## Introduction

Measurement noninvariance; also labeled measurement bias, differential item functioning (DIF), and measurement heterogeneity; is present on survey instruments when individuals equivalent on a measured construct but from different groups do not have identical probabilities of producing observed scores (Mellenbergh 1989). For example, consider a set of survey items examining whether or not individuals were able to meet a number of household expenses. Measurement noninvariance would be present if men and women who were both factually unable to meet an expense nevertheless answered the question differently. Perhaps women would answer in the negative (they were unable to meet the expense) and men the affirmative (they were able to meet the expense). Thus, despite equivalence in their factual values, observed self-report responses, conditioned on group membership, would not reflect equivalence.

As a source of measurement error and non-sampling error (Biemer and Lyberg 2003; Biemer, Groves, Lyberg, Mathiowetz, and Sudman 1991; Fuller 1987), measurement noninvariance has the potential to negatively affect the quality of survey data. Survey quality refers to a variety of attributes across data collection, analysis, and dissemination (Biemer and Lyberg 2003), but generally refers to data, analyses, and reports that are objective, and thus accurate and unbiased (Federal Register 2002). To the extent that the surveys used to gather data are comprised of biased items across subpopulations, differences or similarities in the observed scores across subpopulations may reflect differential item functioning rather than true findings, and the quality of the data will be compromised. Likewise, to the extent that measurement *invariance* is present, *i.e.*, individuals equivalent on a given construct but from different groups *do* have identical probabilities of producing observed scores, greater faith can be placed in statistical estimates generated from the data.

Not surprisingly, Federal statistical agencies are interested in and required to establish the quality of data collected in Federal surveys (Federal Register 2002). The U.S. Census Bureau is no exception, and it has developed data quality standards, regularly undertakes data quality studies, and includes reports on the quality of data it collects. For example, the American Community Survey (ACS) is a relatively new survey under development at the Census Bureau and is intended to replace the decennial census long form (U.S. Census Bureau 2001). It will include items collecting data on a number of social, demographic, economic, and housing characteristics of the U.S. population, and will become the primary source of inter-censal information describing assorted characteristics of the U.S. population.

Given the magnitude of influence that estimates from this survey will have, the Census Bureau has conducted numerous studies seeking to establish the quality of data collected on the ACS, across both broad and specific criteria. These studies have generally compared the ACS and Census 2000, and, although some small differences have been noted, generally these studies have demonstrated excellent data quality for the ACS as compared to Census 2000 (U.S. Census Bureau 2001, 2002a, 2003a, 2004a, 2004b, 2004c, 2004d, 2004e, 2004f, 2004g). Nevertheless, the Census Bureau is continually interested in verifying the accuracy of data collected on the ACS, and data quality studies have not abated.

Of particular interest are assessments concerning the quality of data collected across English and non-English speaking households. Estimates from the Census 2000 suggest that  approximately 47 million people aged five years and older speak a language other English at home, and that approximately 12 million of these individuals are linguistically isolated, living in households in which no individual aged 14 or over speaks English "very well" (U.S. Census Bureau 2002b; U.S. Census Bureau 2002c). Given the potential problems that individuals whose primary language is not English may experience in responding to the ACS, the Census Bureau is expressly concerned with ascertaining the quality of data collected across individuals living in linguistically isolated (LI) and non-LI households (McGovern 2004). To this end, measurement invariance studies can be employed to examine whether measurement invariance exists across these groups, assessing whether individuals equivalent on the constructs measured by the ACS and living in LI and non-LI households are equally likely to produce identical item responses. Thus, the goal of the current study was to exploit an advanced psychometric technique, measurement modeling, to probe for measurement bias across individuals living in LI and non-LI households on a set of items selected from the ACS, thereby addressing internal validity, one dimension of data quality, on the survey.

Measurement models, in the form of latent variable models (LVMs) such as confirmatory factor analysis (CFA) for continuous or ordered-categorical measures, item response theory (IRT) and latent class analysis (LCA), offer a powerful method for examining invariance (Waller, Thompson, and Wenk 2000). Resulting from recent advances in the psychometric tradition (McDonald 1999), these methods posit equations to describe the relations among a set of items and examine whether the relevant parameters in the these equations are statistically equivalent across the groups (Byrne 1998; Clogg and Goodman 1985; Cheung and Rensvold 2002; Muthen 2002). To the extent that the parameters in these equations are different, measurement bias is present.

Although rarely employed in large sample survey research, assessments of this kind have seen increased use in psychological settings. These studies have shown that noninvariance can attenuate or accentuate group differences (Cole 1999; Huang, Church, and Katigbak 1997; Pentz and Chou 1994; Smith and Reise 1998; Waller, et al. 2000), lead to inaccurate diagnoses (Cole, Martin, Peeke, Henderson, and Harwell 1998; Gallo, James, and Muthen 1994; Reid, DuPaul, George, Power, Thomas, Anastopoulos, Rogers-Adkinson, Noll, and Riccio 1998, Waller, et al. 2000), and generally decrease reliability and validity (Byrne and Baron 1993; Byrne, Baron, and Campbell 1993; Byrne, Baron, and Campbell 1994; Byrne, Baron, and Baley 1996; Byrne, Baron, and Baley 1998; Byrne and Campbell 1999; Knight and Hill 1998; Schaefer and Caetano 1996). Studies have also uncovered bias so profound as to render cross group comparisons virtually impossible (Knight, Yun-Tein, Shell, and Roosa 1992; Prelow, Yun-Tein, Roosa, and Wood 2000).

CFA for continuous measures (CFA-CM) is perhaps the most commonly used of the LVMs to assess measurement invariance (Byrne 1998) and can be appropriate for survey data. Unfortunately, few researchers or survey methodologists receive training in LVMs (Embretson and Hershberger 1999). As such, although the general factor analytic models and the subsequent method for testing invariance are covered in detail elsewhere

(Bollen 1989; Byrne, 1998; Cheung and Rensvold 2002; Millsap and Yun-Tein 2004), given the general lack of familiarity, they are briefly reviewed here to aid readers' understanding and interpretation of the current study.

CFA, either for continuous or ordered-categorical measures, indicates a set of equations to describe the relations among items and provides a focused set of hypotheses to test measurement invariance. In developing CFA-CM, let $X_{ij}$ equal the *i*th individual's score on the *j*th item, let the number of items be *p* (*j* = 1, 2, .., p), and suppose some set of *r* factors, ξ, is responsible for the observed scores. The model suggests that $X_{ij}$ is related to the factor(s) as follows:

$$X_{ij} = \tau_j + \lambda_j' \xi_i + u_{ij}, \qquad (1)$$

where $\tau_j$ is a latent intercept parameter, $\lambda_j'$ is an *r* x 1 vector of factor loadings for the *j*th variable on *r* factors, $\xi_i$ is the *r* x 1 vector of factor scores for the *i*th person, and $u_{ij}$ is the *j*th unique factor score for that person.

The intercept parameter is similar to the intercept in simple regression and gives the expected value of the item when the value of $\xi$ is zero. The loadings are similar to the correlation coefficient in simple regression and represent the degree to which an item is related to the factor(s); the greater the value of the factor loading, the greater the relation between the item and the latent variable of interest. Finally, the unique factors include sources of variance not attributable to the factor(s), and include measurement error (Bollen, 1989).

Following distributional assumptions (Bollen, 1989), two equations give the factor structure:

$$\mathrm{E}(X_i) = \mu = \tau + \Lambda\kappa \text{ , and} \quad (2)$$

$$\mathrm{Cov}(X_i) = \Sigma = \Lambda\Phi\Lambda' + \Theta \,. \quad (3)$$

where $X_i$ is a $p$ x 1 vector of measured variables, $\tau' = \{\tau_1, \tau_2, ..., \tau_p\}$ is a vector of factor intercepts, Λ is the $p$ x $r$ factor pattern matrix whose *j*th row is $\lambda_j'$, κ is an $r$ x 1 vector of factor means, Φ is an $r$ x $r$ factor covariance matrix, and Θ is a $p$ x $p$ covariance matrix of the unique factors. When the data are of a continuous nature, these parameters are sufficient to model the mean and covariance matrix of the observed items. However, psychometric and survey data are frequently of an ordered-categorical nature (McDonald, 1999), and applying the continuous model to discrete data can lead to spurious multidimensionality (Bernstein and Teng 1989), as well as incorrect standard error estimates and tests of model fit (Joreskog 1990; Muthen 1984).This has resulted in CFA-CM's extension to CFA for ordered-categorical measures (CFA-OCM; Christoffersson 1975; Muthen 1978; Bartholomew 1980; Bartholomew 1984; Muthen 1984; Mislevy 1986).

In the CFA-OCM model, $X_{ij}$ is an ordered-categorical item with scores that range {0, 1, …, *c*}, where *c* is the largest possible score[1]. The discrete observed responses

[1] For the sake of simplicity, the discussion that follows motivates a model where the values of *c* are invariant across items. However, the general ordered-categorical model allows values of *c* to range across items.

are assumed to be determined by scores on a continuous latent response variate, $X^*_{ij}$, such that threshold values on $X^*_{ij}$ determine responses. If an individual's value on $X^{*}{}_{ij}$ is less than the threshold, they will respond in one category, but, if $X^*_{ij}$ is greater than the threshold, they will respond in at least the next highest category. Formally:

$$X_{ij} = m \text{ if } \nu_{jm} \le X^*_{ij} \le \nu_{j(m+1)}, \qquad (4)$$

where, $m = 0, 1, .., c$ and $\{\nu_{j0}, \nu_{j1}, .., \nu_{j(c+1)}\}$ are latent threshold parameters for the $j$th item. The probabilities associated with observed values of $X_{ij}$ are determined by the probability distribution assumed for $X^*_{ij}$. Once $X^*_{ij}$ is defined, the factor model described above emerges, and, after similar distributional and relational assumptions, the usual factor structure appears, modeling values of the latent response variate:

$$\mathrm{E}\left(X^*_i\right) = \mu^* = \tau + \Lambda\kappa \text{ , and} \qquad (5)$$

$$\mathrm{Cov}\left(X^*_i\right) = \Sigma^* = \Lambda\Phi\Lambda' + \Theta \qquad (6).$$

The threshold parameters have the interpretation just noted, and the interpretation of the remaining parameters mirrors that in the continuous model. In this way, OCM models the mean and covariance matrix of the observed items. Model suitability is addressed by comparing the difference between the covariance structure implied by the model to the covariance structure of the sample data through the use of goodness-of-fit-indices (GFIs; Bollen 1989; Byrne 1998; Hu and Bentler 1998; Cheung and Rensvold 2002; Millsap and Yun-Tein 2004). To examine measurement invariance, the measurement parameters are subscripted ($g$) to allow group differences, and, to the extent that cross group constraints in the model parameters $\left(\tau_g, \Lambda_g, \Theta_g, \left\{\nu_{jg0}, \nu_{jg1}, \ldots, \nu_{jg(c+1))}\right\}\right)$ lead to problematic GFIs, measurement bias is present.

In practice, invariance in the full set of measurement parameters is not tested simultaneously; rather, a series of hierarchically nested models with increasing cross-group constraints in the measurement parameters is examined (Bollen 1989; Byrne 1998; Cheung and Rensvold 2002; Millsap and Yun-Tein 2004). In this method, the researcher first begins with the least restricted cross-group model and successively adds cross-group equivalence constraints in subsequent models, assessing bias in each set of measurement parameters separately. Specifically, a baseline model first statistically identifies the latent variable across groups and tests the hypothesis of configural invariance. Configural invariance specifies a similar cross-group model form, such that the items are associated with and yield the same number of factors across groups (Meredith 1993), and assesses whether participants belonging to different groups conceptualize the construct in a similar way. If the set of GFIs associated with the model is unacceptable (Bolen 1989; Byrne 1998; Cheung and Rensvold 2002; Hu and Bentler 1998), the hypothesis of configural invariance is rejected, and measurement bias is present. Rejection of this hypothesis is often taken as evidence that basic conception of the variable under study differs across the groups (Millsap and Everson 1991). However, if the GFIs indicate acceptable fit, the hypothesis of configural invariance is not rejected and a more restricted model stipulating metric invariance may be examined.

Metric invariance constrains the factor loadings to equivalence across groups. This test allows the possibility that, although the items measure similar constructs and are associated with the same factors across groups, the extent to which individual items are related to a specific factor may differ across the groups. Again, the GFIs are examined, and changes in the GFIs (ΔGFIs) are investigated as well. ΔGFIs describe the change in model fit between the more restricted model and the less restricted model. ΔGFIs that are sufficiently large suggest model misfit as a result of the increased constraints. A problematic set of GFI and ΔGFI values suggests that the equivalence constraints are not tenable and measurement bias is present in the loadings (Cheung and Rensvold 2002). To the extent that the increased constraints do not lead to problematic changes in fit, the hypothesis of metric invariance is not rejected and a yet more constrained model may be examined.

In the next model the analyst may specify cross-group invariance in the thresholds, and examine the GFIs and ΔGFIs comparing the constrained thresholds model to the constrained loadings model. Again, to the extent that the GFIs and ΔGFIs do not indicate problematic changes in fit, the hypothesis of measurement invariance in these parameters is not rejected and a yet more constrained model may be considered. Likewise, if the set of GFIs and ΔGFIs indicate misfit, measurement noninvariance is present. This process continues in a similar manner until invariance in the entire set of measurement parameters of interest has been examined.

The method does not require a specific order in which constraints be added (Bollen 1989; Cheung and Rensvold 2002; Millsap and Yun-Tein 2004). For example, invariance in the thresholds may be examined prior to invariance in the loadings. The approach also does not require that each of the parameters in a set be constrained to equivalence before a more constrained model can be considered (Byrne 1998). This situation, partial measurement invariance, allows the possibility that, while bias exists, it is not present across the entire set of parameters under inspection. Perhaps only one of the loadings is biased across the groups for example.

The technique does require that each succeeding model be nested within the previous model, such that the more restricted model is simply a more constrained version of model in the previous step and includes no additional or previously unspecified parameters (Bollen 1989). To the extent that the constraints in any given step lead to problematic misfit, measurement bias is present; individuals equivalent on the construct of interest do not have identical probabilities of producing observed scores. As a function of group membership, the data are not collecting information in an objective, accurate, and unbiased manner, and data quality is compromised. Through this method psychometric measurement models offer a powerful technique to address data quality concerns, and the current study adopted such an approach.

In summary, the purpose of the current study was to address a data quality concern on the ACS. Previous research has not examined the cross-group measurement properties of data collected on the ACS across individuals living in LI and non-LI households, leaving unclear whether users of the ACS need be apprehensive that the quality of data collected across these groups is more accurate for one group as compared to the other. To this end, CFA-OCM was used to probe for measurement bias on a set of items selected from the ACS across individuals living in LI and non-LI households.

## Methods

### *Participants*

Participants, whose sociodemographic characteristics are described in detail elsewhere (U.S. Census Bureau 2004h), were a subset of the larger 2002 ACS Supplemental Survey, an operational feasibility test of the ACS. The complete set of individuals who reported on themselves and for whom LI data were available were included in the current study. In the sample of 482,684 reference persons, 428,268 individuals living in non-LI households and 15,245 living in LI households met these criteria ( $n_{total}$ = 443,513). Given that missingness on LI solely accounted for exclusion in the subsample, it was not possible to address item nonresponse across LI. Previous research, however, has suggested that differences in overall item nonresponse rates across LI are generally slight (McGovern 2004).

### *Procedures*

The ACS 2002 Supplemental Survey; an operational feasibility test of the planned ACS using the ACS survey design, method, and questionnaire; collected housing and sociodemographic information for up to five household residents at a sampled address. Selected from the Master Address File (MAF) maintained by the Census Bureau to represent each county in the US, the ACS will use a rolling unclustered one-stage systematic sample design (Alexander 2001; Kish 1998; U.S. Census Bureau 2003b). Areas with lower mail response rates will be oversampled to minimize the impact of differential response (U.S. Census Bureau 2003b). With some small exceptions, among them the exclusion of group quarters, the ACS 2002 Supplemental Survey reflected this design (U.S. Census Bureau 2003b).

Three distinct modes collected data in the sample: self-enumeration, computer assisted telephone interviewing (CATI), and computer assisted personal interviewing (CAPI). Mailable addresses were first contacted through an English language self report questionnaire, and unmailable addresses were initially contacted through CAPI. Approximately six weeks after the questionnaire's mailing, nonresponding sample addresses were contacted via CATI, which included Spanish language assistance but no support for other non-English languages. Following failed CATI contact attempts, one in three of the remaining uninterviewed addresses were contacted through CAPI, the last nonresponse follow-up. These methods resulted in an overall response rate of 97.7% U.S. (Census Bureau 2003i).

### *Measures*

### *Disability*

Given the assumptions of the CFA model (described above), it was not possible to assess invariance in the total set of ACS items simultaneously. As such, six items assessing disability, described in detail in Appendix A, were selected from the ACS for use in the current study. Participants responded to one item assessing whether or not they had a long-lasting condition that resulted in a vision or hearing impairment and a second item examining whether they had a long-lasting condition that substantially limited a variety of physical activities. Participants also answered four questions addressing whether or not a physical, mental, or emotional condition lasting 6 months or more resulted in difficulty: 1) learning, remembering or concentrating; 2) dressing, bathing, or getting around inside the home; 3) going outside the home alone to shop or visit a doctor's office; 4) and/or working at a job or business.

### *Linguistic Isolation*

Consistent with the Census Bureau's official definition, LI was computed as a function of two questions. Respondents were first asked whether they spoke a language other than English at home. When a language other than English was spoken at home, participants were asked how well they and other members of the household aged 14 or over spoke English: "Very Well", "Well", "Not Well", or "Not at All". Individuals who spoke a language other than English at home and lived in households were all individuals aged 14 and over spoke English less than very well were coded as LI, all other individuals were coded as non-LI.

## Results

### *General Analytic Strategy*

Analyses proceeded in two stages. First, given that previous research had not examined the factor structure of the items under study, it was necessary to establish the adequacy of the baseline model (Byrne 1998). To this end, a random sample of 5000 interviews was selected from the original sample and CFA-OCM examined the fit of a single factor baseline model. In the second stage, CFA-OCM was used to investigate the invariance of the measurement properties of the baseline model across individuals living in LI and non-LI households in the full sample.

All analyses were conducted using Mplus (3.1; Muthen and Muthen 2004). Measurement invariance was examined following the methods described by Millsap and Yun-Tein (2004), Byrne (1998), and Cheung and Rensvold (2002). Preferred levels of fit for indices of global and local model fit were adopted *a priori* and followed those suggested by Hu and Bentler (1998), Muthen and Muthen (2002; 2004), Steiger (1998), and Cheung and Rensvold (2002). The $\Delta\chi^2$ value was determined using the method described by the Carle (2005). Given the $\chi^2$ and $\Delta\chi^2$'s functional dependence on *N* and sensitivity to trivial deviations of fit (Cheung and Rensvold, 2002), overall model and invariance model assessments were conducted using the set of indices. Models were rejected when the majority of indices indicated inadequate fit. Means and covariances were included at each step, the models were identified as described by Millsap and Yun-Tein (2004), and the theta parameterization and robust weighted least squares (WLSMV) estimator were used in all analyses. To address Type I error given the number of comparisons in the study, an $\alpha$ of 0.01 was adopted for all comparisons.

### *Baseline Model Establishment*

The appropriateness of a single factor model, hypothesizing that the covariance among items was accounted for by a single underlying disability factor, was examined in a randomly selected sample of 5000 interviews. For statistical identification, the factor mean was fixed to zero and the factor variance was fixed to one. No other constraints were placed on the model. Although the $\chi^2$ test of exact fit was significant ($\chi^2$= 62.66, 8, $n$ = 5000, $p$ < 0.001), the remaining fit indices demonstrated excellent fit to the data: RMSEA = 0.037, CFI = 0.99, McDonald's NCI = 0.99, and Gamma Hat = 0.996. Given the set of indices, the single factor baseline model was not rejected and analyses turned to assessing measurement invariance.

### ***Measurement Invariance Analyses***

The configural invariance of the single factor baseline model was tested across the LI and non-LI groups. For statistical identification: the factor mean was fixed at zero for the non-LI group, the factor variance was fixed at one for the non-LI group, item intercepts were constrained to zero in each group, the loading for the "work" item was constrained to equality across the groups, the threshold for the "work" item was constrained to equality across the groups, and the diagonal elements of $\Theta$ were fixed to a value of one in each group. Tables 1 and 2 give the relevant descriptive statistics for each group. Table 3 summarizes the fit results of the baseline and succeeding models. For the baseline model, the $\chi^2$ test of exact fit was significant ($\chi^2$ = 6282.03, 16, n = 443,513, $p$ < 0.001). However, the remaining set of fit indices suggested the data were well fit by the model specifying configural invariance (RMSEA = 0.042, CFI = 0.99, McDonald's NCI = 0.99, and Gamma Hat = 0.995); the hypothesis of configural invariance was not rejected; and metric invariance was examined next.

This model retained the configural invariance restraints in the previous model, constrained the loadings to equality across the groups, and allowed variation in the remaining parameters. The set of fit indices continued to suggest excellent fit: RMSEA = 0.034, CFI = 0.99, ΔCFI < 0.01, McDonald's NCI = 0.99, Δ McDonald's NCI < 0.01, Gamma Hat = 0.996, ΔGamma Hat = 0.001, $\chi^2$ = 4866.41 (19, n = 443,513, *p* < 0.001), and $\Delta\chi^2$ = 12.964 (4, n = 5400, *p* = 0.011). Consequently, the hypothesis of metric invariance was not rejected and analyses moved to examining invariance in the thresholds.

The invariant thresholds model retained the constraints in the invariant loadings model and constrained the thresholds to equality. Again, the set of fit indices demonstrated good fit to the data: RMSEA = 0.032, CFI = 0.99, ΔCFI < 0.01, McDonald's NCI = 0.99, Δ McDonald's NCI < 0.01, Gamma Hat = 0.996, ΔGamma Hat < 0.001, $\chi^2$ = 4881.82 (22, n = 443,513, *p* < 0.001), and $\Delta\chi^2$ = 37.84 (5, n = 5400, *p* < 0.001). Modification indices (MIs) and expected parameter change indices (EPCs) suggested that constraining the threshold for the "going out" item was predominantly responsible for the increase in misfit observed in the $\Delta\chi^2$ (MI = 16.10), overestimating the level of the disability needed before LI individuals would endorse this item as compared to non-LI individuals. However, as noted, the $\Delta\chi^2$'s is sensitive to small deviations of fit, and the remaining ΔGFI did not indicate problematically increased misfit. As such, constraining the threshold for the "going out" item was considered tenable and the model was not rejected. Analyses next turned to examining invariance in the uniquenesses across the groups.

To examine invariance in the uniquenesses, a new model hierarchy was motivated. For dichotomous models, it is not possible to statistically identify a model that simultaneously allows variation in the loadings, thresholds, and uniquenesses. Thus, the baseline model specified above initially constrained the uniquenesses across the groups and allowed variation in the remaining parameters. By incorporating the constraints in the loadings and thresholds just described, it was possible to establish a new “baseline” model that constrained the loadings and thresholds across groups and allowed variation in the uniquenesses. The fit of this model could then be compared to the fit of a model that constrained the loadings, thresholds, and uniquenesses across the groups.

The variant uniquenesses model relaxed the cross group uniqueness constraints of the invariant thresholds model described above, retained the remaining constraints, and fit the data well: RMSEA = 0.036, CFI = 0.99, McDonald’s NCI = 0.99, Gamma Hat = 0.996, and $\chi^2$ = 5295.77 (18, n = 443,513, $p < 0.001$). The fit of this model was compared to a fully invariant model that included cross group constraints in the uniquenesses. The set of fit indices demonstrated that the additional cross group constraints in the uniquenesses did not lead to problematic misfit: RMSEA = 0.032, CFI = 0.99, ΔCFI < 0.01, McDonald’s NCI = 0.99, Δ McDonald’s NCI < 0.01, Gamma Hat = 0.996, ΔGamma Hat < 0.001, $\chi^2$ = 4981.82 (22, n = 443,513, $p < 0.001$), and $\Delta\chi^2$ = 20.64 (5, n = 5400, $p < 0.001$). The MIs suggested that the source of misfit noted in the $\Delta\chi^2$ resulted from the constraints in the thresholds (as discussed above), rather than from the constraints in the uniquenesses. Thus, given the final set of fit indices, the fully invariant measurement model specifying invariance in the loadings, thresholds, and uniquenesses was not rejected. The unstandardized estimates for this model are summarized in Tables 4-8.

INSERT TABLES 1 THROUGH 8 ABOUT HERE

## Discussion

The goal of this study was to use LVM methods to offer one assessment of the internal validity and quality of data collected on the ACS. To this end, CFA-OCM examined whether six items measuring disability provided equivalent measurement across individuals living in LI and non-LI households. With respect to the adopted statistical criteria, results demonstrated that the disability items on the ACS demonstrate invariant measurement across individuals living in LI and non-LI households, supporting the notion that the quality of data collected with this item set is equivalent across these groups and that measurement bias is not present.

In the analyses, the baseline model fit well in both groups and the majority of GFIs and ΔGFIs found no differences in the loadings, thresholds, and uniquenesses across LI and non-LI individuals. Substantively this suggests that, regardless of LI status, the items should reflect similar degrees of disability, item responses should correspond to the same category at any given level of disability, and similar amounts of unique variance should be associated with items. These findings refute the hypothesis that LI substantially affects the ability of individuals living in LI households to answer these items on the ACS in a meaningful way. Rather, the findings suggest that individuals living in LI households interpret and respond to the disability items in a manner similar to individuals living in non-LI households. Thus, when making comparisons, researchers need not be concerned about possible negative effects of LI on the data. Despite what would seem an impediment to understanding the survey questions at hand, individuals living in LI households appear able to gather the resources necessary to provide valid answers as compared to individuals living in non-LI households.

However, it is important to note some of the study's limitations. First, CFA methods assume that an underlying, unobserved variable or dimension accounts for observed item responses, making it impossible to examine the entire set of ACS items simultaneously. Invariance was examined only for the disability set of items and the ability of the study to address to the quality of data across individuals living in LI and non-LI households on the remaining set of ACS items in this study was limited. Future assessments of this sort should be conducted on other item sets. Second, as a first step, the study investigated only individuals' responses about themselves and did not examine their responses describing other household members. Given the current findings of invariance, it is desirable to explore whether these findings extend to individuals' reports on other household members across LI status. By incorporating a multilevel procedure that addresses non-independence among the observations (Muthen 2002), future studies may examine this issue. Third, the study did not examine whether these findings hold across mode of data collection. The possibility exists that the quality of data collected in mail-out, mail-back surveys differs as compared to CATI and/or CAPI collected data for LI individuals, such that mail-out, mail-back methods do not provide equivalent measurement as compared to CATI and/or CAPI methods. Fourth, the study examined only LI status, and did not examine other groups across which data quality concerns might arise. Future studies may benefit by addressing these concerns.

In conclusion, the results of the current study affirm the quality of disability data collected across individuals living in LI and non-LI households on the ACS. When making comparisons regarding disability, investigators using ACS data can be less concerned that LI status impacts the validity of comparisons and can place greater faith in the quality of their comparisons. Importantly, these findings contribute to a growing literature generally establishing the quality of data collected on the ACS, across LI status or otherwise. Although other important aspects of data quality have yet to be addressed, the current investigation demonstrates the utility of latent variable models for assessing data quality in large scale survey research and provides support that the ACS likely provides an internally valid measurement of disability across individuals living in LI and non-LI households.

## References

Alexander, Charles. H. 2001. “Still rolling: Leslie Kish’s “Rolling Samples” and the American Community Survey.” *Proceedings of the Statistics Canada Symposium 2001: Achieving Data Quality in a Statistical Agency, A Methodological Perspective.* Ottawa, Canada.

Bartholomew, David. J. 1980. “Factor analysis for categorical data.” *Journal of the Royal Statistical Society*, Series B, 42, 293-321.

Bartholomew, David. J. 1984. “Scaling binary data using a factor model.” *Journal of the Royal Statistical Society*, Series B, 46, 120-123.

Bernstein, Ira. H. and Gary Teng 1989. “Factoring items and factoring scales are different: Evidence for multidimensionality due to item categorization.” *Psychological Bulletin*, 105, 465-477.

Biemer, Paul. P, and Lars. E. Lyberg, 2003. *Introduction to Survey Quality*. Hoboken, NJ: Wiley.

Biemer, Paul. P, Robert. M. Groves, Lars. E. Lyberg, Nancy. A. Mathiowetz, and Seymour Sudman. eds. 1991. *Measurement Error in Surveys*. New York: Wiley.

Bollen, Kenneth. A. 1989. *Structural Equations with Latent Variables*. John Wiley and Sons, New York.

Byrne, Barbara. 1998. *Structural Equation Modeling with LISREL, PRELIS, and SIMPLIS: Basic applications and programs.* Lawrence Erlbaum, New Jersey.

Byrne, Barbara. M., and Pierre Baron. 1993. The Beck Depression Inventory: Testing and cross-validating a hierarchical factor structure for nonclinical adolescents. *Measurement and Evaluation in Counseling and Development*, 26, 164-178.

Byrne, Barbara. M., Pierre Baron, and Jorj Baley. 1998. The Beck Depression Inventory A cross-validated test of second-order factorial structure for Bulgarian adolescents. *Educational and Psychological Measurement*, 58, 241-251.

Byrne, Barbara. M., Pierre Baron, and Jorj Baley. 1996. The Beck Depression Inventory: Testing for its factorial validity and invariance across gender for Bulgarian non-clinical adolescents. *Personality and Individual Differences*, 21, 641-651.

Byrne, Barbara. M., Pierre Baron and T. Leanne Campbell. 1993. Measuring adolescent depression: Factorial validity and invariance of the Beck Depression Inventory across gender. *Journal of Research on Adolescence*, 3, 127-143.

Byrne, Barbara. M., Pierre Baron and T. Leanne Campbell. 1994. The Beck Depression Inventory French version: Testing for gender-invariant factorial structure for nonclinical adolescents. *Journal of Adolescent Research*, 9, 166-179.

Byrne, Barbara. M., and T. Leanne Campbell. 1999. Cross-cultural comparisons and the presumption of equivalent measurement and theoretical structure: A look beneath the surface. *Journal of Cross Cultural Psychology*, 30, 555-574.

Carle, Adam C. (2005). Measurement invariance, sample size, and the chi-square difference test: A Monte Carlo based method for determining minimum sample size. Paper presented at the annual meeting of the American Psychological Association, Washington, DC

Cheung, Gordon. W. and Roger B. Rensvold. 2002 Evaluating Goodness-of-Fit Indexes for Testing Measurement Invariance. *Structural Equation Modeling*, 9, 233-255.

Christoffersson, Anders. 1975. Factor analysis of dichotomized variables. *Psychometrika*, 40, 5-32.

Clogg, Clifford. C. and Leo A. Goodman. 1985. Simultaneous latent structure analysis in several groups. In N. B. Tuma Ed. *Sociological Methodology*. San Francisco: Josey-Bass.

Cole, David. A., Joan. M. Martin, Lachlan Peeke, Annette Henderson, and Jennifer Harwell. 1998. Validation of depression and anxiety measures in White and Black youths: Multitrait-multimethod analyses. *Psychological Assessment*, 10, 261-276.

Cole, Stephen. R. 1999. Assessment of differential item functioning in the Perceived Stress Scale-10. *Journal of Epidemiology and Community Health*, 53, 319-320.

Embretson, Susan. and Scott Hershberger. 1999. The new rules of measurement: what every psychologist and educator should know. Lawrence Erlbaum Associates, NJ.

Federal Register 2002. Vol. 67, No. 36, February 22, 2002.

Fuller, Wayne. A. 1987. *Measurement Error Models*. New York, Wiley.

Gallo, Joseph. J., Anthony C. James, and Bengt. O. Muthen. 1994. Age differences in the symptoms of depression: A latent trait analysis. *Journal of Gerontology*, 49*,* 251-264.

Hu, Li-Tze, and Peter M. Bentler. 1998. Fit indices in covariance structure modeling: Sensitivity to underparameterized model misspecification. *Psychological Methods*, 3, 6, 424-453.

Huang, C. David, Anthony T. Church, and Marcia S. Katigbak. 1997. Identifying cultural differences in items and traits: Differential item functioning in the NEO Personality Inventory. *Journal of Cross Cultural Psychology*, 28, 192-218.

Joreskog, Karl J. 1990. New developments in LISREL: Analysis of ordinal variables using polychoric and weighted least squares. *Quality and Quantity*, 24, 387-404.

Kish, Leslie 1998. Space/Time variation and rolling samples. *Journal of Official Statistics*, *14*, 31-46.

Knight, George P., and Nancy E. Hill. 1998. Measurement equivalence in research involving adolescents. In V. C. McLoyd, L. Steinberg. Eds. *Studying minority adolescents: Conceptual, methodological, and health issues.* Laurence Erlbaum, New Jersey.

Knight, George, Jenn Yun-Tein, Rita Shell, and Mark Roosa. 1992. The cross-ethnic equivalence of parenting and family interaction measures among Hispanic and Anglo-American families. *Child Development*, 63, 1392-1403.

McDonald, Roderick P. 1999. *Test Theory: A Unified Treatment*. Mahwah, NJ: Erlbaum.

McGovern, Pamela D. 2004. *Quality Assessment of Data Collected in the American Community Survey from Households with Low English Proficiency*. U.S. Census Bureau, Statistical Research Division, Research Report, Series 2004-01: Washington, DC.

Mellenbergh, Gideon J. 1989. Item bias and item response theory. *International Journal of Educational Research,* 13, 127-143.

Millsap, Roger. E., and Howard Everson. 1991. Confirmatory measurement model comparisons using latent means. *Multivariate Behavioral Research*, 26, 479-494.

Millsap, Roger E. and Jenn Yun-Tein. 2004. Assessing factorial invariance in ordered-categorical measures. *Journal of Multivariate Behavioral Research*, 39, 479-515.

Mislevy, Robert. J. 1986. Recent developments in the factor analysis of categorical variables. *Journal of Educational Statistics*, 11, 3-31.

Muthen, Bengt O. 1978. Contributions to factor analysis of dichotomized variables. *Psychometrika*, 43, 551-560.

Muthen, Bengt O. 1984. A general structure equation model with dichotomous, ordered categorical and continuous latent variable indicators. *Psychometrika*, 49, 115-132.

Muthen, Bengt O. 2002. Beyond SEM: General latent variable modeling, *Behaviormetrika*, 29, 1, 81-117.

Muthen, Linda K. and Bengt O. Muthen. 2002; 2004. *Mplus User's Guide*. Los Angeles: Muthen and Muthen.

Pentz, Mary Ann, and Chih-Ping Chou. 1994. Measurement invariance in longitudinal clinical research assuming change from development and intervention. *Journal of Consulting and Clinical Psychology*, 62, 450-462.

Prelow, Hazel, Jenn Yun-Tein, Mark W. Roosa, and Jennifer Wood. 2000. Do coping styles differ across sociocultural groups? The role of measurement equivalence in making this judgment. *American Journal of Community Psychology*, 28, 225-244.

Reid, Robert, George J. DuPaul, Thomas J. Power, Arthur. D. Anastopoulos, Diana Rogers-Adkinson, Mary-Beth Noll, and Cynthia Riccio. 1998. Assessing culturally different students for attention deficit hyperactivity disorder using behavior rating scales. *Journal of Abnormal Child Psychology*, 26, 187-198.

Schafer, John, and Raul Caetano. 1996. The DSM-IV construct of cocaine dependence in a treatment sample of Black, Mexican American, and White men. *Psychological Assessment*, 8, 304-311.

Smith, Larissa, and Steven P. Reise. 1998. Gender differences on negative affectivity: An IRT study of differential item functioning on the Multidimensional Personality Questionnaire Stress Reaction scale. *Journal of Personality and Social Psychology*, 75, 1350-1362.

Steiger, James H. 1998. A note on multiple sample extensions of the RMSEA fit index. *Structural Equation Modeling*, 5, 411-419.

U.S. Census Bureau 2001. *Meeting 21st Century Demographic Data Needs-Implementing the American Community Survey: July 2001. Report 1: Demonstrating Operational Feasibility.* U.S. Census Bureau: Washington, DC.

U.S. Census Bureau 2002a. *Meeting 21st Century Demographic Data Needs-Implementing the American Community Survey: May 2002. Report 2: Demonstrating Survey Quality.* U.S. Census Bureau: Washington, DC.

U.S. Census Bureau 2002b. *Census 2000 Summary File 3*. U.S. Census Bureau: Washington DC.

U.S. Census Bureau 2002c. *Census 2000 Summary File 3: Technical Documentation*. U.S. Census Bureau, Washington, DC.

U.S. Census Bureau 2003a. *Meeting 21$^{st}$ Century Demographic Data Needs-Implementing the American Community Survey. Report 3: Testing the Use of Voluntary Methods.* U.S. Census Bureau: Washington, DC.

U.S. Census Bureau 2003b. *American Community Survey Operations Plan: Release 1: March 2003*. U.S. Census Bureau: Washington, DC.

U.S. Census Bureau 2004a. *Meeting 21$^{st}$ Century Demographic Data Needs-Implementing the American Community Survey Report 4: Comparing General Demographic and Housing Characteristics with Census 2000.* U.S. Census Bureau: Washington, DC.

U.S. Census Bureau 2004b. *Meeting 21$^{st}$ Century Demographic Data Needs-Implementing the American Community Survey. Report 5: Comparing Economic Characteristics with Census 2000.* U.S. Census Bureau: Washington, DC.

U.S. Census Bureau 2004c. *Meeting 21$^{st}$ Century Demographic Data Needs-Implementing the American Community Survey. Report 6: The 2001-2002 Operational Feasibility Report of the American Community Survey*. U.S. Census Bureau: Washington, DC.

U.S. Census Bureau 2004d. *Meeting 21$^{st}$ Century Demographic Data Needs-Implementing the American Community Survey. Report 7: Comparing Quality Measures: The American Community Survey's Three Year Averages and Census 2000 Long Form Sample Estimates*. U.S. Census Bureau: Washington, DC.

U.S. Census Bureau 2004e. *Meeting 21$^{st}$ Century Demographic Data Needs-Implementing the American Community Survey. Report 8: Comparison of the*

*American Community Survey Three Year Averages and the Census Sample for a Sample of Counties and Tracts*. U.S. Census Bureau: Washington, DC.

U.S. Census Bureau 2004f. *Meeting 21st Century Demographic Data Needs-Implementing the American Community Survey. Report 9: Comparing Social Characteristics with Census 2000*. U.S. Census Bureau: Washington, DC.

U.S. Census Bureau 2004g. *Meeting 21st Century Demographic Data Needs-Implementing the American Community Survey. Report 10: Comparing Selected Physical and Financial Characteristics of Housing with Census 2000*. U.S. Census Bureau: Washington, DC.

U.S. Census Bureau 2004h. *2002 American Community Survey Profile*. Retrieved October 1, 2004, from U.S. Census Bureau, American Community Survey Web site: http//www.census.gov/acs/www/Products/Profiles/Single/2002/ACS/Narrative/010/NP01000US.htm

U.S. Census Bureau 2004i. *2002 Accuracy of the Data*. Retrieved October 1, 2004, from U.S. Census Bureau, American Community Survey Web site: http://www.census.gov/acs/www/Downloads/ACS/accuracy2002.pdf

Waller, Niels G., Jane S. Thompson, and Ernst Wenk. (2000). Using IRT to separate measurement bias from true group differences on homogeneous and heterogeneous scales: An illustration with the MMPI. *Psychological Methods*, 5, 125-146.

Table 1:
*Non-LI and LI Item Response Proportions*

| | Response | Non-LI | LI |
|---|---|---|---|
| "Dressing" | "Yes" | 0.026 | 0.042 |
| | "No" | 0.974 | 0.958 |
| | | | |
| "Vision" | "Yes" | 0.056 | 0.065 |
| | "No" | 0.944 | 0.935 |
| | | | |
| "Going Out" | "Yes" | 0.063 | 0.119 |
| | "No" | 0.937 | 0.881 |
| | | | |
| "Physical" | "Yes" | 0.115 | 0.131 |
| | "No" | 0.885 | 0.869 |
| | | | |
| "Memory" | "Yes" | 0.047 | 0.073 |
| | "No" | 0.953 | 0.927 |
| | | | |
| "Work" | "Yes" | 0.111 | 0.15 |
| | "No" | 0.889 | 0.85 |

Table 2:

*Tetrachoric Item Correlations (LI in the Lower Triangle, non-LI in the Upper Triangle)*

| | “Dressing” | “Vision” | “Going Out” | “Physical” | “Memory” | “Work” |
|---|---|---|---|---|---|---|
| “Dressing” | 1 | 0.527 | 0.830 | 0.860 | 0.698 | 0.742 |
| “Vision” | 0.586 | 1 | 0.522 | 0.624 | 0.528 | 0.514 |
| “Going Out” | 0.780 | 0.481 | 1 | 0.694 | 0.645 | 0.822 |
| “Physical” | 0.824 | 0.683 | 0.622 | 1 | | 0.769 |
| “Memory” | 0.753 | 0.609 | 0.600 | 0.745 | 1 | 0.674 |
| “Work” | 0.693 | 0.536 | 0.868 | 0.702 | 0.622 | 1 |

Table 3:

*CFA for Ordered-Categorical Measures Goodness of Fit Indices and ΔGFI Comparisons ($\Delta\chi^2$ [d], ΔCFI, ΔGamma Hat , ΔMcDonald's NCI)*

| | $\chi^2$ | *df* | *p* | $\Delta\chi^2$ | Δ*df* | $\Delta\chi^2$ *p* | RMSEA | CFI | ΔCFI | McDonald's NCI | ΔMcDonald's NCI | Gamma Hat | ΔGamma Hat |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Baseline | 6282.03* | 16 | < 0.001 | - | - | - | 0.042 | 0.99 | - | 0.99 | - | 0.995 | - |
| Invariant Λ | 4866.41* | 19 | < 0.001 | 12.964 [a] | 4 [a] | 0.011 [a] | 0.034 | 0.99 | < 0.01 [a] | 0.99 | < 0.01 [a] | 0.996 | 0.001 [a] |
| Invariant Λ & ν | 4981.82* | 22 | < 0.001 | 37.840* [b] | 5 [b] | < 0.001 [b] | 0.032 | 0.99 | < 0.01 [b] | 0.99 | < 0.01 [b] | 0.996 | < 0.001 [b] |
| | | | | | | | | | | | | | |
| Invariant Λ & ν; Θ Free | 5295.77* | 18 | < 0.001 | - | - | - | 0.036 | 0.99 | - | 0.99 | - | 0.996 | - |
| Invariant Λ, ν, & Θ | 4981.82* | 22 | < 0.001 | 20.640* [c] | 6 [c] | 0.002 [c] | 0.032 | 0.99 | < 0.01 [c] | 0.99 | < 0.01 [c] | 0.996 | < 0.001 [c] |

a. Compares the Invariant Λ Model to the Baseline Model.

b. Compares the Invariant Λ and ν Model to the Invariant Λ Model.

c Compares the Invariant Λ, ν, and free Θ Model to the Invariant Λ, ν, and Θ Model.

d Cross validation sample $\Delta\chi^2$ method (Carle, 2005).

* Significant $\chi^2$ values: α=0.01.

Table 4:
*Factor Loading Estimates for the Final Measurement Model*

| | Non-LI | LI |
|---|---|---|
| "Dressing" | 2.466 | 2.466 |
| "Vision" | 0.827 | 0.827 |
| "Going Out" | 1.859 | 1.859 |
| "Physical" | 1.842 | 1.842 |
| "Memory" | 1.172 | 1.172 |
| "Work" | 1.877 | 1.877 |

Table 5:
*Threshold Estimates for the Final Measurement Model*

| | | Non-LI | LI |
|---|---|---|---|
| "Dressing" | $\nu_1$ | -5.154 | -5.154 |
| "Vision" | $\nu_1$ | -2.066 | -2.066 |
| "Going Out" | $\nu_1$ | -3.210 | -3.210 |
| "Physical" | $\nu_1$ | -2.521 | -2.521 |
| "Memory" | $\nu_1$ | -2.570 | -2.570 |
| "Work" | $\nu_1$ | -2.601 | -2.601 |

Table 6:
*Intercept Parameter Estimates for the Final Measurement Model*

| | Non-LI | LI |
|---|---|---|
| "Dressing" | 0.00 | 0.00 |
| "Vision" | 0.00 | 0.00 |
| "Going Out" | 0.00 | 0.00 |
| "Physical" | 0.00 | 0.00 |
| "Memory" | 0.00 | 0.00 |
| "Work" | 0.00 | 0.00 |

Table 7:
*Unique Factor Variance Estimates for the Final Measurement Model*

| | Non-LI | LI |
|---|---|---|
| "Dressing" | 1.00 | 1.00 |
| "Vision" | 1.00 | 1.00 |
| "Going Out" | 1.00 | 1.00 |
| "Physical" | 1.00 | 1.00 |
| "Memory" | 1.00 | 1.00 |
| "Work" | 1.00 | 1.00 |

Table 8:
*Factor Mean and Variance Estimates for the Final Measurement Model*

| | Non-LI | LI |
|---|---|---|
| κ | 0.00 | -0.171 |
| Φ | 1 | 1.103 |

## Appendix A

### *Disability Items*[2]

Does this person have any of the following long-lasting conditions:

a. Blindness, deafness, or a sever vision or hearing impairment? (Yes/No)
b. A condition that substantially limits one or more basic physical activities such as walking, climbing stairs, reaching, lifting, or carrying? (Yes/No)

Because of a physical, mental, or emotional condition lasting 6 months or more, does this person have any difficulty doing any of the following activities:

a. Learning, remembering, or concentrating? (Yes/No)
b. Dressing, bathing, or getting around inside the home? (Yes/No)
c. Going outside the home alone to shop or visit a doctor's office? (Yes/No)
d. Working at a job or business? (Yes/No)

[2] The items as they appear in the questionnaire can be viewed at (http://www.census.gov/acs/www/Downloads/SQuest.pdf).