Running Head: HOPE Scale Invariance

Evaluation of Differential Item and Test Functioning of the *HOPE Teacher Rating Scale*

Scott J. Peters
University of Wisconsin – Whitewater

Marcia Gentry
Purdue University

Word Count: 1959 (w/o abstract)

The *HOPE Scale* was introduced with initial validity evidence at the 2009 AERA conference. Based on these results a new validity study was conducted on revisions made to the *Scale*. Items were added and the *HOPE Scale* was completed by a new sample of 71 teachers on their respective 1700 students. Using these data, the *HOPE Scale* was evaluated using confirmatory factor analysis and multi-group CFA to evaluate differential item and test-level functioning with regard to gender, ethnic / racial groupings, and family income. Results were mixed in that differential item functioning was not found with regard to ethnic / racial group membership, but was found when gender groups were compared. Different ethnic / racial groups did also not show any overall mean scale differences, suggesting the *HOPE Scale* items and overall instrument are not affected by racial or ethnic group membership, but that gender does have an inappropriate influence.

Purpose

The purpose of this study was to evaluate a newly developed teacher rating instrument specifically designed to help educators identify proportional numbers of low-income and ethnically / racially diverse students for gifted and talented programs. Such an instrument attempts to address the issue of certain groups, namely African American, Native American, Hispanic, and low-income students, being underrepresented in programs for the gifted and talented. In addition, both the *Code of Fair Testing Practices in Education* (JCTP, 2005) and Lohman (2006) have called for test developers and researchers to evaluate across-group equivalence in any new or pre-existing instrument. Such equivalence is important in that it determines if an instrument can be used with confidence across multiple groups. Without such evaluation, scores may be due to extraneous variables and can lead to biased results and non-valid conclusions.

Literature Review

Underrepresented Students

Traditionally underrepresented students typically include those from African American, Native American, Hispanic and low-income families and are all disproportionately represented in programs for the gifted and talented (Wyner, Bridgeland, & Diiulio, 2007; Yoon & Gentry, 2009). Although this underrepresentation has been a problem in the field of gifted education since the mid 1970s and some groups have made progress toward proportional representation, African American and Hispanic students have seen little progress (Donovan & Cross, 2002). Caucasian and Asian American students continue to be overrepresented and low-income students have only recently received as much attention as those who are underrepresented for ethnic or racial reasons.

Several factors have been proposed as roots of the current underrepresentation problem. Cultural bias in testing and personal bias on the part of educators (Briggs, Reis, & Sullivan, 2008) are among the issues that prevent high-ability students who are also from low-income or minority families from being noticed and can prevent them from receiving high test scores. Teachers who lack adequate understanding of minority cultures often serve as gatekeepers to gifted and talented programs and can fail to recognize talent in minority populations (Siegle & Powell, 2004). When using standardized tests, the use of national norms over local norms has also been suggested as a reason for the failure to recognize some students as gifted and talented (Lohman, 2006). Finally, cultural issues relating to qualities that are seen as important in a gifted and talented student do not often take minority cultures into consideration, but rather focus on Caucasian values and culture (Oakland & Rossen, 2005; Peterson & Margolin, 1997). Whatever the reasons, the fact remains that many students who come from low-income or minority families are overlooked when placement decisions are made. Such an occurrence is not only contrary to the current federal definition of giftedness and talent (USDOE, 1993) but also stands in opposition to the idea of an equal education for all students.

HOPE Scale Development

The initial creation, exploratory factor analysis, confirmatory factor analysis, and invariance testing with regard to income groups were presented at the 2009 AERA conference (Peters, Gates, Gentry, Peterson, & Mann, 2009) and are the subject of a scholarly article currently under review. These analyses found that a two-factor model best fit the Revised *HOPE Scale* data which was collected on 5995 K-5 students by their 349 respective teachers in five metropolitan and rural school corporations in Indiana. In addition, *Scale* items were found to be invariant with regard to family income. This finding indicated that the *HOPE Scale* is not

affected by differential item functioning with regard to income groups. However, despite this finding, students from low-income families did receive lower mean subscale scores than those students not from low-income families. This could be due to actual lower levels of the underlying factors – academic and social components of giftedness and talent – or could have been due to consistent lower ratings by teachers. Because the *HOPE Scale* is a teacher rating form, the results will always only be as good as the individual teacher's perspective.

One of the findings from the analyses of the *HOPE Scale* was that only three items loaded on the Social factor. Therefore, six additional items that helped to define the construct were added to the *Scale*, creating an instrument on which further testing was needed because of the addition of the new items. The 17-item *HOPE Scale* is presented in the Appendix.

Methods and Data Analysis

Participants

The *HOPE Scale* was completed by 71 teachers on their respective 1700 K-5 students from Indiana and Illinois. The students all came from one of two metropolitan or one rural school corporation. The sample demographics are presented in Table 1.

[Table 1 Here]

Methods

A CFA was run using the *HOPE Scale* data described above. The Academic factor had nine items (1, 3, 8, 9, 10, 13, 15, 16, 17) and the Social factor had eight (2, 4, 5, 6, 7, 11, 12, 14). Although a two-factor model (Academic and Social) was expected to best fir the data a single-factor model was also evaluated for comparison. These base models were evaluated using chi-square values, fit indices, modification indices, and R^2 values. Revisions were made based on these statistics and revised models were evaluated with additional CFA testing. Once *Scale*

revisions were complete it was evaluated for measurement invariance using MCFA on the sample described in Table 1. Some of the CFA covariance matrices and parameter estimates are omitted here due to space restrictions, but will be included in the final paper and presentation.

Results

Confirmatory Factor Analysis

The CFA fit indices and chi-square values for the *HOPE Scale* are presented in Table 2.

[Table 2 Here]

Although it is clear that the two-factor model fits the data better than the single-factor model, all of the fit statistics and the chi-square value indicate a worse fitting model than was found in previous research of the 13-item *HOPE Scale* (Peters, et al., 2009). Because of this, additional revisions were necessary. Items 4 and 14 had the lowest R^2 values at .504 and .214 respectively. Because these two items contributed so little to accounting for the variance in the model, both were removed. All other items had R^2 values of between .665 and .859. The majority of the modification indices had to do with correlated error (theta-delta) terms. In general, correlated errors should be avoided because these correlations signify an overlap in content coverage and the existence of non-unidimensional items. By far the largest modification index suggested that the errors of Items 6 (*Shows compassion for others*) and 12 (*Is empathetic*) be allowed to correlate. This single modification decreased the model chi-square value by more than 624. In this case, Item 12 was removed from the *Scale* instead of allowing its error to correlate with that of Item 6. Item 12 was chosen because its removal contributed to better overall model fit than did the removal of Item 6. This choice made sense from the standpoint of content because the items addressed very similar behaviors, but also allowed for better fit by requiring one less parameter to be estimated. The same issue was present for Items 8 and 9. The errors of these two items

were originally allowed to correlate as part of the *HOPE Scale* model. However, at this point in the *Scale's* development, such a pairing was not ideal. In this case, removing Item 8 contributed to better model fit than did removing Item 9. The final suggested error correlation was between Items 16 and 15. As with the other two suggested pairings, one item, in this case Item 15, was removed.

Only one item appeared to cross-load on both factors. Item 7 (*Is a leader within his/her group of peers*) appeared to be split on both factors. The modification index suggested a chi-square decrease by approximately 200 if this item was cross-loaded onto the Academic factor instead of contributing only to the Social factor. Because cross-loading items are undesirable, this item was removed. Although retaining Item 7 as a cross-loading item would have improved model fit, it would not have made sense based on the theory and therefore would have been less likely to be reproduced in additional samples. Two other modification indices were above 100, but were not made since they did not make sense from the theoretical perspective of the two-factor model. Fit statistics for the revised model are presented in Table 3.

[Table 3 Here]

Once the revisions were made to the *Scale,* the model was then evaluated with MCFA procedures to evaluate measurement invariance.

HOPE Scale Invariance

Although invariance was evaluated with regard to income in previous research, it was re-evaluated here with additional items and a new sample. In addition, invariance due to racial / ethnic group membership and gender were evaluated. Although gender equity was not the primary focus of the *HOPE Scale,* such invariance is important in any standardized measure. The first step in the invariance testing involved the model being evaluated on each group separately,

followed by testing seven increasingly restrictive models. This was done for the two gender groups, the four race / ethnic groups, and the two income groups (free or reduced lunch and non free or reduced lunch). Tables 4, 5, and 6 presents the results of the gender, race / ethnicity, and income group invariance testing.

[Table 4 Here]

[Table 5 Here]

[Table 6 Here]

Tables 5 and 6 indicate that no differential item functioning was present at the item level when race / ethnic and income groups were compared. For income groups this result is the same as that which was found in previous research with the *HOPE Scale* (See Peters et al., 2009). This result means that family background, with regard to income and race / ethnicity, did not result in a worse fitting model for the *HOPE Scale*. Unfortunately, the same could not be said for gender (Table 4). When invariance was tested across gender groups there was a significant increase in chi-square value, indicating non-invariance at the intercepts of the individual items. Because of this finding, further interpretation of the gender invariance results was not possible as the tests of equal latent means and variances would be at least partially due to item-level non-invariance.

The latent variance and latent mean results for race / ethnicity and income group were mixed. Although no differential functioning was found at the subscale level due to race / ethnicity, the same was not true for income group. Again, this result is similar to that which was found with previous research of the *HOPE Scale*. This finding indicates that although no differential item functioning was found to be due to income group membership, low-income students still received significantly lower scores than did students from non low-income families.

Discussion

The findings from this study further support the use of the *HOPE Scale* in identifying traditionally underrepresented students for gifted and talented programs. However, the invariance results suggest a major caveats with regard to practice. Due to some level of non-invariance, student comparisons should only be made within their specific subgroup. For example, male students from low-income families should be compared to other male students from low-income families in order to avoid misattributing scores to the student when they are actually due to non-invariance. This specific norm group comparison is supported by Lohman's (2006) argument that such comparisons will result in a more accurate understanding of a student's aptitude than if that student was compared to national norms or other less-specific norm group.

Implications and Recommendations

One major implication is that no instrument, regardless of the design or validation procedures, can be assumed to be free of differential item or test-level functioning. However, when such issues were evaluated and understood, such as with the *HOPE Scale*, accurate comparisons can be made, within specific norm groups, without the fear that scores are inappropriately influenced by extraneous variables.

Limitations

The greatest limitation of this study is that the sample was relatively small and not representative of the full US population. As such additional research is needed before widespread generalizations can be made. In addition, teacher rating scales are always limited by the knowledge, background, and personal biases of the person doing the rating. As such some degree of training in gifted education and underrepresented populations may help yield the most accurate ratings of such students.

References

Briggs, C. J., Reis, S. M., & Sullivan, E. E. (2008). A national view of promising programs and practices for culturally, linguistically, and ethnically diverse gifted and talented students. *Gifted Child Quarterly*, 52, 131-145.

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

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

Lohman, D. F. (2006). *Identifying academically talented minority students* (Research Monograph RM05216). Storrs, CT: The National Research Center on the Gifted and Talented, University of Connecticut.

Oakland, T., & Rossen, E. (2005). A 21st-century model for identifying students for gifted and talented programs on light of national conditions: An emphasis on race and ethnicity. *Gifted Child Today*, 28(4), 56-65.

Peters, S. J., Gates, J. C., Gentry, M., Peterson, J. S., & Mann, R. L. (2009, April). Exploratory and confirmatory validation of the *HOPE Scale*: Instrumentation to identify low-income K-5 students. Annual Meeting of the American Educational Research Association, San Diego, CA.

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

Piechowski, M. M. (2006). *'Mellow out,' they say. If I only could: Intensities and sensitivities of the young and bright.* Madison, WI: Yunasa Books.

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

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

Table 1. Revised *HOPE Scale* Sample Demographics

	West	Southern	Tiller
Designation	Metro	Rural	Metro
K-5 Sample	486	675	539
Free/Reduced Lunch Students	59%	14%	43%
Caucasian	22%	96%	33%
African American	33%	1%	17%
Hispanic	16%	1%	36%
Asian	20%	<1%	12%
Multi-racial	5%	1%	0%
Native American	1%	<1%	2%
Unknown Ethnicity	3%	0%	0%
Gender	51% Male	51% Male	48% Male

Table 2. Revised *HOPE Scale* Base Model Fit Statistics

	Two-Factor Model		Single-Factor Model	
Index	Value	Notes	Value	Notes
Chi-square	2862.764	*p*-value : <.001	4336.941	*p*-value : <.001
	df=117		df=118	
RMSEA	.119	90% CI: .115-.123	.147	90% CI: .143-.151
CFI	.917		.872	
TLI	.903		.853	
SRMR	.052		.055	

Table 3. *Revised Model Fit Statistics*

Index	Value	Notes
Chi-square	1051.705	*p*-value : <.001
	df=53	
RMSEA	.107	90% CI: .101-.113
CFI	.953	
TLI	.941	
SRMR	.033	

Table 4. *Invariance Tests for Gender*

	$\chi 2$	df	$\chi 2$diff	Δdf	RMSEA (90% CI)	SRMR	CFI	TLI
Single Group Solutions								
Male (*n*=758)	394.934*	43	-	-	.104 (.095 - .113)	.03	.96	.95
Female (*n*=753)	464.294*	43	-	-	.114 (.105 - .124)	.03	.95	.94
Measurement Invariance								
Equal Form	859.228*	86	-	-	.109 (.103 - .116)	.03	.96	.95
Equal Factor Loading	876.031*	95	16.803	9	.104 (.098 - .111)	.04	.96	.95
Equal Indicator Intercepts	918.633*	104	42.602*	9	.102 (.096 - .108)	.04	.95	.95
Equal Indicator Error Variances	946.210*	115	27.577*	11	.098 (.092 - .104)	.04	.95	.95
Population Heterogeneity								
Equal Factor Variance	949.146*	117	2.936	2	.097 (.091 - .103)	.04	.95	.96
Equal Latent Mean	1027.202*	119	78.056*	2	.101 (.095 - .106)	.06	.95	.95

Note. * significant at $p<.001$

Table 5. *Invariance Tests for Ethnic / Racial Groups*

		$\chi 2$	df	$\chi 2$diff	Δdf	RMSEA (90% CI)	SRMR	CFI	TLI
Single Group Solutions									
Caucasian	(*n*=876)	528.487*	43	-	-	.114 (.105 - .122)	.03	.96	.94
Asian	(*n*=157)	193.284*	43	-	-	.149 (.128 - .171)	.04	.92	.90
African American	(*n*=202)	140.207*	43	-	-	.106 (.087 - .126)	.04	.96	.94
Hispanic	(*n*=223)	222.616*	43	-	-	.137 (.119 - .155)	.04	.92	.90
Measurement Invariance									
Equal Form		1084.594*	172	-	-	.121 (.114 - .128)	.04	.95	.93
Equal Factor Loading		1117.481*	199	32.887	27	.113 (.106 - .119)	.05	.95	.94
Equal Indicator Intercepts		1148.74*	226	31.259	27	.106 (.100 - .112)	.05	.95	.95
Equal Indicator Error Variances		1327.819*	259	179.079*	33	.106 (.101 - .112)	.05	.94	.95
Population Heterogeneity									
Equal Factor Variance		1346.148*	265	18.329*	6	.106 (.100 - .111)	.113	.94	.95
Equal Latent Mean		1351.703*	271	5.555	6	.105 (.099 - .110)	.113	.94	.95

Note. * significant at $p<.001$

Table 6. *Invariance Tests for Low-Income vs. Non Low-Income Students*

		$\chi 2$	df	$\chi 2$diff	Δdf	RMSEA (90% CI)	SRMR	CFI	TLI
Single Group Solutions									
Paid	(*n*=1007)	549.291*	43	-	-	.108 (.100 - .116)	.03	.96	.95
Free / Reduced	(*n*=499)	403.717*	43	-	-	.13 (.118 - .141)	.04	.93	.91
Measurement Invariance									
Equal Form		953.007*	86	-	-	.116 (.109 - .122)	.04	.95	.94
Equal Factor Loading		973.78*	95	20.773	9	.111 (.105 - .117)	.04	.95	.94
Equal Indicator Intercepts		983.737*	104	9.957	9	.106 (.100 - .112)	.04	.95	.95
Equal Indicator Error Variances		1054.560*	115	70.823*	11	.104 (.098 - .110)	.04	.95	.95
Population Heterogeneity									
Equal Factor Variance		1072.624*	117	18.082*	2	.104 (.098 - .110)	.10	.95	.95
Equal Latent Mean		1113.776*	119	41.136*	2	.105 (.100 - .111)	.114	.94	.95

Note. * significant at $p<.001$

Appendix: Revised *HOPE Teacher Rating Scale*

Teacher's Name/Code: _____

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 ☐ Mixed Race

☐ Hispanic

When rating students on each item below please think about the student *compared to other children similar in age, experience, and/or environment.*

Use the following scale to indicate how frequently you observe the traits and behaviors listed in items 1 – 18.

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 sensitive to larger or deeper issues of human concern.						
3. Is curious, questioning.						
4. Shows emotional intensity.						
5. Is self-aware.						
6. Shows compassion for others.						
7. Is a leader within his/her group of peers.						
8. Has *desire* to work with advanced concepts and materials.						
9. Is eager to explore new concepts.						
10. Exhibits intellectual intensity.						
11. Effectively interacts with adults or older students.						
12. Is empathetic.						
13. Uses alternative processes.						
14. Is more energetic than most people his/her age.						
15. Is insightful and intuitive.						
16. Thinks "outside the box."						
17. Has intense interests.						
18. Please indicate all content areas where the student shows talent. ☐ Math ☐ Reading ☐ Creative Writing ☐ Social Studies ☐ Science ☐ Foreign Language ☐ Arts ☐ Other _____						

Please provide additional information concerning this child's potential:

_____

_____

_____

_____

_____

[1]Developed with funding from Jack Kent Cooke 2007

Revised 4/16/2008