Qualitative or Quantitative Differences?

Latent Class Analysis of Mathematical Ability for Special Education Students

Xiangdong Yang, Julia Shaftel, Douglas Glasnapp, and **John Poggio**
University of Kansas

The current article investigates whether the mathematics achievement of students in special education can be used to identify those who share common cognitive skills that may not be in concordance with their disability labels. Latent class analysis of a comprehensive test of mathematics taken by fourth-grade students with various disabilities reveals that a model with 2 latent classes is adequate to characterize the latent structure of the data. A parallel relationship of response profiles across the 2 classes suggests differences in the levels of mathematical ability (quantitative), rather than differences in the type of mathematical ability (qualitative), between the 2 latent classes in terms of generic mathematical proficiency. Cross-validation on a separate data set with careful matching of content areas within the math test verified this conclusion. Although a significant relationship exists between the identified latent classes and various disabilities, the analysis also found common mathematical problem-solving behaviors across disability categories. Implications for intervention and limitations of the current study are discussed.

Students with disabilities comprise a protected population for which particular educational and instructional interventions are expected to be provided. Ideally, remedial interventions should be individualized to each student's need. This is the goal of the Individuals with Disabilities Education Act of 1990 (IDEA) and embodied in the Individualized Education Program (IEP) process, but complete individualization is usually impossible due to limited resources and the need for teachers trained to provide services tailored to specific groups, such as mobility training or Braille for students with visual impairment. As a result, students who meet categorical eligibility criteria specified in IDEA and further delineated by relevant state special education laws and procedures are sometimes grouped for instruction rather than receiving individual attention.

The disability categories defined by IDEA and state laws were formed on historical and political, rather than empirical, foundations. Categories are typically defined on the basis of such characteristics as type of impairment (e.g., perceptual, language, physical, cognitive) or etiology (e.g., traumatic brain injury). Emotional/behavioral disorders are defined by maladaptive behaviors and moods that differ from the norm for age peers, and specific learning disabilities are defined by differences among abilities and achievement within the individual. Some categories (e.g., perceptual impairments, autism, specific learning disabilities, traumatic brain injury) match medical or psychological diagnoses, whereas others (e.g., other health impairment, physical disabilities) cover a range of diagnostic conditions. Several broad categories defined by IDEA and state law (e.g., noncategorical identification, developmental delay, other health impairment) are used as catch-alls for students who clearly display impairment but cannot be assigned a definitive label because age makes specific diagnosis unreliable or because the nature of the problem does not fit into another category.

Because most of the disability categories are broadly defined, characteristics often intersect, and a student may meet the criteria for more than one type of disability. For example, students with behavioral/emotional impairment and specific learning disabilities frequently have dual diagnoses. Students with perceptual or physical impairments span the normal range of cognitive and learning abilities. Although below-average cognitive functioning is an exclusionary factor to identify specific learning disabilities, studies investigating special education placement decision-making have revealed failures to differentiate students with mental retardation and a consequent overuse of the learning disabilities category as more socially acceptable for students who show cognitive deficiencies (Gottlieb, Alter, Gottlieb, & Wishner, 1994; MacMillan, Gresham, & Bocian, 1998). Recent research also has identified significant overlaps between students with language impairments and reading disabilities (McArthur, Hogben, Edwards, Heath, & Mengler, 2000), language impairments and emotional/

Address: Xiangdong Yang, Center for Educational Testing and Evaluation, Joseph R. Pearson Hall, University of Kansas, 1122 West Campus Road, Lawrence, KS 66045-3101; e-mail: xiangdon@ku.edu

behavioral disorders (Caulfield, Fischel, DeBaryshe, & Whitehurst, 1989; Griffith, Rogers-Adkinson, & Cusick, 1997; Toppelberg & Shapiro, 2000), and comorbidity among these three disabilities (Tomblin, Zhang, & Buckwalter, 2000).

In one well-investigated area, the distinction between special education and general education has been called into question. In reading research spanning the past 2 decades, repeated studies have attempted to distinguish between students with specific reading disabilities who have IQ–achievement discrepancies and those without discrepancies. The reading skills profiles of these two groups are remarkably similar, as are the interventions that are most effective (Francis, Shaywitz, Stuebing, Shaywitz, & Fletcher, 1996; Shaywitz, Fletcher, Holahan, & Shaywitz, 1994; Siegel, 1989; Stage, Abbott, Jenkins, & Berninger, 2003; Stuebing et al., 2002). A general conclusion of this body of research is that reading disabilities occur in children at all levels of cognitive functioning who may or may not meet special education eligibility criteria. Instructional interventions should target the specific reading difficulties or deficiencies that students experience, regardless of their eligibility for special education services.

Special education categories, then, are likely to be heterogeneous rather than homogenous, and the dividing lines are blurred even between some special education and general education groups. According to Brinker (1990), special education categories are primarily administrative units: “We have persisted in our maintenance of categories of exceptionality that encompass individual differences that vary as greatly as the differences between categories” (p. 182). Further, special education classification does not stipulate services or placement options for individual students (Brinker, 1990). If special education classification does not provide sufficient reliable information to prescribe effective services, what does?

Another research thread involving commonalities among student groups has been the search for aptitude–treatment interactions (ATIs). ATIs are expected to occur when instructional programs or remedial interventions target the individual needs or weaknesses of certain students while not addressing those of other students. ATIs are at the heart of instruction for groups of students who share common characteristics. Among the aptitudes studied in ATI research have been attributional style, language and cognitive abilities, strategy use, and reading difficulty (Speece, 1990), as well as psycholinguistic processing and modality preferences among students with learning disabilities (Vaughn & Linan-Thompson, 2003).

ATI research has had only limited success in identifying useful interventions (Deno, 1990; Lloyd, 1984; Vaughn & Linan-Thompson, 2003). For example, research on cooperative learning in mathematics has supported the intervention’s effectiveness for two groups—special needs and general education students—without any interaction effect differentiating the groups (Slavin, Madden, & Leavey, 1984). A study on the effectiveness of integrated and segregated preschool classrooms confirmed earlier findings that integrated environments were more beneficial for higher functioning students, whereas segregated settings were better for lower functioning students (Cole, Mills, Dale, & Jenkins, 1991). That study measured only the severity of disability among young children identified as developmentally delayed, a heterogeneous category used for deficits in cognitive, language, social, gross motor, or fine motor development. For students with learning disabilities, individualized instruction based on processing problems, modality matching, and multisensory instruction has failed to show positive effects on academic achievement (Vaughn & Linan-Thompson, 2003).

In sum, ATI research methods have not generally proved useful for the study of individual differences. Aptitude differences may exist, but current ATI methodology is so fraught with sources of error in both conceptualization and measurement as to be rendered useless. Many specific difficulties with ATI research have been catalogued. Although Cronbach defined and proposed the ATI model for research on individual differences in 1957, he had abandoned the ATI model by 1975 due to intractable measurement problems (Reschly & Ysseldyke, 2002). Fuchs and Fuchs (1986) delineated four problems with ATI research:

1. Knowledge of student characteristics and aptitudes is incomplete.
2. Tests of student characteristics are not technically adequate.
3. Test administration may not be unbiased for children with different characteristics.
4. We cannot specify all the possible interactions among students, teachers, environment, and possible interventions.

Speece (1990) pointed out the “weak conceptualizations” of learning problems and, by extension, aptitudes defined and studied by researchers (p. 140). In addition to echoing Fuchs and Fuchs’ (1986) concern about the adequacy of the measurement instruments used in ATI studies, Speece suggested that the empirical failure of much ATI research is due to the heterogeneous makeup of the students within an aptitude group, in which other meaningful attributes are not defined or measured. Measurement difficulties are compounded by methodological problems that occur when high intraclass correlations, such as for students nested within classes or schools, are not considered. This may lead either to masking of true interaction effects or to spurious findings that are in fact due to other causes, such as group effects that violate the assumption of independence of measurement at the level of the individual (Sheehan & Han, 1996). Finally, ATIs are based on correlational effects that apply to groups of students who presumably share an aptitude, but the effects are not necessarily prescriptive at the individual student level and do not consider the individual student’s interaction with a particular classroom environment (Brinker, 1990; Deno, 1990; Sheehan & Han, 1996; Speece, 1990). ATIs have relied on the a priori classification of students into groups, using expensive diagnostic procedures involving instruments with questionable predictive powers and then extending generalizations, which are no more than probable predictions about the characteristics of the group, to individual group members with unfounded certainty. As Deno (1990) pointed out, "General policy decisions can be made on such group information, but specific programming decisions cannot" (p. 165).

ATI-based thinking has not entirely lost its grip on special education practice. Vaughn and Linan-Thompson (2003) stated that interventions for students with learning disabilities based on the cognitive process deficits that were hypothesized to underlie their learning problems failed to influence academic outcomes and detracted from instruction in areas of academic weakness. Modality-based interventions, preferred learning styles, and multisensory approaches such as tactile-kinesthetic reading instruction have not been shown to be effective in improving outcomes for students with learning problems. Nonetheless, these remedial methodologies for learning disabilities have become broadly associated with the field, and their use persists.

Given the heterogeneity of special education classifications and the failure of research based on individual aptitudes to show differential treatment effects as a function either of disability type or of other student characteristics, questions arise about whether groups classified by disability are indeed similar and how these similarities might be revealed. Do students within disability classifications have aptitude or achievement commonalities? Can distinct profiles of ability be identified within or across categories? Can special education students be classified into groups that share common characteristics of intellectual skills or learning abilities? Do such groups form along special education categorical lines or according to some other pattern? The current study will use outcome data to investigate whether students with disabilities form distinct groups on the basis of achievement.

The purpose of this study was to investigate the differentiation of mathematical abilities of students with different categories of disability using latent class analysis (LCA). The primary research hypothesis is that some latent classes, or groups, exist in the special education student population. Each of the classes represents a distinctive pattern of mathematical problem-solving behaviors. Members of the same class share the same cognitive characteristics of mathematical problem solving. The membership of these special education students in the latent classes may or may not be consistent with their disability categories. Specifically, two questions were of interest:

1. Do special education categories differentiate students with disabilities from each other in terms of the ability to solve different categories of mathematics problems?
2. If so, are the differences in qualitatively different mathematical skill profiles or quantitatively different levels of overall mathematics ability?

Theoretical Framework

This study explores the heterogeneity of a population of fourth-grade students with disabilities. The authors applied LCA, as well as item response theory (IRT), to achieve this purpose. Justification and description of both methodologies are given in the next section.

Latent Class Analysis

Heterogeneous populations are common in such social sciences as education and psychology. The heterogeneity is usually caused by existing subpopulations in the total population. Participants within a subpopulation are more similar to each other, whereas participants across subpopulations are usually less similar. In this case, statistical inferences obtained from procedures that assume homogeneous populations may be misleading, a phenomenon commonly referred to as Simpson's paradox (Agresti, 1996; Simpson, 1951).

Heterogeneity of the target population can be handled by one of two approaches, depending on whether the source of heterogeneity is known (Luke & Muthén, in press). If a source of heterogeneity is known, the subpopulation can be identified on the basis of that source. For example, if gender is known as a source of heterogeneity, the characteristics being investigated in the study can be statistically compared across the subpopulations defined by gender. Common statistical methods that may be applied in this case include the *t*-test, ANOVA or MANOVA, regression, and multigroup common factor analysis (MG-CFA; Jöreskog, 1971; Stevens, 1992). If the source of heterogeneity is unknown, subpopulations cannot be defined explicitly. As Luke and Muthén stated, in this case, subpopulations are called *latent classes* because they are not distinguishable on the basis of observed features. Statistical methods that may be suitable in this case include LCA (Lazarsfeld & Henry, 1968; McCutcheon, 1987; Vermunt & Magidson, 2002a), cluster analysis (Everitt, 1993; Kaufman & Rousseeuw, 1990), latent profile analysis (Vermunt & Magidson, 2002a), and finite mixture models (McLachlan & Peel, 2000). Luke and Muthén provide a more comprehensive survey of these methods.

Both cluster analysis and LCA can be applied to classify similar participants into classes, but the latter has several advantages over the former (Vermunt & Magidson, 2002a, 2002b). First, in traditional cluster analysis, the number of clusters is arbitrary. In LCA, on the other hand, theoretical formulations for each cluster can be specified directly and tested empirically by the dataset under investigation. Moreover, LCA allows more rigorous methods to be applied in comparing alternative models, such as likelihood-ratio tests, Akaike's information criteria (AIC), Bayesian information criteria (BIC), or Consistent AIC (CAIC). Second, LCA is robust to different scaling of the observed variables, which is always an issue in traditional cluster analysis. Third, LCA takes into account the uncertainty of a subject's membership in a latent class; traditional cluster analysis cannot do this. Results from a simulation study showed that LCA outperformed traditional cluster analysis even in the settings in which the simulated data were favorable to cluster analysis (Vermunt & Magidson, 2002b). On the other hand, if applied in an exploratory fashion, LCA is similar to cluster analysis so that LCA is often labeled as mixture likelihood approach to clustering (McLachlan & Basford, 1988; Everitt, 1993), model-based clustering (Banfield & Raftery, 1993; Bensmail, Celeux, Raftery, & Robert, 1997; Fraley & Raftery, 1998), mixture-model clustering (Jorgensen & Hunt, 1996; McLachlan, Peel, Basford, Adams, 1999), and latent class clustering analysis (Vermunt & Magidson, 2000, 2002a).

In short, LCA allows the researcher to identify a set of mutually exclusive and exhaustive latent classes from the measurement of a set of discrete variables (McCutcheon, 1987). Classical LCA includes three basic assumptions:

1. Each participant can be classified into one and only one latent class.
2. The set of observed variables is a set of imperfect measures of these latent classes.
3. Given a participant's class membership, the probability of answering one observed variable is independent of the probabilities of answering other observed variables, an important assumption called *local independence* (see Vermunt & Magidson, in press, for various relaxations of this assumption).

Based on the three assumptions, the latent class (LC) model for dichotomous data can be formulated as the following: Suppose a set of n dichotomous items were administered to a sample of N participants and $x_i = 1$ when item i, $i = 1, 2 \ldots n$, is correctly solved and $x_i = 0$ otherwise. P_i denotes the probability of getting item i correct, and, therefore, the probability of getting item i incorrect is $1 - P_i$. Assume there are k latent classes in the sample, and denote P_{ik} as the probability of getting item i correct given a participant in the *kth* latent class. If further assuming that responses are independent within a latent class, then P_r, the probability of getting a given response pattern (r) is given as

$$P(r) = \sum_{k=1}^{K} \pi_k P(r \mid k) = \sum_{k=1}^{K} \pi_k \prod_{i=1}^{n} P_{ik}^{x_i} (1 - P_{ik})^{1-x_i}$$

where $P(r|k)$ is the probability of getting the response pattern r given the *kth* class. π_k is the probability of being in the kth latent class. Equivalently, π_k can be referred to as the latent class proportion or the size of the *kth* latent class. When applied to real data, both π_k and P_{ik} are parameters that need to be estimated.

If π_k and P_{ik} are specified as free parameters to be estimated, the corresponding LC model is called unrestrained. In some situations, however, the unrestrained LC model is not identifiable, which means that either a unique set of parameter estimates is not present (Vermunt & Magidson, in press) or too many parameters need to be estimated so that no degrees of freedom are available. A common practice in this case is to constrain certain model parameters to a fixed value or to be equal to each other to achieve model identification (Formann, 1985). Once the model is identified and parameters are estimated, the nature of each class can be determined by plotting P_{ik} for the given class and examining the corresponding profile.

Important issues in LCA are model fit and model selection. Two approaches are usually applied to evaluate model fit. One approach is to directly compare the differences between observed and predicted. Two commonly used statistics within this approach are the Pearson chi-squared statistic χ^2 and the likelihood-ratio chi-squared statistic G^2 (Agresti, 1996). Both statistics are special cases of a family of test statistics known as power divergence statistics (Cressie & Read, 1984; Read & Cressie, 1988). Cressie and Read also proposed a test statistic that compromises between χ^2 and G^2. For all three test statistics, a small value of the statistic relative to its degrees of freedom, which corresponds to a large p value, indicates a good model fit. A special case of this approach is when the sample size is relatively small compared to the number of measured variables—specifically, when the expected frequencies in some possible combinations of the variables are small (< 5) or zero. In this case, the data are called sparse data. Then the aforementioned test statistics do not work well and the obtained p values of model fit cannot be trusted (Dayton, 1998; Vermunt & Magidson, in press). Instead, empirical p values can be obtained by evaluating empirical distributions of the test statistics using the bootstrapping method (Efron & Tibshirani, 1993).

The second approach is to compare the relative goodness-of-fit among different models. There are two different situations within the second approach. One situation is when the two models being compared are nested, one within the other. Two models are nested when the simpler model between the two is obtained by imposing certain constraints on one or more parameters in the more complex model (Loehlin, 1998). For example, the LC model specified in the prior equation is an unrestrained model with k latent classes, in which all the P_{ik} are different parameters. We can then obtain a simpler model by constraining the P_{ik} in, say, class 3, to be equal to each other, denoted as $P_{13} = P_{23} = \ldots = P_{n3}$. In this case, the simpler model is nested in the unrestrained model with $n - 1$ fewer parameters. Because the more complex model contains more parameters than the simpler one, it will fit the data better than or as well as the simpler model. For each model, the corresponding test statistic G^2 can be calculated. Statistical theory states that the difference between the two G^2s has an approximate chi-square distribution, with degrees of freedom of the distribution equaling the difference in the number of parameters between the two models. In the above example, this degree of freedom is $n - 1$ because the complex model has $n - 1$ more parameters than the simpler one. Large values for the difference between the two G^{2}s relative to the degree of freedom, corresponding to a small p value, indicate that the goodness-of-fit of the more complex model is statistically better than that of the simpler one. It should be pointed out that the two models in our example have the same number of latent classes, with constraints being imposed on within-class parameters. Although counterintuitive, LC models with different numbers of classes are *not* nested one within the other (Dayton, 1998; McCutcheon, 1987; Vermunt & Magidson, 2002a). This is, in fact, the second situation for comparing goodness-of-fit between models (i.e., model selection when they are not nested). The statistic using G^2 difference does not work well in this case. Popular approaches in this situation are to use information criteria, such as AIC, BIC, or CAIC (Vermunt & Magidson, 2002a). The idea behind all three information criteria is the same: The goodness-of-fit of a given model is penalized for its complexity (Loehlin, 1998). For models with the same level of goodness-of-fit, the one with fewer parameters is favored because it is more parsimonious. Smaller values of the information criteria usually mean better fit. Results from different information criteria may contradict each other. In that case, BIC is preferred because research has shown that BIC is more consistent (Li & Nyholt, 2001) and tends to select a more parsimonious model than AIC (Lin & Dayton, 1997).

Item Response Theory

Since the 1950s, item response theory (IRT) has become the new theoretical basis for educational and psychological measurement (Embretson & Reise, 2000). The primary question in IRT focuses on how the probability of a correct response to an item relates to examinees' mental characteristics, such as intelligence, academic proficiency, or attitude, and the item's properties, such as item difficulty, item discrimination, and so forth. Such a relationship is usually formulated through an item response function (IRF) or item characteristic curve (ICC). As many as three aspects of an item's characteristics can be captured by the ICC (e.g., item difficulty, item discrimination, guessing). *Item difficulty* is how hard an item is. Higher ability is needed to solve more difficult items. *Item discrimination* indicates how sensitive the probability of answering an item correctly is to changes in an examinee's ability level. Items with high discrimination are favored over items with low discrimination because the former is more sensitive to the differences among examinees of various ability levels; poorly discriminating items should be eliminated from the test. The *guessing* parameter indicates the probability of guessing the item correctly by examinees with very low ability. Items with high values of the guessing parameter c should also be eliminated from the test (Embretson & Yang, in press). Specialized computer programs such BILOG-MG3 (Zimowski, Muraki, Mislevy, & Bock, 2002) can be used to obtain estimates for these parameters.

Method

Measurement

Mathematical ability was measured by responses to a state-mandated mathematics assessment in a midwestern state. Items in the assessment were designed to measure the state's curricular standards for mathematics. The mathematics curriculum standards are composed of three levels: standard, benchmark, and indicator. A standard is a "general statement of what a student should know and be able to do in academic subjects" (Kansas State Board of Education, 1999). The state's standards include four mathematics content areas: number and computation, algebra, geometry, and data. Benchmarks are more specific than standards. They are used to define what the student must be able to do to meet the standard. Indicators list the detailed knowledge or skills that students should demonstrate in order to meet the benchmark (see Table 1 for the benchmarks included in each standard). Within the content specification, two categories of items, knowledge and application, are included in the test. Items that measure knowledge

TABLE 1. The Kansas Curricular Standards for Mathematics

		Item	
Standard	**Benchmark**	**Set-1**	**Set-2**
Number & computation	Number sense	NP1I3	NP4I12
	Number systems and their properties	NP3I9	NP1I5
	Estimation	NP2I12	NP3I12
	Computation	NP1I12	NP1I1
Algebra	Pattern	NP2I1	NP2I10
	Variables, equation, and inequality	NP2I2	NP2I2
	Functions	NP2I4	NP2I6
	Models	NP2I8	NP2I7
Geometry	Geometry figures and their properties	NP3I4	NP3I4
	Measurement and estimation	NP3I1	NP3I2
	Transformational geometry	NP3I11	NP3I10
	Geometry from algebraic perspective	NP3I8	NP3I8
Data	Probability	NP4I9	NP4I9
	Statistics	NP4I2	NP4I8

are those that measure mathematical facts, concepts, or solutions of one-step story problems; application items measure how to apply mathematical knowledge to a real-world situation or to carry out a procedure such as computation.

The state mathematics assessment for fourth grade was used in this study. This test was designed by following accepted principles of standardized test development, including item preparation by content area teachers at the grade levels tested, large sample field tests, CTT and IRT item analysis, differential item functioning (DIF) analysis, and bias review. There are 52 items in the test. Table 2 lists examples of items that were designed to measure each of the four content areas and gives a short description of each item in terms of the specific content area it is supposed to measure.

Sample and Data

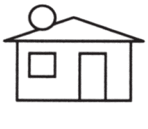
The goal of this study was to explore within the population of special education students the existence of different latent classes in which students share the same cognitive characteristics with respect to mathematical proficiency. Therefore, only identified special education students were included. All students with disabilities in fourth grade across the state during the 2001–2002 academic year were included in the analysis. Twelve categories of disability were included in the data set: hearing impairment, visual impairment, speech/language impairment, physical impairment, specific learning disability, emotional disorder, mental retardation, autism, traumatic brain injury, deaf-blindness, noncategorical identification, and other health impairment. The students in this study were identified as having disabilities according to state special education procedures and criteria (Kansas State Department of Education, 2001). The State Department of Education monitors identification and placement procedures, as well as special education services, for compliance on a routine basis. The disability category for each student was coded on the student's answer sheet. The number of students within each of the disability categories is given in Table 7.

There were 3,289 students with disabilities in the data set: 2,224 boys and 1,065 girls. Responses from all students with disabilities who completed all four parts of the 52-item test form were evaluated for missing data. Among these observations, 608 cases had responses to at least 1 of 52 items that were either missing or multiply marked. These cases were not included in the present analysis. As a result, 2,681 cases were used in the actual analysis.

Rationale for the Analysis

The research questions for this study were answered through statistical analysis in three stages. In Stage 1, because the hypothetical sources for the existence of different latent classes

TABLE 2. Example Items for Each Content Area

Standard	Example item	Item description
Number & computation	Susan bought 24 books for $29.99 each. About how much did she spend? A) $400 B) $550 C) $750 D) $900	This item measures the student's ability to determine if a problem calls for an exact answer or an approximate answer and perform the appropriate computation.
Algebra	A dog can run 30 feet in 2 seconds, 45 feet in 3 seconds, and 75 feet in 5 seconds. If this relation continues, how many feet will the dog run in 9 seconds? A) 100 B) 135 C) 145 D) 180	This item measures the student's ability to recognize relationships between whole numbers.
Geometry	Which shape was not used to form this house? A) rectangle B) triangle C) parallelogram D) circle 	This item measures the student's ability to recognize and describe geometric figures and their basic properties.
Data	Five students in John's class have the following weights: 45, 40, 38, 41, 46. What is the range among them? A) 6 B) 8 C) 38 D) 46	This item measures the student's ability to use basic statistical measures for a whole number data set, such as mean, mode, range, and so forth.

are common cognitive characteristics, which are unknown a priori, the LC modeling procedure was applied to identify such subpopulations. Various LC models were fitted to the data. Rigorous model fit and selection procedures were used to choose the best model. Then, the number of the subpopulations, as well as the probabilities of their occurrences, were estimated as model parameters.

In Stage 2, the nature of each subpopulation was examined by inspecting the characteristics that the subpopulation exhibited in terms of mathematical problem solving. This was done by plotting the class-specific item response probabilities P_{ik} against each item across the classes. Many possibilities exist for profile differences across classes. One possibility is that the profile for one class would cross over the profile for another class. For example, one class might show high probabilities of answering computational items correctly and low probabilities of answering geometric items correctly, whereas for another class the reverse pattern might be found. Because computational and geometric items usually tap different aspects of mathematical ability (the former sources on numerical operation, whereas the latter sources on spatial manipulation), such a difference is qualitative. It indicates that these two classes exhibit different combinations of various components of mathematical ability. If, on the other hand, the profiles from the two classes are parallel to each other and the probabilities of answering all items correctly for one class are consistently higher than those for the other class, the difference can be regarded as quantitative. It would indicate that one class exhibits higher levels on all components of mathematical ability.

In Stage 3, the relation between the identified latent subpopulations and various categories of disability was examined through statistical testing procedures, specifically the chi-square test. This is possible because the membership of each participant among the set of latent classes identified can be calculated through the LCA. If the cognitive profile of each student was completely determined by his or her disability category and different categories showed completely different profiles, the number of subpopulations identified through LCA would equal the number of disability categories. A perfect relationship would exist between the two. The corresponding theoretical value of chi-square statistic would be positively infinite. If, on the other extreme, the cognitive profile of each student had nothing to do with his or her disability category, the identified latent classes would be independent of students' disability categories and the corresponding theoretical value of chi-square statistic would be zero. More realistically, we may expect some relation between the identified latent classes and the set of disability categories. For example, students within certain disability categories may tend to be in certain classes.

Model Selection and Cross Validation

The data were submitted to the program WINMIRA2001 (Von Davier, 2001) to carry out the LCA. Because the number of possible response patterns was much larger than the number of examinees, the data were relatively sparse. As mentioned earlier, the chi-square p-value approximation for the goodness-of-fit statistics is not appropriate as a model selection criterion in this case. Instead, a parametric bootstrap approach was used to calculate the fit statistics, and the empirical values of both Creiss-Reid and Pearson chi-square statistics were considered if nested LC models were compared.

As mentioned earlier, for LC models with different numbers of classes, the chi-square difference test is not appropriate. Instead, various information criteria (e.g., AIC, BIC, CAIC) were used as model selection criteria, combined with the bootstrap goodness-of-fit statistics. The model with the minimum values of the information criteria is considered to be the model of choice. When the minimal values among the three information criteria are inconsistent, the model with minimal value of BIC is selected as the preferred model.

No previous research was available in terms of differentiation of mathematical ability among different categories of disability; thus, the whole test was split into two equivalent halves to carry out the cross validation. The split was based on the three levels of standards. In the following discussion, the two data sets are named *half-1* and *half-2,* respectively.

Results

Initial Results From Latent Class Analysis

Unrestrained LC models with one, two, three, four, and five classes were fitted to the data in *half-1.* Table 3 presents the information criteria and the goodness-of-fit statistics obtained from the parametric bootstrapping method. Minimum values of AIC, BIC, and CAIC are shown in bold. Note that the model with one class provides poor fit to the data (both empirical p-values of Cress Reid and Pearson chi-square from parametric bootstrapping method are less than 0.05). This suggests that the population of the students with disabilities is not homogenous. Further identification of the true structure in the data is needed. Although the goodness-of-fit statistics for all other models in Table 3 appear adequate, the model with three classes has the minimal value of BIC and CAIC. Note that the model with five classes, which is a more complex model, has the minimal value of AIC. Because the more parsimonious model is preferred, the model with three latent classes is considered the model of choice.

Item difficulty profiles for the LC model with three classes are depicted in Figure 1. Each line in the figure represents the item difficulties for students within that specific class. Note that the same item shows different difficulties for different classes. Items appear more difficult for class 1 and less difficult for class 3, but moderately difficult for class 2. An interesting feature of the item profiles is that they are almost parallel to each other across the three classes, which means that the differences among classes are not qualitative but quantitative. The classes are different in their abilities: Class 1 has

TABLE 3. Models Fitted to the Data *half-1*

Model	AIC	BIC	CAIC	Empirical p value using parametric bootstrap method	
				Cress Reid	Pearson χ^2
LCM with 1 class	91800.27	91954.21	91980.21	0.000	0.025
LCM with 2 classes	87654.11	87967.92	88020.92	1.000	1.000
LCM with 3 classes	86958.24	**87431.9**	**87511.9**	0.825	0.925
LCM with 4 classes	86886.93	87520.45	87627.45	0.950	0.925
LCM with 5 classes	**86852.1**	87645.53	87779.53	0.725	0.900

Note. AIC = Akaike's information criteria; BIC = Bayesian information criteria; CAIC = Consistent Akaike's information criteria; LCM = Latent class model. Boldface indicates minimum values.

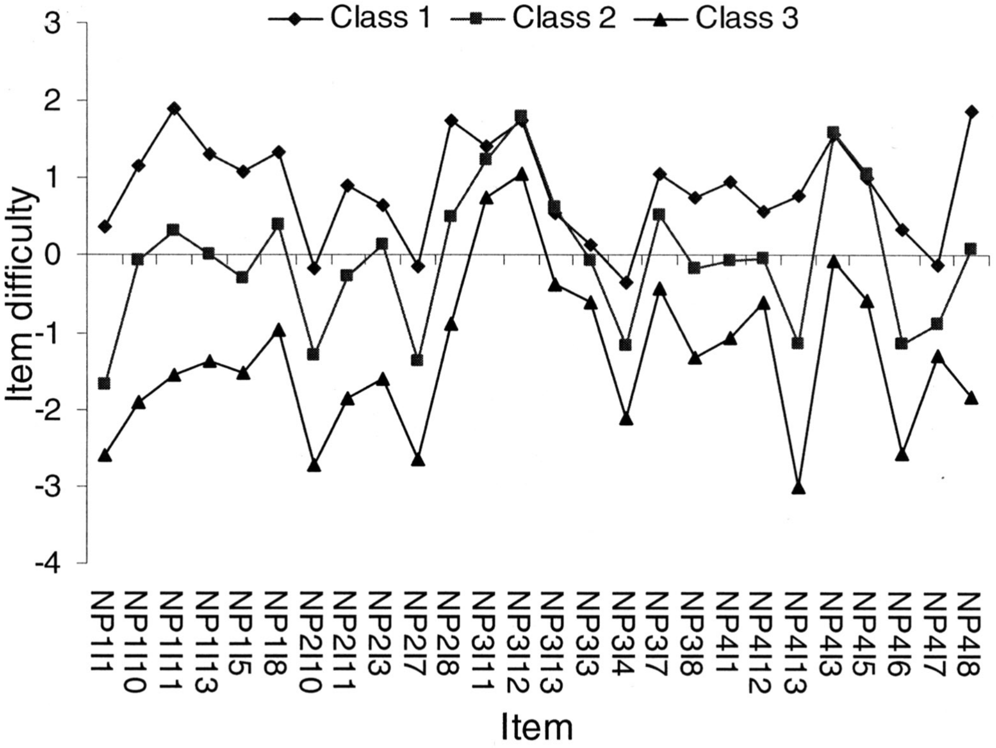


FIGURE 1. Item difficulty profiles for model with three latent classes.

the lowest ability because items have relatively high difficulty parameters for this class, and class 3 has the highest ability.

Item Response Theory Analysis

Before interpreting the nature of each class and making a conclusion based on Figure 1, a more careful inspection reveals that several items do not function well. For example, the item labeled *np3i12,* as well as several other items within that neighborhood, has very high difficulty parameters for all the classes. Further inspection reveals that these items do not fit the model in terms of the Q index, which is an item-fit index calculated in the program WINMIRA2001. To investigate the properties of these items, the data in *half-1* were submitted to BILOG-MG3 using two- and three-parameter logistic models. Results from the two-parameter logistic model confirmed that these items have very low discrimination parameters (most of their discrimination parameters are about 0.16 compared to the mean values of the parameters, which is near 1.0). Estimates from the three-parameter logistic model, however, showed that their discrimination parameters were acceptable, but with relatively high guessing parameters (about 0.3). The item *np3i13* has a guessing parameter as high as 0.46, which means that nearly one half of the students can correctly answer this item just by guessing.

Latent Class Analysis After Scale Examination

Scale Reconstruction. Results from item response theory analysis of *half-1* items reveal that some of the items in the test do not function well for this population. Thus, rather than continue with the planned cross-validation using *half-2* items, all 52 items in the test were reexamined and two different sets of items were selected. Each item in the two new sets of items was selected on the basis of its psychometric properties (i.e., item difficulty, discrimination, and guessing parameters) and its correspondence to the four mathematics content areas. The last two columns of Table 1 show the two sets of items, each having 14 items, denoted as *set-1* and *set-2*. Note that four items are identical in the two sets: *np2i2, np3i4, np3i8,* and *np4i9*. When only one item measured a specific content area, the item was used in both item sets.

Content-Specific Latent Class Analysis. Several LC models were fitted to the data for each of the four content areas. Only items that measured the given content area in *set-1* were included in each analysis, except for the last content area. In the fourth content area, *data,* only one item, *np4i9,* measures probability, and two items measure statistics. Item *np4i9* is a very difficult item. For completeness of the content coverage, however, all three items in *set-1* and *set-2* are included in the analysis. The models fitted are summarized in Table 4. Models were selected for each mathematics content area on the basis of the goodness-of-fit index and the minimal value of information criteria, which is given in Table 4. In Table 4, models that show best fit are in boldface. For the models of choice, all the fit indices and information criteria gave consistent results. Generally speaking, an LC model with two classes seems adequate to characterize the data for each of the four content areas. For the content area *data,* only three items are involved, so some constraints have to be imposed on the parameters to identify the model. Specifically, the response probabilities for item *np4i9* were constrained to be equal across classes. The class-specific response probabilities for the other two items were constrained as equal within class 1, whereas those for class 2 were set free. The rationale for choosing this pattern of constraints is that students in class 1 are viewed as *nonmasters* of the content area, with very low probabilities of getting items right by guessing. Therefore, no differences are reflected in item difficulty because guessing does not relate to item difficulty. The model showed an adequate fit to the data, as shown in Table 4.

Class probabilities and specific response probabilities of items given class membership are presented in Table 5. For most of the items, students in class 1 have probabilities that are near the chance level. Given that each item has five options, the chances of getting an item correct by guessing is about 0.20. This suggests that students in class 1 do not master the skills or knowledge to solve these items in the four content areas. As such, we can call students in class 1 *nonmasters*.

TABLE 4. Models Fitted to Each Content Area of the Data

Model	AIC	BIC	CAIC	G^2	*df*	*p*
Number and computation (4 items)						
Unrestrained LCM: 1	13661.39	13684.97	13688.97	145.1	11	0.00
Unrestrained LCM: 2	**13536.8**	**13589.9**	**13598.9**	**10.53**	**6**	**0.10**
Unrestrained LCM: 3[a]	13540.55	13623.06	13637.06	4.26	2	0.04
Algebra (4 items)						
Unrestrained LCM: 1	13789.23	13812.81	13816.81	452.78	11	0.00
Unrestrained LCM: 2	**13353.1**	**13406.2**	**13415.2**	**6.68**	**6**	**0.35**
Unrestrained LCM: 3	13359.09	13441.6	13455.6	2.64	2	0.10
Geometry (4 items)						
Unrestrained LCM: 1	13445.54	13469.12	13473.12	50.85	11	0.00
Unrestrained LCM: 2	**13409.5**	**13462.5**	**13471.5**	**4.8**	**6**	**0.57**
Unrestrained LCM: 3	13418.44	13500.95	13514.95	3.75	2	0.05
Data[b] (3 items)						
Unrestrained LCM: 1	9695.28	9712.96	9715.96	86.74	4	0.00
Restrained LCM[c]: 2	**9613.51**	**9642.98**	**9647.98**	**0.97**	**2**	**0.61**

Note. AIC = Akaike's information criteria; BIC = Bayesian information criteria; CAIC = Consistent Akaike's information criteria; LCM = Latent class model. Models in bold type were selected for each content area on the basis of the goodness-of-fit index and the minimal value of information criteria.
[a]According to Dayton (1998), an unrestrained LCM with 3 latent classes for 4 items has 2 degrees of freedom, rather than 1 degree, as calculated in the model. [b]One item from data *set-2*, the other two items from data *set-1*. [c]Response probabilities of item *np4i9* were constrained as equal across classes, whereas response probability of item *np4i2* and *np4i8* were constrained as equal within class 1.

Correspondingly, students in class 2 demonstrate higher probabilities in correctly solving the items. We call students in class 2 *masters.* Across all four content areas, students in class 2 have consistently higher probabilities of correctly solving these items than students in class 1 do. This suggests a generic difference in their mathematical abilities, rather than differences in specific content areas.

The size of each class within each of the four content areas is also displayed in Table 5. Note that the sum of the classes within a content area is 1, because all the students have to fall into a class. More students seem to belong to class 1, which shows low mathematical ability. This result is expected, given that the population under investigation consists of students with disabilities. For the areas of number and computation, about 80% of students belong to the lower ability category, suggesting this is a difficult area for these students in general. About 55% of the students have lower ability in algebra, and about one half of the students have difficulty in solving statistics. Note that for the content area *data,* the two class-specific response probabilities for item *np4i9* are equal (0.163) across classes because they were constrained to be so. The estimated response probability is very low, however, suggesting that the item is truly a very hard item. Another constraint is set for the two response probabilities of items *np4i2* and *np4i8* within class 1. It is interesting that more students belong to class 2 in terms of solving geometric items. Inspecting each item's class-specific response probabilities reveals, however, that the probabilities of getting a geometric item correct vary greatly across items, in terms of both the difficulty level and the differences between the two classes. It appears that the differences of response probabilities for the geometric items between class 1 and class 2 are generally smaller than the differences in other content areas, which may suggest that these items do not differentiate the students well.

Overall Analysis and Cross Validation. Combining results from all four content areas, data on all 12 items of *set-1* were submitted to the program WINMIRA2001 to get an overall picture of the latent structure that may exist in the data. Unrestrained LC models with one to seven classes were fitted to the data. Table 6 summarizes the goodness-of-fit indices obtained from the parametric bootstrapping method and the values of the information criteria for these models. Based on the criteria discussed earlier, the model with two latent classes is the preferred model. Table 6 also shows the corresponding models fitted to data in *set-2*. The same model was selected on the basis of the minimal value of the information criteria.

Class proportions and item class-specific response probabilities are depicted in Figure 2. Similar patterns of class-specific response probabilities can be observed from the upper and lower panels of Figure 2, which were derived from data *set-1* and *set-2,* respectively. Overall, the same structure exists in the two data sets, although some variations exist, most likely due to the differences in specific item properties. Again, note that the item response probability profiles across the two classes are essentially parallel to each other, which suggests that the differences between the two classes are primarily due to the quantitative differences in their generic mathematical abilities. Also, consistent with previous results, more students (63%) belong to the class with low response probabilities, given the special population in this analysis.

Relation Between Categories of Disability and Class Membership

Accepting the model with two latent classes as the model of choice, each student was classified into one of the two latent classes. Table 7 shows the cross-tabulated data between different categories of disability and the classes. This is the re-

TABLE 5. Class Proportion and Item Class-Specific Probabilities for Models Fitted to Each Content Area

Content area		Item class-specific probability				Class proportion
Number and computation		NP1I3	NP3I9	NP2I12	NP1I12	
	Class 1	0.280	0.243	0.276	0.244	0.793
	Class 2	0.504	0.611	0.635	0.714	0.207
Algebra		NP2I1	NP2I2	NP2I4	NP2I8	
	Class 1	0.611	0.356	0.267	0.172	0.551
	Class 2	0.961	0.737	0.751	0.560	0.449
Geometry		NP3I4	NP3I1	NP3I11	NP3I8	
	Class 1	0.533	0.356	0.181	0.36	0.483
	Class 2	0.89	0.488	0.272	0.618	0.517
Data		NP4I9	NP4I2	NP4I8		
	Class 1	0.163	0.218	0.218		0.506
	Class 2	0.163	0.658	0.612		0.494

TABLE 6. Models Fitted to the Data Set-1 and Set-2

				Empirical *p* value using parametric bootstrap method	
Model	AIC	BIC	CAIC	Cress Reid	Pearson χ^2
		Set-1			
LCM with 1 class	46955.26	47037.77	47051.77	0.000	0.000
LCM with 2 classes	45680.30	**45851.22**	**45880.22**	0.875	0.875
LCM with 3 classes	45600.11	45859.4	45903.4	0.625	0.575
LCM with 4 classes	45596.32	45944.06	46003.06	0.300	0.475
LCM with 5 classes	45599.85	46036	46110	0.35	0.425
LCM with 6 classes	45603.95	46128.51	46217.51	0.45	0.475
LCM with 7 classes	**45586.16**	46199.13	46303.13	0.225	0.475
		Set-2			
LCM with 1 class	44416.76	44499.28	44513.28	0.000	0.000
LCM with 2 classes	42967.22	**43138.14**	**43167.14**	0.275	0.375
LCM with 3 classes	42882.48	43141.8	43185.8	0.175	0.275
LCM with 4 classes	42869.75	43217.49	43276.49	0.025	0.250
LCM with 5 classes	42869.38	43305.53	43379.53	0.05	0.275
LCM with 6 classes	**42867.29**	43391.85	43480.85	0.075	0.250
LCM with 7 classes	42868.1	43481.07	43585.07	0.125	0.350

Note. AIC = Akaike's information criteria; BIC = Bayesian information criteria; CAIC = Consistent Akaike's information criteria; LCM = Latent class model. Boldface indicates minimum values.

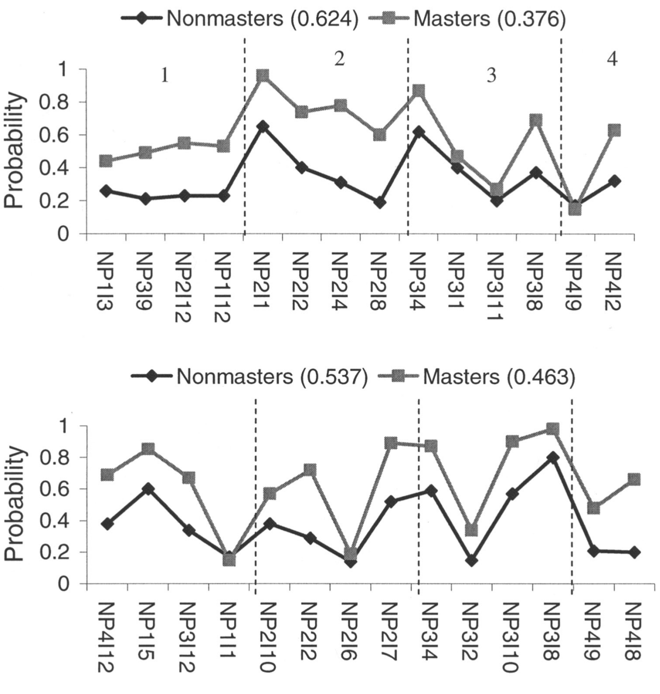


FIGURE 2. Class-specific response probabilities of items for each content area of *set-1* and *set-2*. *Note.* Upper panel = *set-1;* lower panel = *set-2;* 1 = Number and computation; 2 = Algebra; 3 = Geometry; 4 = Data.

sult from the LC model with two latent classes, based on the data *set-1*. Consistent with earlier results, class 1, which shows low mathematical ability, consists of 1,697 students, about 63% of the whole data set. The remaining 984 students in class 2 show relatively higher mathematical abilities. Most of the disability categories have more students in class 1. This suggests that this population of students with disabilities shows low mathematical ability in general. The distributions between lower and higher mathematical ability are not uniform across different categories of disability. The distribution across the two classes has a significant relationship with the type of disability, $\chi^2(1, N = 2{,}681) = 69.08, p < 0.0001$; after adjusting for cell counts, $\chi^2(9, N = 2{,}681) = 62.431, p < 0.0001$. The adjusted residuals for cell counts are also presented in Table 7. These residuals are the differences between observed frequencies and their expected values within each cell, adjusted in such a way that each adjusted residual has a large-sample standard normal distribution (see Note). Therefore, an adjusted residual that exceeds about 2 or 3 in absolute value indicates a great discrepancy between the observed frequency and its expected value. The sign of an adjusted residual reflects the direction of this discrepancy. A positive value means that there are more observed counts than expected within a given cell, and a negative value means the opposite.

A few categories of disability show large discrepancies between the observed and expected frequencies, such as mental retardation, speech and language impairment, and visual

TABLE 7. Special Education Category and Class Membership

Special education category	Membership Class 1 (0.624)	Adjusted residual	Class 2 (0.376)	Adjusted residual	Totals
Hearing impairment	13	–0.41	9	0.41	22
Visual impairment	3	–2.48	8	2.48	11
Speech and language impairment	353	–5.41	296	5.41	649
Physical impairment	11	0.45	5	–0.45	16
Specific learning disability	818	1.69	441	–1.69	1,259
Emotional disorder	91	–1.33	65	1.33	156
Mental retardation	63	5.29	4	–5.29	67
Autism	13	–0.41	9	0.41	22
Traumatic brain injury	5	–0.87	5	0.87	10
Deaf/blindness	1	0.76	0	–0.76	1
Noncategorical	190	1.93	87	–1.93	277
Other health impairment	136	2.35	55	–2.35	191
Totals	1,697		984		2,681

impairment. Students with mental retardation are more likely to show low mathematical proficiency as compared to students with other special needs (the adjusted residual is 5.29). On the other hand, mathematical ability for students with speech and language impairment or visual impairment is not affected as severely as other impairments (the adjusted residuals are −5.41 and −2.48, respectively). Specific learning disability also seems to have a relatively severe impact on students' mathematical ability (the adjusted residual is 1.69). Other health impairment also shows a negative impact on students' math abilities. Further information would be needed, however, about the specific characteristics of students in this category before any substantive explanation could be offered.

Discussion

In this study, LCA of a comprehensive test of mathematics taken by students with various disabilities reveals that the model with two latent classes is adequate to characterize the latent structure of the data. Although variations on class-specific response probabilities of items are observed, the same conclusion can be made on the basis of the results from cross validation on a separate data set with careful matching of content areas within the math test. A parallel relationship was observed between the class-specific response probabilities of each item across the two classes, which suggests that students within one class have consistently higher levels of mathematical performances than students in the other class and thus do not exhibit nonparallel profiles of mathematical skills. This leads to the conclusion that the differences in the two latent classes are quantitative rather than qualitative in nature. Although students in class 1 do not possess the proficiency to solve these math items, students in class 2 do, which reflects general differences in mathematical abilities between the two classes.

These results were not initially expected. The initial hypothesis was that qualitative differences in terms of math skills or cognitive abilities would be delineated in the profiles of the various classes. For example, one class might be good at solving some categories of items, whereas other classes would be better at other categories. Results obtained from this analysis indicate that this is not the case, at least for the data set under investigation. Students in some groups, such as those with cognitive impairments and specific learning disabilities, performed worse overall, whereas those from other groups, such as speech/language or perceptual impairments, tended to be distributed across both classes. No distinctive profiles among different categories of special education students were found, however, in either set of items, other than a generic difference in terms of overall mathematics proficiency.

Further investigation of the relationship between the different categories of disability and their class membership showed some general patterns of achievement for math abilities across disability categories. Students with mental retardation tend to have proportionally lower mathematical abilities than other categories of disability, and students with speech and language impairment and visual impairment tend to have higher mathematical abilities. These results are consistent with, and confirm, the identification criteria for the disability categories. Students with mental retardation are identified, in part, by their general low cognitive functioning with respect to peers. Therefore, greater representation in the lower math achievement group would be expected. Identification of speech/language and perceptual impairments does not imply lower cognitive functioning, and greater proportional placement in the higher achievement group of math ability would also be expected. The key finding is that lower achieving students in math perform like other low-achieving students in math regardless of disability classification.

Although results from this analysis do not conform to initial expectations, they provide important information about the initial research questions. Do different categories of disability differentiate students from each other in terms of mathematical ability? What is the nature of such differences? Results from this study show some evidence of the differentiating effects of mathematical abilities in different categories of disability. The differences among different categories of disability, however, are not in terms of types of skills but more in terms of general skill level. It appears from these data that students in different disability categories who need remedial mathematics instruction are more alike than different in their general levels of mathematical proficiency and in their responses to specific types of mathematics problems. Although this study did not address instruction, this outcome seems to confirm earlier ATI research that did not support different instructional interventions for special education students with aptitude differences except with respect to overall cognitive functioning (Cole, Mills, Dale, & Jenkins, 1991). This result also echoes research in reading, which has identified the same remedial interventions as most effective for students who lag in reading proficiency, regardless of special education status (Francis, Shaywitz, Stuebing, Shaywitz, & Fletcher, 1996; Shaywitz, Fletcher, Holahan, & Shaywitz, 1994; Siegel, 1989; Stage, Abbott, Jenkins, & Berninger, 2003; Stuebing et al., 2002).

When considering the results, some limitations of the study should be taken into account. First, the current study is limited to mathematics performance of students with disabilities in the fourth grade. Although the data set is adequately large and the numbers for most categories of special education within the data set are acceptable, similar analyses on data sets at various grade levels are desirable before these conclusions can be confidently asserted. Second, although the hypothesis that different cognitive characteristics, rather than categories of disability, serve as primary sources for heterogeneity in the special educational student population is a viable one, more measures of cognitive functioning are needed to understand adequately the relationship between cognitive characteristics and categories of disability. The current study is part of a larger project designed to investigate these issues more comprehensively. Replication in other academic and cognitive areas will be required.

AUTHOR'S NOTE

This study was supported by funding provided by the Kansas State Department of Education. Opinions expressed herein are those of the authors and do not necessarily reflect those of the sponsoring agency.

NOTE

The adjusted residuals are calculated by the following formulae: Let Res_{ij} stand for the adjusted residual in the cell at the intersection of row *i* and column *j* of the contingency table, then

$$\mathrm{Res}_{ij} = \frac{n_{ij} - \hat{u}_{ij}}{\sqrt{\hat{u}_{ij}(1 - p_{i+})(1 - p_{+j})}}$$

where n_{ij} and $\hat{u}_{ij}$ are the observed and expected frequency in the cell, respectively. p_{i+} is the marginal proportion of frequency in row *i,* across all the columns of *j, j* = 1, 2, . . . J, and $p_{+\,j}$ are the marginal proportion of frequency in column *j,* across all of the rows of *i, i* = 1, 2, . . . I.

REFERENCES

Agresti, A. (1996). *An introduction to categorical data analysis.* New York: Wiley.

Banfield, J. D., & Raftery, A. E. (1993). Model-based Gaussian and non-Gaussian clustering. *Biometrics, 49,* 803–821.

Bensmail, H., Celeux, C., Raftery, A. E., & Robert, C. P. (1997). Inference in model-based clustering. *Statistics and Computing, 7,* 1–10.

Brinker, R. P. (1990). In search of the foundation of special education: Who are the individuals and what are the differences? *The Journal of Special Education, 24,* 174–184.

Caulfield, M. B., Fischel, J. E., DeBaryshe, B. D., & Whitehurst, G. J. (1989). Behavioral correlates of developmental expressive language disorder. *Journal of Abnormal Child Psychology, 17,* 187–201.

Cole, K. N., Mills, P. E., Dale, P. S., & Jenkins, J. R. (1991). Effects of preschool integration for children with disabilities. *Exceptional Children, 58,* 36–45.

Cressie, N. A. C., & Read, T. R. C. (1984). Multinomial goodness-of-fit tests. *Journal of the Royal Statistical Society, B, 46,* 440–464.

Dayton, C. M. (1998). Latent class scaling analysis. *Sage University papers series on quantitative applications in the social sciences.* Thousand Oaks, CA: Sage.

Deno, S. L. (1990). Individual differences and individual difference: The essential difference of special education. *The Journal of Special Education, 24,* 160–173.

Efron, B., & Tibshirani, R. J. (1993). *An introduction to the bootstrap.* New York: Chapman & Hall.

Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists.* Mahwah, NJ: Erlbaum.

Embretson, S. E., & Yang, X. (in press). Item response theory. In G. Camilli, P. Elmore, & J. Green (Eds.). *Complementary research methods in education, 3rd edition.* Washington, DC: American Educational Research Association.

Everitt, B. S., & Hand, D. J. (1981). *Finite mixture models.* New York: Chapman and Hall.

Everitt, B. S. (1993). *Cluster analysis.* London: Edward Arnold.

Formann, A. K. (1985). Constrained latent class models: Theory and applications. *British Journal of Mathematical and Statistical Psychology, 38,* 87–111.

Fraley, C., & Raftery, A. E. (1998). *MCLUST: Software for model-based cluster and discriminant analysis* (Tech. Rep. No. 342). University of Washington, Seattle: Department of Statistics.

Francis, D. J., Shaywitz, S. E., Stuebing, K. K., Shaywitz, B. A., & Fletcher, J. M. (1996). Developmental lag versus deficit models of reading disability: A longitudinal, individual growth curves analysis. *Journal of Educational Psychology, 88,* 3–17.

Fuchs, L. S., & Fuchs, D. (1990). Introduction to special section: The importance of individual differences to special educator effectiveness. *The Journal of Special Education, 24,* 135–138.

Gottlieb, J., Alter, M., Gottlieb, B. W., & Wishner, J. (1994). Special education in urban America: It's not justifiable for many. *The Journal of Special Education, 27,* 453–465.

Griffith, P. L., Rogers-Adkinson, D. L., & Cusick, G. M. (1997). Comparing language disorders in two groups of students with severe behavioral disorders. *Behavioral Disorders, 22,* 160–166.

Gulliksen. H. (1950). *Theory of mental test.* New York: Wiley.

Jöreskog, K. G. (1971). Simultaneous factor analysis in several populations. *Psychometrika, 36,* 409–426.

Jorgensen, M., & Hunt, L. (1996). Mixture model clustering of data sets with categorical and continuous variables. *Proceedings of the Conference ISIS '96, Australia,* 375–384.

Kansas State Board of Education. (1999). *Kansas curricular standards for mathematics.* Topeka, KS: Author. Available at www.ksde.org

Kansas State Department of Education. (2001). *Kansas State regulations for special education.* Topeka, KS: Author. Available at http://www.kansped.org/ksde/processchg/spfc/apE.pdf

Kaufman, L., & Rousseeuw, P. J. (1990). *Finding groups in data: An introduction to cluster analysis.* New York: Wiley.

Lazarsfeld, P. F., & Henry, N. W. (1968). *Latent structure analysis.* Boston: Houghton Mifflin.

Li, W., & Nyholt, D. R. (2001). Marker selection by Akaike information criterion and Bayesian information criterion. *Genetic Epidemiology, 21*(Suppl. 1), 272–277.

Lin, T. H., & Dayton, C. M. (1997). Model-selection information criteria for non-nested latent class models. *Journal of Educational and Behavioral Statistics, 22,* 249–264.

Lloyd, J. W. (1984). How shall we individualize instruction—or should we? *Remedial and Special Education, 5,* 7–15.

Loehlin, J. C. (1998). *Latent variable models: An introduction to factor, path and structural analysis* (3rd ed.) New Jersey: Erlbaum.

Luke, G., & Muthén, B. (in press). Investigating population heterogeneity with factor mixture models. *Psychological Methods.*

Lyon, G. R. (1989). IQ is irrelevant to the definition of learning disabilities: A position in search of logic and data. *Journal of Learning Disabilities, 22,* 504–512.

MacMillan, D. L., Gresham, F. M., & Bocian, K. M. (1998). Discrepancy between definitions of learning disabilities and school practices: An empirical investigation. *Journal of Learning Disabilities, 31,* 314–326.

McArthur, G. M., Hogben, J. H., Edwards, V. T., Heath, S. M., & Mengler, E. D. (2000). On the "specifics" of specific reading disability and specific language impairment. *Journal of Child Psychology and Psychiatry, 41,* 869–874.

McCutcheon, A.L. (1987). Latent class analysis. *Sage university papers series on quantitative applications in the social sciences.* Thousand Oaks, CA: Sage.

McLachlan, G. J., & Basford, K. E. (1988). *Mixture models: Inference and application to clustering.* New York: Marcel Dekker.

McLachlan, G. J., & Peel, D. (2000). *Finite mixture models.* New York: Wiley.

McLachlan, G. J., Peel, D., Basford, K. E., & Adams, P. (1999). The EMMIX software for the fitting of mixtures of normal and t-components. *Journal of Statistical Software, 4*(2).

Read, T. R. C., & Cressie, N. A. C. (1988). *Goodness-of-fit statistics for discrete multivariate data.* New York: Springer Verlag.

Reschly, D. J., & Ysseldyke, J. E. (2002). Paradigm shift: The past is not the future. In A. Thomas & J. Grimes (Eds.), *Best practices in school psychology IV,* (pp. 3–20). Bethesda, MD: National Association of School Psychologists.

Shaywitz, B. A., Fletcher, J. M., Holahan, J. M., & Shaywitz, S. E. (1992). Discrepancy compared to low achievement definitions of reading disability: Results from the Connecticut Longitudinal Study. *Journal of Learning Disabilities, 25,* 639–648.

Siegel, L. S. (1989). IQ is irrelevant to the definition of learning disabilities. *Journal of Learning Disabilities, 22,* 469–478.

Sheehan, J. K., & Han, T. (1996). Hierarchical modeling techniques to analyze contextual effects: What happened to the aptitude by treatment design? *Mid-Western Educational Researcher, 9*(4), 4–7.

Simpson, E. H. (1951). The interpretation of interactions in contingency tables. *Journal of the Royal Statistical Society, B* (13), 238–241.

Slavin, R. E., Madden, N. A., & Leavey, M. (1984). Effects of team assisted individualization on the mathematics achievement of academically handicapped and nonhandicapped students. *Journal of Educational Psychology, 76,* 813–819.

Speece, D. L. (1990). Aptitude-treatment interactions: Bad rap or bad idea? *The Journal of Special Education, 24,* 139–149.

Stage, S. A., Abbott, R. D., Jenkins, J. R., & Berninger, V. W. (2003). Predicting response to early reading intervention from verbal IQ, reading-related language abilities, attention ratings, and verbal IQ-word reading discrepancy: Failure to validate discrepancy method. *Journal of Learning Disabilities, 36,* 24–33.

Stevens, J. (1992). *Applied multivariate statistics for the social sciences.* Mahwah, New Jersey: Erlbaum.

Stuebing, K. K., Fletcher, J. M., LeDoux, J. M., Lyon, G. R., Shaywitz, S. E., & Shaywitz, B. A. (2002). Validity of IQ-discrepancy classifications of reading disabilities: A meta-analysis. *American Educational Research Journal, 39,* 469–518.

Tomblin, J. B., Zhang, X., & Buckwalter, P. (2000). The association of reading disability, behavioral disorders, and language impairment among second-grade children. *Journal of Child Psychology and Psychiatry, 41,* 473–482.

Toppelberg, C. O., & Shapiro, T. (2000). Language disorders: A 10-year research update review. *Journal of the American Academy of Child and Adolescent Psychiatry, 39,* 143–152.

Vaughn, S., & Linan-Thompson, S. (2003). What is special about special education for students with learning disabilities? *The Journal of Special Education, 37,* 140–147.

Vermunt, J. K., & Magidson, J. (2000). *Latent GOLD's user's guide.* Boston: Statistical Innovations, Inc.

Vermunt, J. K., & Magidson, J. (2002a). Latent class cluster analysis. In J. A. Hagenaars & A. L. McCutcheon (Eds.), *Applied latent class analysis* (pp. 89–106). Cambridge, UK: Cambridge University Press.

Vermunt, J. K., & Magidson, J. (2002b). Latent class models for clustering: A comparison with K-means. *Canadian Journal of Marketing Research, 20,* 37–44.

Vermunt, J. K., & Magidson, J. (in press). Latent class analysis. In M. Lewis-Beck, A. Bryman, & T. F. Liao (Eds.), *Encyclopedia of research methods for the social sciences.* Newbury Park: Sage.

Von Davier, M. (2001). *WINMIRA2001* [Computer software]. St. Paul, MN: Assessment Systems Corporation.

Zimowski, M., Muraki, E., Mislevy, R. J., & Bock, R. D. (2002). *BILOG-MG 3: Item analysis and test scoring with binary logistic models for multiple groups* [Computer software]. Mooresville, IN: Scientific Software.