the fourth restriction can be further expressed by setting

$$v_{q+1,q} = -\sum_{p=1}^{q-1} v_{qp} v_{q+1,p} / v_{qq} = -\sum_{p=1}^{q-1} v_{qp} v_{q+1,p} / \sqrt{1 - r^2_{(q-1)/2}},$$

where $q = 3, 5, \ldots, 2J-1$.
Note that $|\Sigma_\epsilon| = |V|^2 = \prod_{q=1}^{2J+K+1} v_{qq}^2 = \prod_{q=1}^{J} v_{2q,2q}^2 \prod_{q=2J+1}^{2J+K+1} v_{q,q}^2 \prod_{q=1}^{J-1}(1 - r_q^2)$.

3.3. *The Use of Sampling Weights.* As described in the Appendix, we used the survey sample weights from NHANES both in the model fitting procedure and, after having fit the model, in estimating the distributions of usual intake.

While not displayed here, we redid the model fitting calculations without weighting, because the covariates we use are major players in determining the sampling weights, hence it is reasonable to believe that the model in Section 3 holds both in the sample and in the population. When we did this, the parameter estimates were essentially unchanged.

Thus, we use the sampling weights only for estimation of the population distributions. We actually did this for the purpose of handling the clustering in the sample design. For such a complex statistical procedure as ours, we knew we could not do theoretical standard errors, so we thought about the bootstrap, and realized that putting together a bootstrap for the complex survey would be nearly impossible. However, we already had developed a set of Balanced Repeated Replication (BRR) weights (Wolter, 1995), see

Section 5.7 for details. These BRR weights have the property that, in the frequentist survey sampling sense, they appropriately reflect the clustering in the standard error calculations.

Of course, the use of sampling weights in the modeling provide unbiased estimates of the (super) population parameters of interest. In addition, the use of sampling weights in the distribution estimation provides an estimated distribution that is representative of the US population, not just the sample.

3.4. *Distribution of Usual Intake and the HEI-2005 Scores.* We assume here that estimates of Σ_u, Σ_ϵ and β_j for $j = 1, ..., 19$ have been constructed, see Section 4. Here we discuss what we mean by usual intake for an individual, how to estimate the distribution of usual intakes, how to convert usual intakes into HEI-2005 scores, and how to assess uncertainty.

Consider the first episodically consumed dietary component, a food group, with reporting being done on a weekend. Set $X_{i1,\text{wkend}}$ and $X_{i2,\text{wkend}}$ to be the versions of X_{i1k} and X_{i2k} where the dummy variable has the indicator of the weekend and that the recall is the first one. Following Kipnis, et al. (2009), we define the usual intake for an individual on the weekend to be the expectation of the reported intake conditional on the person's random effects $\widetilde{U}_i$. Let the (q, p) element of Σ_ϵ be denoted as $\Sigma_{\epsilon,q,p}$. As in Kipnis, et al. define

$$g^*_{\text{tr}}\{v, \lambda, \Sigma_{\epsilon,q,p}\} \quad = \quad g^{-1}_{\text{tr}}(v, \lambda) + \frac{1}{2}\Sigma_{\epsilon,q,p}\frac{\partial^2 g^{-1}_{\text{tr}}(v, \lambda)}{\partial v^2}. \tag{3.7}$$

Detailed formulas for this are given in Appendix A.11. Then, following the convention of Kipnis, et al. (2009), the person's usual intake of the first episodically consumed dietary component on the weekend is defined as

$$T_{i1,\text{wkend}} = \Phi(X^{\text{T}}_{i1,\text{wkend}}\beta_1 + U_{i1})g^*_{\text{tr}}\left(X^{\text{T}}_{i2,\text{wkend}}\beta_2 + U_{i2}, \lambda_1, \Sigma_{\epsilon,2,2}\right).$$

Similarly, let $X_{i1,\text{wkday}}$ and $X_{i2,\text{wkday}}$ be as above but the dummy variable is appropriate for a weekday. Then the person's usual intake of the first episodically consumed food group on weekdays is defined as

$$T_{i1,\text{wkday}} = \Phi(X^{\text{T}}_{i1,\text{wkday}}\beta_1 + U_{i1})g^*_{\text{tr}}\left(X^{\text{T}}_{i2,\text{wkday}}\beta_2 + U_{i2}, \lambda_1, \Sigma_{\epsilon,2,2}\right).$$

Finally, the usual intake of the first episodically consumed food for the individual is

$$T_{i1} = (4T_{i1,\text{wkday}} + 3T_{i1,\text{wkend}})/7,$$

since Fridays, Saturdays and Sundays are considered to be weekend days. Usual intake for the other episodically consumed food groups is defined similarly.

A person's usual intake of a daily-consumed food group/nutrient and energy on the original scale is defined similarly. Consider, for example, energy, which is the 13^{th} dietary component and the 19^{th} set of terms in the model. Let $X_{i,19,\text{wkend}}$ and $X_{i,19,\text{wkday}}$ be the versions of $X_{i,19,k}$ where the dummy variable has the indicator of the weekend or weekday, respectively, and that the recall is the first one. Then

$$\begin{aligned} T_{i,13,\text{wkend}} &= g^*_{\text{tr}}\left(X^{\text{T}}_{i,19,\text{wkend}}\beta_{19} + U_{i,19}, \lambda_{13}, \Sigma_{\epsilon,19,19}\right); \\ T_{i,13,\text{wkday}} &= g^*_{\text{tr}}\left(X^{\text{T}}_{i,19,\text{wkday}}\beta_{19} + U_{i,19}, \lambda_{13}, \Sigma_{\epsilon,19,19}\right); \\ T_{i,13} &= (4T_{i,13,\text{wkday}} + 3T_{i,13,\text{wkend}})/7. \end{aligned}$$

Similar formulae are used for the other daily-consumed foods and nutrients.

Finally, the energy-adjusted usual intakes and the HEI-2005 scores are then obtained as in Table 1, using the estimated usual intakes of the dietary components.

To find the joint distribution of usual intakes of the HEI-2005 scores, it is convenient to use Monte-Carlo methods. Recall that w_i is the sampling weight for individual i. Let B be a large number: we set $B = 5,000$. Generate $b = 1, ..., B$ observations $\widetilde{U}_{bi} = \text{Normal}(0, \Sigma_u)$ and then obtain $\widetilde{T}_{bi} = (T_{bi\ell})_{\ell=1}^{13}$ by replacing U_{ij} in their formulae by U_{bij}. With appropriate sample weighting, the $\widetilde{T}_{bi}$ can be used to estimate joint and marginal distributions. Thus, for example, consider the total HEI-2005 score, which is a deterministic function of the usual intakes, say $G(\widetilde{T}_i)$. Its cumulative distribution function is estimated as

$$\widehat{F}(x) = \frac{\sum_{i=1}^{n}\sum_{b=1}^{B} I\{G(\widetilde{T}_{bi}) \le x\} w_i}{\sum_{i=1}^{n}\sum_{b=1}^{B} w_i}. \tag{3.8}$$

Frequentist standard errors of derived quantities such a mean, median and quantiles can be estimated using the Balanced Repeated Replication (BRR) method (Wolter, 1995), see Section 5.7 for details.

4. Comments on the Approach to Estimation. Our model (3.3)-(3.4) is a highly nonlinear, mixed effects model with many latent variables and nonlinear restrictions on the covariance matrix Σ_ϵ. As seen in Section 3.4, we can estimate relevant distributions of usual intake in the population if we can estimate Σ_u, Σ_ϵ and β_j for $j = 1, ..., 19$. We have found that working within a pseudo-likelihood Bayesian paradigm is a convenient way to do this computation. We emphasize, however, that we are doing this only to get frequentist parameter estimates based on the well-known asymptotic equivalence of frequentist likelihood estimators and Bayesian posterior means,

and especially the consistency of both (Lehmann and Casella, 1998). We are specifically not doing Bayesian posterior inference, since valid Bayesian inference in a complex survey such as NHANES is an immensely challenging task, and because frequentist estimation and inference are the standard in the nutrition community.

Kipnis, et al. (2009) were able to get estimates of parameters separately for each food group using the nonlinear mixed effects program NLMIXED in SAS with sampling weights. While this gives estimates of β_j for $j = 1, ..., 19$, it only gives us parts of the covariance matrices Σ_u and Σ_ϵ, and not all the entries. Using the 2001-2004 NHANES data, we have verified that our estimates and the subset of the parameters that can be estimated by one food group at a time using NLMIXED are in close agreement, and that estimates of the distributions of usual intake and HEI-2005 component scores are also in close agreement. We expect this because of the rather large sample size in our data set. Zhang, et al. (2010) have shown that even considering a single food group plus energy is a challenge for the NLMIXED procedure, both in time and in convergence, and using this method for the entire HEI-2005 constellation of dietary components is impossible.

Full technical details of the model fitting procedure are given in Appendices A.1-A.10.

Of course, our model has assumptions, e.g., additivity and homoscedasticity on a transformed scale for observed and latent variables, normality of person-specific random effects and normality of day-to-day variability on the transformed scale. These assumptions are clearly not exactly correct, although our marginal model-checking suggests to us that they are mostly not disastrously wrong. Some reasons for this conclusion include the facts that we reproduce the marginal distributions of the components, that comparison with 24hr recalls shows differences that decrease when moving from one 24hr recall to two 24hr recalls, that q-q plots of the data are fairly satisfactory, etc. Thinking, as we do, of our work as a first step, and not a last step, it would be extremely interesting to make the model more general, e.g., skew-normal, skew-t or Dirichlet process distributions after transformation, and possibly directly modeling heteroscedasticity. Such generalizations will require effort to implement, but will speak to the robustness of the results and would be a useful future step.

5. Empirical Work.

5.1. *Basic Analysis.* We analyzed data from the 2001-2004 National Health and Nutrition Examination Survey (NHANES) for children age 2-8. The study sample consisted of $2,638$ children, among whom $1,103$ children have

two 24hr recalls and the rest have only one. We used the dietary intake data to calculate the 12 HEI-2005 components plus energy. In addition, besides age, gender, race and interaction terms, two covariates were employed, along with an intercept. The first was a dummy variable indicating whether or not the recall was for a weekend day (Friday, Saturday, or Sunday) because food intakes are known to differ systematically on weekends and weekdays. The second was a dummy variable indicating whether the 24hr recall was the first or second such recall, the idea being that there may be systematic differences attributable to the repeated administration of the instrument.

5.2. *Contextual Information.* When we ran our program based on the variables in Table 1, the results were disastrous. Mixing of the MCMC sampler was very poor, with long sojourns in different regions.

The reason for this failure to converge depends on the context of the dietary variables. For example, whole grains are a subset of total grains. Thus, if someone consumes any whole grains, then necessarily, with probability 1.0, that person also consumes total grains. Such a restriction cannot be handled by our model, because it would force one of the random effects U to equal infinity. A similar thing happens for energy. Calories coming from saturated fat are a subset of total calories as are calories from SoFAAS, so there is a restriction that total calories must be greater than calories from saturated fat and also greater than calories from SoFAAS. Since the latter sum makes up a significant portion of calories, this restriction is not something that our model can handle well.

Luckily, there is an easy and natural context-based solution. Instead of using total grains in the model, we used grains that are not whole grains, i.e., refined grains, thus decoupling whole grains and total grains, and removing the restriction mentioned above. Similarly, instead of using total fruit, we use fruit that is not whole fruits, i.e., fruit juices. Additionally, instead of using total vegetables, we use total vegetables excluding dark green and orange vegetables and legumes. Finally, instead of total energy, we use total energy minus the sum of energy from saturated fat (11% of mean energy) and from SoFAAS (35% of mean energy). We recognize that there is overlap of energy from saturated fat and energy from solid fat, but this has no impact on our analysis since total energy has sources other than these two. An alternative of course, would have been to simply use total energy minus energy from SoFAAS,

This is sufficient to estimate the distributions of interest. If, for example, in the new data set T_{i1} represents usual intake of non-whole fruits, and T_{i2} is usual intake of whole fruits, then the usual intake of total fruits is $T_{i1} + T_{i2}$.

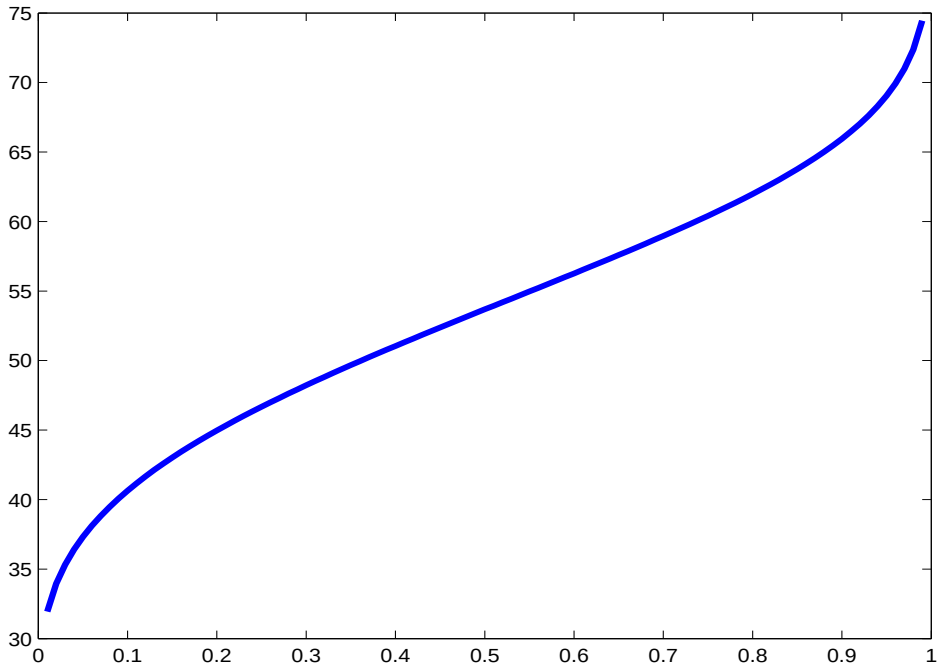

FIG 1. *The estimated percentiles of the HEI-2005 total score. The horizontal axis is the percentile of interest, e.g.,* 0.5 *refers to the median, while the vertical axis gives percentile of the HEI-2005 scores. Standard error estimates are given in Table 2.*

Similar remarks apply for total grains and total vegetables.

With these new variables, our model mixed well and gave reasonable looking answers that, as mentioned in Section 4, give similar results to other methods employed with smaller parts of the data set.

5.3. *Estimation of the HEI-2005 Scores.* In the introduction, we posed 4 questions to which answers had not been possible previously. The first open question concerned the distribution of the HEI total score. Along the way towards this, Table 2 presents the energy-adjusted distributions of the dietary components used in the HEI-2005. Table 3 presents the distributions of the HEI-2005 individual component scores and the total score, with a graphical view given in Figure 1.

Table 3 presents the first estimates of the distribution of HEI-2005 scores for a vulnerable subgroup of the population, namely children aged 2-8 years.

TABLE 2. *Estimated distributions of energy-adjusted usual intakes for children aged 2-8; NHANES, 2001-2004. For each dietary component, the first line = estimate from our model, while the second line is its BRR-estimated standard error. Here, "DOL" is dark green and orange vegetables and legumes. Also, "SoFAAS" is calories from solid fats, alcoholic beverages and added sugars. Total Fruit, Whole Fruit, Total Vegetables, DOL and Milk are in cups. Total Grains, Whole Grains and Meat and Beans are in ounces. Oil and Sodium are in grams. Saturated Fat and SoFAAS are in % of energy. Further discussion of the size of the BRR-estimated standard errors is given in the supplementary material.*

			Percentile						
Component	Units	Mean	5^{th}	10^{th}	25^{th}	50^{th}	75^{th}	90^{th}	95^{th}
Total Fruit	cups/(1000 kcal)	0.70	0.14	0.21	0.37	0.62	0.95	1.30	1.54
		0.02	0.02	0.02	0.02	0.02	0.03	0.05	0.07
Whole Fruit	cups/(1000 kcal)	0.31	0.04	0.07	0.14	0.26	0.42	0.61	0.73
		0.02	0.01	0.01	0.02	0.02	0.03	0.04	0.06
Total Vegetables	cups/(1000 kcal)	0.47	0.23	0.27	0.36	0.46	0.58	0.69	0.77
		0.01	0.02	0.02	0.02	0.01	0.02	0.03	0.03
DOL	cups/(1000 kcal)	0.05	0.00	0.01	0.02	0.03	0.07	0.11	0.15
		0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.01
Total Grains	ounces/(1000 kcal)	3.32	2.35	2.54	2.87	3.28	3.72	4.16	4.45
		0.05	0.08	0.07	0.06	0.05	0.06	0.08	0.10
Whole Grains	ounces/(1000 kcal)	0.27	0.05	0.07	0.13	0.23	0.36	0.52	0.64
		0.01	0.01	0.01	0.02	0.01	0.02	0.03	0.04
Milk	cups/(1000 kcal)	0.97	0.28	0.38	0.60	0.90	1.26	1.64	1.90
		0.02	0.03	0.03	0.02	0.02	0.03	0.05	0.07
Meat and Beans	ounces/(1000 kcal)	1.84	1.06	1.21	1.48	1.80	2.16	2.51	2.73
		0.04	0.09	0.08	0.06	0.04	0.04	0.05	0.07
Oil	grams/(1000 kcal)	7.13	4.05	4.60	5.63	6.93	8.41	9.90	10.89
		0.23	0.24	0.21	0.17	0.20	0.35	0.54	0.68
Saturated Fat	% of Energy	11.71	8.56	9.20	10.33	11.64	13.01	14.32	15.13
		0.15	0.25	0.20	0.15	0.15	0.22	0.32	0.38
Sodium	grams/(1000 kcal)	1.49	1.16	1.23	1.34	1.48	1.63	1.77	1.86
		0.01	0.02	0.02	0.01	0.01	0.02	0.03	0.03
SoFAAS	% of Energy	36.93	27.19	29.28	32.87	36.90	40.96	44.61	46.77
		0.48	0.93	0.81	0.63	0.48	0.49	0.64	0.75

TABLE 3

Estimated distributions of the usual intake HEI-2005 scores. For each component score, the first line = estimate from our model, while the second line is its BRR-estimated standard error. The total score is the sum of the individual scores. Here, "DOL" is dark green and orange vegetables and legumes. Also, "SoFAAS" is calories from solid fats, alcoholic beverages and added sugars. Further discussion of the size of the BRR-estimated standard errors is given in the supplementary material.

		Percentile						
Component	Mean	5^{th}	10^{th}	25^{th}	50^{th}	75^{th}	90^{th}	95^{th}
Total Fruit	3.55	0.87	1.31	2.33	3.90	5.00	5.00	5.00
	0.09	0.13	0.14	0.15	0.15	0.00	0.00	0.00
Whole Fruit	3.14	0.49	0.82	1.71	3.24	5.00	5.00	5.00
	0.14	0.12	0.16	0.21	0.26	0.03	0.00	0.00
Total Vegetables	2.16	1.02	1.24	1.63	2.10	2.62	3.15	3.48
	0.06	0.10	0.10	0.07	0.06	0.07	0.12	0.16
DOL	0.62	0.05	0.09	0.21	0.45	0.86	1.38	1.76
	0.04	0.02	0.03	0.04	0.05	0.06	0.08	0.13
Total Grains	4.81	3.92	4.23	4.79	5.00	5.00	5.00	5.00
	0.03	0.13	0.12	0.09	0.00	0.00	0.00	0.00
Whole Grains	0.90	0.16	0.24	0.43	0.75	1.21	1.74	2.13
	0.04	0.04	0.05	0.05	0.05	0.05	0.10	0.14
Milk	6.77	2.15	2.96	4.62	6.91	9.67	10.00	10.00
	0.12	0.23	0.22	0.18	0.17	0.25	0.00	0.00
Meat and Beans	7.22	4.23	4.83	5.91	7.21	8.64	10.00	10.00
	0.16	0.34	0.30	0.23	0.17	0.15	0.11	0.00
Oil	5.92	3.37	3.83	4.69	5.77	7.01	8.25	9.07
	0.18	0.20	0.18	0.14	0.17	0.29	0.45	0.57
Saturated Fat	5.16	0.00	1.09	3.18	5.38	7.48	8.53	8.96
	0.21	0.35	0.51	0.35	0.24	0.23	0.13	0.16
Sodium	4.52	1.25	2.05	3.31	4.62	5.83	6.85	7.44
	0.09	0.30	0.24	0.15	0.09	0.11	0.16	0.19
SoFAAS	8.73	2.15	3.60	6.02	8.73	11.42	13.81	15.21
	0.32	0.50	0.42	0.33	0.32	0.42	0.54	0.62
Total Score	53.50	37.42	40.74	46.73	53.68	60.36	65.87	68.96
	0.81	1.45	1.34	1.09	0.83	0.82	0.96	1.08

A previous analysis of 2003-04 NHANES data, looking separately at 2-5 year olds and 6-11 year olds, was limited to estimates of mean usual HEI-2005 scores (59.6 and 54.7, respectively, see Fungwe, et al., 2009). The mean scores noted here are comparable to those and reinforce the notion that children's diets, on average, are far from ideal. However, this analysis provides a more complete picture of the state of US children's diets. By including the scores at various percentiles, we estimate that only 5% of children have a score of 69 or greater and another 10% have scores of 41 or lower. While not in the Table, we also estimate that the 99^{th} percentile is 74. This analysis suggests that virtually all children in the US have suboptimal diets and that a sizeable fraction (10%) have alarmingly low scores (41 or lower.)

We have also considered whether our multivariate model fitting procedure gives reasonable marginal answers. To check this, we note that it is possible to use the SAS procedure NLMIXED *separately for each component* to fit a model with one episodically consumed food group or daily consumed dietary component together with energy. The marginal distributions of each such component done separately are quite close to what we have reported in Table 3, as is our mean, which is 53.50 compared to the mean of 53.25 based on analyzing one HEI-2005 component at a time with the NLMIXED procedure. The only case where there is a mild discrepancy is in the estimated variability of the energy-adjusted usual intake of oils, likely caused by the NLMIXED procedure itself, which has an estimated variance 9 times greater than our estimated variance.

Of course, it is the distribution of the HEI-2005 total score that cannot be estimated by analysis of one component at a time.

There are other things that have not been computed previously that are simple by-products of our analysis. For example, the correlations among energy-adjusted usual intakes involving episodically consumed foods have not been estimated previously, but this is easy for us, see Table 4. The estimated correlation of -0.64 between energy-adjusted total fruit and energy-adjusted SoFAAS, and the -0.47 correlation between DOL and SoFAAS are surprisingly high.

5.4. *Component Scores and Other Scores.* As described in the introduction, an open problem has been to estimate the correlation between the individual score on each dietary component and the scores of all other dietary components. In their Table 3, Guenther, et al. (2008b) consider this problem, but of course they did not have a model for usual energy adjusted intakes, and instead they used a single 24hr recall. In Table 5, we show the resulting correlations using (a) a single 24hr recall; (b) the mean of

TABLE 4. *Estimated correlation matrix for energy-adjusted usual intakes. Here TF = Total Fruits, WF = Whole Fruits, TV = Total Vegetables, WG = Whole Grains, TG = Total Grains, SatFat = Saturated Fat. Here, "DOL" is dark green and orange vegetables and legumes. Also, "SoFAAS" is calories from solid fats, alcoholic beverages and added sugars.*

Component	TF	WF	TV	DOL	TG	WG	Milk	Meat	Oil	SatFat	Sodium	SoFAAS
TF	1	0.76	0.07	0.41	-0.10	0.33	0.16	0.08	-0.35	-0.38	-0.25	-0.64
WF		1	0.14	0.49	0.03	0.35	0.10	0.05	-0.17	-0.30	-0.20	-0.51
TV			1	0.51	-0.25	-0.23	-0.09	0.51	-0.08	0.08	0.42	-0.16
DOL				1	-0.08	0.11	0.14	0.25	-0.06	-0.23	0.01	-0.47
TG					1	0.30	-0.30	-0.13	0.44	-0.36	0.17	-0.22
WG						1	0.18	-0.18	-0.11	-0.29	-0.17	-0.46
Milk							1	-0.37	-0.21	0.21	-0.27	-0.21
Meat & Beans								1	-0.06	-0.08	0.39	-0.19
Oil									1	-0.06	0.11	0.05
SatFat										1	0.09	0.46
Sodium											1	0.04
SoFAAS												1

TABLE 5
Estimated correlations between each individual HEI-2005 component score and the sum of the other HEI component scores, i.e., the difference of the total score and each individual component. The column labeled "Two 24hr" is the naive analysis that uses the mean of the two 24hr recalls, while the column labeled "First 24hr" is the naive analysis that uses the first 24hr recall. The column labeled "Model" is our analysis, and the column labeled "BRR s.e." is the estimated standard error of our estimates. Here, "DOL" is dark green and orange vegetables and legumes. Also, "SoFAAS" is calories from solid fats, alcoholic beverages and added sugars.

	First 24hr	Two 24hr	Model	BRR s.e.
Total Fruit	0.38	0.44	0.62	0.05
Whole Fruit	0.31	0.37	0.59	0.10
Total Vegetables	0.09	0.11	0.10	0.11
DOL	0.18	0.24	0.41	0.07
Total Grains	0.00	0.00	0.06	0.11
Whole Grains	0.12	0.16	0.53	0.08
Milk	-0.07	-0.01	0.01	0.08
Mean and Beans	-0.03	-0.01	-0.03	0.15
Oil	0.08	0.05	-0.17	0.08
Saturated Fat	0.21	0.23	0.36	0.06
Sodium	-0.03	0.05	0.07	0.12
SoFAAS	0.52	0.59	0.72	0.04

two 24hr recalls for those who have two 24hr recalls; and (c) our model for usual intake. The numbers for the former differ from that of Guenther, et al. (2008b) because we are considering here a different population than do they. A striking and not unexpected aspect of this table is that for those components with non-trivial correlations, the correlations all increase as one moves from a single 24hr recall to the mean of two 24hr recalls and then finally to estimated usual intake. Thus, for example, the correlation between the HEI-2005 score for total fruit and its difference with the total score is 0.38 for a single 24hr recall, 0.44 for the mean of two 24hr recalls and then finally 0.62 for usual intake.

5.5. *Distributions of Intakes for Subsets of HEI Total Scores.* A third open question is: among those whose total HEI-2005 score is > 50 or ≤ 50, what is the distribution of energy-adjusted usual intake of whole grains, whole fruits, dark green and orange vegetables and legumes (DOL) and calories from solid fats, alcoholic beverages and added sugars (SoFAAS)? This follows naturally from our method. Following (3.8), let $G_1(\widetilde{T}_{bi})$ be energy adjusted usual intake and let $G_2(\widetilde{T}_{bi})$ be the HEI total score. Then the distributions in question for when the total HEI-2005 score is > 50 can be estimated as $\widehat{F}(x) = \sum_{i=1}^{n}\sum_{b=1}^{B} w_i I\{G_1(\widetilde{T}_{bi}) \leq x\} I\{G_2(\widetilde{T}_{bi}) > 50\} / \sum_{i=1}^{n}\sum_{b=1}^{B} w_i I\{G_2(\widetilde{T}_{bi}) > 50\}$.

The results are provided in Table 6, with a graphical view in Figure 2.

TABLE 6
Estimated distributions of energy-adjusted usual intake for those whose total HEI-2005 total scores are ≤ 50 and > 50. Here, "DOL" is dark green and orange vegetables and legumes. Also, "SoFAAS" is calories from solid fats, alcoholic beverages and added sugars. Units of measurement are given in Table 2.

			Percentile						
Component	Mean	s.d	5^{th}	10^{th}	25^{th}	50^{th}	75^{th}	90^{th}	95^{th}
Whole Fruit									
Total Score ≤ 50	0.15	0.12	0.02	0.03	0.07	0.12	0.21	0.30	0.38
Total Score > 50	0.39	0.22	0.11	0.15	0.23	0.35	0.51	0.68	0.80
Whole Grains									
Total Score ≤ 50	0.18	0.13	0.03	0.05	0.09	0.15	0.25	0.36	0.44
Total Score > 50	0.32	0.20	0.07	0.10	0.17	0.28	0.42	0.59	0.70
DOL									
Total Score ≤ 50	0.02	0.02	0.00	0.00	0.01	0.02	0.03	0.05	0.07
Total Score > 50	0.06	0.05	0.01	0.01	0.03	0.05	0.09	0.13	0.17
SoFAAS									
Total Score ≤ 50	42.43	3.97	36.40	37.59	39.66	42.16	44.92	47.67	49.42
Total Score > 50	33.83	4.44	26.01	27.89	30.97	34.15	36.98	39.28	40.57
Total Score	53.50	9.58	37.42	40.74	46.73	53.68	60.36	65.87	68.96

The results show that those who have poorer diets with usual HEI-2005 total score ≤ 50 are consistently eating poorer diets, i.e., less whole fruits, less whole grains and less DOL, but higher SoFAAS.

5.6. *Dietary Consistency.* We stated in the introduction that it is interesting to understand the percentage of children whose usual intake HEI score exceeds the median HEI score on all 12 HEI components. Those median scores, say $(\kappa_1, ..., \kappa_{12})$, are estimated in Table 3. If $G_j(\widetilde{T}_{bi})$ is the HEI component score for episodically consumed food j, then following (3.8) the quantity in question can be estimated as $\sum_{i=1}^{n}\sum_{b=1}^{B} w_i \prod_{j=1}^{6} I\{G_j(\widetilde{T}_{bi}) \geq \kappa_j\}/\sum_{i=1}^{n}\sum_{b=1}^{B} w_i$. We estimate that the percentage is 6%, woefully small. The percentage of children whose usual intake HEI score exceeds the median HEI score on all 12 HEI components is 0.24%. Figure 3 gives the estimated probabilities of exceeding the κ percentile on all 12 HEI components simultaneously, for $\kappa = 1, 2, ..., 99$.

5.7. *Uncertainty Quantification.* The BRR standard errors of HEI-2005 components' adjusted usual intakes and scores are shown in Tables 2 and 3. The BRR weights are only used in variance calculations. Once we have estimated some quantity, say $\widehat{\theta}$, from the sample using sample weight, we will need to compute the same quantity using, in succession, the 32 BRR weights. This will give us 32 estimates $\widehat{\theta}_1, \widehat{\theta}_2, \ldots, \widehat{\theta}_{32}$. The BRR estimate for

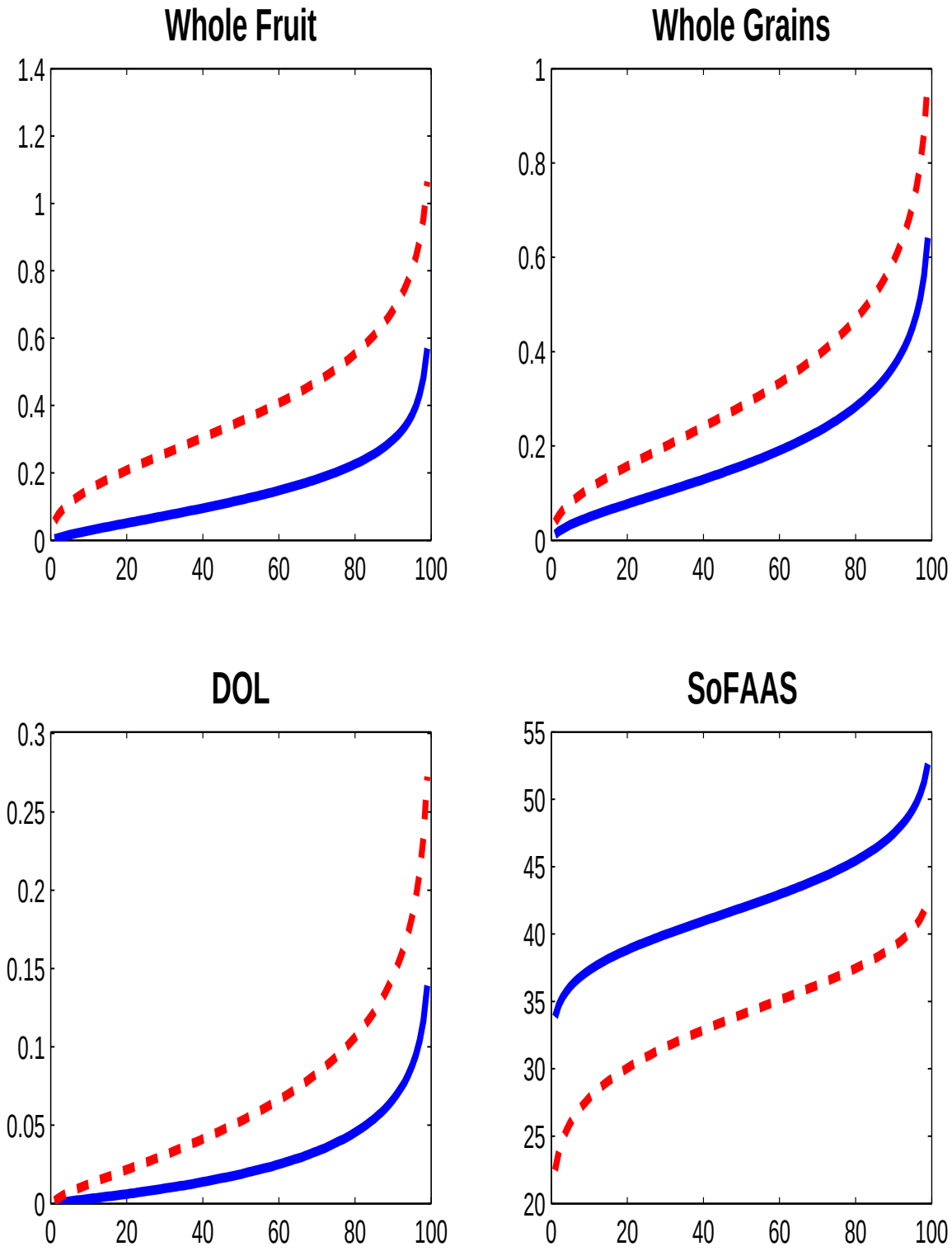

FIG 2. *The estimated percentiles of the energy-adjusted usual intakes for Whole fruits (Top left) in cups/(1000 kcal), Whole grains (Top right) in ounces/(1000 kcal), DOL (bottom left) in cups/(1000 kcal) and calories from SoFAAS (bottom right) in % of Energy. The solid lines are for those whose usual HEI-2005 total score is* ≤ 50*, i.e., poorer diets, while the dashed lines are for those whose usual HEI-2005 total score is* > 50*, i.e., better diets.*

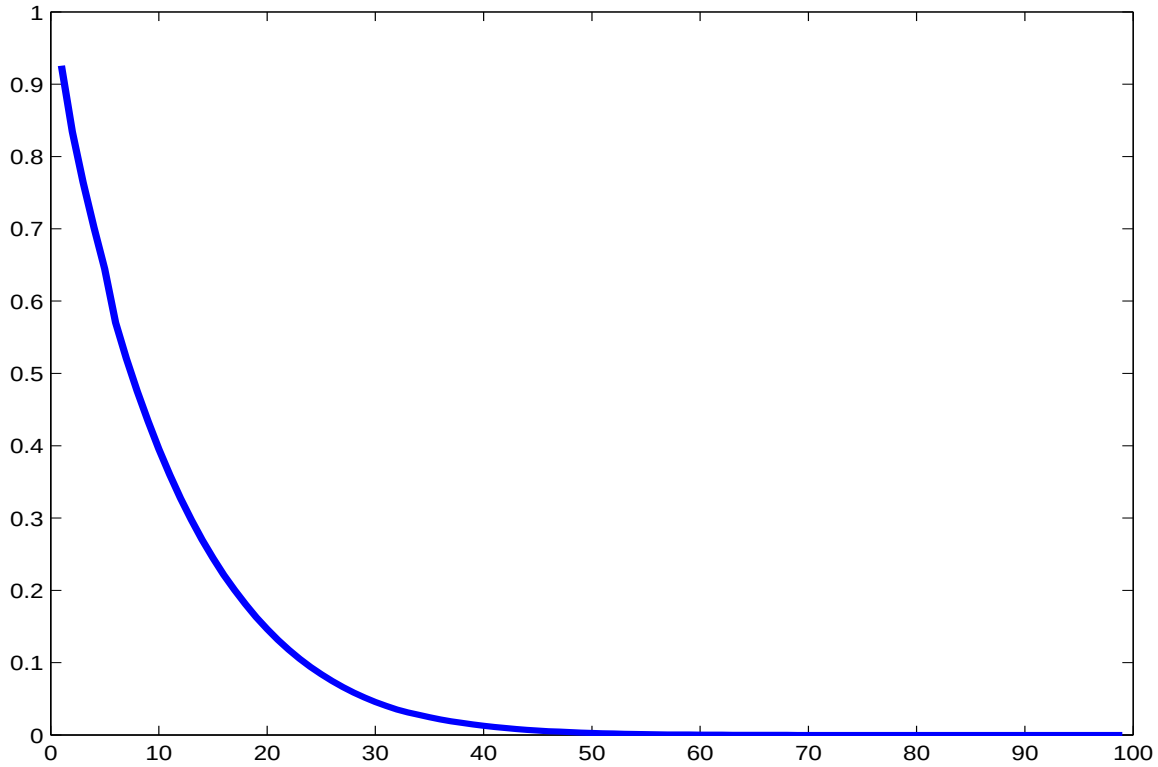

FIG 3. *The Y-axis gives the estimated probabilities of exceeding the κ (X-axis) percentile on all 12 HEI components, for $\kappa = 1, 2, ..., 99$, see Section 5.6.*

the variance of $\widehat{\theta}$ is $(32 \times 0.49)^{-1} \sum_{p=1}^{32} (\widehat{\theta}_p - \widehat{\theta})^2$. The 32 in the denominator is for the 32 different estimates from the 32 different sets of weights, and the 0.49 is the square of the perturbation factor used to construct the BRR weight sets (Wolter, 1995).

6. Further Discussion of the Analysis.

6.1. *Never Consumers.* An aspect of the modeling that we have not discussed is the possibility that some people never, ever consume an episodically consumed dietary component. Our model does not allow for this, for general reasons and for reasons that are specific to our data analysis.

It is in principle possible to add an additional modeling step for non-consumers, via fixed effects probit regression, but we do not think this is a practical issue in our case, for two reasons.

- The first is that the HEI-2005 is based on 6 episodically consumed dietary components, namely total fruit, whole fruit, whole grains, total vegetables, DOL, and milk, the latter of which includes cheese, yogurt and soy beverages. None of these are "lifestyle adverse", unlike say alcohol. While 40% of the responses for whole fruits, for example, equal zero, the percentage of children who never eat any whole fruits at all is likely to be minuscule.
- Even if one disputes whether there are very few individuals who never consume one of the dietary components, then it necessarily follows that we have overestimated the HEI-2005 total scores, and hence the estimates of the proportion of individuals with alarmingly low HEI scores are deflated, and not inflated. The reason is that our model suggests everyone has a positive usual intake of the 6 episodically consumed dietary components. Since the HEI-2005 score components are nondecreasing functions of usual intake of the episodically consumed dietary components, this would mean that we overestimate the HEI-2005 total score.

6.2. *Computing and Data.* Our programs were written in Matlab. The programs, along with the NHANES data we used, are available in the *Annals of Applied Statistics* online archive. Although a much smaller amount of computing effort yields similar results, using $70,000$ MCMC steps with a burn-in of $20,000$ takes approximately 10 hours on a Linux server.

We also estimated the Monte Carlo standard error which is defined by Flegal, et al. (2008) as $\widehat{\sigma}_g/\sqrt{n}$, where n is the total of iterations, and $n = ab$,

where a is the number of blocks and b is the block size, and where

$$\bar{Y}_j = b^{-1} \sum_{i=(j-1)b+1}^{jb} g(X_i) \quad for \quad j = 1, \ldots, a.$$

The batch means estimate of σ_g^2 is

$$\widehat{\sigma}_g^2 = \frac{b}{a-1} \sum_{j=1}^{a} (\bar{Y}_j - \bar{g}_n)^2.$$

The ratio of the Monte Carlo standard error to the estimated standard deviation of the estimated parameters averages 3.4% for Σ_u and 1.7% for β.

Because of the public health importance of the problem, the National Cancer Institute has contracted for the creation of a SAS program that performs our analysis. It will allow any number of episodically and daily consumed dietary components. The first draft of this program, written independently in a different programming language, gives almost identical results to what we have obtained, at least suggesting that our results are not the product of a programming error.

7. Discussion.

7.1. *Transformations.* In Section A.12, we describe how we estimated the transformation parameters as a separate component-wise calculation. We have done some analyses where we simultaneously transform each component, and found very little difference with our results. However, the computing time to implement this is extremely high, because of the fact that different transformations make data on different scales, so we have to compute the usual intakes at each step in the MCMC, and not just at the end.

7.2. *What Have We Learned That Is New.* There are many important questions in dietary assessment that have not been able to be answered because of a lack of multivariate models for complex, zero-inflated data with measurement errors and a lack of ability to fit such multivariate models. Nutrients and foods are not consumed in isolation, but rather as part of a broader pattern of eating. There is reason to believe that these various dietary components interact with one another in their effect on health, sometimes working synergistically and sometimes in opposition. Nonetheless, simply characterizing various patterns of eating has presented enormous statistical challenge. Until now, descriptive statistics on the HEI-2005 have been limited to examination of either the total scores or only a single

energy-adjusted component at a time. This has precluded characterization of various patterns of dietary quality as well as any subsequent analyses of how such patterns might relate to health.

This methodology presented in this paper presents a workable solution to these problems which has already proven valuable. In May 2010, just as we were submitting the paper, a White House Task Force on Childhood Obesity created a report. They had wanted to set a goal of all children having a total HEI score of 80 or more by 2030, but when they learned we estimated only 10% of the children ages 2-8 had a score of 66 or higher, they decided to set a more realistic target. The facility to estimate distributions of the multiple component scores simultaneously will be important in tracking progress toward that goal.

7.3. *In What Other Arenas Will Our Work Have Impact?.* There are many other important problems where multivariate models such as ours will be important. One such problem arises when studying the relationship between multiple dietary components or dietary patterns and health outcomes. Traditionally, for cost reasons, large cohort studies have used a food frequency questionnaire (FFQ) to measure dietary intake, sometimes with a small calibration study including short-term measures such as 24hr recalls. However, there is a new web-based instrument called the Automated Self-administered 24-hour Dietary Recall (ASA24$^{\mathrm{TM}}$),
see http://riskfactor.cancer.gov/tools/instruments/asa24, which has been proposed to replace or at least supplement the FFQ and which is currently undergoing extensive testing. The dietary data we will see then is what we have called Y_{ijk}, i.e., 24hr recall data. In order to correct relative risk estimates for the measurement error inherent in the ASA24$^{\mathrm{TM}}$, regression calibration (Carroll, et al., 2006) will almost certainly be the method of choice, as it is in most of nutritional epidemiology. This method attempts to produce an estimate of the regression of usual intake on the observed intakes, and then to use these estimates in Cox and logistic regression for the health outcome. In order to perform this regression, a multivariate measurement error model will be required, since the regression is on all the observed dietary intake components in the regression model measured by the ASA24$^{\mathrm{TM}}$, and not on each individual component. Our methodology is easily extended to address this problem.

Supplemental Material. We have created supplemental material that has additional tables related to various issues in the analysis, as well as the programs and data that were used.

Acknowledgments. This paper forms part of Zhang's Ph.D. dissertation at Texas A&M University. Zhang and Carroll's research was supported by a grant from the National Cancer Institute (CA57030). This work was also supported by National Science Foundation Instrumentation grant number 0922866.

APPENDIX A: DETAILS OF THE FITTING PROCEDURE

In this Appendix we give the full details of the model fitting procedure.

A.1. Notational Convention. In our example, age was standardized to have mean 0.0 and variance 1.0, to improve numerical stability.

As described in Section 3.1, the observed, transformed non-zero 24hr recalls were standardized to have mean 0.0 and variance 2.0. More precisely, for $\ell = 1, 2, ..., 6$, we first transformed the non-zero food group data as $Z_{i,2\ell,k} = g(Y_{i,2\ell,k}, \lambda_\ell)$, and then we standardized these data as $Q_{i,2\ell,k} = \sqrt{2}\{Z_{i,2\ell,k} - \mu(\lambda_\ell)\}/\sigma(\lambda_\ell)$, where $\{\mu(\lambda_\ell), \sigma(\lambda_\ell)\}$ are the mean and standard deviation of the non-zero food intakes $Z_{i,2\ell,k}$. Similarly, for non-episodically consumed dietary components and energy we transformed to $Z_{i,6+\ell,k} = g(Y_{i,6+\ell,k}, \lambda_\ell)$ for $\ell = 7, ..., 13$, and then standardized to $Q_{i,6+\ell,k} = \sqrt{2}\{Z_{i,6+\ell,k} - \mu(\lambda_\ell)\}/\sigma(\lambda_\ell)$. Of course, whether the food group is consumed or not is $Q_{i,2\ell-1,k} = Y_{i,2\ell-1,k}$ for $\ell = 1, ..., 6$. Collected, the data are $\widetilde{Q}_{ik} = (Q_{ijk})_{j=1}^{19}$. The terms $\{\mu(\lambda_\ell), \sigma(\lambda_\ell)\}$ are not random variables but are merely constants used for standardization, and we need not consider inference for them. Back-transformation is discussed in Appendix A.11.

A.2. Prior Distributions. Because the data were standardized, we used the following conventions.

- The prior for all β_j were normal with mean zero and variance 100.
- The prior for Σ_u was exchangeable with diagonal entries all equal to 1.0 and correlations all equal to 0.50. There were 21 degrees of freedom in the inverse Wishart prior, i.e., $m_u = 21$. Thus, the prior is $\text{IW}\{(m_u - 19 - 1)\Sigma_{u,\text{prior}}, m_u\}$. We experimented with this prior by using zero correlation, and the results were essentially unchanged.
- The prior for r_k is Uniform[-1, 1]. Set the initial value: $r_k = 0$, $k = 1, \ldots, 5$.
- The prior for θ_k is Uniform$[-\pi, \pi]$. Set the initial value: $\theta_k = 0$, $k = 1, \ldots, 25$.
- The priors for $v_{22}, v_{44}, \ldots, v_{12,12}$ and $v_{13,13}, \ldots, v_{19,19}$ were Uniform[-3,3]. Set the initial values: $v_{22} = v_{44} = \ldots = v_{12,12} = v_{13,13} = \ldots =$

$v_{19,19} = 1$.

- For the rest of the non-diagonal v_{ij}'s which could not be determined by the restrictions, we used Uniform[-3,3] priors. Set the initial values to be 0.

The constraints on Σ_ϵ are nonlinear, and our parameterization enforces them easily without having to have prior distributions for the original parameterization that satisfy the nonlinear constraints.

The key thing that makes things work well with the other components of the matrix V with $\Sigma_\epsilon = VV^{\mathrm{T}}$ is that we have standardized the data as described in Section A.1. With this standardization, things become much nicer. For example, the variance of the ϵ's for energy is $\sum_{j=1}^{19} v_{19,j}^2$. However, since the sample variance for energy is standardized to equal 2.0, we simply just need to make priors for $v_{19,j}$ be uniform on a modest range to have real flexibility.

A.3. Generating Starting Values for the Latent Variables. While we observe $\widetilde{Q}_{ik}$, in the MCMC we need to generate starting values for the latent variables $\widetilde{W}_{ik} = (W_{ijk})_{j=1}^{19}$ to initiate the MCMC.

- For nutrients and energy, $Q_{ijk} = W_{ijk}$, no data need be generated, $j = 13, \ldots, 19$.
- For the amounts, Q_{i2k}, Q_{i4k}, Q_{i6k}, Q_{i8k}, $Q_{i,10,k}$ and $Q_{i,12,k}$, we set $W_{i2k} = Q_{i2k}$, $W_{i4k} = Q_{i4k}$, $W_{i6k} = Q_{i6k}$, $W_{i8k} = Q_{i8k}$, $W_{i,10,k} = Q_{i,10,k}$ and $W_{i,12,k} = Q_{i,12,k}$.
- For consumption, we generate $\widetilde{U}_i$ as normally distributed with mean zero and covariance matrix given as the prior covariance matrix for Σ_u. For $\ell = 1, \ldots, 6$, we also compute $z_{ik} = |X_{i,2\ell-1,k}^{\mathrm{T}} \beta_{2\ell-1,\mathrm{prior}} + U_{i,2\ell-1} + \mathcal{Z}_{ik}|$, where $\mathcal{Z}_{ik} = \mathrm{Normal}(0,1)$ are generated independently. We then set $W_{i,2\ell-1,k} = z_{ik} Q_{i,2\ell-1,k} - z_{ik}(1 - Q_{i,2\ell-1,k})$.
- Finally, we then updated $\widetilde{W}_{ik}$ by a single application of the updates given in Appendix A.9.

A.4. Complete Data Loglikelihood. Let $J = 19$. The complete data include the indicators of whether a food was consumed, the W variables, and

the random effect U variables. The loglikelihood of the complete data is

$$\begin{aligned}
&\textstyle\sum_{\ell=1}^{6}\sum_{i=1}^{n}\sum_{k=1}^{m_i}\log\{Q_{i,2\ell-1,k}I(W_{i,2\ell-1,k}>0)+\\
&\qquad\qquad(1-Q_{i,2\ell-1,k})I(W_{i,2\ell-1,k}<0)\}\\
&+(\textstyle\sum_{i=1}^{n}w_i/2)\log(|\Sigma_u^{-1}|)-(1/2)\sum_{i=1}^{n}w_i\widetilde{U}_i^{\mathrm{T}}\Sigma_u^{-1}\widetilde{U}_i\\
&-(1/2)\textstyle\sum_{j=1}^{J}(\beta_j-\beta_{j,\mathrm{prior}})^{\mathrm{T}}\Omega_{\beta,j}^{-1}(\beta_j-\beta_{j,\mathrm{prior}})\\
&+\{(m_u+J+1)/2\}\log(|\Sigma_u^{-1}|)-\{(m_u-J-1)/2\}\mathrm{trace}(\Sigma_{u,\mathrm{prior}}\Sigma_u^{-1})\\
&-(1/2)\textstyle\sum_{i=1}^{n}w_im_i\log\{(v_{22}^2v_{44}^2v_{66}^2v_{88}^2v_{10,10}^2v_{12,12}^2v_{13,13}^2\dots v_{JJ}^2)\prod_{q=1}^{5}(1-r_q^2)\}\\
&-(1/2)\textstyle\sum_{i=1}^{n}w_i\sum_{k=1}^{m_i}\{\widetilde{W}_{ik}-(X_{i1k}^{\mathrm{T}}\beta_1,...,X_{iJk}^{\mathrm{T}}\beta_J)^{\mathrm{T}}-\widetilde{U}_i\}^{\mathrm{T}}\Sigma_\epsilon^{-1}\\
&\qquad\qquad\times\{\widetilde{W}_{ik}-(X_{i1k}^{\mathrm{T}}\beta_1,...,X_{iJk}^{\mathrm{T}}\beta_J)^{\mathrm{T}}-\widetilde{U}_i\}.
\end{aligned}$$

We used Gibbs sampling to update this complete data loglikelihood, the details for which are given in subsequent appendices. The weights w_i are integers and are used here in a pseudo-likelihood fashion. One can also think of this as expanding each individual into w_i individuals, each with the same observed data but different latent variables. For computational convenience, since we are only asking for a frequentist estimator and not doing full Bayesian inference, the latent variables in the process are generated once for each individual. Estimates of Σ_u, Σ_ϵ and β_j for $j=1,...,J$ were computed as the means from the Gibbs samples. Once again, we emphasize that we are not doing a proper Bayesian analysis, but only using MCMC techniques to obtain a frequentist estimate, with uncertainty assessed using the frequentist BRR method.

A.5. Complete Conditionals for r_q, θ_q and v_{pq}. Except for irrelevant constants, the complete conditional for r_q $(q=1,\dots,5)$ is

$$\begin{aligned}
\log[r_q|\mathrm{rest}] &= -\frac{1}{2}\textstyle\sum_{i=1}^{n}w_im_i\log(1-r_q^2)\\
&\quad-\frac{1}{2}\textstyle\sum_{i=1}^{n}w_i\sum_{k=1}^{m_i}\{\widetilde{W}_{ik}-(X_{i1k}^{\mathrm{T}}\beta_1,...,X_{i,19,k}^{\mathrm{T}}\beta_{19})^{\mathrm{T}}-\widetilde{U}_i\}^{\mathrm{T}}\\
&\qquad\qquad\times\Sigma_\epsilon^{-1}\{\widetilde{W}_{ik}-(X_{i1k}^{\mathrm{T}}\beta_1,...,X_{i,19,k}^{\mathrm{T}}\beta_{19})^{\mathrm{T}}-\widetilde{U}_i\}.
\end{aligned}$$

Except for irrelevant constants, the complete conditionals for v_{qq} $(q=2,4,6,8,10,12,13,\dots,19)$ are

$$\begin{aligned}
\log[v_{qq}|\mathrm{rest}] &= -\frac{1}{2}\textstyle\sum_{i=1}^{n}w_im_i\log(v_{qq}^2)\\
&\quad-\frac{1}{2}\textstyle\sum_{i=1}^{n}w_i\sum_{k=1}^{m_i}\{\widetilde{W}_{ik}-(X_{i1k}^{\mathrm{T}}\beta_1,...,X_{i,19,k}^{\mathrm{T}}\beta_{19})^{\mathrm{T}}-\widetilde{U}_i\}^{\mathrm{T}}\\
&\qquad\qquad\times\Sigma_\epsilon^{-1}\{\widetilde{W}_{ik}-(X_{i1k}^{\mathrm{T}}\beta_1,...,X_{i,19,k}^{\mathrm{T}}\beta_{19})^{\mathrm{T}}-\widetilde{U}_i\}.
\end{aligned}$$

Except for irrelevant constants, the compete conditionals for θ_q, ($q = 1, \ldots, 25$) and non-diagonal free parameters v_{pq} are

$$\begin{aligned}\log[x|\text{rest}] &= -\frac{1}{2}\sum_{i=1}^{n} w_i \sum_{k=1}^{m_i} \{\widetilde{W}_{ik} - (X_{i1k}^{\mathrm{T}}\beta_1, ..., X_{i,19,k}^{\mathrm{T}}\beta_{19})^{\mathrm{T}} - \widetilde{U}_i\}^{\mathrm{T}} \\ &\quad \times \Sigma_\epsilon^{-1} \{\widetilde{W}_{ik} - (X_{i1k}^{\mathrm{T}}\beta_1, ..., X_{i,19,k}^{\mathrm{T}}\beta_{19})^{\mathrm{T}} - \widetilde{U}_i\}.\end{aligned}$$

The full conditionals do not have an explicit form, so we use a Metropolis-Hastings within a Gibbs sampler to generate it.

- r_q ($q = 1, \ldots, 5$)
  We discretize the values of r_q to the set $\{-0.99 + 2 \times 0.99(j-1)/(M-1)\}$, where $j = 1, ..., M$ and we choose $M = 41$.
  Proposal: The current value is $r_{q,t}$. The proposed value of $r_{q,t+1}$ is selected randomly from the current value and the two nearest neighbors of $r_{q,t}$. Then $r_{q,t+1}$ is accepted with probability $\min\{1, g(r_{q,t+1})/g(r_{q,t})\}$, where

$$\begin{aligned}&g(y) \propto (1-y^2)^{-\frac{1}{2}\sum_{i=1}^{n} w_i m_i} \\ &\times \exp\left[-\frac{1}{2}\sum_{i=1}^{n} w_i \sum_{k=1}^{m_i} \{\widetilde{W}_{ik} - (X_{i1k}^{\mathrm{T}}\beta_1, ..., X_{i,19,k}^{\mathrm{T}}\beta_{19})^{\mathrm{T}} - \widetilde{U}_i\}^{\mathrm{T}} \Sigma_\epsilon^{-1}(\bullet)\right],\end{aligned}$$

  where here and in what follows, for any A, $A^{\mathrm{T}}\Sigma_\epsilon^{-1}(\bullet) = A^{\mathrm{T}}\Sigma_\epsilon^{-1}A$.
- θ_q ($q = 1, \ldots, 25$)
  We discretize similarly as above.
  Proposal: The current value is $\theta_{q,t}$. The proposed value $\theta_{q,t+1}$ is selected randomly from the current value and the two nearest neighbors of $\theta_{q,t}$. Then $\theta_{q,t+1}$ is accepted with probability $\min\{1, g(\theta_{q,t+1})/g(\theta_{q,t})\}$, where

$$\begin{aligned}&g(y) \\ \propto\ &\exp\left[-\frac{1}{2}\sum_{i=1}^{n} w_i \sum_{k=1}^{m_i} \{\widetilde{W}_{ik} - (X_{i1k}^{\mathrm{T}}\beta_1, ..., X_{i,19,k}^{\mathrm{T}}\beta_{19})^{\mathrm{T}} - \widetilde{U}_i\}^{\mathrm{T}} \Sigma_\epsilon^{-1}(\bullet)\right].\end{aligned}$$

- v_{qq} ($q = 2, 4, 6, 8, 10, 12, 13, \ldots, 19$)
  Proposal: The current value is $v_{qq,t}$. A candidate $v_{qq,t+1}$ is generated from the Uniform distribution of length 0.4 with mean $v_{qq,t}$. The candidate value $v_{qq,t+1}$ is accepted with probability $\min\{1, g(v_{qq,t+1})/g(v_{qq,t})\}$, where

$$\begin{aligned}&g(y) \propto y^{-\sum_{i=1}^{n} w_i m_i} \\ &\times \exp\left[-\frac{1}{2}\sum_{i=1}^{n} w_i \sum_{k=1}^{m_i} \{\widetilde{W}_{ik} - (X_{i1k}^{\mathrm{T}}\beta_1, ..., X_{i,19,k}^{\mathrm{T}}\beta_{19})^{\mathrm{T}} - \widetilde{U}_i\}^{\mathrm{T}} \Sigma_\epsilon^{-1}(\bullet)\right].\end{aligned}$$

- non-diagonal free parameters v_{pq}
  Proposal: The current value is $v_{pq,t}$. The candidate value $v_{pq,t+1}$ is generated from the Uniform distribution of length 0.4 with mean $v_{pq,t}$. The candidate value is accepted with probability $\min\{1, g(v_{pq,t+1})/g(v_{pq,t})\}$, where

$$g(y) \propto \exp\left[-\frac{1}{2}\sum_{i=1}^{n} w_i \sum_{k=1}^{m_i} \{\widetilde{W}_{ik} - (X_{i1k}^{\mathrm{T}}\beta_1, ..., X_{i,19,k}^{\mathrm{T}}\beta_{19})^{\mathrm{T}} - \widetilde{U}_i\}^{\mathrm{T}} \Sigma_\epsilon^{-1}(\bullet)\right].$$

A.6. Complete Conditionals for Σ_u. The dimension of the covariance matrices is $J = 19$. By inspection, the complete conditional for Σ_u is

$$[\Sigma_u|\text{rest}] = \text{IW}\{(m_u - J - 1)\Sigma_{u,\text{prior}} + \sum_{i=1}^{n} w_i \widetilde{U}_i \widetilde{U}_i^{\mathrm{T}}, n + m_u\}$$

where here IW = the Inverse-Wishart distribution. The density of $\text{IW}(\Omega, m)$ for a $J \times J$ random variable is

$$\text{IW}(\Omega, m) = f(Q|\Omega, m) \propto |Q|^{-(m+J+1)/2} \exp\{-\frac{1}{2}\text{trace}(\Omega Q^{-1})\}.$$

This has expectation $\Omega/(m - J - 1)$.

A.7. Complete Conditionals for β. Let the elements of Σ_ϵ^{-1} be $\sigma_\epsilon^{j\ell}$. For any j, except for irrelevant constants,

$$\begin{aligned}
\log[\beta_j|\text{rest}] &= -\frac{1}{2}(\beta_j - \beta_{j,\text{prior}})^{\mathrm{T}} \Omega_{\beta,j}^{-1} (\beta_j - \beta_{j,\text{prior}}) \\
&\quad -\frac{1}{2}\sum_{i=1}^{n} w_i \sum_{k=1}^{m_i} (W_{ijk} - X_{ijk}^{\mathrm{T}}\beta_j - U_{ij})^2 \sigma_\epsilon^{jj} \\
&\quad -\sum_{i=1}^{n} w_i \sum_{k=1}^{m_i} \sum_{\ell \neq j} \sigma_\epsilon^{j\ell} (W_{ijk} - X_{ijk}^{\mathrm{T}}\beta_j - U_{ij}) \\
&\qquad\qquad \times (W_{i\ell k} - X_{i\ell k}^{\mathrm{T}}\beta_\ell - U_{i\ell}) \\
&= \mathcal{C}_1^{\mathrm{T}}\beta_j - \frac{1}{2}\beta_j^{\mathrm{T}}\mathcal{C}_2^{-1}\beta_j,
\end{aligned}$$

which implies $[\beta_j|\text{rest}] = \text{Normal}(\mathcal{C}_2\mathcal{C}_1, \mathcal{C}_2)$, where

$$\begin{aligned}
\mathcal{C}_2 &= \left(\Omega_{\beta,j}^{-1} + \sum_{i=1}^{n} w_i \sigma_\epsilon^{jj} \sum_{k=1}^{m_i} X_{ijk} X_{ijk}^{\mathrm{T}}\right)^{-1}; \\
\mathcal{C}_1 &= \Omega_{\beta,j}^{-1}\beta_{j,\text{prior}} + \sum_{i=1}^{n} w_i \sum_{k=1}^{m_i} \sigma_\epsilon^{jj} X_{ijk}(W_{ijk} - U_{ij}) \\
&\quad + \sum_{i=1}^{n} w_i \sum_{k=1}^{m_i} \sum_{\ell \neq j} \sigma_\epsilon^{j\ell} (W_{i\ell k} - X_{i\ell k}^{\mathrm{T}}\beta_\ell - U_{i\ell}) X_{ijk}.
\end{aligned}$$

A.8. Complete Conditionals for $\widetilde{U}_i$. The NHANES 2001-2004 weights are integers, representing the number of children that each sampled child represents. Thus, as described therein, the loglikelihood in Section A.4 could also be rewritten equivalently by developing w_i pseudo-children, each with the same observed data values. It thus does not make sense to use the weights to generate an individual $\widetilde{U}_i$. Instead, as described in Section A.4, for computational convenience for generating a $\widetilde{U}_i$ to represent w_i children, we set the weight for that child temporarily = 1.0. Then, except for irrelevant constants,

$$\begin{aligned}
\log[\widetilde{U}_i|\text{rest}] &= -\frac{1}{2}w_i\widetilde{U}_i^{\mathrm{T}}\Sigma_u^{-1}\widetilde{U}_i \\
&\quad -\frac{1}{2}w_i\textstyle\sum_{k=1}^{m_i}\{\widetilde{W}_{ik} - (X_{i1k}^{\mathrm{T}}\beta_1, ..., X_{i,19,k}^{\mathrm{T}}\beta_{19})^{\mathrm{T}} - \widetilde{U}_i\}^{\mathrm{T}}\Sigma_\epsilon^{-1} \\
&\qquad\qquad \times\{\widetilde{W}_{ik} - (X_{i1k}^{\mathrm{T}}\beta_1, ..., X_{i,19,k}^{\mathrm{T}}\beta_{19})^{\mathrm{T}} - \widetilde{U}_i\} \\
&= \mathcal{C}_1^{\mathrm{T}}\widetilde{U}_i - \frac{1}{2}\widetilde{U}_i^{\mathrm{T}}\mathcal{C}_2^{-1}\widetilde{U}_i.
\end{aligned}$$

Remembering that for purposes of this section we are setting $w_i = 1.0$, this implies that $[\widetilde{U}_i|\text{rest}] = \text{Normal}(\mathcal{C}_2\mathcal{C}_1, \mathcal{C}_2)$, where

$$\begin{aligned}
\mathcal{C}_2 &= (\Sigma_u^{-1} + m_i\Sigma_\epsilon^{-1})^{-1}; \\
\mathcal{C}_1 &= \textstyle\sum_{k=1}^{m_i}\Sigma_\epsilon^{-1}\{\widetilde{W}_{ik} - (X_{i1k}^{\mathrm{T}}\beta_1, ..., X_{i,19,k}^{\mathrm{T}}\beta_{19})^{\mathrm{T}}\}.
\end{aligned}$$

A.9. Complete Conditional for $W_{i\ell k}$, $\ell = 1, 3, 5, 7, 9, 11$. Here we do the complete conditional for $W_{i\ell k}$ with $\ell = 1, 3, 5, 7, 9, 11$. Except for irrelevant constants,

$$\begin{aligned}
\log[W_{i\ell k}|\text{rest}] &= \log\{Q_{i\ell k}I(W_{i\ell k} > 0) + (1 - Q_{i\ell k})I(W_{i\ell k} < 0)\} \\
&\quad -\frac{1}{2}w_i(W_{i1k} - X_{i1k}^{\mathrm{T}}\beta_1 - U_{i1}, ..., W_{i,19,k} - X_{i,19,k}^{\mathrm{T}}\beta_{19} - U_{i,19}) \\
&\qquad\qquad \times\Sigma_\epsilon^{-1}(\bullet)^{\mathrm{T}} \\
&= \log\{Q_{i\ell k}I(W_{i\ell k} > 0) + (1 - Q_{i\ell k})I(W_{i\ell k} < 0)\} \\
&\quad -\frac{1}{2}w_i\sigma_\epsilon^{\ell\ell}(W_{i\ell k} - X_{i\ell k}^{\mathrm{T}}\beta_\ell - U_{i\ell})^2 \\
&\quad -w_i\textstyle\sum_{j\neq\ell}\sigma_\epsilon^{\ell j}(W_{i\ell k} - X_{i\ell k}^{\mathrm{T}}\beta_\ell - U_{i\ell})(W_{ijk} - X_{ijk}^{\mathrm{T}}\beta_j - U_{ij}) \\
&= \log\{Q_{i\ell k}I(W_{i\ell k} > 0) + (1 - Q_{i\ell k})I(W_{i\ell k} < 0)\} + \mathcal{C}_1 W_{i\ell k} \\
&\qquad\qquad -\frac{1}{2}W_{i\ell k}^2\mathcal{C}_2^{-1},
\end{aligned}$$

where, using the convention of Section A.8,

$$\begin{aligned}\mathcal{C}_2 &= 1/(\sigma_\epsilon^{\ell\ell})\\ \mathcal{C}_1 &= \sigma_\epsilon^{\ell\ell}(X_{i\ell k}^{\mathrm{T}}\beta_\ell + U_{i\ell}) - \textstyle\sum_{j\neq\ell}\sigma_\epsilon^{\ell j}(W_{ijk} - X_{ijk}^{\mathrm{T}}\beta_j - U_{ij}).\end{aligned}$$

If we use the notation $\mathrm{TN}_+(\mu,\sigma,c)$ for a normal random variable with mean μ and standard deviation σ that is truncated from the left at c, and similarly use $\mathrm{TN}_-(\mu,\sigma,c)$ when truncation is from the right at c, then it follows that with $\mu = \mathcal{C}_2\mathcal{C}_1$ and $\sigma = \mathcal{C}_2^{1/2}$,

$$\begin{aligned}[W_{i\ell k}|\mathrm{rest}] &= Q_{i\ell k}\mathrm{TN}_+(\mu,\sigma,0) + (1-Q_{i\ell k})\mathrm{TN}_-(\mu,\sigma,0)\\ &= \mu + Q_{i\ell k}\mathrm{TN}_+(0,\sigma,-\mu) + (1-Q_{i\ell k})\mathrm{TN}_-(0,\sigma,-\mu)\\ &= \mu + Q_{i\ell k}\mathrm{TN}_+(0,\sigma,-\mu) - (1-Q_{i\ell k})\mathrm{TN}_+(0,\sigma,\mu)\\ &= \mu + \sigma\{Q_{i\ell k}\mathrm{TN}_+(0,1,-\mu/\sigma) - (1-Q_{i\ell k})\mathrm{TN}_+(0,1,\mu/\sigma)\}.\end{aligned}$$

Generating $\mathrm{TN}_+(0,1,c)$ is easy: if $c<0$, simply do rejection sampling of a Normal$(0,1)$ until you get one that is $> c$. If $c>0$, there is an adaptive rejection scheme (Robert, 1995).

A.10. Complete Conditionals for W_{i2k}, W_{i4k}, W_{i6k}, W_{i8k}, $W_{i,10,k}$ and $W_{i,12,k}$ When Not Observed. For $p = 2,4,6,8,10,12$, the variable W_{ipk} is not observed when $Q_{i,p-1,k} = 0$, or, equivalently, when $W_{i,p-1,k} < 0$. Except for irrelevant constants,

$$\begin{aligned}\log[W_{ipk}|\mathrm{rest}] &= -\frac{1}{2}w_i\textstyle\sum_j\sum_\ell\sigma_\epsilon^{j\ell}(W_{ijk} - X_{ijk}^{\mathrm{T}}\beta_j - U_{ij})(W_{i\ell k} - X_{i\ell k}^{\mathrm{T}}\beta_\ell - U_{i\ell})\\ &= -\frac{1}{2}W_{ipk}^2\mathcal{C}_2^{-1} + \mathcal{C}_1 W_{ipk},\end{aligned}$$

where, using the convention of Section A.8,

$$\begin{aligned}\mathcal{C}_2 &= 1/(\sigma_\epsilon^{pp});\\ \mathcal{C}_1 &= \sigma_\epsilon^{pp}(X_{ipk}^{\mathrm{T}}\beta_p + U_{ip}) - \textstyle\sum_{\ell\neq p}\sigma_\epsilon^{p\ell}(W_{i\ell k} - X_{i\ell k}^{\mathrm{T}}\beta_\ell - U_{i\ell}).\end{aligned}$$

Therefore,

$$[W_{ipk}|\mathrm{rest}] = Q_{ipk}Q_{i,p-1,k} + (1-Q_{i,p-1,k})\mathrm{Normal}(\mathcal{C}_2\mathcal{C}_1,\mathcal{C}_2).$$

A.11. Usual Intake, Standardization and Transformation. Here we present detailed formulas for functions defined in Section 3.4. When $\lambda = 0$, the back-transformation is

$$\begin{aligned}g_{\mathrm{tr}}^{-1}(z,0) &= \exp\left\{\mu(0) + \sigma(0)z/\sqrt{2}\right\};\\ \partial^2 g_{\mathrm{tr}}^{-1}(z,0)/\partial z^2 &= \frac{\sigma^2(0)}{2}g_{\mathrm{tr}}^{-1}(z,0).\end{aligned}$$

When $\lambda \neq 0$, the back-transformation is

$$\begin{aligned} g_{\mathrm{tr}}^{-1}(z,\lambda) &= \left[1+\lambda\left\{\mu(\lambda)+\sigma(\lambda)z/\sqrt{2}\right\}\right]^{1/\lambda}; \\ \partial^2 g_{\mathrm{tr}}^{-1}(z,\lambda)/\partial z^2 &= \frac{\sigma^2(\lambda)}{2}(1-\lambda)\left[1+\lambda\left\{\mu(\lambda)+\sigma(\lambda)z/\sqrt{2}\right\}\right]^{-2+1/\lambda}. \end{aligned}$$

A.12. Transformation Estimation. As part of an earlier project (Freedman, et al., 2009), we estimated the transformations for one food/nutrient at a time using the method of Kipnis, et al. (2009), both for the data and also for each BRR weighted data set. To facilitate comparison with the one food/nutrient at a time analysis, in our analysis of all HEI-2005 components, we used these transformations as well. Of course, our methods can be generalized to allow for estimation of the transformations as well. By allowing a different transformation for each BRR weighted data set, we have captured the variation due to estimation of the transformations.

SUPPLEMENTARY MATERIAL

Supplementary Material for A New Multivariate Measurement Error Model with Zero-Inflated Dietary Data, and its Application to Dietary Assessment:
(http://???/???). ???

REFERENCES

Buonaccorsi, J. (2010). *Measurement Error: Models, Methods and Applications*. Chapman and Hall/CVRC Press.

Carriquiry, A. L. (1999) Assessing the prevalence of nutrient inadequacy. *Public Health Nutrition*, 2, 23-33.

Carriquiry, A. L. (2003). Estimation of usual intake distributions of nutrients and foods. *Journal of Nutrition*, 133, 601-608.

Carroll, R. J., Ruppert, D., Stefanski, L. A. and Crainiceanu, C. M. (2006). *Measurement Error in Nonlinear Models: A Modern Perspective*, Second Edition. Chapman and Hall CRC Press.

Delaigle, A. (2008). An alternative view of the deconvolution problem. *Statistica Sinica*, 18, 1025-1045.

Delaigle, A. and Hall, P. (2010). Estimation of observation-error variance in errors-in-variables regression. *Statistica Sinica*, to appear.

Delaigle, A. and Hall, P. (2008). Using SIMEX for smoothing-parameter choice in errors-in-variables problems. *Journal of the American Statistical Association*, 103, 280-287.

Delaigle, A., Hall, P. and Meister, A. (2008). On deconvolution with repeated measurements. *Annals of Statistics*, 36, 665-685.

Delaigle, A. and Meister, A. (2008). Density estimation with heteroscedastic error. *Bernoulli*, 14, 562-579.

Ferrari, P, Roddam, A, Fahey, M. T., Jenab, M., Bamia, C., Ock, M., Amiano, P., Hjartker, A., Biessy, C., Rinaldi, S., Huybrechts, I., Tjnneland, A., Dethlefsen, C., Niravong, M., Clavel-Chapelon, F., Linseisen, J., Boeing, H., Oikonomou, E., Orfanos, P., Palli, D., Santucci de Magistris, M., Bueno-de-Mesquita, H. B., Peeters, P. H., Parr, C. L., Braaten, T., Dorronsoro, M., Berenguer, T., Gullberg, B., Johansson, I., Welch, A. A., Riboli, E., Bingham, S. and Slimani, N. (2009). A bivariate measurement error model for nitrogen and potassium intakes to evaluate the performance of regression calibration in the European Prospective Investigation into Cancer and Nutrition study. *European Journal of Clinical Nutrition*, 63, Supplement 4, S179-187.

Flegal, J. M., Haran, M. and Jones, G. L. (2008). Markov Chain Monte Carlo: can we trust the third significant figure? *Statistical Science*, 23, 250-260.

Fraser, G. E. and Shavlik, D. J. (2004). Correlations between estimated and true dietary intakes. *Annals of Epidemiology*, 14, 287-95.

Freedman, L. S., Guenther, P. M., Krebs-Smith, S. M., Dodd, K. W. and Midthune D. (2010). A population's distribution of Healthy Eating Index-2005 component scores can be estimated when more than one 24-hour recall is available. *Journal of Nutrition*, 140, 1529-1534.

Fuller, W. A. (1987). *Measurement Error Models*. Wiley, New York.

Fungwe, T., Guenther, P. M., Juan, W. Y., Hiza, H, and Lino, M. (2009). The quality of children's diets in 2003-04 as measured by the Healthy Eating Index-2005. *Nutrition Insight*, 43, USDA Center for Nutrition Policy and Promotion.

Guolo, A. (2008). A flexible approach to measurement error correction in casecontrol studies. *Biometrics* 64, 1207-1214.

Gustafson, P. (2003). *Measurement Error and Misclassification in Statistics*

and Epidemiology: Impacts and Bayesian Adjustments. Chapman and Hall/CRC Press.

Guenther, P. M., Reedy, J. and Krebs-Smith, S. M. (2008a). Development of the Healthy Eating Index-2005. *Journal of the American Dietetic Association*, 108, 1896-1901.

Guenther, P. M., Reedy, J., Krebs-Smith, S. M. and Reeve, B. B. (2008b). Evaluation of the Healthy Eating Index-2005. *Journal of the American Dietetic Association*, 108, 1854-1864.

Kipnis, V., Midthune, D., Buckman, D. W., Dodd, K. W., Guenther, P. M., Krebs-Smith, S. M., Subar, A. F., Tooze, J. A., Carroll, R. J. and Freedman, L. S. (2009). Modeling data with excess zeros and measurement error: application to evaluating relationships between episodically consumed foods and health outcomes. *Biometrics*, 65, 1003-1010.

Kipnis, V., Freedman, L. S., Carroll, R. J. and Midthune, D. (2010). A measurement error model for episodically consumed foods and energy. Preprint.

Kott, P. S., Guenther, P. M., Wagstaff, D. A., Juan W. Y. and Kranz, S. (2009). Fitting a linear model to survey data when the long-term average daily intake of a dietary component is an explanatory variable. *Survey Research Methods*, Vol 3, No 3, 157-165.

Küchenhoff, H., Mwalili, S. M., & Lesaffre, E. (2006). A general method for dealing with misclassification in regression: The misclassification SIMEX. *Biometrics*, 62, 85-96.

Lehmann, E. L. and Casella, G. (1998). *Theory of Point Estimation*. Springer, New York.

Liang, H., Thurston, S., Ruppert, D., Apanasovich, T. and Hauser, R. (2008). Additive partial linear models with measurement errors. *Biometrika*, 95, 667-678.

Messer, K. and Natarajan, L. (2008). Maximum likelihood, multiple imputation and regression calibration for measurement error adjustment. *Statistics in Medicine*, 27, 6332-6350.

Natarajan, L. (2009). Regression Calibration for Dichotomized Mismeasured Predictors. *International Journal of Biostatistics*, 5, nihpa121098.

Nusser, S. M., Carriquiry, A. L., Dodd, K. W., and Fuller, W. A. (1996). A semiparametric approach to estimating usual intake distributions.*Journal*

of the American Statistical Association, 91, 1440-1449.

Nusser, S. M., Fuller, W. A., and Guenther, P. M. (1997). Estimating usual dietary intake distributions: Adjusting for measurement error and non-normality in 24-hour food intake data. In Lyberg, L, Biemer, P, Collins, M, Deleeuw, E, Dippo, C, Schwartz, N, and Trewin, D (editors). *Survey Measurement and Process Quality*, pp.670-689, New York: Wiley, 1997.

Prentice, R. L. (1996). Measurement error and results from analytic epidemiology: dietary fat and breast cancer. *Journal of the National Cancer Institute*, 88, 1738-47.

Prentice, R.L. (2003). Dietary assessment and the reliability of nutritional epidemiology reports. *Lancet*, 362, 182-183.

Staudenmayer, J., Ruppert, D., and Buonaccorsi, J.P. (2008). Density estimation in the presence of heteroskedastic measurement error. *Journal of the American Statistical Association*, 103, 726-736.

Tooze, J. A., Grunwald, G. K., and Jones, R. H. (2002). Analysis of repeated measures data clumping at zero. *Statistical Methods in Medical Research* ,11, 341-355.

Tooze, J. A., Midthune, D., Dodd, K. W., Freedman, L. S., Krebs-Smith, S. M., Subar, A. F., Guenther, P. M., Carroll, R. J. and Kipnis, V. (2006). A new statistical method for estimating the distribution of usual intake of episodically consumed foods. *Journal of the American Dietetic Association*, 106, 1575-1587.

Wand, M. P. (1998). Finite sample performance of deconvolving kernel density estimators. *Statistics and Probability Letters*, 37, 131-139.

Wolter, K. M. (1995). *Introduction to Variance Estimation*. Springer-Verlag, New York.

Zhang, S. Midthune, D., Pérez, A, Buckman, D. W., Kipnis, V., Freedman, L. S., Dodd, K. W., Krebs-Smith, S. M. and Carroll, R. J. (2010). A bivariate measurement error model for episodically consumed dietary components.

S. Zhang
R.J. Carroll
Department of Statistics
Texas A&M University
3143 TAMU
College Station, Texas 77843-3143
U.S.A.
E-mail: sjzhang@stat.tamu.edu
E-mail: carroll@stat.tamu.edu

D. Midthune
V. Kipsnis
K. Dodd
Biometry Research Group
Division of Cancer Prevention
National Cancer Institute
6130 Executive Boulevard EPN-3131
Bethesda, Maryland 20892-7354
U.S.A.
E-mail: midthund@mail.nih.gov
E-mail: kipnisv@mail.nih.gov
E-mail: doddk@mail.nih.gov

P.M. Guenther
Center for Nutrition Policy and Promotion
U.S. Department of Agriculture
3101 Park Center Drive, Ste. 1034
Alexandria, Virginia 22302
U.S.A.
E-mail: Patricia.Guenther@cnpp.usda.gov

S. Krebs-Smith
Applied Research Program
Division of Cancer Control and Population Sciences
National Cancer Institute
6130 Executive Boulevard, EPN-4005
Bethesda, Maryland 20892, U.S.A.
E-mail: krebssms@mail.nih.gov

D. Buckman
Information Management Services, Inc.
12501 Prosperity Drive
Silver Spring, Maryland 20904, U.S.A.
E-mail: BuckmanD@imsweb.com

J. Tooze
Department of Biostatistical Sciences
Wake Forest University, School of Medicine
Medical Center Boulevard
Winston-Salem, North Carolina 27157, U.S.A.
E-mail: jtooze@wfubmc.edu

L. Freedman
Gertner Institute for Epidemiology and Health Policy Research
Sheba Medical Center
Tel Hashomer 52161, Israel
E-mail: lsf@actcom.co.il