Comparing Fit of Nonsubsuming Probability Models

Gregory Alvord and George B. Macready
University of Maryland

A "mixture" probability model that incorporates two component models defined by nonsubsuming sets of parameters is introduced, and a strategy for using this model in the selection of a preferred component model is developed. Example applications of the suggested strategy are considered for the special case in which the Rasch item response model and a Latent State Mastery model are the component models compared. Simulated data sets generated under each of these models were used to provide example applications of the proposed model selection strategy.

Within the realm of probability modeling, investigators are frequently interested in ascertaining how well specific probability models account for manifest data. For example, it may be of interest to an investigator to determine how well a structural equations model (see Jöreskog & Sörbom, 1979) that assumes certain specified causal links among a set of cognitive traits can explain individuals' behavior on those traits. Or, in the case of the measurement of some attribute, it may be of interest to a test developer to ascertain whether a single underlying latent trait (see Hambleton & Swaminathan, 1983) is adequate for explaining examinees' test performance (i.e., the latent trait in question is unidimensional). A frequently used statistic for assessing fit of probability models is the likelihood ratio (LR) statistic defined as

$$\chi^2_{LR} = -2 \sum_{i=1}^{n} O_i \ln\left(\frac{\hat{E}_i}{O_i}\right) \quad , \tag{1}$$

where O_i = observed frequency for the ith response category,
$\hat{E}_i$ = estimated expected frequency for the ith response category, and
n = number of response categories.

The estimated expected frequency for the ith response category is, in general, defined as

$$\hat{E}_i = N\hat{P}(u_i) \quad , \tag{2}$$

where N is the number of respondents and $\hat{P}(u_i)$ is the estimated probability of the occurrence of the ith response category, u_i. When local independence is assumed (see Lord & Novick, 1968) and u_i represents a *pattern of item responses* across m dichotomously scored items, then its probability may be designated

$$P(u_i) = \prod_{g=1}^{m} P_g^{h_{gi}}(1-P_g)^{1-h_{gi}} \quad , \tag{3}$$

where h_{gi} = 1 when a correct response is encountered on the gth item within the ith response pattern, and 0 otherwise, and
P_g = the probability of a positive response to the gth item.

APPLIED PSYCHOLOGICAL MEASUREMENT
Vol. 9, No. 3, September 1985, pp. 233–240

0146-6216/85/030233-08$1.65

Alternatively, when u_i represents a *total score* i across the m items, $P(u_i)$ may be designated:

$$P(u_i) = \sum_{j=1}^{K_i} \left[\prod_{g=1}^{m} P_g^{h_{gj \cdot i}} (1 - P_g)^{1 - h_{gj \cdot i}} \right] , \qquad (4)$$

where K_i = the number of item response patterns containing a total of i positive responses, and

$$h_{gj \cdot i} = \begin{cases} 1 & \text{when a correct response is encountered on the } g\text{th item within the } j\text{th response pattern (having a total of } i \text{ positive responses), and} \\ 0 & \text{otherwise.} \end{cases}$$

The preferability of these two types of response categories, u_i, (i.e., total scores or item response patterns) will depend on such factors as (1) number of respondents, (2) number of items, and (3) number of independent parameters to be estimated. In either case, however, the LR statistic is asymptotically distributed as chi-square (given that $n > k + 1$) with degrees of freedom equal to $n - k - 1$, where k is the number of independently estimated parameters related to the model in question.

In addition to assessing the absolute fit of models to data, researchers are frequently interested in comparing fit provided by different models. For example, in trying to determine whether a one-, two- or three-parameter logistic model is preferable for use with a given test, relative tests of fit may be employed. Similarly, when log-linear models are used to assess the nature of the dependency structure existing among a set of categorical variables, relative assessment of competing models may prove useful (see Baker, 1981). This same kind of relative comparison among models is useful when linear logistic latent trait models (see Whitely & Schneider, 1981) are used to assess the relations between test item characteristics and the difficulties of those items. When comparisons between pairs of models are of interest, two situations may arise: (1) the models have a *subsuming* relation, or, (2) the models have a *nonsubsuming* relation.

A subsuming relation exists between two models when a set of one or more linear constraints imposed upon the parameters defining one model (here called the subsuming or full model, F) results in the set of more restricted parameters defining the second more constrained model (here called the reduced model, R). When models have such a subsuming relation, it is possible to assess relative fit of the reduced model to the full model, that is, to determine whether the reduced model fits as well as the full model. This may be accomplished by using the difference in the LR statistics defined in Equation 1 for the full (F) and the reduced (R) models,

$$\chi^2_D = \chi^2_{LR(R)} - \chi^2_{LR(F)} \quad . \qquad (5)$$

This difference statistic, χ^2_D, is asymptotically distributed as chi-square under certain regularity conditions (see Wilks, 1938) with degrees of freedom equal to the difference in the number of independently estimated parameters related to the two models.

Two models have a nonsubsuming relation when the defining parameters in neither model may be specified by imposing one or more linear constraints on the parameters of the other model. Such a nonsubsuming relation may exist between two models that are (1) members of the same family or (2) members of different families of models. It may be noted that for either of these cases a problem arises if a researcher is interested in comparing two models that have such a nonsubsuming relation, since the difference statistic defined in Equation 5 has not been shown to be distributed as chi-square. Thus, alternative procedures are needed for assessing the relative adequacy of such pairs of models with respect to fit.

The purpose of this paper is to present a strategy for assessing relative fit provided by pairs of nonsubsuming models. This approach is based on the use of a mixture model that incorporates the nonsubsuming models of interest. The class of mixture models, on which the strategy that is presented in this paper is based, was originally suggested by Cox (1961, 1962) and developed by Atkinson (1970).

The Mixture Model

The general model, Mx, presented in this paper incorporates two alternative nonsubsuming models

(here designated as Models I and II) as weighted components of the full model. Under the Mixture model, the probability of the occurrence of the ith response category is defined as

$$P(u_i)_{Mx} = \frac{P(u_i)_I^{1-\lambda} P(u_i)_{II}^{\lambda}}{\sum_{i=1}^{n} P(u_i)_I^{1-\lambda} P(u_i)_{II}^{\lambda}} \quad . \tag{6}$$

Within this model, $P(u_i)_I$ and $P(u_i)_{II}$ respectively represent probabilities for the ith response category under the nonsubsuming component Models I and II, while λ represents a mixture parameter that designates the relative contributions provided by each of the component models to the full Mixture model. This mixture parameter, λ, is defined on the interval $(-\infty, \infty)$. The likelihood function for the mixture model may be specified as

$$L = \frac{N!}{\prod_{i=1}^{n} f_i!} \prod_{i=1}^{n} P(u_i)_{Mx}^{f_i} \quad , \tag{7}$$

where f_i is the number of cases observed in the ith multinomial category for $i = 1, \ldots, n$.

When all parameters in the Mixture model are estimated simultaneously by standard maximum likelihood estimation procedures (see Rao, 1973), the number of independent parameters, k_{Mx}, would be equal to the number of the independent parameters in Models I and II plus one for the mixture parameter (λ). Given that the models of interest (i.e., Models I, II, and Mx) are identified for the data in question and maximum likelihood estimates are obtained (separately for each model), then tests of absolute fit may be performed in accordance with Equation 1. In addition, tests of relative fit of the component models to the Mixture model may be implemented through the use of Equation 5, since both Models I and II have a subsuming relation with Mx. The hypotheses that are addressed in assessing the relative fit of Model I to Mx may be specified as H_O: $\lambda = 0$ and H_A: $\lambda \neq 0$. For this case, the mixture model reduces to Model I under H_O. Similarly, in assessing the relative fit of Model II to Mx, the hypotheses that are considered are H_O: $\lambda = 1$ and H_A: $\lambda \neq 1$. Here the mixture model reduces to Model II under H_O.

This approach, though statistically appealing may nevertheless result in insurmountable problems in application. First, the number of independent parameters in Mx may be so large that simultaneous estimation of all the parameters is not practical. Second, the Mixture model may not be identified (see Goodman, 1979), thus preventing the unconstrained use of the maximum likelihood approach.

An alternative two-stage estimation/fit assessment approach that may be employed is to estimate parameters for the component models of interest, using some data set 1, and then conditionally estimate the mixture parameter and assess both relative and absolute fit of the models using a second independent data set 2. Although this approach provides a less efficient use of the data, it greatly reduces the likelihood of the above problems arising in estimation.

Under this approach, a two-stage maximum likelihood estimation strategy is used. In the first stage maximum likelihood estimates (MLEs) are obtained separately for each component model and used to compute the estimated probabilities of response categories for the two models, $\hat{P}(u_i)_I$ and $\hat{P}(u_i)_{II}$, which are substituted into Equation 6. In the second stage, a second set of data is used to obtain a MLE of the mixture parameter, λ, by maximizing Equation 7. Since an independent data set is being used at this stage, the procedure involves the estimation of a single parameter, λ. Thus, the LR statistic as defined in Equation 1 may be calculated for Mx with $E_i = N\hat{P}(u_i)_{Mx}$ and degrees of freedom $n - 2$.

There is possible a third alternative approach to parameter estimation that is based on conditional procedures but only requires the use of a single data set. Under this approach, the mixture parameter λ along with the parameters for Model I are conditionally estimated (i.e., given the estimated parameters related to Model II). The estimates for Model I are then fixed and λ and the parameters of Model II are conditionally estimated. This sequence of conditional estimation is continued until some specified level of convergence is obtained across two successive sets of estimates. Under this approach, the degrees of freedom related to the full model may be estimated to equal the difference

between the number of independent observations and the rank of the matrix of first-order partial derivatives related to the full model (see Goodman, 1974).

Strategy for Assessing Fit and Selecting a Preferred Model

In this section a stagewise strategy for selecting a preferred model is presented. Under this strategy, it is possible to reach any one of the following conclusions related to model preference based on fit:

1. Model I provides acceptable fit, and is preferred,
2. Model II provides acceptable fit, and is preferred,
3. Both models provide acceptable fit, but neither is preferred, or
4. Both models provide unacceptable fit and thus both are rejected.

A detailed description of the strategy for selecting a preferred model follows, with the flow diagram in Figure 1 reflecting the steps within this strategy.

I. Assess the relative fit of Model I and Model II to Mx through the use of the difference chi-square defined in Equation 5.

A. If both Model I and Model II provide acceptable relative fit, go to II.
B. If only Model II provides acceptable relative fit, go to III.
C. If only Model I provides acceptable relative fit, go to IV.
D. If neither Model I nor Model II provide acceptable relative fit, stop. Conclude that neither Model I nor Model II are adequate to account for the data.

II. (From I-A.) Assess Model I and Model II with respect to their absolute fits using the LR statistic defined in Equation 1.

A. If both Model I and Model II provide acceptable fit, stop. Conclude that both Model I and Model II are acceptable models but neither is "preferred" over the other model (at least based on the assessment of fit).
B. If only Model II provides acceptable fit, stop. Select Model II.
C. If only Model I provides acceptable fit, stop. Select Model I.
D. If neither model provides acceptable absolute fit, stop. Conclude that neither model is acceptable.

III. (From I-B.) Assess the absolute fit of Model II using the LR statistic defined in Equation 1.

A. If absolute fit is obtained, stop. Select Model II.
B. If absolute fit is not obtained, stop. Conclude that neither model is acceptable.

IV. (From I-C.) Assess the absolute fit of Model I using the LR statistic defined in Equation 1.

A. If absolute fit is obtained, stop. Select Model I.
B. If absolute fit is not obtained, stop. Conclude that neither model is acceptable.

If either Model I or II is selected as preferred according to the model selection strategy, the investigator may reasonably conclude that the preferred model provides a more adequate explanation for the data than the alternative model to which it was compared.

Another outcome that may be reached with this strategy is that both models considered may be found to be acceptable. This outcome leaves the determination of a "preferred" model in doubt. When such an outcome occurs, further analyses and/or data collection may be desirable. Such analyses might include one or more of the following components:

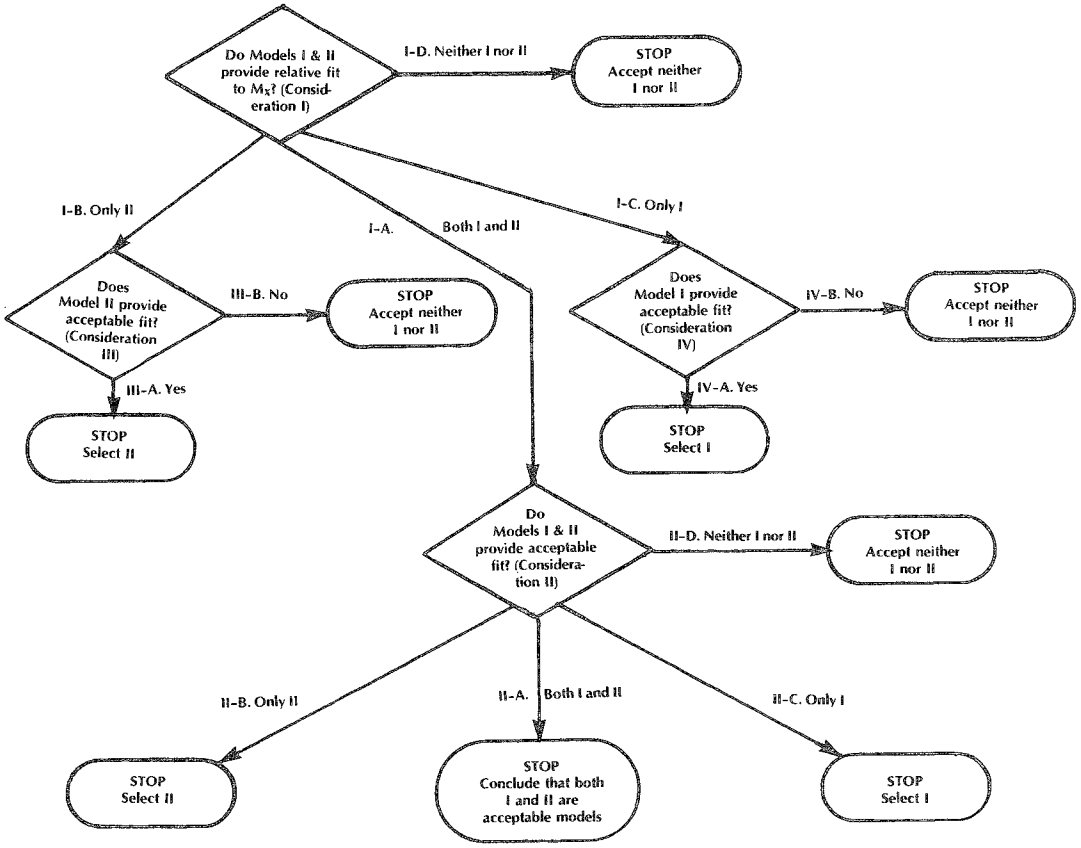
1. Assessment of model parsimony,
2. Study replication, and/or
3. Consideration of alternative models.

These kinds of explorations may result in a clearly preferred model or lead to the development of a new model that incorporates the best aspects of both of the original models being compared.

Example Applications

In the area of mastery assessment, two general classes of latent structure models have been proposed. These classes have been called Continuum

Figure 1
Flow Diagram of the Proposed Strategy for Selecting a Preferred Model



models and State models (see Meskauskas, 1976). For the Continuum models, trait acquisition is assumed to be continuous in nature and defined by one or more latent dimensions, whereas for State models, trait acquisition is perceived as an "all-or-none" process with mastery viewed as the presence of trait acquisition. Two models that fall respectively within the Continuum and State model classes (and that do not have a subsuming relation) are the Rasch model (see Wright & Douglas, 1977) and the Latent State Mastery (LSM) model (see Dayton & Macready, 1976; Macready & Dayton, 1977, 1980). This section provides two example applications of the proposed decision strategy for selection between the Rasch and LSM models. These two example applications are based on simulated data sets generated from the Rasch and LSM models, respectively.

Models

For the Rasch model, the probability of a correct response to item g for a respondent with latent ability x_j is

$$P_{g|x_j} = \frac{\exp[\bar{a}(x_j - b_g)]}{1 + \exp[\bar{a}(x_j - b_g)]}, \quad (8)$$

where x_j = the latent ability of the jth examinee,
b_g = the difficulty of the gth item, and
$\bar{a}$ = a common discrimination factor for all items.

If the density function of the ability parameter x_j is denoted as $f(x)$, then the (unconditional) prob-

ability of a correct response to item g is

$$P_g = \int_{-\infty}^{\infty} P_{g|x} f(x) dx \quad . \tag{9}$$

In turn, the probability of the response vector u_i across the m items of interest is defined by Equation 3, which may be considered as $P(u_i)_{\mathrm{I}}$ in Equation 6.

For the LSM model, the probability of a correct response to item g is defined as

$$P_g = (P_{g|M})(\Delta) + (P_{g|\overline{M}})(1 - \Delta) \quad , \tag{10}$$

where $P_{g|\overline{M}}$ = the conditional probability that a nonmaster ($\overline{M}$) answers item g correctly,

$P_{g|M}$ = the conditional probability that a master (M) answers item g correctly, and

Δ = the latent proportion of masters.

The probability of the ith response vector (u_i) under the LSM model is defined by Equation 3 incorporating P_g as defined in Equation 10. Since this is the second model considered here, the above definition of $P(u_i)$ may be substituted for $P(u_i)_{\mathrm{II}}$ in Equation 6.

Estimation of Parameters

Iterative maximum likelihood procedures were used for estimation of parameters underlying the Rasch, LSM, and Mixture models. Specifically, the Newton-Raphson method was used for both the Rasch model and the LSM model (see Rao, 1973).

Estimation of the mixture parameter, λ, was obtained in a second stage in which the estimated values of $P(u_i)_{\mathrm{I}}$ and $P(u_i)_{\mathrm{II}}$ were used. This second stage of estimation was carried out with an independent data set. Maximum likelihood estimates of λ were obtained in this second stage using the method of bisection (see Conte & de Boor, 1972).

Data Generation

Four sets of simulated data, based on two generating models, were considered. The first two sets of data were generated from the Rasch model with each set containing $N = 600$ simulated cases over $m = 5$ items. The first set of data was used for obtaining estimates of the parameters under both the Rasch and LSM models, whereas the second set was then used to conditionally estimate λ and to assess the absolute and relative fits of the two models of interest. The item difficulties employed in data generation were uniformly distributed; $b_g = -.90, -.45, 0.00, .45$, and $.90$ for items $g = 1, \ldots, 5$, respectively, while the discrimination factor ($\bar{a}$) was set at .85. The latent trait values corresponding to each simulated individual were randomly generated from a normal distribution with a mean of 0.0 and a standard deviation of 1.0.

For data generation under the LSM model, a similar procedure to that described above was used with the two generation data sets each containing information on $N = 600$ simulated cases over $m = 5$ items. Under this generating model, the five conditional probabilities for positive item responses by nonmasters, $P_{g|\overline{M}}$, were .25, .23, .21, .19, and .17 for items $g = 1, \ldots, 5$, respectively, while the corresponding conditional probabilities for positive responses by masters, $P_{g|M}$, were .87, .88, .89, .90, and .91. The proportion of latent masters, Δ, was set at .5. As in the previous example, the two generated data sets were used in the same manner to estimate parameters and to assess fit.

Results

In this section, the Rasch model is designated as Model "I" and the LSM model as Model "II." This designation may prove useful in following the progression of assessment detailed in Figure 1. For all tests considered, models were judged as providing acceptable absolute (or relative) fit if $P(\chi^2) \leq .05$. Using this criterion, results related to the first example in which the data were generated from the Rasch model are presented in Table 1.

Following the strategy for selecting the preferred model detailed in Figure 1, the relative fits for both the Rasch (I) and LSM (II) models to the Mixture model were assessed at "consideration I." As presented in Table 1, only the Rasch model provided acceptable relative fit to the Mixture model (thus, path I-C in Figure 1 was followed). Hence, only

Table 1
Results for Tests of Fit Based on Data Generated from the Rasch Model

Assessed Condition	χ^2	df	p
Relative fit of Rasch model to Mx	.462	1	.50
Relative fit of LSM model to Mx	5.750	1	.001
Absolute fit of Rasch model	25.702	26	.48

the absolute fit of the Rasch model was tested (see consideration IV in Figure 1). The value of chi-square obtained for this assessment (see Table 1) was found to provide acceptable fit and thus (following path IV-A in Figure 1), the Rasch model was selected as the preferred model.

Although the absolute fit of the Mixture model was not assessed in the decision strategy, it is interesting to note that the obtained χ^2_{LR} for Mx was 30.525 (p = .44), which closely approximates the expected value of 30. For this case, the obtained estimate of λ was .219 indicating that the estimated relative weighting given the Rasch model ($1 - \lambda$ = .781) was far greater than that for the LSM model.

The results for the second example in which the data were generated from the LSM model are presented in Table 2. As in the previous example, at consideration I both the Rasch (I) and LSM (II) models were assessed with respect to their relative fits to the Mixture model. For this case, however, only the LSM model provided acceptable relative fit (thus, path I-B was followed). This outcome at consideration I resulted in only the LSM model being assessed with respect to its absolute fit (see consideration III in Figure 1). Since absolute fit was obtained for the LSM model, path III-A was followed and a final decision to select LSM as the preferred model was made.

Here again, it may be noted that the Mixture model provided very good fit to the data since the χ^2_{LR} for fit of Mx was 29.517 (p = .50). In addition, it may be noted that the obtained estimate of λ was .770, reflecting a relatively large estimated contribution to Mx by the LSM model. In each of the examples considered, it may be seen that the decision strategy led to the selection of the generating model.

Discussion

In this paper a procedure has been developed for empirically assessing which of two nonsubsuming probability models is to be preferred for explaining data. This procedure is seen as having potential value in dealing with a selection problem for which there is presently no viable alternative. Although the results presented in this paper show that the selection strategy here presented can work effectively, more work is obviously required to identify under what conditions the suggested approach may be effective.

Table 2
Results of Tests of Fit Based on Data Generated from the LSM Model

Assessed Condition	χ^2	df	p
Relative fit of Rasch model to Mx	23.353	1	.001
Relative fit of LSM to Mx	1.749	1	.19
Absolute fit of LSM model	19.452	20	.50

References

Atkinson, A. C. (1970). A method for discriminating between models. *Journal of the Royal Statistical Society, 3,* 323–345.

Baker, F. B. (1981). Log-linear, Logit-linear models: A didactic. *Journal of Educational Statistics, 6,* 75–102.

Conte, S. D., & deBoor, C. (1972). *Elementary numerical analysis*. New York: McGraw-Hill.

Cox, D. R. (1961). Tests on separate families of hypotheses. *Proceedings of the 4th Berkeley Symposium, 1,* 105–123.

Cox, D. R. (1962). Further results on tests of separate families of hypotheses. *Journal of the Royal Statistical Society, Series B, 24,* 406–424.

Dayton, C. M., & Macready, G. B. (1976). A probabilistic model for validation of behavioral hierarchies. *Psychometrika, 41,* 189–204.

Goodman, L. A. (1974). Exploratory latent structure analysis using both identified and unidentified models. *Biometrika, 61,* 215–231.

Goodman, L. A. (1979). On the estimation of parameters in latent structure analysis. *Psychometrika, 44,* 123–128.

Hambleton, R. K., & Swaminathan, H. (1983). *Item response theory: Principles and applications*. Boston: Kluwer-Nijhoff Publishers.

Jöreskog, K. G., & Sörbom, D. (1979). *Advances in factor analysis and structural equations models*. Cambridge MA: Abt Books.

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. Reading MA: Addison-Wesley.

Macready, G. B., & Dayton, C. M. (1977). The use of probabilistic models in the assessment of mastery. *Journal of Educational Statistics, 2,* 99–120.

Macready, G. B., & Dayton, C. M. (1980). The nature and use of state mastery models. *Applied Psychological Measurement, 4,* 493–516.

Meskauskas, J. A. (1976). Evaluation models for criterion referenced testing: Views regarding mastery and standard setting. *Review of Educational Research, 46,* 133–158.

Rao, C. R. (1973). *Linear statistical inference and its applications* (2nd ed.). New York: Wiley.

Whitely, S. E., & Schneider, L. M. (1981). Information structure for geometric analogies: A test theory approach. *Applied Psychological Measurement, 5,* 383–397.

Wilks, S. S. (1938). The large sample distribution of the likelihood ratio for testing composite hypotheses. *Annals of Mathematical Statistics, 9,* 60–62.

Wright, B. D., & Douglas, G. A. (1977). Best procedures for sample free item analysis. *Applied Psychological Measurement, 37,* 573–586.

Acknowledgment

The authors are indebted to the University of Maryland Computer Science Center, for providing the computer time necessary for implementing this study.

Author's Address

Send requests for reprints or further information to George B. Macready, Department of Measurement, Statistics and Evaluation, College of Education, University of Maryland, College Park MD 20742, U.S.A.