Published in final edited form as:

Struct Equ Modeling. 2010 January 1; 17(1): 110–133. doi:10.1080/10705510903439094.

Measurement Uncertainty in Racial and Ethnic Identification Among Adolescents of Mixed Ancestry: A Latent Variable Approach

Allison J. Tracy, Sumru Erkut, Michelle V. Porche, Jo Kim, Linda Charmaraman, Jennifer M. Grossman, Ineke Ceder, and Heidie Vázquez García

Wellesley Centers for Women

Abstract

In this article, we operationalize identification of mixed racial and ethnic ancestry among adolescents as a latent variable to (a) account for measurement uncertainty, and (b) compare alternative wording formats for racial and ethnic self-categorization in surveys. Two latent variable models were fit to multiple mixed-ancestry indicator data from 1,738 adolescents in New England. The first, a mixture factor model, accounts for the zero-inflated mixture distribution underlying mixed-ancestry identification. Alternatively, a latent class model allows classification distinction between relatively ambiguous versus unambiguous mixed-ancestry responses. Comparison of individual indicators reveals that the Census 2000 survey version estimates higher prevalence of mixed ancestry but is less sensitive to relative certainty of identification than are alternate survey versions (i.e., offering a “mixed” check box option, allowing a written response). Ease of coding and missing data are also considered in discussing the relative merit of individual mixed-ancestry indicators among adolescents.

Identification of mixed ancestry, even when individuals might not identify themselves as mixed, has important implications. Recognition of the need to develop effective assessments of mixed racial and ethnic identification among adolescents grows out of socio-historical developments that resulted in a change in the 2000 U.S. Census, when it became possible to indicate belonging to more than one race. The change in the Census form has itself added impetus to a growing recognition of legitimacy of claiming more than one racial ancestry. Although this change has moved the Census forward in documenting the prevalence of mixed-ancestry individuals in the United States, tests of the adequacy of alternative wordings building up to the change in the Census were carried out on adults (U.S. Census Bureau, 1996), raising the question of whether the same wording is effective with adolescents.

Copyright © Taylor & Francis Group, LLC

Correspondence should be addressed to Allison J. Tracy, Wellesley Centers for Women, 106 Central Street, Wellesley, MA 02481, USA. atracy@wellesley.edu.

Publisher's Disclaimer: Full terms and conditions of use: <http://www.informaworld.com/terms-and-conditions-of-access.pdf>

This article may be used for research, teaching and private study purposes. Any substantial or systematic reproduction, re-distribution, re-selling, loan or sub-licensing, systematic supply or distribution in any form to anyone is expressly forbidden.

The publisher does not give any warranty express or implied or make any representation that the contents will be complete or accurate or up to date. The accuracy of any instructions, formulae and drug doses should be independently verified with primary sources. The publisher shall not be liable for any loss, actions, claims, proceedings, demand or costs or damages whatsoever or howsoever caused arising directly or indirectly in connection with or arising out of the use of this material.

We view the terms *race* and *ethnicity* to denote socially constructed categories with definitions that have varied over time. In this article, we define mixed ancestry as any combination of pan-ethnic categories of White, African American, Asian American, Native American, and Latino. In other words, we add Latino ethnicity to the four generally recognized racial categories, but do not include ethnic variations within races in our categorization scheme. Latino ethnicity has been referred to as a “pseudo-racial group in daily life” (Trimble, Helms, & Root, 2003, p. 245). Throughout this article, we refer to these five categories more generally as race or race or ethnicity unless otherwise noted.

Researchers who have studied biracial or multiracial populations have used a variety of methods to define their samples, such as adolescent self-reports with an option to indicate more than one race (e.g., Milan & Keiley, 2000; Udry, Li, & Hendrickson-Smith, 2003), reports of parental heritage, or a combination of self-categorization and parental heritage (Cauce et al., 1992; Goldstein & Morning, 2000; Kao, 1999). Inconsistencies occur because not all offspring of interracial unions consider themselves to be biracial or multiracial. Racial or ethnic classification using self-identification is further complicated by the fluidity of adolescent responses: racial and ethnic self-identification can vary over time and place. Comparing answers to racial identification questions in the National Longitudinal Study of Adolescent Health (Add Health) administered in school and a year later at home, Harris (2002) and Harris and Sim (2002) and Doyle and Kao (2007) found that a number of adolescents reported their racial identification differently across time and setting, especially when the answer could be overheard by a parent. Hitlin, Brown, and Elder's (2006) analysis of the first and third waves of the Add Health data underscored the fluidity of racial identification over time. Such fluidity suggests that even among individuals with the same racial or ethnic lineage, the likelihood of claiming a mixed racial or ethnic ancestry will vary across individual characteristics, assessment situations, and broader social contexts. Analysis for this article rests on the question of *identification* at a single point in time, whereas exploration of processes of racial and ethnic *identity development*, which is reflected by this fluidity, is beyond the scope of this investigation (see Helms, 2007, for a discussion of measurement issues in the study of racial and ethnic identity).

Deducing mixed heritage from either adolescent or parent report of parental racial or ethnic heritage has been employed, both in small-sample studies that have focused on biracial children born into interracial unions (e.g., Gibbs & Hines, 1992; Kerwin, Ponterotto, Jackson, & Harris, 1993; Kich, 1992) and in studies of large samples (e.g., Cooney & Radina, 2000; Herman, 2004; Xie & Goyette, 1997). When relying on parental heritage as the method of identification, researchers have to grapple with how far back into the ancestry one goes and how one classifies the offspring of parents, one or both of whom might themselves be multiracial. To add to the complexity, parents' own and adolescents' reports of their parents' ancestry do not always agree (Kao, 1999) and, in the case of adoptees or an absent parent in single-parent families, some adolescents might not know the heritage of their birth parents.

The wording and format of items assessing adolescents' racial or ethnic self-identification appears to influence whether an individual identifies with one ethnic or racial group or multiple groups. For example, Hirschman, Alba, and Farley (2000) found that broadening the response options (e.g., “mark one or more” or inclusion of a multiracial category) increased the number of people identifying as being of mixed race; this was especially true for Native Americans and Hawaiians. In addition, the wording and ordering of ethnic and racial identification questions can influence the likelihood of respondents' identifying with more than one racial category, especially among Latinos. Other research suggests that adults who self-identify as multiracial tend to prefer a racial categorization format that (a) includes a multiracial category as a response option, (b) allows them to elaborate in a free response

format on each of their racial backgrounds, and (c) provides for the opportunity to check all of the responses that apply (see Johnson et al., 1999).

CONCEPTUALIZING MIXED-ANCESTRY IDENTIFICATION AS A LATENT VARIABLE

Because of the variability and fluidity in adolescent racial or ethnic identity (Harris & Sim, 2002), traditional ways of eliciting racial and ethnic self-identification contain a fair amount of measurement uncertainty. An examination of the concordance of *multiple* indicators of mixed ancestry might allow researchers to identify individuals of mixed ancestry with more certainty than through a single self-report. Latent variable modeling is a convenient way of empirically combining such information. However, it is vital to note that, in this context, latent identification is merely a mathematical representation of a set of indicators. One's implied score on such a latent variable does not reveal anything meaningful about internal identity processes or psychological identification with mixed ancestry.

Two potentially useful ways of conceptualizing the variability in mixed-ancestry identification are (a) to assume a *continuum* of identification certainty where individuals have a probability of being identified as mono- or multiracial, given their responses to the mixed-ancestry indicators or (b) to assume the presence of qualitatively distinct *subpopulations* of adolescents with regard to mixed-ancestry identification. In the first option, we might assume that every individual has some racial mixture in her or his background and therefore we might seek to describe a single underlying population. In the second option, we might assume that there are meaningful distinctions among individuals with regard to racial identification. For instance, factors such as proximity in the family tree of a recognized racial union, the similarity in appearance and culture of the constituent races, and historical and social dynamics lending meaning to racial intermixing might result in qualitatively distinct subpopulations of individuals with regard to mixed-ancestry identification. Each way of conceptualizing mixed-ancestry identification implies a different latent variable model: The first option implies a continuous latent variable and the second, a categorical latent variable. In practice, a researcher should select a single conceptualization or modeling approach to answer a given research question. In this article, we illustrate both approaches to provide a heuristic example of each and to explore their similarities and differences.

A critical consideration in selecting an appropriate latent variable modeling paradigm is the expected distribution of the latent construct of interest. A set of indicators with a non-normal multivariate distribution implies a latent variable with a “lumpy” distribution. The shape of the latent identification variable may reflect a non-normal distribution in a single homogeneous population (Option 1) or normal distributions from multiple subpopulations (Option 2). This expected clumping of latent scores suggests the use of mixture modeling.

Mixture modeling has a long tradition (see McLachlan & Peel, 2000; Titterton, Smith, & Makov, 1985, for an overview of the development of this branch of modeling) and recent advances in latent variable modeling have introduced a new set of mixture models (B. O. Muthén, 2008). A full survey of mixture models and their mathematical derivations is beyond the scope of this article, but excellent sources are available for readers interested in the statistical bases and assumptions associated with each (e.g., Marin, Mengersen, & Robert, 2005; McLachlan & Peel, 2000). Conceptually, it is helpful to think of probability mixture modeling as a missing data problem in which the data under consideration have “membership” in one of several distributions. Mixture modeling allows the researcher to estimate parameters that describe these distributions.

We have selected two subtypes of mixture models, mixture factor analysis (MFA) and latent class analysis (LCA), to illustrate alternative latent variable operationalizations of mixed ancestry. Different research questions regarding mixed-ancestry identification can be explored using these latent variable modeling approaches. MFA seeks to structure variables, whereas LCA seeks to structure cases. A mixture factor model affords the opportunity to examine the associations between the degree of concordance among the indicators and hypothesized covariates. For instance, the degree to which an adolescent can be identified as multiracial might be predictive of his or her attitudes toward social inclusiveness or tolerance of ambiguity. On the other hand, latent class models can be used to test qualitative distinctions between covariates and mixed-ancestry identification. For example, adolescents who give unambiguous responses reflecting a clearly multiracial identification might have different antecedent experiences, personal characteristics, and outcomes than do adolescents who give ambiguous mixed-ancestry responses.

Mixture Factor Analysis (Option 1)

A convenient way to describe a lumpy distribution is to assume it is comprised of a mixture of smaller distributions, each having a different factor mean (Lubke & Muthén, 2005; see Figure 1). These distributions may have the same variance, as in latent distributional analysis (Mislevy, 1984; Uebersax & Grove, 1993), but this restriction is not necessary.

In the case of mixed-ancestry identification, we expect a nonnormal distribution of factor scores with a preponderance of cases where none of the responses to the mixed-ancestry indicators imply a multiracial identification (i.e., those who would be identified as monoracial). Beyond this spike on the leftmost side of the expected distribution, the remaining variability in mixed-ancestry identification might or might not be normally distributed and, indeed, could be multimodal. The preponderance of monoracial cases forces us to stray somewhat from conventional mixture factor modeling, as this component “distribution” has a variance of zero, a parameter that is located at the boundary of admissible values. Kim and Muthén (2009) offer a two-part modeling paradigm to account for a preponderance of zeros in the indicators of a common factor. Because we have dichotomous rather than ordinal indicators and wish to account for a preponderance of “zeros” in the factor itself, we take a different approach. In this article, we model the preponderance of zeros in the latent mixed-ancestry identification distribution as a latent class in which there is a zero probability of all mixed-ancestry responses and, as a result, no common factor. The distribution is modeled as one or more latent classes in which variability in the indicators inform a common factor.

Latent Class Analysis (Option 2)

Alternatively, we might expect that the presence of empirically and conceptually distinct subpopulations underlies the variation in mixed-ancestry responses. LCA, which draws from categorical indicators (Lazarsfeld & Henry, 1968) and its cousin, latent profile analysis, which draws from continuous indicators, are conceptually equivalent to conventional cluster analysis, especially the class of cluster analysis models known as finite mixture estimation (Day, 1969; Titterton et al., 1985; Wolfe, 1970). In classic LCA, observed variables are often assumed to be conditionally independent (see Figure 2). This means that the observed correspondence among indicators can be attributed solely to latent class membership. For example, one class might have a high probability of endorsing a pair of indicators, whereas another has a low probability. The observed contingency table in this example would show a linear correspondence between these two items but members of these two cells in the contingency table may be distinct in other important ways.

To summarize, this article demonstrates how, using latent variable modeling, mixed-ancestry identification can be operationalized in a way that accounts for measurement uncertainty and inconsistencies across multiple indicators. We operationalize mixed-ancestry identification as a latent variable to (a) compare three alternative wording formats for asking racial identification questions of adolescents; and (b) demonstrate the modeling of this measurement uncertainty using two alternative latent variable models of mixed-ancestry identification: a mixture factor model approach and a latent class model approach.

METHOD

Procedure

Adolescents were recruited from three public high schools (Grades 9–12) in the Northeast, selected for the variability of the racial composition of the student body, particularly the proportion of identified mixed-ancestry adolescents based on parent report (see Appendix A). Based on the participating schools' guidelines, a passive consent approach was used where parents and guardians were sent a letter (in their native non-English language, as necessary) notifying them of the study and asking them to indicate refusal if they opposed their adolescent's participation. Eight parents withheld consent. Homeroom teachers in two schools and the English Language Arts teachers in the third administered the surveys in classrooms. Teachers were given written instructions that included directions and a suggested time limit of 20 min to standardize administration. Students were told that this was a survey about "how people form ideas of who they are." They were informed that different versions of the survey would be distributed and their responses would be kept confidential. Students were asked to give their assent before completing the survey. In all, 1,752 students from 141 classrooms across the three schools returned usable surveys, 1,738 of whom had information on one or more of the mixed-ancestry indicators. Characteristics of the respondents in this study are given in Table 1.

Measures

Self-categorization of race or ethnicity—Racial or ethnic self-identification is assessed using three alternative experimental wordings (see Appendix B) that were randomly administered within classrooms.

One version was taken directly from the Census 2000 form. It consists largely of multiple-choice check boxes with the instruction, "*Mark one or more races* to indicate what you consider yourself to be" (italics in original). The opportunity to fill in categories of the respondent's choice is available only in reference to specific umbrella categories or "other." The Census asks respondents to indicate Hispanic/Latino separately from race. A separate "mixed" check box is not present in the Census version.

The second, which we refer to as the multiple-choice + fill-in version and which had been used by one member of the research team in a prior study, contains check boxes for the four major race categories selected for focus in this study (White, African American, Asian American, Native American) with ethnic Latino category listed alongside the race categories. In addition, a separate check box for claiming "mixed" is offered and the opportunity to fill in text for any of the categories is given. Here, as in the Census version, respondents are instructed to select one or more categories.

The third, which is the fill-in only version, does not have any check boxes but simply asks respondents to fill in their race or ethnicity in the blank space provided. This version was developed by the research team.

Participants' responses to the three versions were coded into one of five pan-ethnic groups that are widely used in the research literature: African American or Black was the category into which all adolescents indicating African ancestry were coded; Native American was used for coding adolescents indicating heritage from the indigenous inhabitants of the Americas; Asian American was used for coding adolescents who indicated East/Southeast/South Asian or Pacific Island heritage; Latino was used to code adolescents whose ancestry was of Spanish-speaking peoples of the Americas; and White referred to European and Middle Eastern heritage. This coding task was minimal for the Census version, slightly more involved for the multiple-choice + fill-in version, and constituted the sole categorization method for the fill-in only version. Sixty-three of the responses to the fill-in only version were unusable (~11% of 554 given this version of the survey), either because the student misunderstood the question ($n = 8$), gave uninterpretable responses (e.g., opinions about race and ethnicity, identity information unrelated to race and ethnicity: $n = 13$), actively refused to respond ($n = 33$), or simply did not complete the question ($n = 6$). In comparison, four responses to the Census 2000 version were unusable (< 1% of students given this survey version: $n = 1$ misunderstood, $n = 2$ active refusal, $n = 1$ not completed) and eight of the multiple-choice + fill-in responses were unusable (~1%: $n = 1$ uninterpretable, $n = 6$ active refusal, $n = 1$ not completed). Invalid or uninterpretable responses were coded as missing. However, because our model estimation framework handles missing data, these respondents were included in the analysis sample.

In addition to the three randomized versions, all students were asked a dichotomous direct question: "Do you consider yourself to be of mixed race or ethnicity? By this we mean do you consider yourself to be biracial or multiracial?" Fifty-seven students had an invalid or no response to this question (~3% of entire sample), which was coded as missing.

Parental heritage—Adolescents were asked to report the racial heritage of their birth mother and birth father (if known). These responses were recorded as belonging to one or more of the five major pan-ethnic categories listed earlier and a dichotomous indicator was created that reflects mixed heritage (multiple categories for one or both parents or different categories across parents). One hundred seventy-three students (~10%) did not return valid information for mothers ($n = 74$ uninterpretable, $n = 8$ misunderstood, $n = 47$ active refusal, $n = 44$ not completed) and 212 (12%) students did not return valid data for fathers ($n = 105$ uninterpretable, $n = 8$ misunderstood, $n = 53$ active refusal, $n = 46$ not completed). No student gave valid father data without also giving valid mother data. In all, 212 students did not have valid data for one or both parents and so were missing information for the parental heritage indicator.

Observed Prevalence Estimates of Mixed Ancestry

If we take the proportion of the mixed-ancestry responses in each of the indicators singly as an estimate of the proportion of mixed-ancestry adolescents in the population, each yields a different prevalence estimate (see Table 2). In our sample, the Census version yields a somewhat higher mixed-ancestry estimate (14%) than do either the multiple-choice + fill-in (11%) or fill-in only (9%) versions. As a comparison, the direct question yields an estimate of 24% and parental heritage yields 15%. Mixed ancestry based on phenotype ambiguity is estimated to be 19%. Although some of these estimates appear similar, students didn't always report consistently across items. For instance, parental heritage and Census 2000 have very nearly the same prevalence estimates, but 53 students gave inconsistent responses. This number translates to slightly more than 10% of the 517 students given the Census 2000 survey version and who had valid data on both indicators.

Analysis Models

A series of models with increasing numbers of latent classes (beyond the parameter-restricted monoracial class) are estimated and overall model fit statistics are compared. Fit indexes include the log-likelihood value, the Bayes Information Criterion (BIC) and adjusted BIC statistics (recommended as among the strongest choices for assessing the fit of latent class models; Nylund, Asparouhov, & Muthen, 2007), and the chi-square test of model fit for binary outcomes. In addition, we examine the Vuong (1989) and Lo–Mendell–Rubin (2001) sample-size adjusted likelihood ratio test to compare the degree of model improvement in nested models (a model with k classes vs. a model with $k - 1$ classes). The solution of choice, based on a consideration of these criteria, is further evaluated by examining the model results and their implications for interpretation.

In all models, missing data are handled by means of full information maximum likelihood (FIML; see Schafer & Graham, 2002, for an overview of missing data handling options). Classroom membership constitutes the clustering variable in the aggregated complex sample models.

In the mixture factor models, we fix the variance of the latent factor to zero (as well as the factor loadings) in the monoracial class and give high starting values (15) for the indicator thresholds, which correspond to probabilities near zero of having a mixed-ancestry response. The mean of the latent factor is fixed to an arbitrary value. In the remaining latent class(es) describing the mixture distribution, a common factor model is used to describe the relationships among the dichotomous mixed-ancestry indicators (see Glöckner-Rist & Hoijtink, 2003). Here, we assume measurement invariance in the factor loading and threshold values while factor means and variances are freely estimated across classes. Numerical integration is required for the estimation of a latent continuous factor with categorical indicators and so the maximum likelihood estimation with robust standard errors (MLR) was used.

In the LCA model, no latent factor is specified; instead, threshold parameters capture the essential variation in mixed-ancestry responses. The monoracial class is specified by fixing the thresholds of most of the dichotomous indicators to 15 (representing a probability near zero of having a mixed-ancestry response), and the thresholds in the remaining class(es) are freely estimated. Because a number of adolescents who returned otherwise clearly monoracial responses also considered themselves of mixed ancestry, we chose to freely estimate the threshold for the direct question in all classes. A conditional independence assumption is imposed on the data: No covariation among indicators is allowed within class. Because it wasn't necessary to use numerical integration to estimate the model, the maximum likelihood (ML) estimator was used.

The *Mplus* programming code for the final MFA and LCA models is included in Appendix C.

RESULTS

Mixture Factor Model (Option 1)

Table 3 gives the fit indexes of the mixture factor model solutions with one, two, and three classes. In all three models, the assumption that the indicators have an underlying common factor appears to be supported; the chi-square fit for the binary outcomes is nonsignificant in each model. Whereas the BIC indexes increase with each successive model, suggesting the model with the fewest number of classes fits best (i.e., the one-class solution), the LRT statistics comparing nested models indicate that the two-class solution is preferable. Further, an examination of a composite of the observed indicators among those with complete

information shows clear evidence of zero-inflation; therefore, we believe that the one-class solution is misspecified. Consequently, we selected the two-class solution for further examination.

Based on the results of the two-class model (see Table 4), 71% of the sample ($n = 1,236$) were estimated to belong to the monoracial latent class. In this latent variable formulation, we are able to estimate the precision of classification, a feature not available were we to simply count the individuals in the sample with a “mixed” response to at least one indicator. Because items contribute to the solution differently based on the degree of concordance with other items in the set, students with certain response patterns may be assigned to Class 1 (very low likelihood of being identified as of mixed ancestry) as the best fitting latent class but also have a nonzero probability of belonging to Class 2. Uncertainty of classification (entropy) can be estimated by comparing the assignment of individuals based on their most likely latent class membership with the model-estimated posterior probability distributions for each class. In this model, individuals assigned to Class 1 had an average of 27% probability of belonging to Class 2. On the other hand, among those assigned to Class 2, the probability was zero of belonging to Class 1. In other words, we are very certain about the classification of adolescents we identify as mixed-ancestry but we are less certain about those we would identify as monoracial. This uncertainty in the identification of adolescents who might appear by some standards to be monoracial confirms our expectations about the hidden nature of mixed-ancestry adolescents.

The threshold values estimated in the model contain information about the probability of a “1” (mixed-ancestry) response.¹ The probability estimates for a mixed-ancestry response in Class 1, by design, are essentially zero. In Class 2, the highest threshold value is associated with the fill-in only indicator ($\tau = 4.66$, $SE = 2.02$, $p < .05$), followed closely by the multiple-choice + fill-in indicator ($\tau = 4.21$, $SE = 1.52$, $p < .01$). This means that adolescents in this latent class have a very low probability of giving mixed-ancestry responses to these items, although this probability is somewhat higher than for adolescents in Class 1. The direct question is associated with a threshold estimate of zero ($SE = 0.31$, ns), meaning that adolescents in Class 2 are just as likely as not to indicate that they consider themselves to be of mixed ancestry. The threshold values estimated for Census 2000 ($\tau = 1.34$, $SE = 0.38$, $p < .001$) and parental heritage ($\tau = 2.12$, $SE = 0.64$, $p < .001$) correspond to moderately small probability estimates (0.21 and 0.11, respectively) for giving a mixed-ancestry response.

Within Class 2, the estimated factor loadings of the three randomized indicators were significant. The Census 2000 version had the weakest factor loading of the three ($\lambda = 1.87$, $SE = 0.36$, $p < .05$ for Census 2000 vs. $\lambda = 5.55$, $SE = 1.27$ and $\lambda = 5.28$, $SE = 2.23$, for multiple-choice + fill-in and fill-in only, respectively—both p values $< .001$). The factor loading for parental heritage was strong ($\lambda = 3.99$, $SE = 0.52$, $p < .001$) and the factor loading for the direct question was weakest ($\lambda = 1.00$, $SE = 0.22$, $p < .001$).

As can be seen in the standard errors of the factor loading estimates, the precision of the multiple-choice + fill-in and the fill-in only loadings are considerably lower than the precision of loadings for the remaining three indicators. Despite this, the multiple-choice + fill-in loading estimate is high enough to be statistically distinguishable (as defined by a comparison of confidence intervals about the estimates) from both the Census 2000 ($p < .05$) and the direct question indicators ($p < .01$). By virtue of both the precision and strength of

¹The threshold value for a dichotomous variable is the logit coefficient corresponding to the probability of a 1 response. The sign of the threshold is the opposite of the sign taken by a mean or intercept (e.g., a negative threshold corresponds to a probability greater than .50).

the estimate, the factor loading for parental heritage is also statistically distinguishable from the estimates for the Census 2000 ($p < .05$) and direct question ($p < .01$).

The threshold estimates and factor loadings together can be used to calculate item characteristic curves (ICCs) for each indicator, in which the probability of a response indicating mixed ancestry can be traced across the underlying latent factor continuum (see Figure 3). These curves show that four of the five indicators (parental heritage, Census 2000, multiple-choice + fill-in, and fill-in only) cross the .50 probability line at roughly the same spot on the underlying mixed-ancestry identification continuum, whereas the direct question indicator appears distinct in that a 1 response becomes more likely than not at a lower value along the continuum. The ICCs of the parental heritage, multiple-choice + fill-in, and fill-in only indicators are striking in that they have similarly steep slopes, whereas the Census 2000 and direct question slopes are notably more gradual.

Latent Class Model (Option 2)

Table 5 gives the model fit statistics for the one-, two-, three-, and four-class models. The two-class solution has the lowest BIC values and highest entropy. This solution mirrors the model results we found for the mixture factor model in that there is a large monoracial class (low probabilities on each of the indicators by design) and a smaller mixed-ancestry class. However, if we consider the likelihood ratio test (LRT) statistics, we see that the addition of a third class significantly improves the model solution. An examination of the parameter estimates in the two- and three-class models shows that the three-class solution might be the more useful in that it increases the certainty of classification among those we identify as monoracial. Further, the three-class model distinguishes adolescents with ambiguous identification responses from those with a clear mixed-ancestry response pattern (high probability of a mixed-ancestry response on all indicators). Because this is an exploratory study into mixed-ancestry identification, we chose to highlight the more detailed three-class solution in this article.

The results of the latent class model are given in Table 6. Our model estimates that 74% of adolescents ($n = 1,281$) belong to a class with essentially zero probabilities of mixed-ancestry responses (monoracial class) and that the remaining two classes account for 13% each ($n = 225$ and $n = 232$, respectively). The second class is associated with moderately low probabilities of a mixed-ancestry response across the indicators, ranging from .03 to .37 (ambiguous class). In the third class, the probability of a mixed-ancestry response is high, ranging from .71 to .94 (mixed-ancestry class).

As we saw in the evaluation of the overall fit of this model, classification certainty is relatively low. However, the overall classification uncertainty for the monoracial and mixed-ancestry classes (15–16%), not surprisingly, is largely a function of the classification uncertainty in the ambiguous class. Among those assigned to the ambiguous class, there is a one in three chance of actually belonging to the monoracial class and a 1 in 10 chance of actually belonging to the mixed-ancestry class. Conversely, the probability among those assigned to the monoracial class of actually belonging to the mixed-ancestry class is zero. This probability is only .02 for the reverse scenario.

In the monoracial class, all threshold estimates were high, most by design. The freely estimated threshold for the direct question corresponds to a low probability (.09) of a mixed response but the precision of the estimate is poor ($\tau = 2.33$, $SE = 1.44$, ns). In the ambiguous class, the threshold estimates for all the mixed-ancestry indicators are no different from chance, in part due to imprecision in the estimates (large standard errors).

In the mixed-ancestry class, the parental heritage indicator has the lowest threshold ($\tau = -2.67$, $SE = 0.82$, $p < .01$) with a likelihood of 0.94% of giving a mixed-ancestry response. The multiple-choice + fill-in also has a low threshold estimate ($\tau = -1.57$, $SE = 0.73$, $p < .05$), associated with an 83% chance of a mixed-ancestry response. Similarly, the estimate for the direct question is significant ($\tau = -1.41$, $SE = 0.27$, $p < .001$) and associated with a high probability of a mixed response (80%). The estimates for the fill-in only and Census 2000 indicators are somewhat weaker (fill-in only: $\tau = -1.01$, $SE = 0.55$, ns ; Census 2000: $\tau = -0.88$, $SE = 0.45$, $p < .05$), though even these point estimates are associated with a high likelihood of a mixed-ancestry response (73% and 71%, respectively).

DISCUSSION

Our findings show that latent variable modeling of mixed-ancestry identification can be a useful way to combine information from multiple mixed-ancestry indicators. This general modeling approach accounts for measurement and classification uncertainty. Operationalized as a continuous-level mixture distribution, we can account for nonnormality in the latent score distribution. Operationalized as a set of latent classes, mixed-ancestry identification offers the opportunity to compare adolescents with different response profiles.

Comparison of Alternative Self-Categorization Options

In the MFA formulation of mixed-ancestry identification, the latent factor provides a common metric by which the measurement properties of single indicators can be evaluated. Descriptively, the Census 2000 version yields the highest count of mixed-ancestry adolescents. This finding is consistent with results of the research with adults leading up to the wording change in the 2000 Census: The Race and Ethnic Targeted Test study (U.S. Census Bureau, 1996). The reexamination of these data by Hirschman et al. (2000) showed that the wording of the 2000 Census question, which included instruction to “mark one or more races,” increased reporting of multiple races among some racial groups.

However, the multiple-choice + fill-in version (followed very closely by the fill-in only version) outperformed the Census version in terms of identifying what appears to be a critical point along the mixed-ancestry identification continuum beyond which the probability of mixed-ancestry identification is likely. In other words, although more adolescents can and do identify multiple categories of race and ethnicity in a check box format (Census 2000), many do not endorse other indicators of mixed ancestry. On the other hand, allowing adolescents to fill in answers to qualify their responses on check boxes and giving the option of a “mixed” check box along with the more traditional categories (multiple-choice + fill-in) identifies fewer mixed-ancestry adolescents, but those so identified are more likely to report other indicators of mixed ancestry. In this way, the multiple-choice + fill-in (and fill-in only) versions lend themselves to a “quick and dirty” dichotomous estimate of who is and who is not of mixed ancestry on a person-to-person basis, whereas the Census 2000 version might more broadly encompass mixed ancestry with regard to a population.

Interestingly, the direct question, which one might suppose would precisely locate an individual along the mixed-ancestry identification continuum, yielded results very similar to those found for the Census 2000 version: A high proportion of adolescents could be identified as of mixed ancestry using this item and identification occurred all along the mixed-ancestry continuum, revealing a rather loose connection with the other indicators of mixed ancestry. We proffer two interpretations of this finding. First, it might simply be easy to give an affirmative response to a direct question. Additionally, for some, an affirmative response might be perceived as more socially desirable, given the specificity of the direct

question. Alternatively, it could be that some students who answer affirmatively are describing being of mixed ethnicity within a single pan-ethnic racial category.

The second alternative explanation involves a more theoretical view. Goldstein and Morning (2000) posit that defining mixed ancestry either by genealogy, awareness, or identification will have implications for the size of multirace population estimates. In their model, a genealogical definition would yield the largest group because it would include all those who have mixed ancestry many generations back. However, not all members of this group will be aware of having a mixed heritage. Among those who are aware of their multiple racial heritages, some might not identify as multiracial. The implication of this is that the direct question might be eliciting an affirmative response from a relatively larger number of adolescents who are aware of their mixed heritage but who might not identify with being a mixed-ancestry individual. Being aware of one's mixed ancestry but not identifying as biracial or multiracial might seem counterintuitive, but it is reasonable given the structural, institutional, relational, and individual factors that influence identity development (Rockquemore, Lasloffy, & Noveske, 2006). The continuing legacy of the "one-drop rule" is an example of the political and racialized process of other-defined versus self-described identification (Davis, 2006).

Implications for Research

MFA and LCA are two alternative ways of representing nonnormality in the multivariate distribution of the mixed-ancestry indicators as latent classes. In both options, individual-level class membership is not directly known but can be inferred using Bayesian posterior probability distributions, drawing from an individual's response pattern and estimated model parameters.

By explicitly modeling measurement uncertainty, studies of the correlates of mixed ancestry identification can better represent the uncertain nature of the construct itself. For instance, the ambiguous class in the LCA solution could represent a subpopulation of adolescents who are subject to conflicting information about their own racial identity or who seek to emphasize different aspects of themselves for different purposes and in different situations. This subpopulation might differ substantially in formative contexts and subsequent behavior and experiences from subpopulations of adolescents with a clear and unambiguous monoracial or mixed-ancestry identification.

Our results suggest that, when the research question makes it more appropriate to assess adolescent self-reports, one's choice of alternate wordings can yield different estimates. We recommend using multiple indicators in a latent variable approach because mixed ancestry as a latent variable draws from a greater density of information about mixed-ancestry status and experience. However, if the survey permits only one format and the goal is to identify the largest number of adolescents who are likely to identify as mixed ancestry, researchers can choose the Census 2000 wording. If the goal is to identify a smaller group that is consistent with multiple indicators, the multiple-choice + fill-in version offers a shortcut. We recommend using this version over the fill-in only version not only because the former is slightly more consistent with multiple indicators, but also because the fill-in only version produces a higher proportion of unusable responses (e.g., opinions about race or ethnicity, identity information unrelated to race or ethnicity) and necessitates an extensive coding process.

This study incorporated a randomized design resulting in a total of five dichotomous indicators. Other studies are not likely to have this component and therefore should seek to include other information about mixed ancestry (fluidity of identification, participation in practices of multiple cultures including language use, etc.) to improve the identifiability of

the model. The inclusion of dependent variables (e.g., phenotype ambiguity), exogenous variables (e.g., parental socialization regarding race and ethnicity), or covariates (e.g., heterogeneity in the school, neighborhood, and friendship contexts) might serve to stabilize an otherwise empirically underidentified model. Also, because adolescents of mixed ancestry might have a relatively low representation in historically racially homogeneous regions, a large sample size might be necessary to have adequate statistical power to identify mixed-ancestry individuals using a latent variable model and to incorporate adolescents of mixed ancestry in hypothesis tests. Another option is to include information about adolescent racial categorization, cultural participation, fluidity of identity, or phenotype ambiguity elicited from parents, peers, or other knowledgeable individuals as well as from the adolescent. Latent variable modeling of mixed ancestry is particularly suited to the analysis of large, nationally representative data sets such as Add Health (Udry, 2003) that include parental categorization information as well as self-categorization of race and Hispanic ethnicity. The longitudinal nature of such a secondary data set can also be used to operationalize fluidity of identification.

Studies of mixed ancestry should consider specific racial combinations comprising the mixes that adolescents report. Correlates of mixed-ancestry identification may differ depending on the cultural and social dynamics in which specific combinations are embedded. With the multiple indicator approach illustrated here, some adolescents may be identified as having mixed ancestry without indicating the categories from which their mixture derives. For instance, an affirmative response to the direct question does not give any information about the racial or ethnic categories from which the mixture derives. Similarly, the multiple-choice + fill-in check box indicating a “mixed” option might not be accompanied by information about membership in specific categories. The fill-in only question has the potential to provide information only about a general mixture, as well as the potential for providing information about specific combinations of categories. For this reason, if one’s research question requires the identification of specific combinations, the Census 2000 wording, in conjunction with other mixed-ancestry indicators, might be prudent. If specific combinations of racial and ethnic mixtures are not central to the research question, however, the identification of mixed ancestry through latent variable modeling allows the comparison of adolescents who self-identify with those who do not.

The approach to modeling mixed ancestry presented in this article is meant to be a general approach that can be used with different sets of indicators, depending on researchers’ interests and aims. Conceivably, latent approaches can be used for other complex social address variables. For example, identification of individuals’ social class status is multiply determined and requires consideration of education, income, self-identification, and so on, of individual students and their families of origin. From a purely practical standpoint, latent variable modeling offers the opportunity to better handle missing data in the indicators (Schafer & Graham, 2002) via FIML estimation. In manifest variable formulations of a construct, adolescents missing responses to any of the selected indicators would be omitted from the analysis sample. This is important when some questions are asked of only a subset of adolescents (e.g., an in-depth assessment of a subsample, skip patterns resulting in data collected from a subset of participants for whom the question applies, etc.).

Mixture models can have solutions plagued by local maxima and generally require multiple sets of starting values to avoid conclusions based on spurious results. However, because we had strong assumptions about the parameter estimates associated with the monoracial class, we set the number of sets of program-generated starting values to zero, choosing to supply our own starting values based on theory instead. Ideally, multiple sets of starting values would be used to guard against the threat of interpreting results based on a local solution. However, we consider the substantive results of our analyses as merely illustrative of the

latent variable approaches we proffer. Confirmation of the findings we present here will rely on replication studies.

Lubke and Neale (2006) and Bauer and Curran (2004) warned against misrepresenting responses generated by an underlying homogeneous population as a latent class model, resulting in a tendency toward overextraction of classes. Similarly, modeling data arising from a heterogeneous population as a factor model could tend to lead to an overextraction of factors. In this case, we did not attempt to extract multiple factors in the MFA modeling context. If the unidimensional assumption holds, the selected three-class LCA solution would appear to corroborate the two-class MFA solution even if the LCA model was a misrepresentation of the true nature of the population. However, future applications of these models in the identification of mixed-ancestry adolescents might explore these assumptions more directly.

Policy Implications

Contemporary events continue to highlight the need for an improved understanding of political, social, and psychological ramifications of mixed ancestry. For example, the No Child Left Behind (2002) legislation mandates a closing of the achievement gap between White and minority students (identified as African American, Hispanic, and Native American). Whether mixed-ancestry students should be included as a target population in this legislation has not been established. Thus, the conundrum of determining whether schools are succeeding in closing this gap is exacerbated by the difficulty of procuring accurate race and ethnicity information from students, especially from students of mixed ancestry (Lopez, 2003; Vaishnav, 2003). Indeed, Harris and Thomas (2002) and Kao (1999) have shown that mixed-ancestry students who have particular racial combinations (e.g., White/Asian with respect to grade point average) perform differently in school relative to students with other combinations, as well as some monoracial students. At minimum, schools and other institutions should move beyond forced choice, single categorization of race and ethnicity, given what we understand of mixed-ancestry identification.

Some mixed-ancestry students might choose to identify themselves as a monoracial minority on official forms if they perceive that such identification could impact their own opportunities or benefits available to the minority group with which they align themselves. This phenomenon has been noted among multiracial adults (Perlmann & Waters, 2002) but no research has confirmed this to be the case among adolescents. A latent variable approach could be a useful tool to use should this dynamic threaten to underestimate mixed-ancestry adolescents identified for intervention.

Limitations

The definition of mixed ancestry we used focused on mixtures across pan-ethnic categories, thereby ignoring within-category mixes. For example, it might well be that within the pan-ethnic Asian American category, Japanese and Chinese ancestry combination can be as salient as a White and Latino mix for adolescents. Ignoring intracategory mix is an omission we share with other research on mixed-ancestry individuals but conceivably a latent variable approach could easily incorporate distinctions within as well as across general categories of race and ethnicity.

Another limitation of this study is that findings obtained from high school students might not be generalizable to younger children or to older adolescents and young adults. Racial and ethnic identity, and therefore the likelihood of identification, is a developmental process (Phinney, 1989; Quintana, 2007) and what we have captured is racial and ethnic identification in middle adolescence. Future research is warranted to examine the racial and

ethnic identification of individuals over time as they mature and as their social contexts change.

This investigation of various approaches to racial and ethnic self-identification was guided by the premise that measures should be quick and easy to administer, so that they might be included in any survey as standard demographic questions without a great deal of additional explanation. Even though we were working with high school students, it is unclear how well some respondents understood terms such as mixed racial and ethnic ancestry, biracial, or multiracial and the extent to which they were able to distinguish inquiry about race and ethnicity from inquiry about citizenship and nationality. Given the unease in the local Latino community about being targeted by the recent anti-immigration climate, we made a decision not to include nativity questions in the survey. Information of adolescents' own and their parents' nativity would have provided clues as to how to interpret such answers and also enriched the quality of the latent models.

Conclusions

With regard to the growing literature on mixed-ancestry individuals, evidence suggests that mixed-ancestry adolescents (variously defined) might have poorer social adjustment outcomes, such as depression, substance use, and health problems, than their single-race-reporting peers (Cooney & Radina, 2000; Milan & Keiley, 2000; Udry, 2003) or they might have particular strengths in cognitive flexibility (Shih & Sanchez, 2005). Future studies about racial and ethnic identity processes will benefit when subpopulations of mixed-ancestry individuals can be reliably identified in the larger population. This enhanced knowledge base will be important in efforts to increase the visibility and understanding of hitherto hidden populations of mixed-ancestry youth who might be underserved. These research and policy directions can only be effectively pursued using a well-conceptualized method of operationalizing mixed-ancestry identification.

Acknowledgments

This research uses data from Add Health, a program project designed by J. Richard Udry, Peter S. Bearman, and Kathleen Mullan Harris, and funded by grant P01-HD31921 from the Eunice Kennedy Shriver National Institute for Child Health and Human Development, with cooperative funding from 23 other federal agencies and foundations. Special acknowledgment is due to Ronald R. Rindfuss and Barbara Entwisle for assistance in the original design. Persons interested in obtaining data files from Add Health should contact Add Health, Carolina Population Center, 123 W. Franklin Street, Chapel Hill, NC 27516-2524 (addhealth.unc.edu). No direct support was received from grant P01-HD31921 for this analysis.

REFERENCES

- Bauer DJ, Curran PJ. The integration of continuous and discrete latent variable models: Potential problems and promising opportunities. *Psychological Methods* 2004;9:3–29. [PubMed: 15053717]
- Cauce, AM.; Hiraga, Y.; Mason, CA.; Aguilar, T.; Ordonez, N.; Gonzales, N. Between a rock and a hard place: Social adjustment of biracial youth. Newbury Park, CA: Wiley; 1992.
- Cooney TM, Radina ME. Adjustment problems in adolescence: Are multiracial children at risk? *American Journal of Orthopsychiatry* 2000;70:433–445. [PubMed: 11086522]
- Davis, FJ. Define race: Comparative perspectives. In: Brunnsma, DL., editor. *Mixed messages: Multiracial identities in the “color-blind” era*. Boulder, CO: Rienner; 2006. p. 15–31.
- Day NE. Estimating the components of a mixture of normal distributions. *Biometrika* 1969;56:463–474.
- Doyle JM, Kao G. Are racial identities of multiracials stable?: Changing racial self-identification among single and multiple race individuals. *Social Psychology Quarterly* 2007;70:405–423. [PubMed: 19823596]

- Gibbs, JT.; Hines, AM. Negotiating ethnic identity: Issues for Black-White adolescents. In: Root, MPP., editor. *Racially mixed people in America*. Thousand Oaks, CA: Sage; 1992. p. 223-238.
- Glöckner-Rist A, Hoijtink H. The best of both worlds: Factor analysis of dichotomous data using item response theory and structural equation modeling. *Structural Equation Modeling* 2003;10:544–565.
- Goldstein JR, Morning AJ. The multiple-race population of the United States: Issues and estimates. *Proceedings of the National Academy of Sciences of the United States of America* 2000;97:6230–6235. [PubMed: 10811886]
- Harris, DR. Does it matter how we measure? Racial classification and the characteristics of multiracial youth. In: Perlmann, J.; Waters, MC., editors. *The new race question: How the census counts multiracial individuals*. New York: Russell Sage Foundation; 2002. p. 62-101.
- Harris DR, Sim JJ. Who is multiracial? Assessing the complexity of lived race. *American Sociological Review* 2002;67:614–627.
- Harris, DR.; Thomas, JL. The educational costs of being multiracial: Evidence from a national survey of adolescents. Ann Arbor, MI: University of Michigan, Population Studies Center at the Institute for Social Research; 2002. (PSC Rep. No. 02-521)
- Helms JE. Some better practices for measuring racial and ethnic identity constructs. *Journal of Counseling Psychology* 2007;54:235–246.
- Herman M. Forced to choose: Some determinants of racial identification in multiracial adolescents. *Child Development* 2004;75:730–748. [PubMed: 15144483]
- Hirschman C, Alba R, Farley R. The meaning and measurement of race in the U.S. Census: Glimpses into the future. *Demography* 2000;37:381–393. [PubMed: 10953811]
- Hitlin S, Brown JS, Elder GH. Racial self-categorization in adolescence: Multiracial development and social pathways. *Child Development* 2006;77:1298–1308. [PubMed: 16999799]
- Johnson TP, Jobe JB, O'Rourke D, Sudman S, Warnecke R, Chapa-Resendez G, et al. Dimensions of self identification among multiracial and multiethnic respondents in survey interviews. *Evaluation Review* 1999;21:671–687.
- Kao G. Racial identity and academic performance: An examination of biracial Asian and African American youth. *Journal of Asian American Studies* 1999;2(3):223–249.
- Kerwin C, Ponterotto JG, Jackson BL, Harris A. Racial identity in biracial children: A qualitative investigation. *Counseling Psychology* 1993;40:221–231.
- Kich, GK. The developmental process of asserting a biracial, bicultural identity. In: Root, MPP., editor. *Racially mixed people in America*. Thousand Oaks, CA: Sage; 1992. p. 304-317.
- Kim YK, Muthén BO. Two-part factor mixture modeling: Application to an aggressive behavior measurement instrument. *Structural Equation Modeling* 2009;16(4):602–624. [PubMed: 20717486]
- Lazarsfeld, PF.; Henry, NW. *Latent structure analysis*. Boston, MA: Houghton Mifflin; 1968.
- Lo Y, Mendell NR, Rubin DB. Testing the number of components in a normal mixture. *Biometrika* 2001;88:767–778.
- Lopez AM. Mixed-race school-age children: A summary of Census 2000 data. *Educational Researcher* 2003;32:25–37.
- Lubke GH, Muthén BO. Investigating population heterogeneity with factor mixture models. *Psychological Methods* 2005;10:21–39. [PubMed: 15810867]
- Lubke GH, Neale MC. Distinguishing between latent classes and continuous factors: Resolution by maximum likelihood? *Multivariate Behavioral Research* 2006;41:499–532.
- Marin, JM.; Mengersen, K.; Robert, CP. Bayesian modelling and inference on mixtures of distributions. In: Dey, D.; Rao, CR., editors. *Handbook of statistics*. Vol. Vol. 25. Amsterdam: Elsevier; 2005. p. 1-50.
- McLachlan, GI.; Peel, D. *Finite mixture models*. New York: Wiley; 2000.
- Milan S, Keiley MK. Biracial youth and families in therapy: Issues and interventions. *Journal of Marital and Family Therapy* 2000;26:305–315. [PubMed: 10934677]
- Mislevy RJ. Estimating latent distributions. *Psychometrika* 1984;49:359–381.

- Muthén, BO. Latent variable hybrids: Overview of old and new models. In: Hancock, GR.; Samuelsen, KM., editors. *Advances in latent variable mixture models*. Charlotte, NC: Information Age; 2008. p. 1-24.
- Muthén, LK.; Muthén, BO. *Mplus user's guide*. Los Angeles: Muthén & Muthén; 1998–2008.
- No Child Left Behind Act of 2001, Pub. L. 107–110, 115 Stat. 2002. p. 1425
- Nylund KL, Asparouhov T, Muthén BO. Deciding on the number of classes in latent class analysis and growth mixture modeling: A Monte Carlo simulation study. *Structural Equation Modeling* 2007;14:535–569.
- Perlmann, J.; Waters, MC. Introduction. In: Perlmann, J.; Waters, MC., editors. *The new race question: How the Census counts multiracial individuals*. New York: Russell Sage Foundation; 2002. p. 1-30.
- Phinney JS. Stages of ethnic identity development in minority group adolescents. *Journal of Early Adolescence* 1989;9:34–49.
- Quintana SM. Racial and ethnic identity: Developmental perspectives and research. *Journal of Counseling Psychology* 2007;54:259–270.
- Rockquemore, K.; Lasloffy, T.; Noveske, J. It all starts at home: Racial socialization in multiracial families. In: Brunson, DL., editor. *Mixed messages: Multiracial identities in the “color-blind” era*. Boulder, CO: Rienner; 2006. p. 203-216.
- Schafer JL, Graham JW. Missing data: Our view of the state of the art. *Psychological Methods* 2002;7:147–177. [PubMed: 12090408]
- Shih M, Sanchez DT. Perspectives and research on the positive and negative implications of having multiple racial identities. *Psychological Bulletin* 2005;131:569–591. [PubMed: 16060803]
- Titterton, DM.; Smith, AFM.; Makov, UE. *Statistical analysis of finite mixture distributions*. Chichester, UK: Wiley; 1985.
- Trimble, JE.; Helms, JE.; Root, MPP. Social and psychological perspectives on ethnic and racial identity. In: Bernal, G.; Trimble, JE.; Burlew, AK.; Leong, FTL., editors. *Handbook of racial and ethnic minority psychology*. Thousand Oaks, CA: Sage; 2003. p. 239-275.
- Udry, JR. The National Longitudinal Study of Adolescent Health (Add Health), Waves I & II, 1994–1996; Wave III, 2001–2002. Chapel Hill, NC: University of North Carolina at Chapel Hill, Carolina Population Center; 2003. [machine-readable data file and documentation]
- Udry JR, Li RM, Hendrickson-Smith J. Health and behavior risks of adolescents with mixed-race identity. *American Journal of Public Health* 2003;93:1865–1870. [PubMed: 14600054]
- Uebersax JS, Grove WM. A latent trait finite mixture model for the analysis of rating agreement. *Biometrics* 1993;49:823–835. [PubMed: 10798855]
- U.S. Census Bureau. Results of the 1996 Race and Ethnic Targeted Test. 1996 [Retrieved January 13, 2009]. from <http://www.census.gov/population/www/documentation/twps0018/twps0018.html>
- Vaishnav, A. Questions raised on MCAS data: Many students did not list their racial identity. 2003 [Retrieved September 23, 2003]. from http://www.boston.com/news/education/k_12/mcas/articles/2003/09/07/questions_raised_on_mcas_data?mode=PF
- Vuong QH. Likelihood ratio test for model selection and non-nested hypotheses. *Econometrica* 1989;57:307–333.
- Wolfe JH. Pattern clustering by multivariate mixture analysis. *Multivariate Behavioral Research* 1970;5:329–350.
- Xie Y, Goyette K. The racial identification of biracial children with one Asian parent: Evidence from the 1990 Census. *Social Forces* 1997;76:547–570.

APPENDIX A

Racial and Ethnic Composition of Schools in the Sample

	Multicultural High School ^a	Latin High School ^b	White High School ^c
African American	52%	15%	4%
Asian	15%	0%	8%
Hispanic	8%	68%	4%
Native American	< 1%	0%	0%
White	24%	16%	82%
Multirace, non-Hispanic	1%	No data available	3%

Note. Because students were allowed to indicate more than one race or ethnic category, totals may sum to greater than 100%.

^a Chosen based on overall school-reported racial diversity of student body.

^b Chosen based on the potential for Hispanic ethnicity/Black or White race mixes, given regional racial composition.

^c Chosen based on proportion of school-reported multiracial students in otherwise monoracially White-dominated context.

APPENDIX B

THREE RANDOMIZED VERSIONS OF SELF-CATEGORIZATION

Census 2000 Version

Are you Spanish/Hispanic/Latino? Mark the “No” box if **not** Spanish/Hispanic/Latino.

- ☐ **No**, not Spanish/Hispanic/Latino
 ☐ Yes, Puerto Rican
☐ Yes, Mexican, Mexican Am. Chicano
 ☐ Yes, Cuban
☐ Yes, other Spanish/Hispanic/Latino - *Print group in box.* ↓

What is your race? Mark one or more races to indicate what you consider yourself to be.

- ☐ White
☐ Black, African American
☐ American Indian or Alaska Native - *Print name of enrolled or principal tribe in box.* ↓

- ☐ Asian Indian
 ☐ Japanese
 ☐ Native Hawaiian
☐ Chinese
 ☐ Korean
 ☐ Guamanian or Chamorro
☐ Filipino
 ☐ Vietnamese
 ☐ Samoan
☐ Other Asian - *Print race in box.* ↓

- ☐ Other Pacific Islander - *Print race in box.* ↓

- ☐ Some other race - *Print race in box.* ↓

Multiple-Choice + Fill-In Version

What is your race or ethnicity? Mark **one or more** boxes, and use the space to provide any additional information you would like us to have.

- ☐ White: _____
- ☐ Black/African American or of African descent : _____
- ☐ Native American/Alaska Native/American Indian: _____
- ☐ Asian or Pacific Islander: _____
- ☐ Hispanic, Latino, or of Spanish origin: _____
- ☐ Mixed race or ethnicity: _____
- ☐ Some other race or ethnicity: _____

Fill-In Only Version

Race and ethnicity are words used to describe a person's origin or set of origins.

What do you consider to be your **race/ethnicity**?

APPENDIX C

MPLUS PROGRAMMING CODE FOR OPERATIONALIZATING MIXED-ANCESTRY IDENTIFICATION

MIXTURE FACTOR MODEL

TITLE: MIXED-ANCESTRY IDENTIFICATION -Mixture factor Model

DATA: FILE=F:\Mixed-ancestry Group\mixture model.dat;

VARIABLE:

NAMES ARE idcode classid parent direct census multiple fillin;

MISSING=parent-fillin (99,999);

CLUSTER=classid;

IDVARIABLE=idcode;

CATEGORICAL=direct census multiple fillin parent;

CLASSES=C(2);

ANALYSIS:

TYPE=COMPLEX MIXTURE;

ALGORITHM=INTEGRATION;

INTEGRATION=MONTECARLO;

STARTS=0;

MODEL:

```
%OVERALL%
[census$1* multiple$1* fillin$1* direct$1* parent$1*];
IDENT BY census* multiple* fillin* direct* parent*;
IDENT@1; [IDENT@0];
%C#1%
[census$1*15 multiple$1*15 fillin$1*15 direct$1*15 parent$1*15];
IDENT BY census@0 multiple@0 fillin@0 direct@0 parent@0;
IDENT@0; [IDENT@0];
```

OUTPUT: CINTERVAL;

PLOT: TYPE=PLOT2;

LATENT CLASS MODEL

TITLE: MIXED-ANCESTRY IDENTIFICATION -Latent Class Model

DATA: FILE=F:\Mixed-ancestry Group\mixture model.dat;

VARIABLE:

```
NAMES ARE idcode classid parent direct census multiple fillin;
MISSING=parent-fillin (99,999);
CLUSTER=classid;
IDVARIABLE=idcode;
CATEGORICAL=direct census multiple fillin parent;
CLASSES=C(3);
```

ANALYSIS:

```
TYPE=COMPLEX MIXTURE;
STARTS=0;
```

MODEL:

```
%OVERALL%
[census$1* multiple$1* fillin$1* direct$1* parent$1*];
%C#1%
[census$1@15 multiple$1@15 fillin$1@15 parent$1@15];
%C#2%
[census$1*0 multiple$1*0 fillin$1*0 parent$1*0];
%C#3%
[census$1*15 multiple$1*15 fillin$1*15 parent$1*15];
```

OUTPUT: CINTERVAL;

FIGURE 1.
Analysis model operationalizing mixed-ancestry identification as a mixture factor model.

FIGURE 2.
Analysis model operationalizing mixed-ancestry identification as a latent class model.

FIGURE 3.

Item characteristic curves for mixed-ancestry identification indicators calculated from the results of the mixture factor model.

TABLE 1

Descriptive Characteristics of Analysis Sample

Student Characteristics	N (% of Valid Responses)	
Male	849 (48%)	
Female	903 (52%)	
Living with both birth parents	1,096 (66%)	
Living with one birth parent	497 (30%)	
Living with neither birth parent	73 (4%)	
9th grade	548 (33%)	
10th grade	537 (32%)	
11th grade	333 (20%)	
12th grade	255 (15%)	
Self-categorization claims (any of three versions)		
Single category: White	777 (47%)	
Single category: Black	263 (16%)	
Single category: Latino	258 (16%)	
Single category: Asian	142 (9%)	
Single category: Native American	< 10 (< 1%)	
Multiple categories	186 (11%)	
Not categorized	15 (1%)	
	Mother	Father
Did not complete high school	202 (13%)	178 (12%)
High school diploma or GED	233 (15%)	234 (16%)
Some college/2-year degree	260 (16%)	181 (12%)
4-year degree	566 (36%)	408 (28%)
Graduate-level degree	322 (20%)	469 (32%)

TABLE 2

Descriptive Statistics for Latent Variable Indicators

Variable	N (Valid Responses)	N and Proportion Coded 1
Self-categorization claims (randomized)		
Census 2000 ^a version	574	80 (14%)
Multiple-choice + fill-in version	564	63 (11%)
Fill-in only version	489	43 (9%)
Direct question	1,695	409 (24%)
Parental heritage	1,541	227 (15%)

^aThe 2000 nationwide census identifies 2.4% of the U.S. population as mixed race. Our estimate reflects the specific composition of the high schools we selected for our study and includes Hispanic ethnicity as well as race to identify individuals of mixed ancestry.

TABLE 3

Comparative Statistics for Alternative Mixture Factor Model Solutions

Statistic	1-Class Solution	2-Class Solution	3-Class Solution
Log-likelihood	-1,770.220	-1,765.025	-1,762.935
H_0 scaling correction factor for MLR	1.115	0.676	0.732
Information criteria			
No. of free parameters	10	16	18
Bayesian (BIC)	3,615.045	3,649.419	3,660.159
Sample size adjusted BIC ($n^* = (n + 2)/24$)	3,583.276	3,598.588	3,602.975
Chi-square test of model fit for the binary outcomes			
Pearson chi-square	12.115	5.506	4.826
df	21	15	13
p value	.9365	.9869	.9788
Likelihood ratio chi-square	6.258	3.993	5.034
df	21	15	13
p value	.9992	.9978	.9744
Vuong–Lo–Mendell–Rubin LRT			
H_0 Log-likelihood value		-1,770.717	-1,759.741
$2 \times \log$ -likelihood difference		11.383	-6.389
Difference in parameters		6	6
M		-0.627	-42.712
SD		3.121	71.922
p value		.0035	.3750
Lo–Mendell–Rubin adjusted LRT			
Value		11.134	-6.249
p value		.0038	.3728

Note. MLR = maximum likelihood estimation with robust standard; BIC = Bayesian Information Criterion; LRT = likelihood ratio test.

TABLE 4

Mixture Factor Model With a Preponderance of Zeros

	Class 1: Preponderance of Zeros	Class 2: Nonzero Distribution
Estimated count and proportion	1,236 (0.71)	502 (0.29)
Average probability estimates	0.73	1.00
<i>Factor Indicators</i>	<i>Factor Loading (SE)</i>	<i>Factor Loading (SE)</i>
Parental heritage	0.00 ^a	3.99 (0.52) ***
Direct question	0.00 ^a	1.00 (0.22) ***
Census 2000	0.00 ^a	1.87 (0.36) ***
Multiple-choice + fill-in	0.00 ^a	5.55 (1.27) ***
Fill-in only	0.00 ^a	5.28 (2.23) *
	<i>Threshold (SE) P(Y = 1)</i>	<i>Threshold (SE) P(Y = 1)</i>
Parental heritage	24.05 (0.13) ***	0.00 2.12 (0.64) *** .11
Direct question	13.70 (0.05) ***	0.00 0.00 (0.31) .50
Census 2000	14.14 (0.04) ***	.00 1.34 (0.38) *** .21
Multiple-choice + fill-in	27.22 (0.11) ***	.00 4.21 (1.52) ** .01
Fill-in only	24.89 (0.08) ***	.00 4.66 (2.02) * .01

Note. $N = 1,738$.

^aParameter fixed to value shown.

* $p < .05$.

** $p < .01$.

*** $p < .001$.

TABLE 5

Comparative Statistics for Alternative Latent Class Model Solutions

Statistic	1-Class Solution	2-Class Solution	3-Class Solution	4-Class Solution
Log-likelihood				
<i>H0</i> value	-2,155.58	-1,760.778	-1,759.401	-1,759.041
<i>H0</i> scaling correction for MLR	1.427	1.155	1.122	0.818
Information criteria				
No. of free parameters	5	11	13	19
Bayesian (BIC)	4,348.463	3,603.621	3,615.788	3,659.831
Sample size adjusted BIC ($n^* = (n + 2)/24$)	4,332.578	3,568.675	3,574.488	3,599.470
Entropy		0.881	0.574	0.583
Vuong–Lo–Mendell–Rubin LRT				
<i>H0</i> Log L value		-2,155.58	-1,778.070	-1,759.401
$2 \times \text{Log L diff}$		789.604	37.339	0.719
Diff no. of parameters		6	6	6
<i>M</i>		-0.511	6.371	1.015
<i>SD</i>		15.317	5.651	3.123
<i>p</i> value		.000	.0004	.5091
Lo–Mendell–Rubin adjusted LRT				
Value		772.530	36.523	0.704
<i>p</i> value		.000	.0005	.5117

Note. MLR = maximum likelihood estimation with robust standard errors; BIC = Bayesian Information Criterion; LRT = likelihood ratio test.

TABLE 6
Latent Class Analysis Operationalizing Three Classes of Mixed-Ancestry Identification

	Monoracial Class		Ambiguous Class		Mixed-Ancestry Class	
Estimated counts and proportions	1281 (0.74)		225 (0.13)		232 (0.13)	
Average probability estimates within most likely class	0.84		0.59		0.85	
Average probability estimates outside most likely class	P(Ambiguous class) = 0.00 P(Monoracial class) = 0.16 P(Mixed-ancestry class) = 0.84		P(Monoracial class) = 0.32 P(Mixed-ancestry class) = 0.10		P(Monoracial class) = 0.02 P(Ambiguous class) = 0.14	
Entropy value	0.57					
Latent Class Indicators	Threshold (SE)	P(Y = 1)	Threshold (SE)	P(Y = 1)	Threshold (SE)	P(Y = 1)
Parental heritage	15.00 ^a	0.00	1.79 (1.62)	0.14	-2.67 (0.82) **	0.94
Census 2000	15.00 ^a	0.00	1.13 (1.82)	0.24	-0.88 (0.45) *	0.71
Multiple-choice + fill-in	15.00 ^a	0.00	3.43 (1.91)	0.03	-1.57 (0.73) *	0.83
Fill-in only	15.00 ^a	0.00	3.59 (1.89)	0.03	-1.01 (0.55)	0.73
Direct question	2.33 (1.44)	0.09	0.54 (0.61)	0.37	-1.41 (0.27) ***	0.80

Note. N = 1,738.

^aParameter fixed to value shown.

* $p < .05$.

** $p < .01$.

*** $p < .001$.