ATTACHMENT C.
ANALYTIC APPROACHES

1. Reliability in Item Response Theory (IRT) Model

Classical test theory-based reliability indices such as Cronbach's Alpha are not appropriate when 1) the length of the test may differ for each respondent, and 2) the test used Item Response Theory (IRT) methods, where the measurement error is a function of the latent construct level (θ). The reliability coefficient for the AJC measure is, therefore, calculated as the *marginal reliability* (Sireci, Thissen, & Wainer, 1999), which is equivalent to internal consistency estimates of reliability.

First we determine the marginal measurement error variance, $\bar{\sigma}_{e^*}^2$, across all respondents,

$$\bar{\sigma}_{e^*}^2 = \int \sigma_{e^*}^2 \, p(\theta) d\theta = \frac{\sum \sigma_{e^*}^2}{N},$$

where $\sigma_{e^*}^2$ is the square of the standard error of the latent construct estimate, $\hat{\theta}$. Thus, the marginal measurement error variance can be estimated as the average of the squared standard error of $\hat{\theta}$. Then, we estimate the marginal reliability as:

$$\bar{\rho} = \left(\sigma_{\hat{\theta}}^2 - \bar{\sigma}_{e^*}^2\right) / \sigma_{\hat{\theta}}^2$$

where $\sigma_{\hat{\theta}}^2$ is the variance of the observed $\hat{\theta}$ estimates.

Coefficient H

The coefficient H[1] is the measure of *stability* of a construct *as reflected in the data* on the chosen indicators, which is the squared correlation between the latent construct and the optimum linear composite from the survey items. The coefficient H indicates the maximum reliability of construct measured by a set of survey items. The coefficient H is defined as:

$$H = \frac{1}{1 + \frac{1}{\sum_{i=1}^{p} \frac{l_i^2}{1 - l_i^2}}}$$

[1] Hancock, G. R., & Mueller, R. O. (2001). Rethinking construct reliability within latent variable systems. In R. Cudeck, S. du Toit, & D. Sörbom (Eds.), Structural Equation Modeling: Present and Future — A Festschrift in honor of Karl Jöreskog. Lincolnwood, IL: Scientific Software International, Inc.

2. Non Response Weighting: Mean and Variance

For the derivation of the non-response weighted mean and variance, the IMPAQ team will utilize the following equations (Little &Vartivarian, 2005).

The weight mean, $\bar{y}_w$, is defined as:

$$\bar{y}_w = \sum_{c=1}^{C} P_c \bar{y}_{0c} = \sum_{c=1}^{C} w_c P_{0c} \bar{y}_{0c}$$

where C indicates respondents and non-respondents, P_c is the proportions of sampled cases for group c, w_c is the nonresponse weights, and $\bar{y}_{0c}$ is the respondent mean for group c. Then the mean of Y (outcome) for respondents, μ_0, and non-respondents, μ_1, becomes:

$$\mu_0 = \sum_{c=1}^{C} \pi_{0c} \mu_{0c}$$

and $\mu_1 = \sum_{c=1}^{C} \pi_{1c} \mu_{1c}$

where π_{0c} is the probability of being in the respondent group and π_{0c} is the probability of being in the non-respondent group.

The variance of weighted mean, $V(\bar{y}_w$¿, is defined as:

$$V(\bar{y}_w) = (1+\lambda)\sigma^2 / n_0 + \sum_{c=1}^{C} \pi_c (\mu_{0c} - \tilde{\mu}_0) / n$$

where λ is the variance of the nonresponse weights, w_c, and $\tilde{\mu}_0$ is the respondent mean adjusted for the covariates.

3. Partial Credit Model

The partial credit model (PCM; Masters, 1982; Wright & Maters, 1982) is a polytomous extension of the Rasch model (Rasch, 1960) or 1-parameter logistic model (Birnbaum, 1968). The AJC survey includes both dichotomously scored questions. For example, for dichotomously-scored questions 0 will be assigned for No and 1 for Yes. For polytomously-scored questions 1 will be assigned for rarely or not at all, 2 for some of the time, 3 for most of the time and 4 for always. The PCM is an appropriate model to estimate the continuous latent construct estimates (i.e. accessibility level) from ordered categorical responses (i.e. both dichotomously and polytomously scored questions. This model assumes that only item response defines the individual proficiency and all item responses are independent conditional on the individual proficiency. The PCM model is typically defined as:

$$p_{nik} = \frac{\exp \sum_{k=0}^{k} (\theta_n - \delta_{ik})}{\sum_{k=0}^{m_i} \exp \sum_{k=0}^{k} (\theta_n - \delta_{ik})} \qquad (1)$$

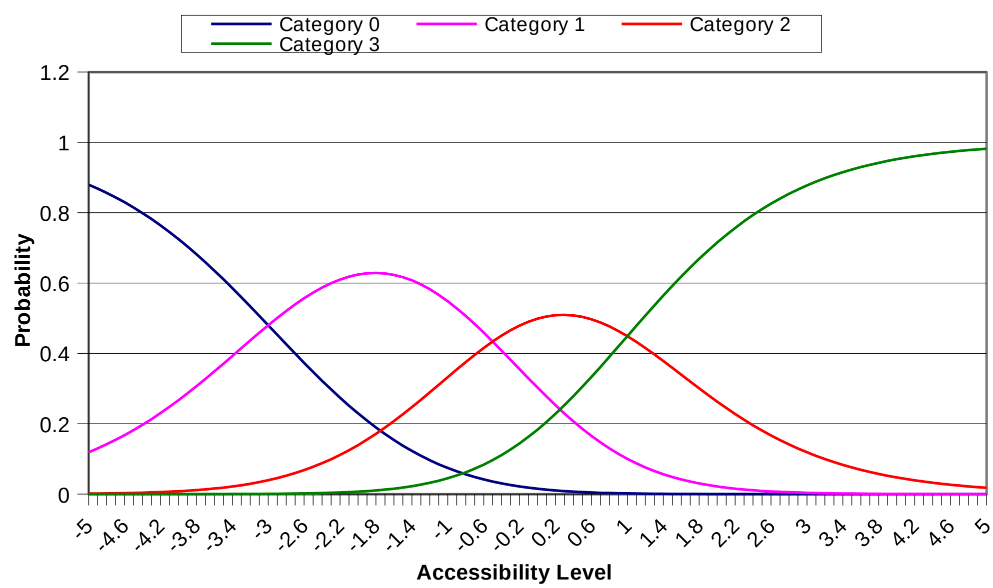
where p_{nik} is the probability of person n scoring k on item i, $\theta_n$is the estimate of the latent construct level, and δ_{ik} refers to the step value (i.e. item difficulties). The better way to interpret the formula is to convert it into log-odds or logit form. The logit form of PCM equation is,

$$logit(p_{nik}) = \ln\left(\frac{p_{nik}}{p_{nik-1}}\right) = \theta_n - \delta_{ik} \qquad (2)$$

wherep_{nik} is the probability of response for person n to the item i in category k, $\theta_n$is a latent ability or proficiency of person/job center n, and δ_{ik} is item estimates on category k in item i.

PCM estimates the expected response probability of each response categories given the latent construct score (see Exhibit C1), characterizing the properties of item. The height of each line is a probability of endorsing each category given the latent construct level. The difficulty or step parameter, δ_{ij}, is where two successive curve intersects.

Exhibit C1: Category Response Curve



4. Facet Model

The IMPAQ team will use the extension of the Partial Credit Model because AJC scale development requires evaluating and accounting for the impact of external factors, such as SDR and SNR. As statistical remedies for SDR and SNR effects, the IMPAQ team will use the Facet model (Linacre, 1989; Eckes, 2011) which allows for the examination of the effect of external factors such as respondent types and survey non-respondents. The Facet model provides an avenue to explore and correct for bias in survey responses. The Facet model estimates the magnitude of the responder's bias (i.e., responder as a facet) within the IRT framework (see Equations 3-6). The Facet model (equations 4-6) and PCM model (equation 3) provides the same accessibility estimates when there is no 'facet' in the model (e.g. SDR and SNR) or if facet effects (e.g. SDR effect) is zero.

$$logit(p) = \theta - \delta \qquad (3)$$

$$logit(p) = \theta - \delta - SDR \qquad (4)$$

$$logit(p) = \theta - \delta - SNR \qquad (5)$$

$$logit(p) = \theta - \delta - SDR - SNR \qquad (6)$$

Facet models will provide the parameter estimates for all SDR and SNR levels and the resulting accessibility construct estimates, θ, which will be adjusted for SDR and/or SNR effects. The magnitude of SDR and SNR parameters will indicate the severity of the SDR and SNR effect.

Other IRT models for polytomous responses such as Rating Scale Model (RSM) or Generalized Partial Credit model (GPCM) are also considered. We have determined neither RSM nor GPCM meets our objectives because: 1) RSM was not appropriate due to the presence of mixed response categories (e.g. binary questions and sliding scale questions) in surveys and 2) inability for GPCM to accommodate other facets such as SDR and SNR effects.

5. Item Parameter Estimation with Missing Responses

IRT models including PCM and Facet modes provides the estimates of a unique set of item parameters measuring a different range of latent constructs of interest. Estimation of item parameters involves maximizing a likelihood function:

$$L(\beta, \theta; x) = P_{\theta,\beta}(x) = \prod_i P_{\theta_i,\beta} = \prod_i \prod_j P_{\theta_i,\beta_j}(x_{ij}) \qquad (7)$$

with respect to the parameters **β** and **θ,** where **β** is a vector of step and facet parameters and **θ** is a vector of latent construct scores. A likelihood function is computed for each respondent given a response vector. The equation 7 implies that the estimation of item parameters utilizes all information available (i.e. missing responses are excluded from the construction of the likelihood function), allowing to respond to different set of survey questions.

This property allows the implementation of the matrix survey deployment plan. Responses from web surveys and site visits will be combined into one dataset with the indicator clearly identify the mode of data collection. This dataset will be estimated jointly providing all parameters on the same scale (i.e. item parameters, latent construct estimates and facet parameters), which eliminates the use of post-hoc adjustment or triangulation of survey scale.

Equation 8 is maximized by the Marginal Maximum Likelihood (MML) estimation (Bock and Aiken, 1981), which produces the consistent estimates of item parameters via the EM-algorithm. The MML integrates out the latent construct parameter, **θ**, to obtain the marginal distribution of the response pattern **X:**

$$P_{\beta,\gamma}(x)=\int P_{\beta,\gamma}(x,\theta)\,d\theta=\prod_i \int P_\beta(x_i \vee \theta_i)\,g_\gamma(\theta_i)\,d\theta_i \quad (8)$$

, where $g_\gamma(\theta)$ is a probability density function of the θ distribution which is distributed normally with $\gamma\ \ (\mu,\sigma^2)$. . Both **β** and γ parameters are simultaneously estimated in MML by maximizing the marginal probability of the observed response pattern x with respect to the parameters **β** andγ**:**

$$L_m(\beta,\gamma;x)=\prod_i \int P_\beta(x_i \vee \theta_i)\,g_\gamma(\theta_i)\,d\theta_i \qquad (9)$$

6. Estimation of Latent Score

Suppose that a AJC staff is presented *N* survey questions, indexed $i=1,2,\ldots,N$, and the AJC staff's response categories on these survey questions are given in the score vector $z=(z_1,\ldots,z_i,\ldots,z_N)$. Let each survey question be characterized by a vector of item parameters $b_i=(b_{i1},\ldots,b_{iK_i})'$, where K_i is the maximum category for survey question i, and collect all item parameters in the vector $b=(b_1^{'},\ldots,b_i^{'},\ldots b_N^{'})$. Then the response likelihood function of PCM model is computed as:

$$L(\theta\,|\,\mathbf{z},\mathbf{b})=\prod_{i\in\{Presented\_items\}}\left[Exp\left(z_i\theta-\sum_{k=1}^{z_i} b_{ik}\right)\Bigg/\left\{1+\sum_{j=1}^{K_i} Exp\left[\sum_{k=1}^{j}(\theta-b_{ik})\right]\right\}\right].$$

The method calculates the IRT construct score, θ, which maximizes the likelihood function for the web survey's observed response vector. The Newton-Raphson method is used to determine the theta value from the likelihood function.

Each unique response pattern and set of presented items gives rise to a distinct shape of likelihood function. Thus, the scoring method takes into account both the number of item score points the respondent endorsed as well as the difficulty of the items the AJC staff was given. This property allows having different versions of surveys, including unique responses from site visits as well as computing subscale scores such as accessibility subscale scores for a different type of disabilities. We will examine if the subscale scores can characterize the accessibility level of AJC for various types of disabilities.

7. Estimation of Error

The standard error estimates (i.e. standard error of parameters, SE) in IRT is a function of set of item responses endorsed by respondents[2], which is a square root of an inverse of sum of information provided by all responded items expressed in the equation below. In IRT, each item provides the *information* about the respondent's construct level or ability, represented by the item information, $I_i(\theta)$. The standard error of estimate, $SE(\theta)$, is the square root of the inverse of sum of information across all responded items:

$$SE(\theta)=\sqrt{\frac{1}{I(\theta)}}=\sqrt{\frac{1}{\sum_i I_i(\theta)}}$$

Where

$$I_i(\theta)=\sum_{k=1}^{m_i} P_{ik}(\theta)\left[\frac{-\partial^2}{\partial\theta^2}\ln P_{ik}(\theta)\right]=\text{¿}\sum_{c=1}^{m_i}\left[T_c-\sum_{c=1}^{m_i}T_c P_{ic}(\theta)\right]^2 P_{ic}(\theta)\text{¿}$$

where **T** is a vector of survey responses unique to each respondent such as **T**=(1, 3, 4, 3, 2, ….,3) and the response probability of survey question *i* endorsing category *k* given the latent score of θ, $P_{ik}(\theta)$, comes from equation C1 on Appendix C. A benefit of the IRT model is that it generates the standard errors of estimates (i.e. precision of estimates) based on the pattern of endorsed survey question responses[2].

8. IRT Fit Indices

Fit-statistics are valuable indices for examining item and person responses. Fit statistics provide information about how well items and respondents fit the IRT model. The un-weighted fit may be referred as Outfit mean-square and weighted fit as Infit mean-square that was originally proposed by Wight and Masters (1982).

These statistics indicated the discrepancy between observed item responses and the predicted item responses based on the IRT model for each item. Both fit statistics had an expected value

[2] Embretson, S.E. (1996). The new rules of measurement. Psychological Assessment, Vol 8(3), 341-349.

of 1. Values substantially greater than 1 indicated unmodeled variation (model underfit), and values less than 1 indicated a lack of stochasticity (model overfit) (Wright & Masters, 1982).
Infit values are generally related to item construction that is sensitive to consistency of item responses. Outfit values when high indicate unexpected responses, such as careless mistakes on the easiest items by students of known higher ability or guessing on hard items by students of known lower ability (Meijer & Sijtsma, 2001). The IRT analysis also provides fit statistics to the latent construct parameters, which potentially identifies erratic respondents (e.g. random responses) or highly deterministic response patterns (e.g. potential sign of systematic biases such as SDR). Adequate range of fit indices for survey is suggested between 0.6 to 1.4 (Bond & Fox, 2001; Smith, Schumacker, & Bush, 1998). WINSTEPS (Linacre, 2013) manual also provides a guideline for fit evaluation.