Initial Validity Evidence for the *HOPE Scale*: New Instrumentation to Identify Low-Income Elementary Students for Gifted Programs

Scott J. Peters
Marcia Gentry
Jillian Gates
Jean Peterson
Rebecca Mann

Purdue University

Word Count (not abstract): 6551 including tables

ABSTRACT (144 Words)

Students with exceptional academic potential who come from low-income families are frequently not identified for and consequently are underrepresented in gifted and talented programs. Because of this, new means of identifying such children must be developed. This paper presents the findings of exploratory and confirmatory factor analyses conducted on the *HOPE Scale*, a 13-item teacher-rating instrument designed to identify academic and social components of giftedness in elementary-aged students. Participants included 349 teachers who completed *HOPE Scales* on 5995 ethnically and economically diverse students from three rural and two metropolitan school districts in the Midwest. MCFA was also used to evaluate measurement invariance between income groups. Findings suggest a two-factor model represents good fit for the data while remaining loyal to the latent constructs of academic and social giftedness. Invariance test results suggested equivalence of model form, factor loading, and factor variances across income groups.

Purpose

In this study we sought to develop a new instrument, the *HOPE Scale*, designed to help teachers more equitably identify K-5 children from low-income families for gifted and talented programs. This project had as its primary goal developing an instrument that could be used with confidence to help teachers identify potential among students from low-income families. Including sufficient numbers of children from low-income and culturally-diverse families in the sample facilitated the investigation of factor structure similarities across income groups and the creation of an instrument less affected by income background than many existing standardized measures. The first step in creating this instrument was to examine the initial factor structure using a sample of teachers and children from schools with a third or more of their children coming from low-income families.

The *HOPE Scale* was not designed to be a stand-alone instrument, but rather to provide additional information beyond that generally provided by standardized achievement or aptitude tests. By combining *HOPE Scale* information with other measures of student achievement and potential, educators can develop a more comprehensive picture of a child's potential as rated by his/her classroom teacher. The *HOPE Scale* uses directions developed by considering the federal definition of gifted and talented students, and asks teachers to rate their students "as compared with others similar in age, experience, or environment" (USDOE, 1993, p. 3). Considering environment and experience are important as students from low-income families may appear less academically advanced if compared with age-peers from non low-income families. For example, children from low-income families may have less access to resources and enrichment experiences, and therefore have less background knowledge, resulting in differences in their test and school performance.

Theoretical Rationale

Income Group Representation

Despite advances in psychological assessment, family income remains one of the highest correlates with academic achievement (Rogers, 1996; Valencia & Suzuki, 2001). Even though factors other than income are involved in this association (e.g., better access to high quality schools), coming from a low-income family remains a disadvantage with regard to school success (Valencia & Suzuki, 2001; Wyner et al., 2007). Low-income students also tend to be underrepresented in programs for the gifted and talented (Stambaugh, 2007; Swanson, 2006). In the 2003 – 2004 school year, more than 40% of all students in American schools were eligible for the federal free and reduced lunch program (NCES, n.d.). This program has consistently been used as a gauge for economic standing and has even been criticized as being too exclusive thereby leaving a number of low-income students without assistance even though they are affected by many of the same problems as students who qualify (Viadero, 2006). Despite this percentage, only 28% of students achieving in the top quartile in first grade were from low-income families (Wyner, Bridgeland, & Diiulio, 2007). Wyner et al. also noted that even more problematic was that of those low-income students in the top quartile in first grade, only 56% maintained this high performance by fifth grade.

The *Achievement Trap* (Wyner et al., 2007) report outlined many problems related to educating students from low-income families. However, it also included several suggestions for how to address their underachievement and underperformance. One of the most important suggestions dealt with finding or identifying such students: “We must adopt a broader vision that recognizes the immense potential of many lower-income students to perform at the highest levels of achievement and consider how to educate them in ways that close the existing high-achievement gap" (p. 29). Because students from low-income families are likely to underperform, they are also less likely to be noticed or nominated for gifted and talented programs (Stambaugh, 2007; Swanson, 2006). This creates a cyclical effect, as high potential students from low-income families remain unnoticed when they may indeed benefit from services in gifted programs.

Stambaugh (2007), summarizing findings from a National Leadership Conference on Low-income, Promising Learners, outlined several practices that could aid in identifying students from low-income families for gifted and talented programs. These practices included beginning identification as early as kindergarten and continuing with ongoing identification to locate low-income students who may not demonstrate gifted and talented behaviors until later grades in school. She also suggested using teacher behavior checklists that have been shown to yield reliable and valid data on giftedness and talent specifically for students from low-income families. Stambaugh emphasized the importance of using more-specific normative groups in assessment than have traditionally been used. Teacher rating scales, as with any other measure, should be used in conjunction with multiple assessments in order to provide a comprehensive view of a student. Finally, participants in the conference identified professional development and teacher training as important to ensure that educators know what behaviors to look for in low-income students who might benefit from gifted and talented programs. One of the clearest conclusions from both the National Leadership Conference (Stambaugh) and the *Achievement Trap* (Wyner et al., 2007) report was that solving the problem of underachievement and under recognition of high potential students from low-income families will require a consorted effort at the local and national levels to better recognize these students in the early elementary years.

Controversy exists in the research literature about the covariance of race / ethnicity and income and their respective effects on school success (Valencia & Suzuki, 2001). In reviewing this issue, Valencia and Suzuki argued that too often the variables of race / ethnicity and socio-economic status (SES) are not examined separately; and because of this, effects cannot be attributed to one particular variable. Accordingly, they recommend that any researchers interested in SES and race / ethnicity as variables should measure each separately in order to analyze their respective influences. Similarly, Callahan (2007) called for any future study of instrument validation to include SES as a variable in addition to race, ethnicity, and gender. If researchers do not to measure SES and race / ethnicity separately in future studies, then the effects of these variables cannot be clearly interpreted.

Racial Representation

Despite the growing numbers of students within virtually every minority group in the U.S. K-12 population, many ethnic and racial minorities continue to be underrepresented in programs for the gifted and talented (Yoon & Gentry, in press). Researchers from the National Research Council reviewed the history of minority representation in special and gifted education (Donovan & Cross, 2002). This report noted that since 1976, the number of minority children identified as gifted and talented students has been steadily increasing. By computing a risk index (RI), the authors compared the different identification proportions for several ethnic / racial groups. For example, from 1976 to 1998 American Indians / Alaskan Natives increased from an RI of .42 to 4.86. This means that in 1976 only .42% of this group's members were identified as gifted and talented; whereas in 1998 4.86% of the members of this group were identified.

However, because of large starting differences in rates of identification, some groups remain underrepresented. Despite the fact that Asian / Pacific Islander representation has become more proportional due to a decrease in the number of students identified, that of black and Hispanic students has seen no sustained improvement. This problem of underrepresentation has been noted elsewhere for African Americans (Ford, 1998) based on the same data, and Hispanics (Plata, Masten, & Trusty, 1999).

Yoon and Gentry (in press) analyzed data from three different sources to determine the extent of under- and over-representation of different racial and ethnic groups: the National Education Longitudinal Study of 1988 (NELS:88), the School and Staff Survey (SASS), and the Office for Civil Rights (OCR) data collection. What made this study unique is that the authors included the most-recent national data available and also disaggregated representation trends by state. Although African American, American Indian, and Hispanic students continue to be underrepresented on the national level and Caucasians and Asians continue to be overrepresented, considerable variation exists among states. For example, American Indian / Alaskan Native representation increased in Arkansas, Georgia, Maryland, Minnesota, New Mexico, Nevada, and Utah. Over the same period of time this group’s representation decreased in Delaware, Iowa, and Pennsylvania. Similar analyses were presented for every racial or ethnic group. Across all of the ethnic and racial groupings, only one or two states per group had close to proportional representation among ethnic / racial groups. Because of the wide range of individuals who might identify themselves under the same ethnic / racial category for the purpose of data collection, the authors argued that national trends are not the ideal means to measure representation. Instead, disaggregation by state and by racial and ethnic sub-group may provide better indicators. Just as with identification procedures, local context and specific population considerations are important.

Teacher Nomination and Rating Scales

According to the 2006-2007 State of the States Report (NAGC, 2007) the most common types of identification procedures for gifted and talented programs include multiple criteria, achievement tests, IQ scores, and nominations, in order of prevalence. Despite multiple criteria being the most reported means of identification, the report also found that the most common time for implementing an identification procedure was after a teacher or parent referral (reported by 30 of 43 responding states). This means that despite any advances in standardized assessment, the initial identification catalyst remains an adult's nomination. Thus, a single teacher can be the gatekeeper to the gifted and talented program. To better focus nominations, an entire genre of teacher rating and nomination forms as well as checklists has emerged over the last 40 years.

Teacher ratings, referrals, and nominations have often been criticized for their lack of validity (e.g., Pegnato & Birch, 1959; Peterson & Margolin, 1997). Pegnato and Birch's (1959) article is one of the most widely cited as having an empirical basis for this conclusion. In this study the authors used teacher nominations for giftedness in several content and non-academic areas (e.g., art, music, social / political). The 154 students who were nominated as "mentally" gifted were then assessed using the Stanford-Binet Intelligence Test. Only 91 of the 154 students who were nominated obtained an IQ score of 136 or higher. This cut-point was used because it yielded the top one-percent of the population. Based on this finding, the authors proposed that teacher nominations were neither efficacious nor efficient. However, when Gagné (1994) reanalyzed the same data, he found that with regard to student identification, "teachers do not come out worse than most other sources of information," (p. 126) including mental ability tests, school grades, and achievement test scores. Gagné argued that Pagnato and Birch (1959) used invalid methods for evaluating teacher rating instruments. He explained that an identification procedure's efficacy and efficiency are not independent of each other and are therefore not appropriate for comparisons of different measures. Effectiveness, or the percentage of students nominated, is actually negatively correlated with efficiency, or the percentage of students later identified as gifted. Instead, Gagné used a 2x2 correlation between the predictor (nomination) and the criterion of interest (IQ score) to make his determination that teacher nominations had phi coefficient of .29, comparable to the other measures used by Pagnato & Birch (1959) in their study. Gagné also noted that using the top one-percent of a full-scale intelligence test as the criteria for "gifted" was unrealistic and would omit many high-ability students.

The call for teachers to be included in the identification process (Gagné, 1994; High & Udall, 1983; Hunsaker, Finley, & Frank, 1997) has led to the creation of a variety of teacher rating forms and behavior checklists. The majority of these instruments can be grouped into two classes. The first group involves instruments that have been subjected to little or no empirical research and were developed using little or inadequate statistical techniques. Examples of such instruments include the *Kingore Observation Inventory* (*KOI*: Kingore, 2001), the *Traits, Attributes, and Behaviors Scale* (*TABS*: Frasier & Passow, 1994; Frasier et al., 1995), and the *Kranz Talent Identification Instrument* (*KTII*: Kranz, 1981). A search of the ERIC and PsychInfo databases revealed no empirical studies supporting any of these three instruments. The *Purdue Academic and Vocational Rating Scales* (Feldhusen, Hoover, & Sayler, 1989) are another set of teacher-rating scales that separate behaviors by academic (English, science, social studies, math, and foreign language) or vocational area. However, these scales were designed only for middle and high school students. Although these scales can be used to help identify students and to make instructional decisions, they have not been subjected to the types of rigorous evaluation necessary in order to yield valid data for diverse populations.

A second group of teacher-as-rater instruments is more promising. Instruments such as the *Gifted Rating Scales* (*GRS*: Pfeiffer & Jarosewich, 2003) were developed using rigorous statistical techniques. Similarly, the *Scales for Rating the Behavioral Characteristics of Superior Students* (*SRBCSS*: Renzulli, Smith, White, Callahan, Hartman, & Westberg, 2002), the *Gifted Evaluation Scales, Second Edition* (*GES-2*: McCarney & Anderson, 1989) and the *Gifted and Talented Evaluation Scales* (*GATES*: Gilliam, Carpenter, & Christensen, 1996) offer psychometric development information in their respective test manuals and/or have been used in empirical research. However, several problems exist. Some of the instrument developers in this group used exploratory factor analysis (EFA) or principal components analysis (PCA) without following up these procedures with a more rigorous confirmatory factor analysis (CFA). There exists a large body of research literature on the problems associated with PCA or with attempting to create an instrument using only exploratory methods (e.g., Thompson, 2004, Widaman, 1993). In addition, some of the instruments have relatively dated or non-representative standardization samples. In what appears to be the most rigorously developed instrument in the class of teacher nomination scales, the *GRS* (Pfeiffer & Jarosewich, 2003) has not been subjected to measurement invariance testing, making its validity for use with underrepresented groups unclear.

Progress toward quality teacher-rating scales has been made. However, at this point, none of the available instruments described above provide all of the information recommended in the Joint Committee on Testing Practices' *Code of Fair Testing Practices in Education* (2005), which suggests authors:

> Obtain and provide evidence on the performance of test takers of diverse subgroups, making sufficient efforts to obtain sample sizes that are adequate for subgroup analyses. Evaluate the evidence to ensure that differences in performance are related to skills being assessed (p. 4).

Interestingly, this is similar to the call made by Callahan (2007) regarding the need for research into the validity of assessment tools used for gifted and talented identification. Specifically, she noted the need for the separate evaluation of race, ethnicity, and income factors when examining outcomes. Although this information may exist for these teacher rating forms, it was not presented in the respective test manuals or scholarly articles. The *Code* was designed to guide the development of instruments and to ensure that results from their development are readily available to consumers. Thus, researchers developing teacher rating instruments or scales should closely adhere to these guidelines.

It is important to note that providing descriptive statistics such as mean scores and standard deviations, although important, does not fully address the *Code* requirements described above. Descriptive information should be followed by multi-group analyses; specifically multi-group confirmatory factor analysis (MCFA) can be used to evaluate performance of subgroups including income groups. This necessary step goes beyond simply describing what scores were received to detail exactly what factors (i.e., income status) may have confounded the actual construct being measured.

Several authors have called for more research into teacher-rating scales in order to make them a more scientific component of a larger identification system (Gagné, 1994; Hodge & Cudmore, 1986; Jarosewich, Pfeiffer & Morris, 2002; McBee, 2006). VanTassel-Baska (2008) argued that teacher rating scales should be considered in the initial screening process to help locate all potential students for further evaluation. With regard to minority students, Plata and Masten (1998) emphasized that teachers need to possess an understanding of their students' cultural backgrounds if nominations of these students are to be successful. In addition, to ensure successful nominations, several authors have emphasized the importance of teacher training (e.g., Siegle & Powell, 2004) and including clearly defined behaviors and characteristics on teacher rating scales (e.g., Hodge & Cudmore, 1986; Jarosewich et al., 2002).

Methods and Data Analysis

Participants

349 teachers from 5 school districts (3 rural, 2 metropolitan) in one Midwestern state completed the *HOPE Scale* on students in their classrooms. Of the 5995 students rated by their teachers, 59% were eligible for the free or reduced lunch program. The 5995 students on whom *HOPE Scales* were completed represent 86% of the total number of students in the five school corporations. Of the five districts, teachers from Anderson, Dennis, and Franklin schools rated nearly 100% of their students; whereas, teachers from Benjamin and Lincoln rated 66% and 82% respectively. Table 1 includes the demographic characteristics of the sample. All corporations are listed using pseudonyms in order to protect corporation privacy. Different sub-samples of this larger sample were used to address different research questions and are described individually in the following sections.

Table 1. Sample Demographic Characteristics by School Corporation

	Anderson	Benjamin	Dennis	Franklin	Lincoln
Designation	Rural	Rural	Rural	Metro	Metro
K-5 population	410	840	705	1561	3425
HOPE Scales Returned	405	557	692	1528	2813
Free/Reduced Lunch Students	36%	38%	34%	62%	58%
Caucasian	96%	90%	91%	59%	60%
African American	0%	<1%	<1%	<1%	10%
Hispanic	2%	5%	8%	37%	21%
Asian	<1%	<1%	0%	<1%	<1%
Multi-racial	<1%	4%	1%	3%	8%
Native American	0%	<1%	0%	<1%	<1%

Instrument development

After reviewing the literature on gifted and talented student behaviors and after reviewing existing instruments, a team of researchers wrote items to define two broad areas, Academic and Social, for the *HOPE Scale*. A sample item from each area follows: “Has desire to work with advanced concepts and materials” (academic); “Shows compassion for others” (social). After multiple revisions for wording, clarity, and content coverage and judgment by content experts, 13 items were retained for data collection from the above-described sample. A six-point rating scale was used based on Comrey’s recommendation (1988) that scales have at least four points and on recommendations made by Brown (2006) that scales with more rating points more closely approximate the normality and continuous data that are necessary for certain statistical techniques. This initial 13-item version of the *HOPE Scale* is included in the Appendix.

Data Collection and Analysis

Data were collected in the fall of 2007 during a six-week time-period using a one-time administration of the *HOPE Scale* in each corporation.

EFA. From the sample of 5995 students, 500 were randomly selected for Exploratory Factor Analysis (EFA). Although sample size recommendations for EFA procedures vary, conservative recommendations place ideal sample sizes at between 400 (Comrey, 1988) and as much as 800 (Fabrigar, Wegener, MacCallum, & Strahan, 1999) when dealing with extremely poor communality estimates and under-defined factors. Due to the correlation among items on the *HOPE Scale,* a Promax oblique rotation was used. Individual items were retained only if they loaded on a single factor at .4 or greater. Scree plots were used as was a parallel analysis (Montanelli & Humphreys, 1976; Thompson, 1996) in order to determine the number of factors to retain. Parallel analysis is the most accurate method of determining the number of factors when the sample size is between 500 and 1000 (Zwick & Velicer, 1986).

CFA. Confirmatory Factor Analyses followed the EFA, thus continuing the investigation of the construct validity of the *HOPE Scale.* Because having an adequately sized sample was not an issue with the current study, and based on the EFA results, 1500 additional students were randomly selected from the remaining students not used in the EFA sample. As with EFA, in CFA there is no fixed formula for the sample size requirements. However, Muthén & Muthén (2002) argued that even for non-normal data in which some missing responses exist, 315 is a sufficient sample-size to detect factor correlation. Additionally, Kieffer (1999) proposed 500 – 1000 participants as an ideal number in order to achieve stability. The CFA model was specified using the model extracted from the EFA. Once the model was fit, three different types of indicators were used for evaluation: chi-square indicators, fit statistics, and standardized residuals (Crowley & Fan, 1997). Because a large sample size almost always guarantees a significant chi-square result (Kline, 2005), several alternative fit indices were also considered. Modification indices were considered if they aligned with gifted education and intelligence theory with respect to the latent factors. Post-modification models were compared with the initial model based on the three above-described criteria.

Once the final model was established from the general CFA, measurement invariance testing was conducted to evaluate the equivalence of different parameters for students from low-income families and those who were not from low-income families. Because the *HOPE Scale* was originally designed to better identify students from low-income families for gifted and talented programs, a multiple-groups CFA (MCFA) was conducted to evaluate model invariance or bias (for or against) when used with students from low-income families. This process is little more than fitting the general model for each group separately followed by tests of increasingly restrictive models in order to establish the degree of between group equivalence (Brown, 2006). Based on this process, an MCFA comparing students who qualify for the free and reduced lunch program with those students who do not would include the following tests: general model for paid students, general model for free or reduced lunch students, equal form, equal factor loadings, equal indicator intercepts, equal error variances, equal factor variances, and equal latent means.

Results

Exploratory Factor Analysis

The scree plot from the EFA indicated a strong elbow after two factors (see Table 2 for eigenvalues). While the Kaiser greater-than-one rule has been popular for determining the number of factors to retain, recent research indicates this rule can overestimate or underestimate the correct number of factors to retain (Kieffer, 1999; Thompson, 1996). A parallel analysis was conducted to determine if the EFA eigenvalues of the first two factors were larger than would be expected if found at random (Table 3). Although these results indicate a single factor model, Fabrigar et al. (1999) indicated it is better to err on the side of extracting too many factors than too few. Based on this recommendation and because factor two (Table 2) is so close to the Parallel Analysis second factor (Table 3), two factors were extracted for further analysis. The second factor contributes an additional 11% of the total variation to the model. The final two-factor model accounts for 99% of the total variation in the data.

Table 2. Eigenvalues

	Eigenvalue	Difference	Proportion	Cumulative
1	8.77528697	7.71558018	0.8845	0.8845
2	1.05970679	0.83116601	0.1068	0.9913
3	0.22854078	0.06715332	0.0230	1.0143
4	0.16138747	0.09504275	0.0163	1.0306

Table 3. Results of Parallel Analysis

Eigenvalue	Random Eigenvalue	Standard Dev
1	1.2723	.0376
2	1.2070	.0027
3	1.1550	.0234
4	1.1104	.0214

The rotated factor pattern coefficients (Table 4) indicated loadings on the first factor for Items 1, 2, 5, 7, 9, 10, 11, and 12. The remaining three items (3, 4, 8) loaded on the second factor. Item 6 was split between the two factors. In addition, Item 13 was removed after further review by the researchers because this question did not directly relate to one factor or the other, but was meant to apply to specific content areas of talent. The Varimax rotated solution yielded an identical factor structure to the Promax solution with a .4 loading criterion, indicating a similar structure regardless of rotation method.

Table 4. Factor Structure and Pattern Coefficients After Promax Rotation

Item	Item Stem	Structure Coefficients		Rotated Factor Pattern	
		Academic	Social	Academic	Social
1	Performs or shows potential for performing at remarkably high levels	0.89673	0.57461	0.87604	0.03349
2	Is curious, questioning	0.82057	0.58464	0.74288	0.12577
3	Is empathetic	0.60347	0.92783	0.04909	0.89751
4	Shows compassion for others	0.55328	0.94072	-0.04495	0.96849
5	Has desire to work with advanced concepts and materials	0.91679	0.61542	0.86771	0.07944
6	Questions authority	0.24637	-0.09899	0.49723	-0.40612
7	Is eager to explore new concepts	0.87949	0.65541	0.76747	0.18135
8	Exhibits a strong sense of social justice and fairness	0.70278	0.74522	0.39205	0.50305
9	Uses alternative processes	0.90985	0.60381	0.86810	0.06759
10	Is insightful and intuitive	0.93353	0.64757	0.86269	0.11470
11	Thinks “outside the box”	0.93595	0.59956	0.91454	0.03465
12	Has intense interests	0.87464	0.59406	0.82091	0.08699
13	Shows outstanding talent in specific content area(s)	0.88905	0.55363	0.88459	0.00723

The model was also run using Maximum Likelihood (ML) estimation methods. Fabrigar et al. (1999) suggested that ML techniques allow for a greater range of fit indices and only have drawbacks if the data do not meet multivariate criteria. However, the resulting factor structure was the same for both methods. In addition, because the *HOPE Scale* responses are scored on a six-point rating scale, both Spearman and Pearson correlations were computed in case of non-normality. However, the results were nearly identical and the subsequent factor structure was the same for either procedure. Therefore, this final two-factor model was established for further testing using CFA.

Confirmatory Factor Analysis

CFA was used to investigate the *HOPE Scale* after items were deleted as informed by the EFA inquiry described above. Thus, this model retained eight items on factor one (1, 2, 5, 7, 9, 10, 11, 12) and three items on factor two (3, 4, 8). Table 5 includes the covariance matrix for the 11 items used in the CFA.

Table 5. Item Covariance Matrix

	1	2	3	4	5	7	8	9	10	11	12
1	1.937										
2	1.446	1.771									
3	0.980	0.995	1.554								
4	0.881	0.903	1.377	1.511							
5	1.633	1.463	1.044	0.956	1.902						
7	1.453	1.398	0.986	0.911	1.584	1.743					
8	1.104	1.122	1.142	1.136	1.198	1.174	1.773				
9	1.372	1.237	0.846	0.764	1.403	1.318	1.086	1.531			
10	1.500	1.372	0.990	0.907	1.528	1.426	1.195	1.427	1.710		
11	1.435	1.313	0.862	0.784	1.437	1.332	1.079	1.416	1.505	1.616	
12	1.285	1.266	0.847	0.784	1.345	1.296	1.079	1.269	1.325	1.329	1.622

The model was analyzed using the new 1500 student sample described earlier. Table 6 presents the standardized parameter estimates for the two factors as well as the inter-factor correlation.

Table 6. Standardized Parameter Estimates – Base Model

	Estimate	S.E.	Est./S.E.	P-Value
Academic Factor				
1	0.878	0.006	137.877	0.000
2	0.847	0.008	109.961	0.000
5	0.900	0.005	167.017	0.000
7	0.877	0.006	136.582	0.000
9	0.928	0.004	229.750	0.000
10	0.947	0.003	299.909	0.000
11	0.941	0.003	271.578	0.000
12	0.858	0.007	118.909	0.000
Social Factor				
3	0.953	0.005	208.508	0.000
4	0.938	0.005	192.596	0.000
8	0.744	0.012	59.909	0.000
Inter-Factor Correlation				
F1 F2	0.664	0.016	42.426	0.000

Of note in Table 6 is the inter-factor correlation of .664. Although this is a moderately strong correlation, a second-order factor would not be appropriate since there would only be two first-order factors. A second-order factor is often useful when two first-order factors are highly correlated or are hypothesized to be related in some fashion to an additional latent construct

(Brown, 2006; Thompson, 2004). However, because the addition of a second-order factor requires additional degrees of freedom in order to estimate, such a second-order factor can only better describe the data if there are four or more first-order factors. In the case of the *HOPE Scale*, a second-order factor would not better explain the data because there are only two first-order factors. Cases of high factor inter-correlation may also indicate the presence of only a single first-order factor. Because of this a single factor model was tested, but fit the data worse as measured by every fit statistic and had a significantly higher chi-square value. Based on these results, the current two single-order factor best fits the data.

Table 7 includes the chi-square values and fit indices for the CFA model as specified by the EFA results. The chi-square value was significant, traditionally indicating a lack of model fit. However, a large sample usually yields significant chi-square values (Brown, 2006; Kline, 2005); therefore other measures should be considered. The Comparative Fit Index (CFI) and Tucker-Lewis Index (TLI) values were .949 and .934 respectively. Values of .95 or greater are recommended as values indicating good fit. Standardized Root-Mean Squared Residual (SRMR) and Root-Mean Square Error of Approximation (RMSEA) values of .07 and .129 are also greater than the .05 recommended values. Thus, the current model has moderate, but not good fit.

Table 7. Indices of Model Fit – Base Model

Index	Value	
Chi-square	1082.369	p-value : <.001
RMSEA	.129	90% CI: .122-.135
CFI	.949	
TLI	.934	
SRMR	.07	

In order to improve model fit, modification indices were considered. Item 8 (Has desire to work with advanced concepts and materials) had the weakest pattern coefficient loading from the EFA and also had the highest residual value in the CFA. This item was also part of the largest modification index suggesting an improvement (decrease) of over 300 in the chi-square value if this item was allowed to cross-load on both factors. Since cross-loading items are undesirable, this item was removed. A second modification index suggested items 5 (Has desire to work with advanced concepts and materials) and 7 (Is eager to explore new concepts) have their errors (theta-deltas) constrained. Not only did this improve model fit, but it also made sense as the two items were similar in wording and content. The modifications were made and the resulting model fit indices are presented in Table 8. This revised model yielded a chi-square value 40% smaller than the original. Although the model chi-square is still significant, the decrease of over 400 represents a statistically significant decrease and improved overall model fit. CFI and TLI indices of .967 and .955 both exceed the recommended minimum of .95 and are improved from the original model. The SRMR value of .025 was also well below the .05 standard. However, the RMSEA value of .113 remained high indicating some model misfit.

Table 8. Indices of Model Fit – Revised Model

Index	Revised Model	
	Value	
Chi-square	664.418	p-value : <.01
RMSEA	.113	90% CI: .106-.121
CFI	.967	
TLI	.955	
SRMR	.025	

Table 9 presents the descriptive statistics and alpha reliability estimates for the two factors. Both scales' reliability estimates are high indicating strong internal consistency. In addition, all of the items have similar means and standard deviations. However, items on the Social factor were generally rated higher than those on the Academic factor.

Table 9. *HOPE Scale* Descriptive Statistics

		Response Percentage										
Factor	Item	1	2	3	4	5	6	Mean	SD	*r* with total[a]	Alpha if removed[b]	Alpha
Academic	1	19	27	24	13	11	6	2.88	1.46	.90	.97	.97
	2	10	23	33	18	10	6	3.11	1.31	.85	.97	
	5	17	30	23	15	11	4	2.87	1.40	.91	.97	
	7	11	25	31	17	11	5	3.05	1.33	.86	.97	
	9	18	36	25	12	5	4	2.62	1.27	.88	.97	
	10	18	31	24	15	7	5	2.77	1.36	.93	.97	
	11	18	38	21	12	7	4	2.61	1.31	.93	.97	
	12	17	36	26	9	9	3	2.65	1.28	.85	.97	
Social	3	7	19	32	23	13	6	3.32	1.28	.90	.90	.95
	4	5	15	33	26	15	6	3.49	1.23	.90	.90	

Note. [a]standardized correlations. [b]standardized coefficients

Table 10 also includes the measures of normal distribution: skewness and kurtosis. Skewness is a measure of asymmetry of a data distribution (Kleinbaum, Kupper, Muller, & Nizam, 1998). Skewness values for the *HOPE Scale* items ranged from .111 to .837 indicating mild departure from normality. Kurtosis values indicate the heaviness of the tails of a distribution with a value of 0 indicating a normal distribution (Kleinbaum et al.). In this case, kurtosis values ranged from -.667 to .293. These values indicate the *HOPE Scale* items have slightly heavier tails, more often heavier in the lower categories, than does a normal distribution.

Table 10. Item Skewness and Kurtosis

Factor	Item	Skewness	Kurtosis
Academic	1	.515	-.667
	2	.396	-.365
	5	.496	-.625
	7	.412	-.453
	9	.825	.293
	10	.624	-.297
	11	.837	.099
	12	.777	.035
Social	3	.187	-.481
	4	.111	-.446

Income group differences

The general CFA was followed by an evaluation of measurement invariance as described above. Although the correlation of item errors is allowed in CFA (Thompson, 2004), the constraint of items 5 and 7 was not included in the invariance testing model, but instead was allowed to vary freely. Items 5 and 7 were retained for the invariance testing, creating a slightly worse fitting model. This was done to facilitate continued development of the *HOPE Scale*. Table 11 presents chi-square values, chi-square difference tests, and fit statistics for the eight tests.

Table 11. Measurement Invariance Tests for Paid Lunch vs. Free or Reduced Lunch Students

		$\chi 2$	df	$\chi 2$diff	Δdf	RMSEA 90% CI	SRMR	CFI	TLI
Single Group Solutions									
Paid	(n=537)	485.089*	43	-	-	.138 (..127 - .139)	.05	.94	.92
Free / Reduced	(n=685)	586.714*	43	-	-	.136 (.126 - .146)	.05	.94	.92
Measurement Invariance									
Equal Form		1071.803*	86	-	-	.137 (.130 - .144)	.05	.94	.92
Equal Factor Loading		1083.223*	95	11.42	9	.130 (.124 - .138)	.05	.94	.93
Equal Indicator Intercepts		1101.671*	104	18.448	9	.125 (.119 - .132)	.05	.94	.93
Equal Indicator Error Variances		1187.304*	114	85.633*	10	.124 (.118 - .131)	.05	.93	.93
Population Heterogeneity									
Equal Factor Variance		1195.936*	116	8.632	2	.123 (.117 - .130)	.08	.93	.94
Equal Latent Mean		1255.300*	118	59.364*	2	.126 (.119 - .132)	.112	.93	.93

Note. * significant at p<.001

Of the 1500 students who were randomly selected for the CFA procedures, free and reduced lunch information was available on 1222. Although this represents 19% missing free and reduced lunch data, the percentages are representative of the degree of missingness in the larger sample. The results presented in Table 11 lead to several important conclusions. Although different sample sizes for student income groups prevent direct chi-square comparison, the fit indices can be compared. SRMS, CFI, and TLI values were identical for both groups. In addition, RMSEA values differed by only .002 in favor on the paid lunch students. In general, this indicates the model fits both groups equally well. The chi-square values for both groups were significant, traditionally indicating poor model fit. However, with such a large sample size, chi-square values are almost always significant (Brown, 2006; Kline, 2005). In addition, although the values for both groups were nearly identical, the CFI and TLI values fall short of the .95 traditional cutoff criteria, as does the RMSEA which ideally should be less than .05. Still, when comparing groups, similarity between groups is of primary interest.

The test of equal factor form is a test similar to the single groups' evaluation that combines all students. This model is then used as the base model for the purposes of comparison. In this case, the test of equal factor loadings resulted in a non-significant increase in the chi-square value. This means that increased equality constraints (equal factor loadings across groups) placed on the data, did not result in a significant chi-square increase. Therefore, the assumption of equal factor loadings holds for the two groups. However, the fit statistics remain just short of traditional cutoff values. The follow-up test of equal indicator (item) intercepts was also non-significant, meaning that students from the two income groups had similar item intercepts. Both of these tests provided evidence that the *HOPE Scale* yielded equally valid scores for both groups of students in assessing Academic and Social components of giftedness. The test of equal indicator error variances is often not done because of its especially stringent nature (Brown, 2006). Assuming that the errors related to item scores are equal between two groups is unlikely to hold true. However, the test was conducted because Brown (2006) recommended it as a necessary step before evaluation of structural parameters of equal factor variances and means. The test of equal indicator error variances resulted in a significant chi-square increase indicating non-equivalence of indicator error variance across the two groups. However, no two groups are likely to have perfectly equal error variances (Brown, 2006; Byrne, 1998).

The final two invariance tests related to population structural parameters. The first step of equal factor variance determines if the amount of within-group variability on the specific construct differs significantly across the two groups. In this case, the test found that the variation between the two groups does not differ significantly. This means that the ranges of scores within both groups were similar, and that teachers used the same range of scores when rating students from either income group. The final test of equal latent means determines if the groups differ significantly on the underlying constructs (factors). The significant chi-square increase indicates that they do differ. An evaluation of the estimated parameters revealed that teachers rated students from low-income families .532 lower on Academic and .241 lower on Social than they rated students from non low-income families. These un-standardized loadings are interpreted within the original 1-6 metric of the items and are substantial from a practical perspective indicating average lower ratings by teachers of students from low-income families.

Discussion

The results from this study suggest that a two-factor model best describes the *HOPE Scale* data. This finding was further supported by a CFA conducted on an additional sample and allowed for refinement of the model. Although these results are encouraging, there remains room for improvement with regard to overall model fit statistics and RMSEA values which currently indicate good or adequate model fit. However, the results from the invariance test suggest that students from low-income families were rated similarly by their respective teachers as those students not from low-income families. Although there was a difference in overall latent mean score, the *HOPE Scale* did not yield differences on tests of equal structure, equal indicator intercepts and equal factor variances. Only the stringent test of equal latent means showed a statistically significant difference indicating lower average factor scores for students from low-income families.

Despite the positive results from the fit statistics of the revised model of the *HOPE Scale*, only two items loaded on the Social factor. Because Brown (2006) indicated that a latent factor with only two indicators will likely yield higher standard errors and biased parameter estimates, additional items will need to be added to the Social scale before it is used for student identification. In addition, as this instrument is revised, additional evaluation of group differences (e.g., race, ethnicity, income status) will be necessary. Thus, the next step in the instrument development process involves adding items for the Social factor and re-administering a *Revised HOPE Scale* to a new sample of students. This revised *HOPE Scale* would then need to be evaluated for bias and group characteristics as a necessary step in instrument development.

Importance of the Study

This study responds to past calls for instruments that are developed and normed using representative populations of low-income and diverse students (Borland, 2008; Ford, 1998; Worrell, 2007). This study also follows recommendations made in the *Code of Fair Testing Practices* that all instruments are evaluated for their usefulness in yielding valid results for multiple groups of test-takers (2005). Previous teacher nomination or rating scales have not been subjected to such analyses. Ideally, these analyses will become more commonplace with regard to instruments developed and used in gifted and talented education.

Worrell (2007) called for culturally sensitive identification methods and the application of invariance testing is one possible statistical step toward the establishment of such sensitivity. The *HOPE Scale* was developed using a sample of students comprised of 59% who are eligible for the federal free and reduced lunch program. We intend to develop norms for this instrument and to conduct comparative analyses of the factor structure for both students who do qualify for free or reduced lunch and those who do not. In summary, this work has important implications in helping educators recognize potential among underserved elementary students.

References

Borland, J. H. (2008). Identification. In J. A. Plucker & C. M. Callahan (Eds.), *Critical issues and practices in gifted education: What the research says* (pp. 261-280). Waco, TX: Prufrock Press.

Brown, T. A. (2006). *Confirmatory factor analysis for applied research*. New York: The Guilford Press.

Byrne, B. M. (1998). *Structural equation modeling with LISREL, PRELIS, and SIMPLIS: Basic concepts, applications, and programming*. Mahwah, NJ: Erlbaum.

Callahan, C. M. (2007). What can we learn from research about promising practices in developing the gifts and talents of low-income students. In. J. VanTassel-Baska & T. Stambaugh (Eds.), *Overlooked gems: A national perspective on low-income promising learners* (pp. 53-56). Washington, DC: National Association for Gifted Children.

Comrey, A. L. (1988). Factor-analytic methods for scale development in personality and clinical psychology. *Journal of Counseling and Clinical Psychology*, 56, 754-761.

Crowley, S. L. & Fan, X. (1997). Structural equation modeling: Basic concepts and applications in personality assessment research. *Journal of Personality Assessment*, 68, 508-531.

Donovan, M. S., & Cross, C. T. (2002). *Minority students in special and gifted education*. Washington, DC: National Academy Press.

Fabrigar, L. R., Wegener, D. T., MacCallum, R. C., & Strahan, E. J. (1999). Evaluating the use of exploratory factor analysis in psychological research. *Psychological Methods*, 4, 272-299.

Feldhusen, J. F., Hoover, S. M., & Sayler, M. F. (1989). *Identification of gifted students at the secondary level.* Monroe, NY: Trillium Press.

Ford, D. Y. (1998). The underrepresentation of minority students in gifted education: Problems and promises in recruitment and retention. *The Journal of Special Education*, 32(1), 4-14.

Frasier, M. M., & Passow, A. H. (1994). *Toward a new paradigm for identifying talent potential* (Research Monograph RM94112). Storrs, CT: The National Research Center on the Gifted and Talented, University of Connecticut.

Frasier, M. M., Martin, D., Garcia, J., Finley, V. S., Frank, E., Krisel, S., et al. (1995). *A new window for looking at gifted children* (Research Monograph RM955222). Storrs, CT: The National Research Center on the Gifted and Talented, University of Connecticut.

Gagné, F. (1994). Are teachers really poor talent detectors? Comments on Pegnato and Birch's (1959) study of the effectiveness and efficacy of various identification techniques. *Gifted Child Quarterly*, 38, 124-126.

Gilliam, J. E., Carpenter, B. O., & Christensen, J. R. (1996). *Gifted and Talented Evaluation Scales*. Waco, TX: Prufrock Press

High, M. H., & Udall, A. J. (1983). Teacher ratings of students in relation to ethnicity of student and school ethnic balance. *Journal for the Education of the Gifted*, 6, 154-166.

Hodge, R. D., & Cudmore, L. (1986). The use of teacher-judgment measures in the identification of gifted pupils. *Teaching & Teacher Education*, 2, 181-196.

Hunsaker, S. L., Finley, V. S., & Frank, E. L. (1997). An analysis of teacher nominations and student performance in gifted programs. *Gifted Child Quarterly*, 41, 19-24.

Jarosewich, T., Pfeiffer, S. I., & Morris, J. (2002). Identifying gifted students using teacher rating scales: A review of existing instruments. *Journal of Psychoeducational Assessment*, 20, 322-336.

Joint Committee on Testing Practices. (2005). *Code of fair testing practices in education*. Washington, DC: American Psychological Association.

Kieffer, K. M. (1999). An introductory primer on the appropriate use of exploratory and confirmatory factor analysis. *Research in the Schools*, 6(2), 75-92.

Kingore, B. (2001). *The Kingore Observation Inventory* (2nd ed.). Austin, TX: The Professional Associates Publishing.

Kleinbaum, D. G., Kupper, L. L., Muller, K. E., & Nizam, A. (1998). *Applied regression analysis and other multivariate methods* (3rd ed.). Pacific Grove, CA: Duxbury Press.

Kline, R. B. (2005). *Principles and practice of structural equation modeling* (2nd ed.). New York: The Guilford Press.

Kranz, B. (1981). *Kranz Talent Identification Instrument*. Moorehead, MN :Morehead State College.

McBee, M. T. (2006). A descriptive analysis of referral sources for gifted identification screening by race and socioeconomic status. *Journal of Secondary Gifted Education*, 2, 103-111.

McCarney, S. B., & Anderson, P. D. (1989). *Gifted Evaluation Scale, Second Edition, Technical Manual*. Columbia, MO: Hawthorne Educational Services.

Montanelli, R. G. & Humphreys, L. G. (1976). Latent roots of random data correlation matrices with squared multiple correlations on the diagonal: A Monte Carlo study. *Psychometrika*, 41, 341-347.

Muthén, L. K., & Muthén, B. O. (2002). How to use a Monte Carlo study to decide on sample size and determine power. *Structural Equation Modeling*, 9, 599-620.

National Association for Gifted Children. (2007). *State of the states 2006-2007.* Washington D.C.

National Center for Educational Statistics. (n.d.). Number and percentage distribution of public elementary and secondary students, by percentage of students in school eligible for free or reduced-price lunch and locale: 2003–04. Retrieved September 10, 2008, from http://nces.ed.gov/pubs2007/ruraled/tables/table1_8.asp?referrer=report

Pegnato, C. W., & Birch, J. W. (1959). Locating gifted children in junior high schools: Comparison of methods. *Exceptional Children*, 25, 300-304.

Peterson, J. S. & Margolin, L. (1997). Naming gifted children: An examination of unintended reproduction. *Journal for the Education of the Gifted*, 21, 82-100.

Pfeiffer, S. I., & Jarosewich, T. (2003). *Gifted Rating Scales*. San Antonio, TX: The Psychological Corporation.

Plata, M., & Masten, W. (1998). Teacher ratings of Hispanic and Anglo students on a behavior rating scale. *Roeper Review*, 21, 139-144.

Plata, M., Masten, W., & Trusty, J. (1999). Teachers' perception and nomination of fifth-grade Hispanic and Anglo students. *Journal of Research and Development in Education*, 32, 113-123.

Renzulli, J. S., Smith, L. S., White, A. J., Callahan, C. M., Hartman, R. K., & Westberg, K. L. (2002). *Scales for Rating the Behavioral Characteristics of Superior Students* (*Revised Edition*). Mansfield Center, CT: Creative Learning Press.

Rogers, K. B. (1996). What the bell curve says and doesn’t say: Is a balanced view possible? *Roeper Review*, 18, 252-255.

Siegle, D., & Powell, T. (2004). Exploring teacher biases when nominating students for gifted programs. *Gifted Child Quarterly*, 48, 21-29.

Stambaugh, T. (2007). Next steps: An impetus for future directions in research, policy, and practice for low-income promising learners. In J. VanTassel-Baska & T. Stambaugh (Eds.), *Overlooked gems: A national perspective on low-income promising learners* (pp. 83-88). Washington, DC: National Association for Gifted Children.

Swanson, J. D. (2006). Breaking through assumptions about low-income, minority gifted students. *Gifted Child Quarterly*, 50, 11-25.

Thompson, B. (2004). *Exploratory and confirmatory factor analysis*. Washington, D.C.: American Psychological Association.

United States Department of Education, Office of Educational Research and Improvement. (1993). *National excellence: A case for developing America's talent*. Washington, DC: U.S. Government Printing Office.

Valencia, R. R., & Suzuki, L. A. (2001). *Intelligence testing and minority students*. Thousand Oaks, CA: Sage Publications.

VanTassel-Baska, J. L. (2008). Epilogue: What do we know about identifying and assessing the learning of gifted students?. In J. L. VanTassel-Baska (Ed.), *Alternative assessment with gifted and talented student* (pp. 309-319). Waco, TX: Prufrock Press.

Viadero, D. (2006, November 8). Scholars test out new yardstick of school poverty. *Education Week*, 26(11), 1, 15. Retrieved September 26, 2008 from http://www.edweek.org/ew/articles/2006/11/08/11poverty.h26.html

Widaman, K. (1993). Common Factor Analysis versus Principal Components Analysis: Differential Bias in Representing Model Parameters?. *Multivariate Behavioral Research*, 28, 263-311.

Worrell, F. C. (2007). Identifying and including low-income learners in programs for the gifted and talented: Multiple complexities. In J. VanTassel-Baska & T. Stambaugh (Eds.), *Overlooked gems: A national perspective on low-income promising learners* (pp. 47-51). Washington, DC: National Association of Gifted Children.

Wyner, J. S., Bridgeland, J. M., & Diiulio, J. J. (2007). *The achievement trap*. Washington D. C.: The Jack Kent Cooke Foundation.

Yoon, S., & Gentry, M. (in press). Racial and ethnic representation in gifted programs: Current status of and implications for gifted Asian American students. *Gifted Child Quarterly*.

Zwick, W. R. & Velicer, W. F. (1986). Comparison of five rules for determining the number of components to retain. *Psychological Bulletin*, 99, 432-442.

Appendix. Original *HOPE Teacher-Rating Scale*

Teacher's Name: ______________________________

HOPE[1] Nomination Scale

Student Name/ID #: ________________________________ Grade: _____ Date: _____________

Date of Birth:______________ Age:___ Sex: ☐ Male ☐ Female ☐ Free/Reduced Lunch

☐ American Indian/Alaska Native ☐ Asian ☐ Black or African American ☐ White

☐ Native Hawaiian or Other Pacific Islander

☐ Hispanic (If the student is Hispanic, please further identify his/her origin)

○ Mexican ○ Cuban ○ Dominican ○ Puerto Rican ○ Central American ○ South American

☐ Asian American (If the student is Asian American, please further identify his/her origin)

○ Chinese ○ Filipino ○ Japanese ○ Korean ○ Southeast Asian ○ South Asian

Please read the following statements and rate how frequently you observe the behaviors using the following scale.

6 = always 5 = almost always 4 = often 3 = sometimes 2 = rarely 1 = never

When completing this form please respond by thinking about the student *compared to other children similar in age, experience, and/or environment.*

6 = always 5 = almost always 4 = often 3 = sometimes 2 = rarely 1 = never

	6	5	4	3	2	1
1. Performs or *shows potential* for performing at remarkably high levels.						
2. Is curious, questioning						
3. Is empathetic.						
4. Shows compassion for others.						
5. Has *desire* to work with advanced concepts and materials.						
6. Questions authority.						
7. Is eager to explore new concepts.						
8. Exhibits a strong sense of social justice and fairness.						
9. Uses alternative processes.						
10. Is insightful and intuitive.						
11. Thinks "outside the box."						
12. Has intense interests.						
13. Shows outstanding talent in specific content area(s).						

13 (a). Please indicate all content areas where the student shows talent.

☐ Math ☐ Reading ☐ Creative Writing ☐ Social Studies
☐ Science ☐ Foreign Language ☐ Other ________________

Please provide additional information concerning this child's potential:

____
____
____
____
____

[1]Developed with funding from Jack Kent Cooke 2007