Online Supporting Material

Appendix A
Details of the Statistical Model for HEI-2005 Components Involving Episodically Consumed Food Groups

For individual i, i = 1, …, n, let

T_{Fi} = true usual intake of (episodically consumed) dietary component,

T_{Ei} = true usual intake of energy,

R_{Fij} = 24-hour recall (24HR)-reported intake of dietary component on day j, j = 1, …, m

R_{Eij} = 24HR-reported intake of energy on day j, j = 1, …, m

X_{ij} = vector of covariates relevant to day j, j = 1, …, m

Often, e.g. for NHANES, m= 2.

We assumed that the 24HR is unbiased on the original scale,

$E(R_{Fij}) = T_{Fi}$,

$E(R_{Eij}) = T_{Ei}$,

and also assumed the following three-part model for 24HR:

$$\{\text{logit}(p_i) \mid X_{ij}, U_{1i}\} = \beta_{10} + \beta_{11}X_{ij} + U_{1i},$$

$$\{g(R_{Fij}; \lambda_F) \mid R_{Fij} > 0, X_{ij}, U_{2i}\} = \beta_{20} + \beta_{21}X_{ij} + U_{2i} + \varepsilon_{2ij},$$

$$\{g(R_{Eij}; \lambda_E) \mid X_{ij}, U_{3i}\} = \beta_{30} + \beta_{31}X_{ij} + U_{3i} + \varepsilon_{3ij}, \qquad (1)$$

where p_i is the probability that $R_{Fij} > 0$, $\text{logit}(p) = \log\{p/(1-p)\}$ is the inverse of the logistic distribution function, $g(t; \lambda) = (t^{\lambda} - 1) / \lambda$ is a Box-Cox transformation, λ_F and λ_E are parameters that transform positive R_{Fij} and R_{Eij} to (approximate) normality, (U_{1i}, U_{2i}, U_{3i}) are random effects that have a joint normal distribution with mean zero, ($\varepsilon_{2ij,}$ ε_{3ij}) are within-person random errors that have a joint normal distribution with mean zero, variances $\sigma^2_{\varepsilon_2}, \sigma^2_{\varepsilon_3}$ and correlation ρ_{12}, and ($\varepsilon_{2ij,}$ ε_{3ij}) are independent of (U_{1i}, U_{2i}, U_{3i}). In

addition, values of (ε_{2ij}, ε_{3ij}) are independent across repeats. The terms β_{10}, β_{20} and β_{30} are scalars. In this application, X is a vector of covariates, including: weekday/weekend, sequence number and dummy variables for age groups. The terms β_{11}, β_{21} and β_{31} are also vectors, with the same number of elements as X.

We assumed that λ_F and λ_E are known, although in practice they have to be estimated. The method used in this application was to choose the value that minimized the mean square error about the best straight line on the QQ plot, using the sampling weights to calculate the distribution points on the plot.

Model (1) was fitted using the SAS NLMIXED procedure (assuming known transformations λ_F and λ_E), with survey sampling weights being incorporated into the estimation procedure.

For this application, we write $X_{ij} = \{W_{ij}, S_{ij}, Z_i\}$, $\beta_{11} = \{\beta_{1W}, \beta_{1S}, \beta_{1Z}\}$, $\beta_{21} = \{\beta_{2W}, \beta_{2S}, \beta_{2Z}\}$ and $\beta_{31} = \{\beta_{3W}, \beta_{3S}, \beta_{3Z}\}$, where W_{ij} is an indicator for whether the reported day was on a weekend, S_{ij} is an indicator for sequence number being equal to 2, and Z_i is a vector of other covariates (for example age-group) that do not depend on j. Then under the model assumptions, usual intake for person i is given by:

$$T_{Fi} = E[H(\beta_{10} + \beta_{1W}W_{ij} + \beta_{1Z}Z_i + U_{1i}) \times g^{-1}(\beta_{20} + \beta_{2W}W_{ij} + \beta_{2Z}Z_i + U_{2i} + \varepsilon_{2ij}; \lambda_F) \mid Z_{ij}, U_{1i}, U_{2i}]$$

$$\approx (3/7)\, H(\beta_{10} + \beta_{1W} + \beta_{1Z}Z_i + U_{1i}) \times g^*(\beta_{20} + \beta_{2W} + \beta_{2Z}Z_i + U_{2i}; \lambda_{F;}\, \sigma^2_{\varepsilon_2}) +$$

$$(4/7)\, H(\beta_{10} + \beta_{1Z}Z_i + U_{1i}) \times g^*(\beta_{20} + \beta_{2Z}Z_i + U_{2i};\ \lambda_F;\ \sigma^2_{\varepsilon_2}),$$

(2)

$$T_{Ei} = E[g^{-1}(\beta_{30} + \beta_{3W}W_{ij} + \beta_{3Z}Z_i + U_{3i} + \varepsilon_{3ij};\ \lambda_E) \mid Z_{ij},\ U_{3i}]$$

$$\approx (3/7)g^*(\beta_{30} + \beta_{3W} + \beta_{3Z}Z_i + U_{3i};\ \lambda_E,\ \sigma^2_{\varepsilon_3}) + (4/7)g^*(\beta_{30} + \beta_{3Z}Z_i + U_{3i};\ \lambda_E,\ \sigma^2_{\varepsilon_3}),$$

where $g^*(\nu;\ \lambda, \sigma^2_\varepsilon)$ is a Taylor series approximation of the expectation $E\{g^{-1}(\nu + \varepsilon_{ij};\ \lambda_R) \mid \nu\}$,

$$g^*(\nu;\ \lambda, \sigma^2_\varepsilon) = g^{-1}(\nu;\ \lambda) + \frac{1}{2}\sigma^2_\varepsilon \frac{\partial^2\{g^{-1}(\nu;\lambda)\}}{\partial \nu^2}.$$

Note that the linear combination weighted by 3/7 and 4/7 respectively refers to an assumption that the weekend comprises Friday, Saturday and Sunday, with the covariate coded so that weekday=0 and weekend=1. The coefficients for sequence number (β_{1S}, β_{2S} and β_{3S}) do not appear because this covariate is coded so that the first administration of the instrument=0 and the second administration=1, and because we use the level of the first report to estimate true intake (i.e., we assume the first report is unbiased).

Formula (2) was used to simulate pairs of true usual intakes of episodically consumed dietary component and energy. For each pseudo-person a vector of random effects (U_{1i}, U_{2i}, U_{3i}) was drawn from a trivariate normal distribution with mean zero and variance-covariance matrix as estimated from the NLMIXED fit. Then the pair of true intakes (T_{Fi}, T_{Ei}) was calculated from (2), using parameters as estimated from that same fit. Finally, the ratio T_{Fi}/T_{Ei} and the HEI-2005 score were calculated. This process was repeated N times for each person i ($i=1,\ldots..,n_k$) in age group k in the sample, and each of these values

received the sampling weight for that person. The empirical distribution of the HEI-2005 score was then formed from the Nn_k values with their sampling weights. Specifically, the Nn_k HEI-2005 scores were placed in ascending order, and the series of cumulative sums of the sampling weights was computed. If S is the total sum of the sampling weights over all Nn_k pseudo-individuals, then the p^{th} percentile was estimated as the HEI-2005 score corresponding to the pseudo-individual whose cumulative sampling weight sum was closest to pS/100. The number of repetitions, N, was chosen to provide a sufficiently large sample to calculate the empirical distribution precisely. For the purposes of the work reported in the paper, we used N=100.