Research Report

Comparison of Subscores Based on Classical Test Theory Methods

Gautam Puhan

Sandip Sinharay

Shelby Haberman

Kevin Larkin

October 2008

ETS RR-08-54

Comparison of Subscores Based on Classical Test Theory Methods

Gautam Puhan, Sandip Sinharay, Shelby Haberman, and Kevin Larkin
ETS, Princeton, NJ

October 2008

As part of its nonprofit mission, ETS conducts and disseminates the results of research to advance quality and equity in education and assessment for the benefit of ETS's constituents and the field.

ETS Research Reports provide preliminary and limited dissemination of ETS research prior to publication. To obtain a PDF or a print copy of a report, please visit:

http://www.ets.org/research/contact.html

Abstract

Will reporting subscores provide any additional information than the total score? Is there a method that can be used to provide more trustworthy subscores than observed subscores? These 2 questions are addressed in this study. To answer the 2nd question, 2 subscore estimation methods (i.e., subscore estimated from the observed total score or subscore estimated using both the observed subscore and observed total score) are compared. Analyses conducted on 8 certification tests indicated that reporting subscores at the examinee level may not be necessary as they do not provide much additional information than the total score. However, at the institutional level (for institution size greater than 30), reporting subscores may not be harmful, although it may be redundant. Finally, results indicated that subscores estimated using both the observed subscore and observed total score were the most trustworthy and may be used if subscores were to be reported.

Key words: Subscores, augmentation, PRMSE, reliability

Acknowledgments

The authors thank Adele (Xuan) Tan, Frank Rijmen, and Dan Eignor for their comments and useful suggestions and Kim Fryer for help with proofreading.

Table of Contents

List of Tables

Page

List of Figures

Page

Introduction

Testing programs commonly report total test scores but not subscores to examinees and academic institutions (e.g., school districts, colleges, and universities). However, the demand for subscores is fast increasing due to at least two important reasons. First, failing candidates want to know their strengths and weaknesses in different content areas to plan future remedial studies. Second, states and academic institutions such as colleges and universities want a profile of performance for their graduates to better evaluate their training and focus on areas that need remediation (Haladyna & Kramer, 2004). The desire to receive subscores at the examinee and/or institutional level is even stronger in certification and licensure testing because a small difference in the total score of these tests can make a difference in the pass or fail status of the examinees. Therefore the general consensus seems to be that examinees attending remedial training may get a slight edge (i.e., improving the total score) by improving on subcontent areas where they may be weaker.

Despite this apparent usefulness of subscores, certain important factors must be considered before making a decision to report subscores at either the individual or institutional level. Although many tests are designed to cover a broad domain and the total test score is considered a composite of different abilities measured by different subsections, a subsection with fewer items than the total test may not be able to precisely measure a unique ability.

Haberman (2005) argued that a subscore may be considered to be of added value only when it provides a more accurate measure of the construct being measured than is provided by the total score. Similarly, Tate (2004) emphasized the importance of ensuring reasonable subscore performance in terms of high reliability and validity to minimize incorrect instructional and remediation decisions. A similar concern is also apparent in Standard 5.12 of the *Standards for Educational and Psychological Testing* (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 1999), which states that “scores should not be reported for individuals unless the validity, comparability, and reliability of such scores have been established” (p. 65). This standard applies to subscores as well.

As evident from the above discussion, a conflict appears to exist between the demand from examinees and institutions to report subscores and the need for testing practitioners to exercise caution in considering whether to report subscores based on a smaller number of items,

which may not be reliable. As Monaghan (2006) pointed out, “While they want to be responsive to the desires of the educational marketplace, assessment organizations are very interested in the appropriate use of subscores” (p.1). For example, institutions may plan a remedial training based on subscore information that may not be reliable and therefore lead to large and needless expense for them. In this context, two important questions arise:

1. Considering that the total test score is already reported to examinees and institutions, will reporting subscores provide any additional information than what is provided in the total score?

2. If testing programs, in their effort to be responsive to the needs of examinees and institutions, were to report subscores, does a method exist that can be used to provide subscores that are more trustworthy or reliable than observed subscores?

This study answers these two questions. First, we evaluated whether a subscore provided added value as compared to the total score for several tests measuring basic skills. If the true subscore can be predicted more accurately from the observed total score than the observed subscore, then the observed subscore may not be of added value and hence may not be worth reporting. Second, we compared three classical test theory–based methods for estimating subscores. If subscores obtained by using a particular method is more trustworthy than the other methods, then it may be preferable to report subscores estimated using that method.

Method

To answer the first question, we used a statistical measure described in Haberman (2005), Haberman, Sinharay, and Puhan (2006), and Sinharay, Haberman, and Puhan (2007) to evaluate when subscores provided any added value over the total scores. We referred to the statistical measure as *trustworthiness*. In statistical terms, trustworthiness refers to proportional reduction of mean squared error (PRMSE). It is an index that is similar to a reliability coefficient and ranges from 0 to 1 with 0 and 1 indicating the lowest and highest degrees of trustworthiness, respectively. A subscore was said to have added value when the trustworthiness or PRMSE for the subscore was larger than that for the total score, which happened when the observed subscore predicted the true subscore more accurately than did the observed total score.

To answer the second question, an additional method that predicted the true subscore using both the observed subscore and the observed total score (see Haberman, 2005, and

Haberman et al., 2006) was used. This method (referred to as the Haberman augmented subscore method throughout this paper) was compared to the other two methods (i.e., true subscore predicted from the observed subscore and true subscore predicted from the observed total score) to assess which of the three methods led to a better prediction of the true subscore (i.e., produced a more trustworthy subscore) as indicated by an increase in the PRMSE.

The analysis was conducted both at the examinee and institutional levels because the results from both analyses might not necessarily have been the same. Examinee level analysis refers to conducting the analyses using the subscores for the individual examinees (i.e., at the total data level) while institutional level analysis refers to analyzing institutional level subscores. Although the analysis at the institutional level in this study constituted school districts, this type of analysis could be extended to other types of meaningful aggregation such as states or subgroups based on gender or ethnicity. In this study, we evaluated the trustworthiness of institutional level subscores for institution sizes of 10, 30, 100, and 150. Although other institution sizes could have been examined, these sizes seemed reasonable in the current context because the institution sizes for the tests under investigation fell into this approximate range.

The earlier discussion evidenced at least three predictors of the true subscore (i.e., the observed subscore, the observed total score, and a combination of the observed subscore and the observed total score). Although the derivations of the formulas and computational details for estimating the true subscores based on these predictors are provided in detail in Haberman (2005), Haberman et al. (2006), and Sinharay et al. (2007), they are also presented briefly in the appendix because these procedures are relatively new and may not be familiar to many testing practitioners.

Tests Used

This study used examinee responses from eight certification tests, including tests containing only multiple choice (MC) items, only constructed response (CR) items, and combinations of MC and CR items. These tests represented a broad range of subject and skill areas such as elementary education, mathematics, and foreign languages. For confidentiality reasons, hypothetical names (e.g., Test A, Test B, etc.) were used for the eight tests. The total number of examinees taking each of these tests and the number of items in each subscore for the eight tests are presented in Tables 1–8. A brief description of the tests and subscores for each test is presented in this section.

Table 1

Proportional Reduction of Mean Squared Error (PRMSE) for Four Subscores: Test A

	N	Subscores			
		1 (30 items)	2 (30 items)	3 (30 items)	4 (30 items)
Examinee level					
$\rho^2(s_t, s)$ [a]	31,001	0.712	0.825	0.735	0.708
$\hat{\rho}^2(s_t, x)$ [b]	31,001	0.768	0.738	0.749	0.818
$\hat{\rho}^2(s_t, sx)$ [c]	31,001	0.819	0.855	0.815	0.839
Institutional level					
$\rho^2(s_I, \overline{s})$ [d]	10	0.691	0.735	0.678	0.656
	30	0.870	0.893	0.864	0.851
	100	0.957	0.965	0.955	0.950
	150	0.971	0.977	0.969	0.966
$\rho^2(s_I, \overline{x})$ [e]	10	0.733	0.733	0.7	0.753
	30	0.864	0.863	0.825	0.887
	100	0.922	0.921	0.880	0.946
	150	0.931	0.930	0.888	0.955
$\rho^2(s_I, \overline{sx})$ [f]	10	0.748	0.760	0.725	0.753
	30	0.890	0.900	0.877	0.891
	100	0.961	0.966	0.957	0.958
	150	0.973	0.977	0.970	0.971

Note. The percent reduction in mean squared error (PRMSE) is similar to the concept of test reliability in that a higher number indicates a higher reliability or lower error.

[a] PRMSE when the true subscore is predicted from the observed subscore. [b] PRMSE when the true subscore is predicted from the observed total score. [c] PRMSE when the true subscore is predicted from the observed subscore and observed total score. [d] PRMSE when the true institution subscore is predicted from the institution average observed subscore. [e] PRMSE when the true subscore is predicted from the institution average observed total score. [f] PRMSE when the true subscore is predicted from the institution average observed subscore and institution average observed total score.

Table 2

Proportional Reduction of Mean Squared Error (PRMSE) for Three Subscores: Test B

	N	Subscores		
		1 (14 items)	2 (11 items)	3 (35 items)
Examinee level				
$\rho^2(s_t, s)$ [a]	7,930	0.464	0.279	0.625
$\rho^2(s_t, x)$ [b]	7,930	0.705	0.726	0.732
$\hat{\rho}^2(s_t, sx)$ [c]	7,930	0.709	0.727	0.733
Institutional level				
$\rho^2(s_I, \bar{s})$ [d]	10	0.527	0.458	0.630
	30	0.770	0.717	0.836
	100	0.918	0.894	0.945
	150	0.944	0.927	0.962
$\rho^2(s_I, \bar{x})$ [e]	10	0.667	0.495	0.658
	30	0.853	0.634	0.842
	100	0.946	0.703	0.933
	150	0.960	0.714	0.948
$\rho^2(s_I, \overline{sx})$ [f]	10	0.667	0.553	0.662
	30	0.854	0.759	0.850
	100	0.947	0.902	0.947
	150	0.963	0.931	0.964

Note. The percent reduction in mean squared error (PRMSE) is similar to the concept of test reliability in that a higher number indicates a higher reliability or lower error.

[a] PRMSE when the true subscore is predicted from the observed subscore. [b] PRMSE when the true subscore is predicted from the observed total score. [c] PRMSE when the true subscore is predicted from the observed subscore and observed total score. [d] PRMSE when the true institution subscore is predicted from the institution average observed subscore. [e] PRMSE when the true subscore is predicted from the institution average observed total score. [f] PRMSE when the true subscore is predicted from the institution average observed subscore and institution average observed total score.

Table 3

Proportional Reduction of Mean Squared Error (PRMSE) for Four Subscores: Test C

	N	Subscores			
		1 (12 items)	2 (12 items)	3 (10 items)	4 (2 items)
Examinee level					
$\rho^2(s_t, s)$ [a]	3,960	0.436	0.443	0.379	0.241
$\rho^2(s_t, x)$ [b]	3,960	0.837	0.820	0.845	0.867
$\hat{\rho}^2(s_t, sx)$ [c]	3,960	0.850	0.847	0.863	0.879
Institutional level					
$\rho^2(s_I, \bar{s})$ [d]	10	0.547	0.554	0.580	0.351
	30	0.783	0.789	0.805	0.618
	100	0.923	0.926	0.932	0.844
	150	0.948	0.949	0.954	0.890
$\rho^2(s_I, \bar{x})$ [e]	10	0.644	0.589	0.626	0.551
	30	0.836	0.764	0.813	0.715
	100	0.934	0.853	0.907	0.798
	150	0.950	0.868	0.923	0.812
$\rho^2(s_I, \overline{sx})$ [f]	10	0.647	0.607	0.646	0.557
	30	0.840	0.810	0.838	0.745
	100	0.941	0.929	0.940	0.878
	150	0.958	0.951	0.958	0.909

Note. The percent reduction in mean squared error (PRMSE) is similar to the concept of test reliability in that a higher number indicates a higher reliability or lower error.

[a] PRMSE when the true subscore is predicted from the observed subscore. [b] PRMSE when the true subscore is predicted from the observed total score. [c] PRMSE when the true subscore is predicted from the observed subscore and observed total score. [d] PRMSE when the true institution subscore is predicted from the institution average observed subscore. [e] PRMSE when the true subscore is predicted from the institution average observed total score. [f] PRMSE when the true subscore is predicted from the institution average observed subscore and institution average observed total score.

Table 4

Proportional Reduction of Mean Squared Error (PRMSE) for Six Subscores: Test D

		Subscores					
	N	1 (29 items)	2 (29 items)	3 (21 items)	4 (19 items)	5 (19 items)	6 (13 items)
Examinee level							
$\rho^2(s_t, s)$ [a]	8,365	0.735	0.756	0.68	0.624	0.519	0.449
$\rho^2(s_t, x)$ [b]	8,365	0.818	0.851	0.854	0.83	0.738	0.744
$\hat{\rho}^2(s_t, sx)$ [c]	8,365	0.847	0.869	0.865	0.845	0.764	0.766
Institutional level							
$\rho^2(s_I, \bar{s})$ [d]	10	0.629	0.649	0.649	0.645	0.540	0.587
	30	0.836	0.847	0.848	0.845	0.779	0.810
	100	0.944	0.949	0.949	0.948	0.922	0.934
	150	0.962	0.965	0.965	0.965	0.946	0.955
$\rho^2(s_I, \bar{x})$ [e]	10	0.674	0.684	0.720	0.645	0.655	0.645
	30	0.826	0.838	0.882	0.790	0.803	0.790
	100	0.896	0.909	0.958	0.858	0.871	0.858
	150	0.907	0.921	0.970	0.869	0.882	0.869
$\rho^2(s_I, \overline{sx})$ [f]	10	0.685	0.693	0.722	0.686	0.667	0.683
	30	0.855	0.862	0.884	0.859	0.835	0.849
	100	0.948	0.951	0.960	0.950	0.933	0.942
	150	0.964	0.966	0.972	0.966	0.952	0.959

Note. The percent reduction in mean squared error (PRMSE) is similar to the concept of test reliability in that a higher number indicates a higher reliability or lower error.

[a] PRMSE when the true subscore is predicted from the observed subscore. [b] PRMSE when the true subscore is predicted from the observed total score. [c] PRMSE when the true subscore is predicted from the observed subscore and observed total score. [d] PRMSE when the true institution subscore is predicted from the institution average observed subscore. [e] PRMSE when the true subscore is predicted from the institution average observed total score. [f] PRMSE when the true subscore is predicted from the institution average observed subscore and institution average observed total score.

Table 5

Proportional Reduction of Mean Squared Error (PRMSE) for Four Subscores: Test E

		Subscores			
	N	1 (34 items)	2 (35 items)	3 (30 items)	4 (21 items)
Examinee level					
$\rho^2(s_t, s)$ [a]	2,154	0.836	0.827	0.862	0.693
$\rho^2(s_t, x)$ [b]	2,154	0.847	0.836	0.876	0.619
$\hat{\rho}^2(s_t, sx)$ [c]	2,154	0.890	0.879	0.910	0.768
Institutional level					
$\rho^2(s_I, \overline{s})$ [d]	10	0.669	0.686	0.624	0.462
	30	0.859	0.868	0.833	0.720
	100	0.953	0.956	0.943	0.896
	150	0.968	0.970	0.961	0.928
$\rho^2(s_I, \overline{x})$ [e]	10	0.711	0.711	0.711	0.515
	30	0.881	0.881	0.880	0.638
	100	0.961	0.961	0.960	0.697
	150	0.974	0.974	0.973	0.706
$\rho^2(s_I, \overline{sx})$ [f]	10	0.711	0.711	0.715	0.559
	30	0.881	0.881	0.882	0.759
	100	0.961	0.961	0.961	0.902
	150	0.974	0.974	0.974	0.931

Note. The percent reduction in mean squared error (PRMSE) is similar to the concept of test reliability in that a higher number indicates a higher reliability or lower error.

[a] PRMSE when the true subscore is predicted from the observed subscore. [b] PRMSE when the true subscore is predicted from the observed total score. [c] PRMSE when the true subscore is predicted from the observed subscore and observed total score. [d] PRMSE when the true institution subscore is predicted from the institution average observed subscore. [e] PRMSE when the true subscore is predicted from the institution average observed total score. [f] PRMSE when the true subscore is predicted from the institution average observed subscore and institution average observed total score.

Table 6

Proportional Reduction of Mean Squared Error (PRMSE) for Four Subscores: Test F

	N	Subscores			
		1 (10 items)	2 (6 items)	3 (2 items)	4 (7 items)
Examinee level					
$\rho^2(s_t, s)$ [a]	3,878	0.541	0.462	0.435	0.476
$\rho^2(s_t, x)$ [b]	3,878	0.597	0.664	0.711	0.690
$\hat{\rho}^2(s_t, sx)$ [c]	3,878	0.681	0.687	0.711	0.694
Institutional level					
$\rho^2(s_I, \bar{s})$ [d]	10	0.604	0.496	0.599	0.488
	30	0.821	0.747	0.792	0.741
	100	0.938	0.908	0.927	0.905
	150	0.958	0.936	0.950	0.935
$\rho^2(s_I, \bar{x})$ [e]	10	0.601	0.631	0.646	0.677
	30	0.766	0.804	0.823	0.863
	100	0.847	0.890	0.910	0.955
	150	0.861	0.903	0.924	0.960
$\rho^2(s_I, \overline{sx})$ [f]	10	0.659	0.637	0.651	0.677
	30	0.843	0.821	0.836	0.863
	100	0.943	0.926	0.937	0.955
	150	0.960	0.947	0.955	0.969

Note. The percent reduction in mean squared error (PRMSE) is similar to the concept of test reliability in that a higher number indicates a higher reliability or lower error.

[a] PRMSE when the true subscore is predicted from the observed subscore. [b] PRMSE when the true subscore is predicted from the observed total score. [c] PRMSE when the true subscore is predicted from the observed subscore and observed total score. [d] PRMSE when the true institution subscore is predicted from the institution average observed subscore. [e] PRMSE when the true subscore is predicted from the institution average observed total score. [f] PRMSE when the true subscore is predicted from the institution average observed subscore and institution average observed total score.

Table 7

Proportional Reduction of Mean Squared Error (PRMSE) for Three Subscores: Test G

	N	Subscores		
		1 (17 items)	2 (12 items)	3 (21 items)
Examinee level				
$\rho^2(s_t, s)$ [a]	6,818	0.608	0.587	0.651
$\rho^2(s_t, x)$ [b]	6,818	0.809	0.780	0.807
$\hat{\rho}^2(s_t, sx)$ [c]	6,818	0.809	0.787	0.808
Institutional level				
$\rho^2(s_I, \bar{s})$ [d]	10	0.672	0.648	0.690
	30	0.860	0.846	0.870
	100	0.953	0.948	0.957
	150	0.968	0.965	0.971
$\rho^2(s_I, \bar{x})$ [e]	10	0.745	0.707	0.73
	30	0.896	0.850	0.877
	100	0.964	0.915	0.944
	150	0.974	0.925	0.954
$\rho^2(s_I, \overline{sx})$ [f]	10	0.746	0.717	0.734
	30	0.896	0.873	0.886
	100	0.965	0.953	0.960
	150	0.975	0.967	0.973

Note. The percent reduction in mean squared error (PRMSE) is similar to the concept of test reliability in that a higher number indicates a higher reliability or lower error.

[a] PRMSE when the true subscore is predicted from the observed subscore. [b] PRMSE when the true subscore is predicted from the observed total score. [c] PRMSE when the true subscore is predicted from the observed subscore and observed total score. [d] PRMSE when the true institution subscore is predicted from the institution average observed subscore. [e] PRMSE when the true subscore is predicted from the institution average observed total score. [f] PRMSE when the true subscore is predicted from the institution average observed subscore and institution average observed total score.

Table 8

Proportional Reduction of Mean-Squared Error (PRMSE) for Three Subscores: Test H

	N	Subscores		
		1 (25 items)	2 (25 items)	3 (25 items)
Examinee level				
$\rho^2(s_t, s)$ [a]	3,637	0.865	0.837	0.852
$\rho^2(s_t, x)$ [b]	3,637	0.895	0.854	0.868
$\hat{\rho}^2(s_t, sx)$ [c]	3,637	0.913	0.889	0.899
Institutional level				
$\rho^2(s_I, \bar{s})$ [d]	10	0.702	0.722	0.586
	30	0.876	0.886	0.809
	100	0.959	0.963	0.934
	150	0.973	0.975	0.955
$\rho^2(s_I, \bar{x})$ [e]	10	0.729	0.729	0.729
	30	0.890	0.890	0.890
	100	0.964	0.964	0.964
	150	0.976	0.976	0.976
$\rho^2(s_I, \overline{sx})$ [f]	10	0.729	0.729	0.729
	30	0.890	0.89	0.890
	100	0.964	0.964	0.964
	150	0.976	0.976	0.976

Note. The percent reduction in mean squared error (PRMSE) is similar to the concept of test reliability in that a higher number indicates a higher reliability or lower error.

[a] PRMSE when the true subscore is predicted from the observed subscore. [b] PRMSE when the true subscore is predicted from the observed total score. [c] PRMSE when the true subscore is predicted from the observed subscore and observed total score. [d] PRMSE when the true institution subscore is predicted from the institution average observed subscore. [e] PRMSE when the true subscore is predicted from the institution average observed total score. [f] PRMSE when the true subscore is predicted from the institution average observed subscore and institution average observed total score.

Test A is designed for prospective teachers in primary through upper elementary school grades. It comprises 120 MC items divided into four broad categories (i.e., subscores), namely language arts/reading (30 items), mathematics (30 items), social studies (30 items), and science (30 items).

Test B is designed for examinees who plan to teach in a special education program at any grade level from preschool through grade 12. It comprises 60 MC questions divided into three broad categories, namely understanding exceptionalities (14 items), legal and societal issues (11 items), and delivery of services to students with disabilities (35 items).

Test C is designed to assess a beginning teacher's knowledge of a variety of job-related criteria. It comprises 24 MC questions and 12 CR questions that are divided into four broad categories, namely students as learners (8 MC items and 4 CR items), instruction and assessment (8 MC items and 4 CR items), teacher professionalism (8 MC items and 2 CR items), and communication techniques (2 CR items).

Test D is designed to assess whether an examinee has the knowledge and skills necessary for a beginning social studies teacher in a secondary school. This test comprises 130 MC items divided into six broad categories, namely United States history (29 items), world history (29 items), government/civics/political science (21 items), geography (19 items), economics (19 items), and the behavioral sciences (13 items).

Test E is designed to assess the knowledge and competencies necessary for a beginning or entry-year teacher of Spanish. This test comprises 120 MC questions divided into four broad categories, namely interpretive listening (34 items), structure of the language (35 items), interpretive reading (30 items), and cultural perspectives (21 items).

Test F is designed to measure whether entry-level principals and other school leaders have standards-relevant knowledge believed necessary for competent professional practice. It comprises 25 CR questions divided into four broad categories, namely evaluation of Actions I (10 items covering situations a principal might encounter), evaluation of Actions II (6 items that present a dilemma based on typical school issues), synthesis of information and problem solving (two items that require the candidate to propose a course of action to address a complex problem), and analysis of information and decision making (7 items that the candidate has to answer based on some documents that relate to teaching and learning issues such as staff evaluations, budget, etc.).

Test G is designed to assess the mathematical knowledge and competencies necessary for a beginning teacher of secondary school mathematics. It comprises 50 MC questions divided into three broad categories, namely mathematical concepts and reasoning (17 items), ability to integrate knowledge of different areas of mathematics (12 items), and the ability to develop mathematical models of real life situations (21 items).

Test H is used to measure skills necessary for prospective and practicing paraprofessionals. It comprises 75 MC questions divided into three broad categories, namely reading (25 items), mathematics (25 items), and writing (25 items).

Results

The results from the study are presented in a series of graphs. The first two graphs (Figures 1 and 2) present the results from the examinee level analyses (i.e., whether subscores are reasonable to report to the examinees). Each figure has four panels with each panel showing the examinee level results for one test (e.g., Test A). Figure 3 shows the partitioning of a single test form into two half forms. The remaining graphs (Figures 4–11) present the results from the institutional level analyses. Each of these figures, showing the results for a particular test, has four panels showing institutional level results for institution sizes 10, 30, 100, and 150. A single panel in any of the figures (except Figure 3) compares the PRMSEs for the true subscore predicted from the observed subscore (TS-OS), true subscore predicted from the observed total score (TS-OTS), and the true subscore predicted from the combination of the observed subscore and observed total score (Haberman augmented subscore). For interested readers, the same information is also provided numerically in Tables 1–8.

As seen in Figures 1 and 2, the PRMSE or the trustworthiness of the TS-OS was lower for most tests as compared to that of the TS-OTS, suggesting that reporting observed subscores to examinees may not be of added value. The only instances when a subscore was favorable occurred for the second subscore in Test A and the fourth subscore in Test E where the PRMSE of the TS-OS was larger than the PRMSE of the TS-OTS.

It was also possible to show that the subscores had little added value for these data by computing simple correlations. As an example, consider the 75-item paraprofessional test (i.e., Test H) split into two parallel, smaller subforms of 38 and 37 items each, known as H1 and H2, respectively. The subform H1 consisted of the 38 odd-numbered items from Test H, whereas the subform H2 consisted of the 37 even-numbered items from Test H. Their composition (in terms

of the number of items corresponding to each subscore) and difficulty were very similar to that of H, so H1 and H2 were treated as parallel test forms. The subscores of form H1 (i.e., reading, math, and writing) were presented as Reading 1, Math 1, and Writing 1 in H1 and Reading 2, Math 2, and Writing 2 in H2 (see Figure 3, which shows this partitioning).

Since the two subforms were created to be parallel, if the subscores are reliable enough, it is reasonable to expect that Reading 1 (the reading subscore on H1) will correlate higher with Reading 2 (the reading subscore in H2) than it does with H2 alone (the total test). However, if Reading 1 correlates higher with H2 than Reading 2, then it would suggest that the total score (H2), not the reading subscore (Reading 2) in a parallel form, is a better predictor of Reading 1. For the data set we analyzed, the correlation between Reading 1 and H2 was 0.80 and that between Reading 1 and Reading 2 was 0.77, suggesting that H2 was a better predictor (than Reading 2) of Reading 1. Similar results were observed for the mathematics and writing subscores.

Other information such as the factor structure of a test (i.e., from a factor analysis) or correlations between subscores can also be used to determine whether subscores provide any additional information than what is already provided by the total score. As an example, consider Test H again. The largest eigenvalue computed from the 3×3 correlation matrix was 2.50, which was much larger than the remaining two eigenvalues of 0.27 and 0.24, suggesting the presence of a single dominant factor. Also, the correlations (after correcting for attenuation) of 0.88 between the reading and math subscore, 0.91 between the reading and writing subscores, and 0.88 between the math and writing subscores were fairly high, suggesting that the subscores are very similar and therefore reporting separate subscores may not be necessary if the total score is already reported.

Figures 1 and 2 also show that the PRMSE was larger for the Haberman augmented score when compared to the TS-OS for all the tests. The PRMSE for the Haberman augmented score is also larger or similar to TS-OTS for all the tests.

The results from the institutional level analyses depended on the institution size and therefore varied according to the institution size. As seen in Figures 4–11, the PRMSE for the smallest institution size condition ($N = 10$) was similar or higher for the TS-OTS as compared to the TS-OS. Similarly, the PRMSE of the Haberman augmented subscore was higher than the TS-OS and similar or higher than the TS-OTS for that sample size condition for all the tests.

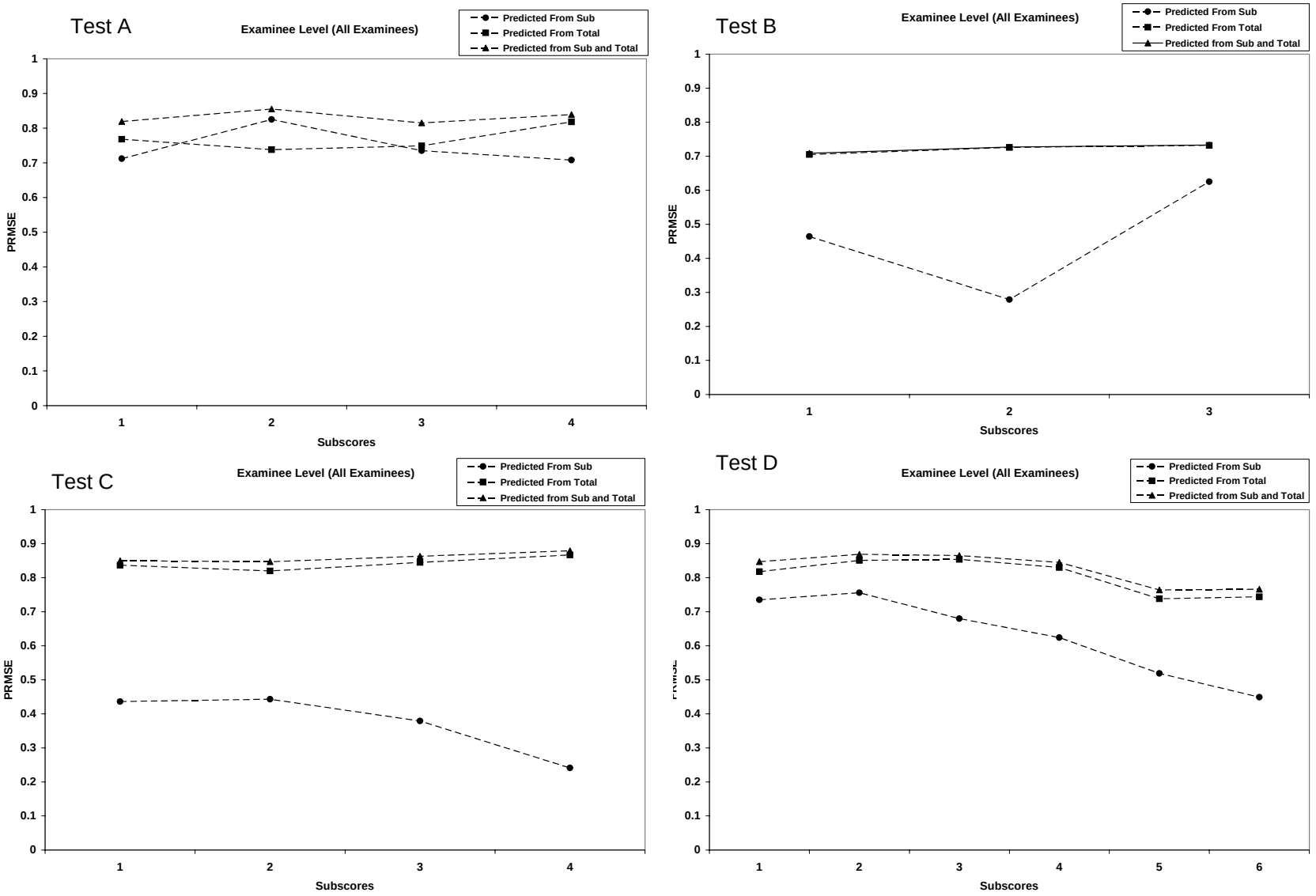

Figure 1. **Proportional reduction of mean squared error (PRMSE) for Tests A–D (examinee level analysis).**

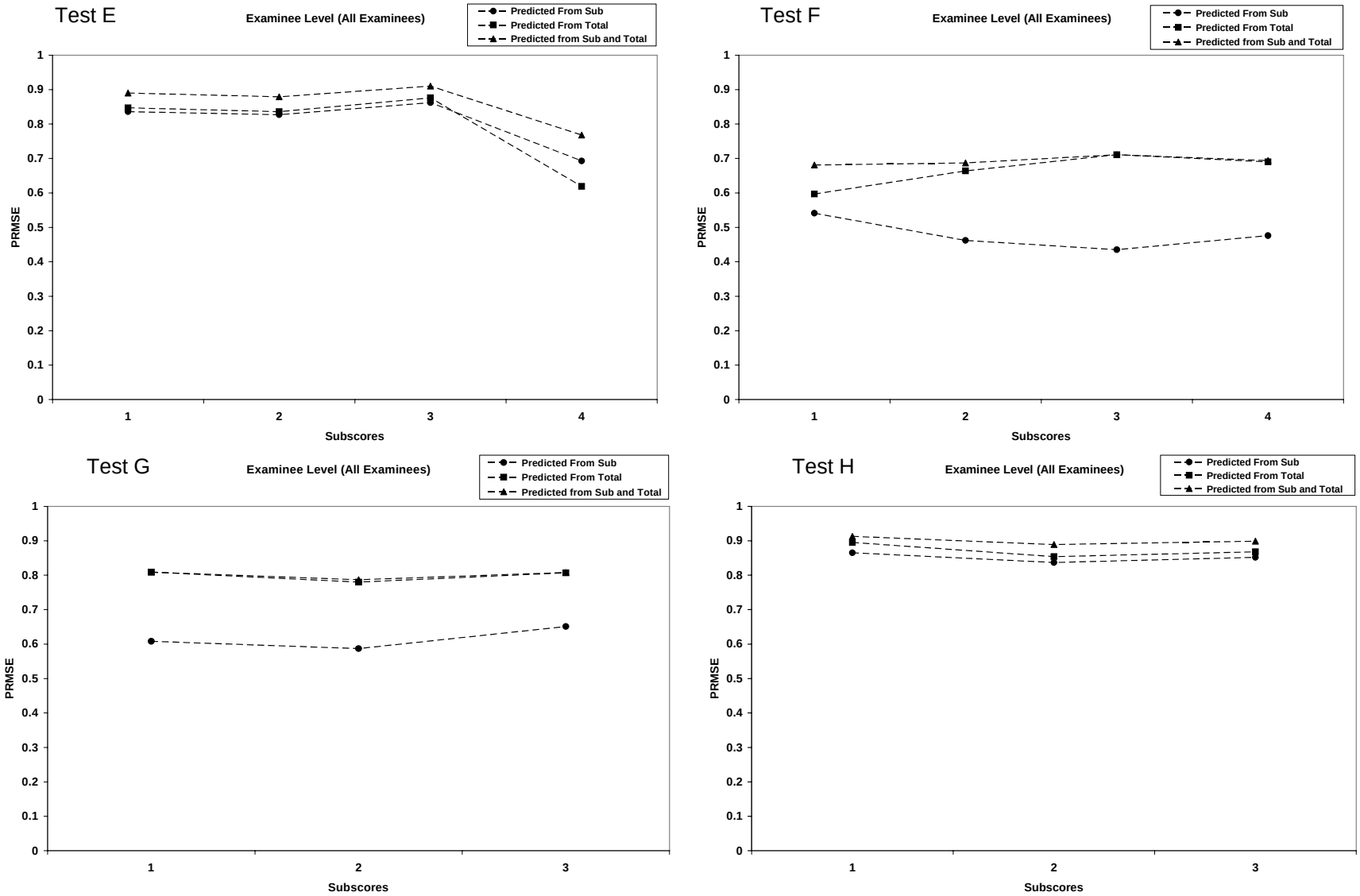

Figure 2. **Proportional reduction of mean squared error (PRMSE) for Tests E–H (examinee level analysis).**

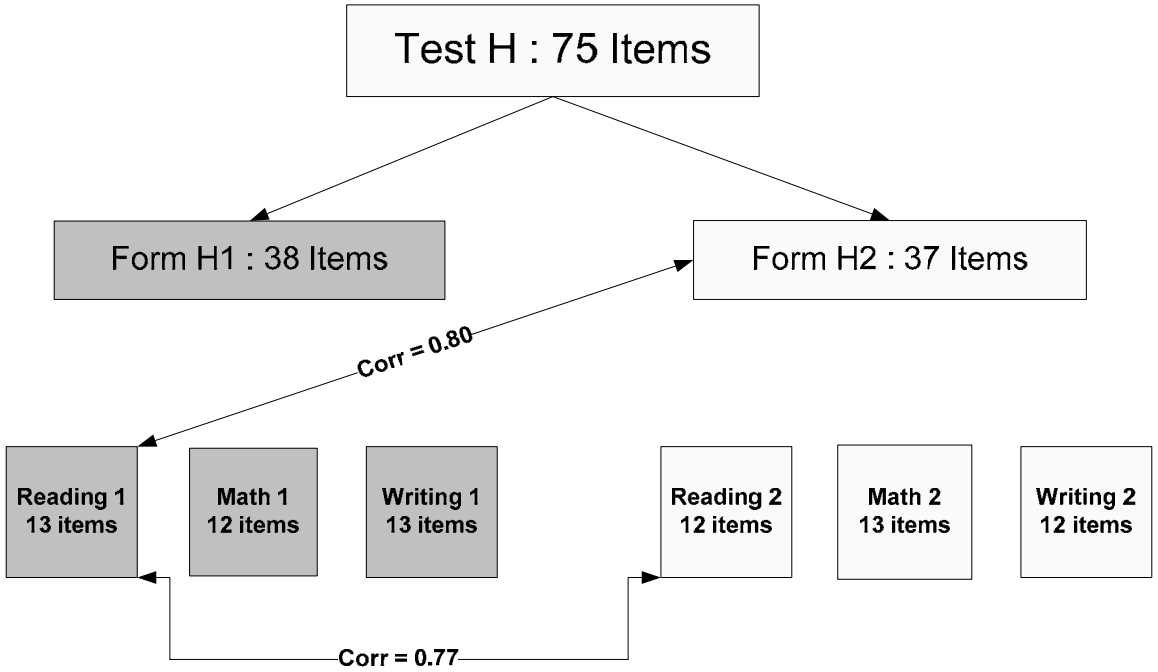

Figure 3. **Construction of two parallel forms from the total Test X.**

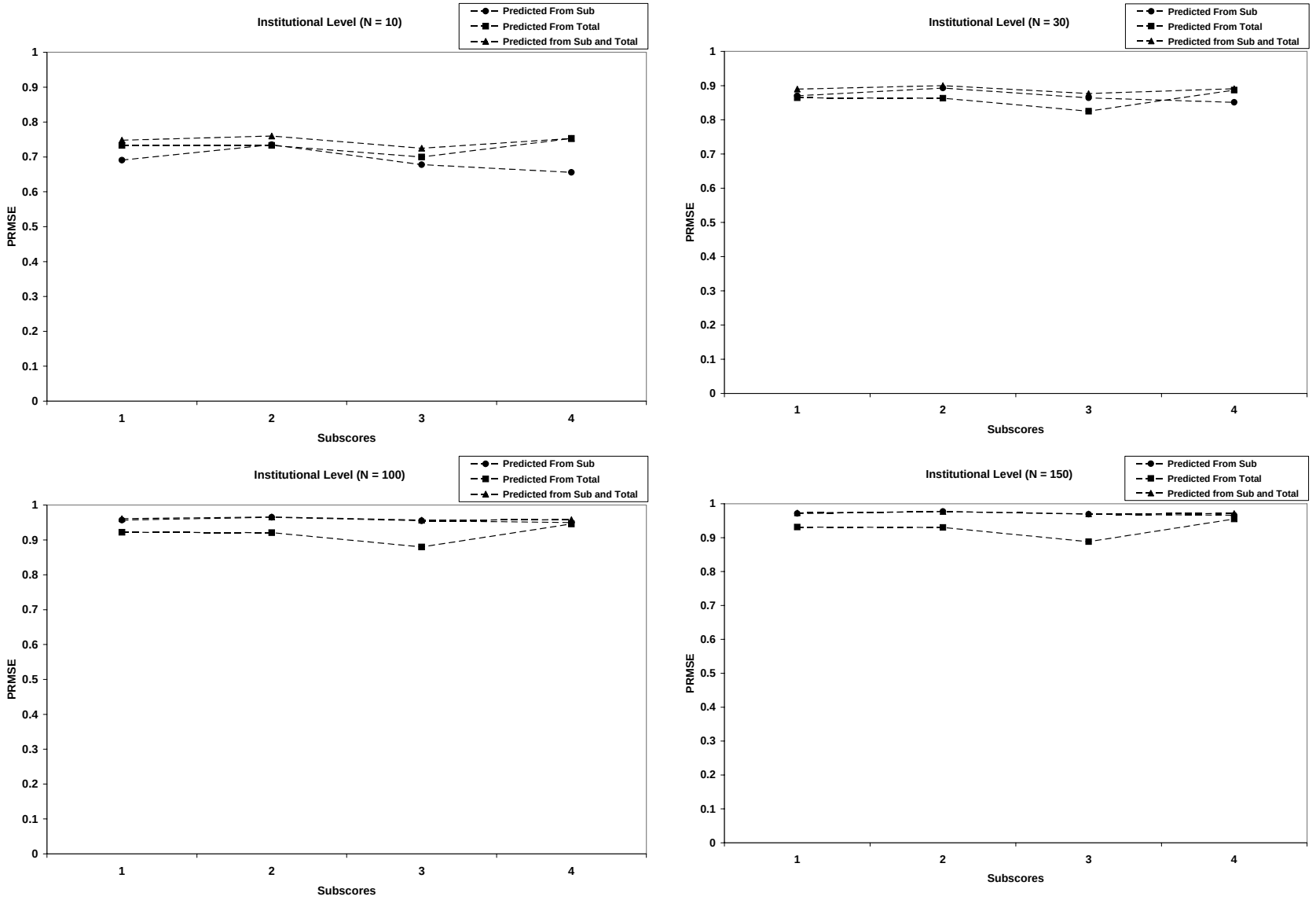

Figure 4. **Proportional reduction of mean squared error (PRMSE) for four subscores for Test A (institutional level analysis).**

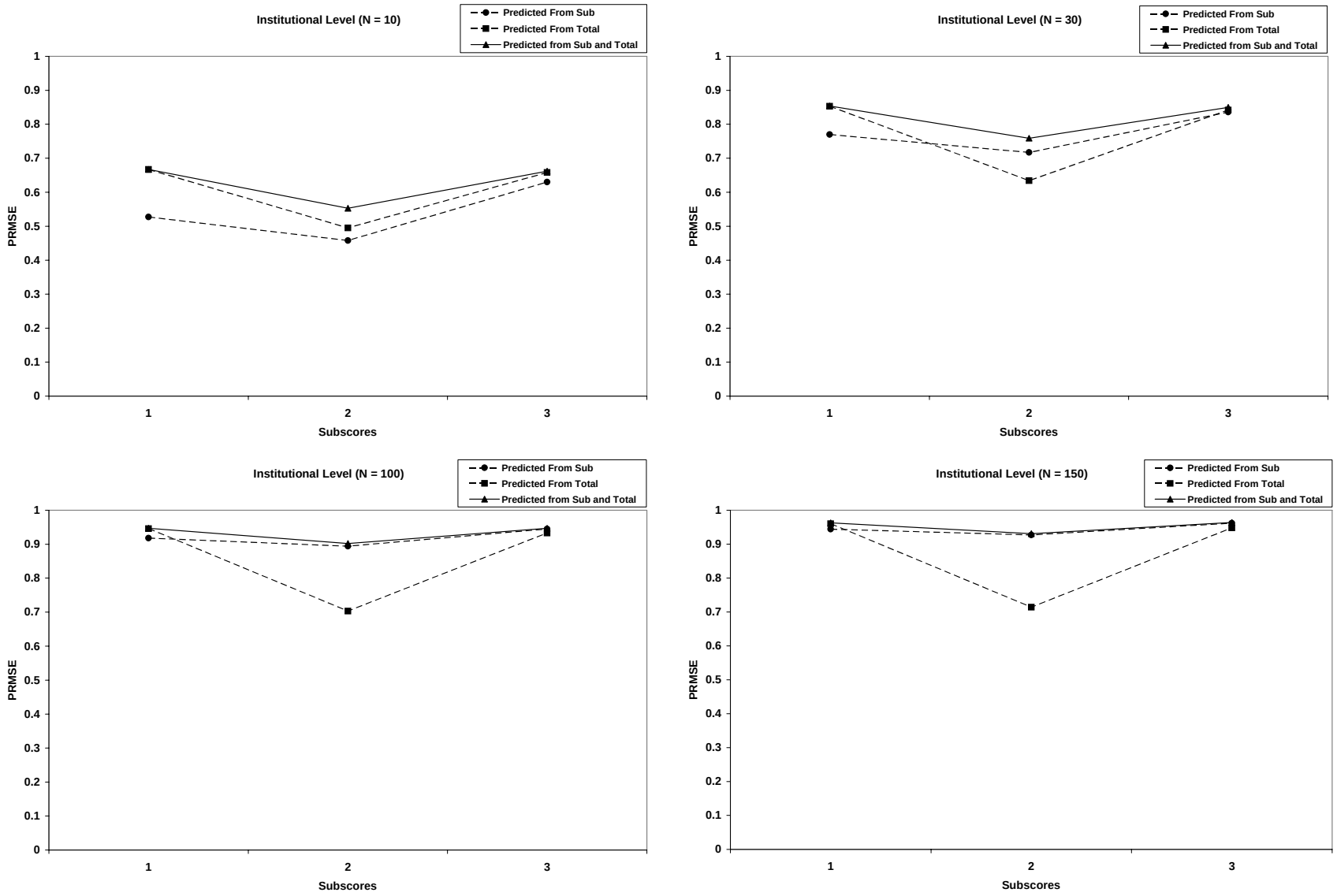

Figure 5. **Proportional reduction of mean squared error (PRMSE) for three subscores for Test B (institutional level analysis).**

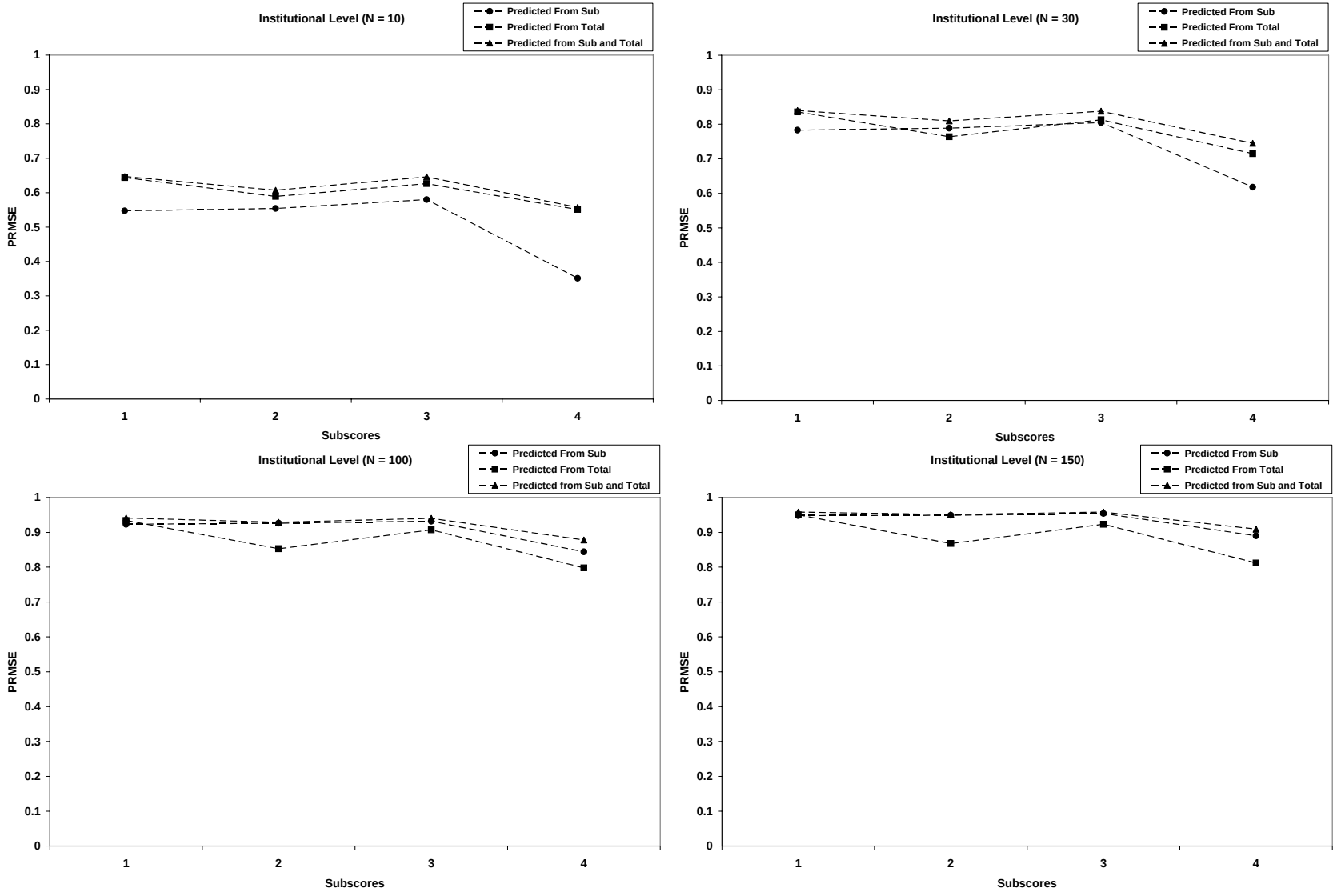

Figure 6. **Proportional reduction of mean squared error (PRMSE) for four subscores for Test C (institutional level analysis).**

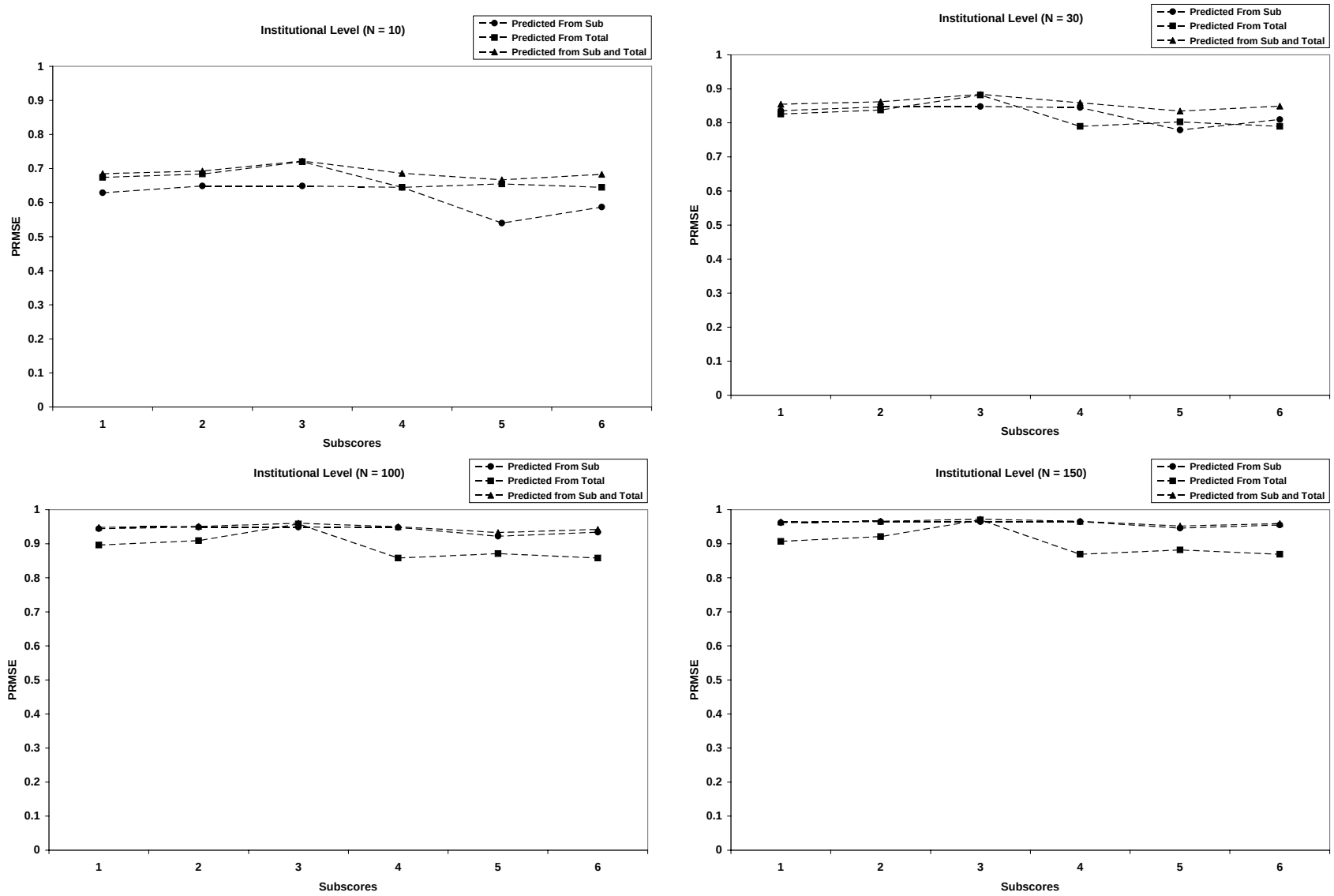

Figure 7. **Proportional reduction of mean squared error (PRMSE) for six subscores for Test D (institutional level analysis).**

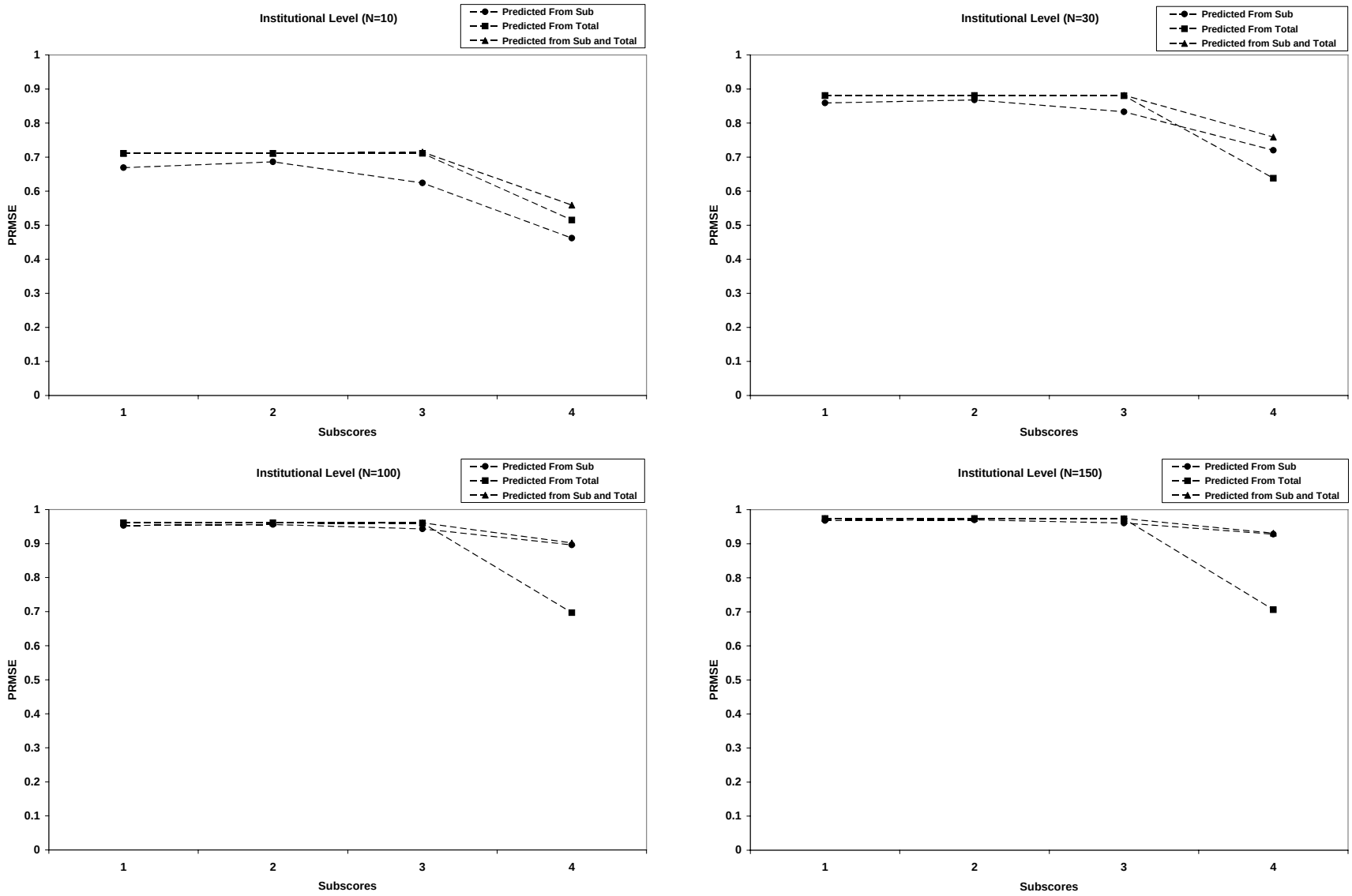

Figure 8. **Proportional reduction of mean squared error (PRMSE) for four subscores for Test E (institutional level analysis).**

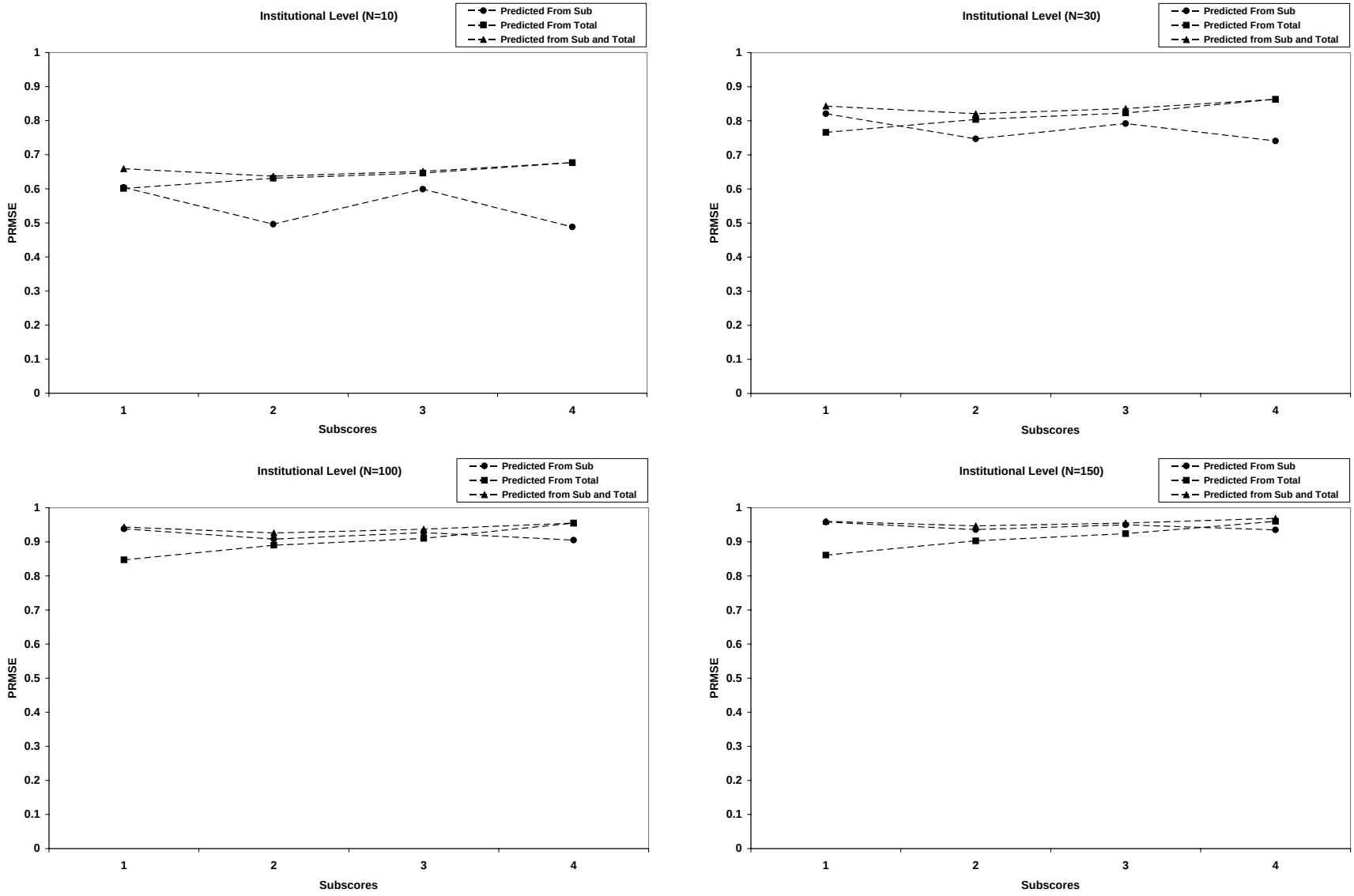

Figure 9. **Proportional reduction of mean squared error (PRMSE) for four subscores for Test F (institutional level analysis).**

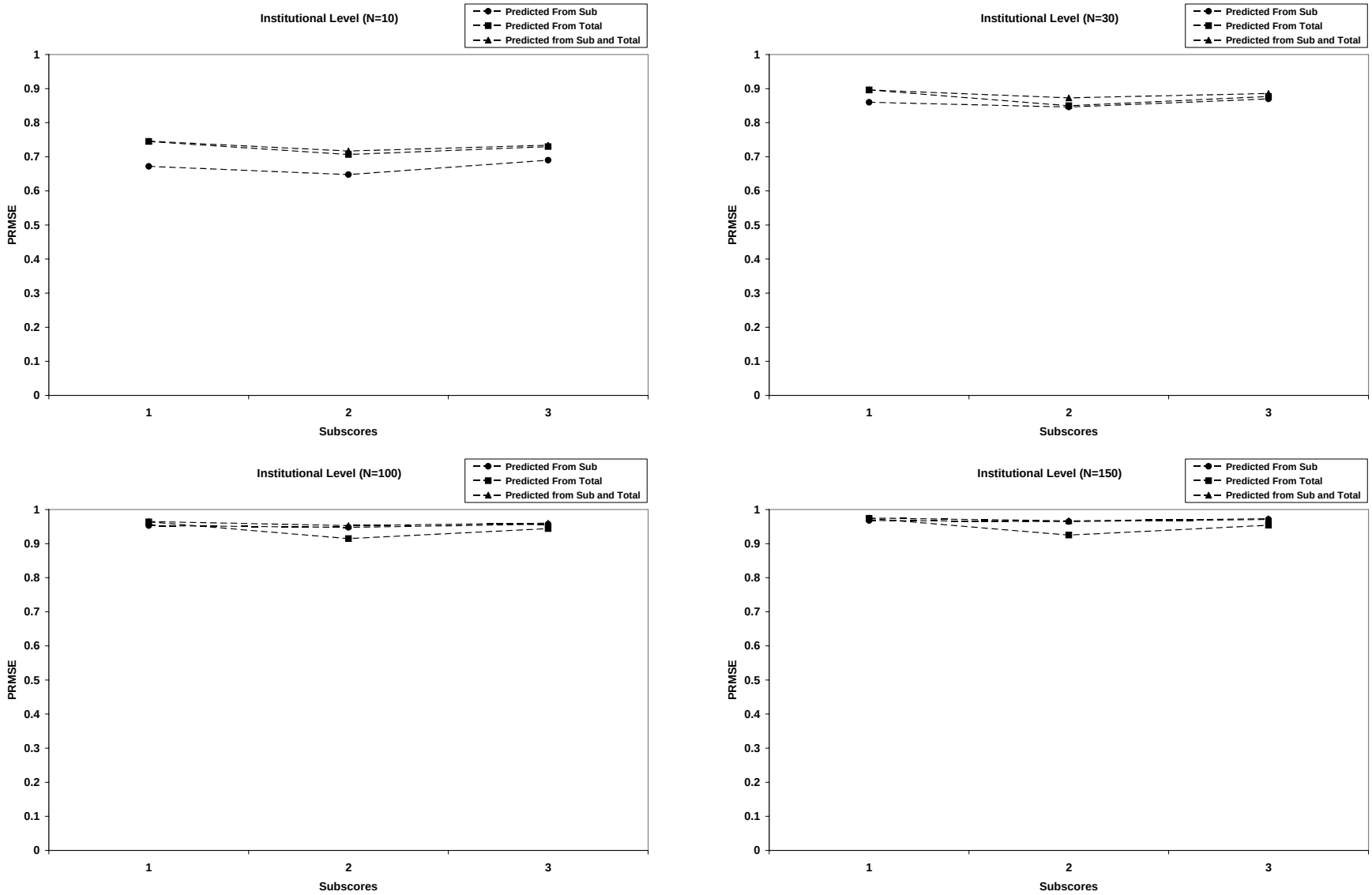

Figure 10. **Proportional reduction of mean squared error (PRMSE) for three subscores for Test G (institutional level analysis).**

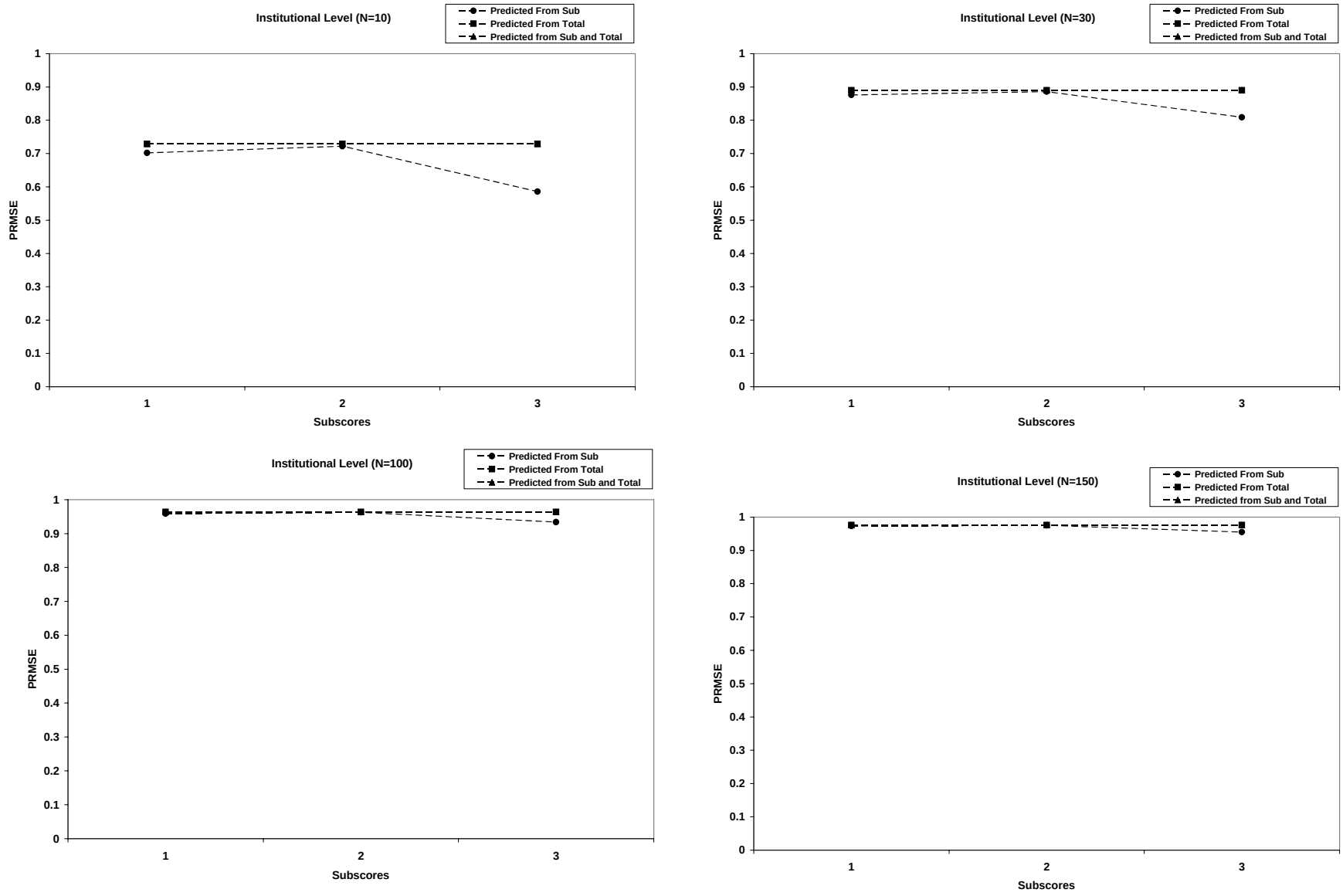

Figure 11. **Proportional reduction of mean squared error (PRMSE) for four subscores for Test H (institutional level analysis).**

For the moderate institution size conditions (N = 30, 100, and 150), the differences in the PRMSEs for TS-OS and TS-OTS became smaller, and for some sample size conditions and subscores, the PRMSEs for the TS-OS actually became slightly larger than the PRMSEs for the TS-OTS, suggesting that reporting the observed subscore for these cases may be reasonable although redundant in many cases (especially in cases where the PRMSEs for the TS-OS is only slightly larger than the TS-OTS). For example, if the PRMSE for the TS-OS versus that of the TS-OTS are 0.87 and 0.85, respectively, reporting the observed subscore may not be harmful but would not provide much added information over the observed total score. Therefore, if the total score is reported to test takers and institutions, reporting subscores may be redundant in these circumstances.

Finally, as seen in Figures 4–11, the PRMSE for the Haberman augmented subscore was either similar or higher than the PRMSE of the TS-OS or TS-OTS for all sample sizes used in this study, suggesting that if one option were used to report subscores, then the Haberman augmented subscore would be a reasonable choice.

Conclusions and Recommendations

The purpose of this study was to investigate whether the reporting of subscores provides any additional information other than what is provided in the total score for several content tests used for teacher certification. The study also compared different classical test theory–based methods that can be used to predict subscores that are more trustworthy or reliable than observed subscores. This study is important because if testing organizations need to report subscores (in their effort to be responsive to the needs of examinees and institutions), then reporting a predicted subscore that results in the highest reliability is preferable. The results of the study led to some tentative conclusions and recommendations that are described below.

For the tests used in this study, reporting subscores at the examinee level might be unnecessary because the subscores did not provide much additional information over what had already been provided by the total score[1] (i.e., the PRMSE for the TS-OTS is higher than that of TS-OS). It should also be noted that a higher PRMSE for the TS-OTS compared to that of the TS-OS did not automatically guarantee that a predicted subscore based on the observed total score was worth reporting. For example, the three subscores for Test B predicted from the observed subscore had PRMSEs of 0.464, 0.279, and 0.732, respectively. Although using the observed total score to predict the three subscores improved the PRMSEs to 0.705, 0.726, and

0.732, these PRMSEs might be still be considered too low to justify subscore reporting (see Table 2).

For the tests used in this study, reporting subscores at the institutional level might not have been harmful (although it might have been redundant) for institution sizes greater than 30 because the PRMSEs for the TS-OS and the TS-OTS were quite similar for sample sizes equal to or greater than 30. Therefore, some members of the educational community such as teachers and policy makers may argue that since reporting subscores in these cases is not harmful, then it may be reasonable to report subscores because subscores do have a perceived usefulness for users.

As evident in this study, results differed depending on the type of test (i.e., results from the examinee and institutional level analyses) and sample size (i.e., results from the institutional level analyses), and it was difficult to predict beforehand which test or sample size would benefit more from subscore reporting. Therefore tests should be evaluated on a case-by-case basis to decide whether subscore reports are useful. Furthermore, although it seems clear that subscores based on relatively few items are likely to provide much value, it does not mean that subscores based on many items or many tasks will be useful. For example, with Test E (see examinee level analysis in Table 5), the first three subscores did not provide any additional information over what is provided by the total score even though these subscores were based on a relatively large number of items (about 30 or more items in each section). However, for the fourth subscore (based on 21 items), the subscore seemed to provide more information than what was provided by the total score. Thus, this result further reinforced the recommendation that tests must be evaluated on a case-by-case basis to decide whether subscore reports are useful.

Finally, as seen from the results of this study, the Haberman augmented subscore performed equally well or outperformed the other methods at both the examinee and institutional level analyses. Therefore, if subscores are to be reported, then the Haberman augmentation subscore is recommended over the subscore based on only the observed subscore or only the observed total score. Since the Haberman augmented subscore method resulted in either similar or more trustworthy (larger PRMSE) subscores for all tests and all sample sizes, it seems to be the better option when considering subscore reporting.

An Additional Contender for Predicting Subscores

When predicting subscores, it is also possible to use at least one additional predictor for both individual level and institutional level subscores: an augmented score suggested by Wainer

et al. (2001). Under the augmented subscore approach of Wainer et al., each of the subscores were regressed on all of the other subscores. Weights were assigned to each of the subscores (e.g., reading, writing, and mathematics), and an examinee's score on a particular sub- area (e.g., reading) was a function of the examinee's ability in reading and also in writing and mathematics. The subscores that had the strongest correlation with the reading subscore had larger weights and thus provided more information for augmenting the reading subscore.

For the data employed in this study, this method produced higher PRMSEs (at the examinee level) than the TS-OS or the TS-OTS. But the results did not show any gain over the Haberman augmented score (i.e., the augmented subscore of Wainer et al., 2001, and the Haberman augmented score produced almost identical PRMSEs). For example, for Test A, the PRMSEs for the Haberman augmented scores were 0.819, 0.855, 0.815, and 0.839, and the PRMSEs for the Wainer et al. augmented scores were 0.819, 0.857, 0.822, and 0.841, respectively for the four subscores, indicating that the two methods produced very similar results. Therefore either method may be used for examinee level subscore reporting, although it should be noted that the Haberman augmented score is computationally less intensive and therefore may be preferable for operational subscore reporting. Since the results were very similar for these two methods at the examinee level, it did not seem necessary to investigate their comparability at the institutional level. It was therefore assumed that both methods will produce similar results at the institutional level.

Limitations and Future Research

The main idea followed in this study and suggested in earlier studies (Haberman, 2005; Haberman et.al., 2006; Sinharay et. al., 2007) is that subscores should be considered for reporting only when the true subscore can be predicted better by the observed subscore than by the observed total score. Although this idea may be appropriate if the purpose of testing is only to discriminate among the examinees at a single point in time, it may not be appropriate if the purpose of testing includes the measurement of examinees' growth in the constructs being measured (Samuel A. Livingston, personal communication, December 2007). For example, consider a test that includes two sections: a math section and a verbal section. When the test is administered at the beginning of the school year, the examinees' subscores (math and verbal) may correlate highly in relation to their reliabilities, so that, by the above mentioned argument, no value is gained in reporting those two subscores. Now suppose the examinees receive extensive math training. When the test is

administered again at the end of the school year, substantial gains in the examinees' math scores but much smaller gains in their verbal scores is likely to be observed. And yet, in this posttesting, the two subscores may again correlate highly, and hence, according to the method employed in this study, there may be no value in reporting the two subscores. Therefore it is highly essential that the purpose for using subscores be clearly defined before implementing the methods used in this study to evaluate the usefulness of subscores.

Subscores, if defined in raw score units, are not directly comparable across different forms of the test. This finding is also true of augmented subscores. Therefore an important issue with reporting subscores, both for individuals and institutions, is that subscores have to be equated and/or scaled for comparability. In typical cases, equating is possible for the total score but may be challenging for subscores. For example, if a common item design is used to equate the total score, only a few of the anchor test items will correspond to a particular subscore, so that anchor test equating of the subscore will probably not be feasible. Scaling the subscores to the total score may be a possibility considering that the correlation between the subscores and the total scores is usually high. However, in cases where the correlation between the subscores and the total scores is not high, the scaling results may be questionable. Furthermore, when a subscore has few possible score points, appropriate scaling also may not be feasible. In addition, the possibility exists that a scaling that may be adequate for individuals may be far from satisfactory if applied to institutions. Therefore, future research for developing methods to effectively equate subscores needs to be conducted.

Finally, with augmented approaches for estimating subscores, the examinee's observed subscore is regressed on his or her scores on other parts of the test (e.g., the total test score or scores on other subscales). Although this approach is useful in obtaining a more stable estimate of examinee's subscore, there may be cases (especially where the subscores are highly correlated to the total test score) where this approach may hide differences between subscores by forcing the different subscores of examinees to appear very similar to each other. For example, if the different subscores on a test are highly correlated to the total score, then the Haberman augmentation approach would assign a large weight to the total score and a small weight to the observed subscore, thereby making the augmented scores for different subscales appear to be very similar to each other. For a contradictory example (i.e., where the augmented subscores would not hide important differences) see Sinharay and Haberman (in press, pp. 11–12).

References

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.

Haberman, S. J. (2005). *When can subscores have value?* (ETS Research Rep. No. RR-05-08). Princeton, NJ: ETS.

Haberman, S. J., Sinharay, S., & Puhan, G. (2006). *Subscores for institutions* (ETS Research Rep. No. RR-06-13). Princeton, NJ: ETS.

Haladyna, T. M., & Kramer, G. A. (2004). The validity of subscores for a credentialing test. *Evaluation and the Health Professions, 24*(7), 349–368.

Kuder, G. F., & Richardson, M. W. (1937). The theory of estimation of test reliability. *Psychometrika, 2,* 151-160.

Monaghan, W. (2006). *The facts about subscores* (ETS R&D Connections No. RDC-04). Princeton, NJ: ETS.

Sinharay, S., Haberman, S. J., & Puhan, G. (2007). Subscores based on classical test theory: To report or not to report. *Educational Measurement: Issues and Practice, 26*(4), 21–28.

Sinharay, S., & Haberman, S. J. (in press). *Reporting subscores: A survey.* Princeton, NJ: ETS.

Tate, R. L. (2004). Implications of multidimensionality for total score and subscore performance. *Applied Measurement in Education, 17*(2), 89–112.

Wainer, H., Vevea, J. L., Camacho, F., Reeve, B. B., Rosa, K., Nelson, L., et al. (2001). Augmented scores—"borrowing strength" to compute scores based on small numbers of items. In D. Thissen & H. Wainer (Eds.), *Test scoring* (pp. 343–387). Hillsdale, NJ: Lawrence Erlbaum Associates.

Notes

[1] It should be acknowledged that there are competing demands when test scores serve as entry requirements to a career. On the one hand, subscore information should be psychometrically sound; on the other hand, examinees who fail should be provided some guidance about areas of weakness. Although the ideal solution would be to have subscores based on a sufficient number of items, this solution is not always feasible given constraints about test length, test cost, client mandated limitations, etc.

Appendix

Computational Details for Estimating the True Subscores Using Three Different Predictors

Here, the approach used in the study to evaluate the trustworthiness of subscores at both the examinee and institutional levels is described. As mentioned earlier, the true subscore can be predicted from the following:

1. The observed subscore s itself.

2. The predictor based on a regression of the true subscore on the observed subscore (which is Kelley's formula applied to the subscore). The predictor, obtained after some algebra (Haberman, 2005), is given by

$$s_s = E(s) + \rho^2(s_t, s)[s - E(s)], \tag{A1}$$

where $\rho^2(s_t, s)$ is the reliability of the subscore. In an application of the above formula, $E(s)$ is estimated by the average observed subscore over all the examinees and $\rho^2(s_t, s)$ is estimated by the KR-20 approach (Kuder & Richardson, 1937).

3. The predictor of the true subscore based on the observed total score is a regression of the true subscore on the observed total score, which, after some algebra, is given by

$$s_x = E(s) + \rho(s_t, x)[\sigma(s_t)/\sigma(x)][x - E(x)], \tag{A2}$$

where $\rho(s_t, x)$ is the correlation between the true subscore s_t and the observed total score x. In an application of the above formula, $\sigma(s_t) = \sigma(s)\sqrt{\rho^2(s_t, s)}$ is computed using the values of the observed variance of the subscore and estimated reliability, $\sigma(x)$ is the observed standard deviation (SD) of the total score, and $\rho(s_t, x)$ is computed using the formula

$$\rho(s_t, x) = \sqrt{\rho^2(s_t, x_t)\rho^2(x_t, x)}, \tag{A3}$$

where $\rho^2(x_t, x)$, the total score reliability, is computed using the KR-20 approach (Kuder & Richardson, 1937), and the computation of $\rho^2(s_t, x_t)$ is described in Haberman (2005).

4. Finally, the predictor of the true subscore s_t based on the linear regression s_{sx} of s_t on both the observed subscore s and the observed total score x. The regression is given by

$$s_{sx} = E(s) + \beta[s - E(s)] + \gamma[x - E(x)], \quad \text{(A4)}$$

where β and γ are constants. For details about computation of β and γ, see Haberman (2005).

For subscores to provide additional information than the total score (which is already reported to test takers and institutions), at least one among s, s_s, and s_{sx} has to be a better predictor of the true subscore than s_x. A natural question is what criterion should we use to judge that a predictor is better than another. An answer to the question is discussed below.

Criterion for Comparing Predictors of True Subscore

Haberman (2005) suggested the use of mean squared error (MSE) of a predictor as the criterion in this situation. The MSE is a popular criterion for comparing the performance of competing estimators. The MSE for a predictor in this context measures the average squared error in predicting the true subscore by the predictor. Practically, larger MSE would lead to more error in instructional and remedial decisions.

For the predictor s above, the MSE is

$$E(s - s_t)^2 = E(s_e^2) = \sigma^2(s_e), \quad \text{(A5)}$$

that is, the subscore error variance.

For the predictor s_s, the MSE can be shown to be

$$E(s_s - s_t)^2 = \sigma^2(s_t)[1 - \rho^2(s_t, s)], \quad \text{(A6)}$$

for the predictor s_x, the MSE can be shown to be

$$E(s_x - s_t)^2 = \sigma^2(s_t)[1 - \rho^2(s_t, x)]\cdot, \quad \text{(A7)}$$

and for the predictor s_{sx}, the MSE can be shown to be

$$E(s_t - s_{sx})^2 = \sigma^2(s_t)[1 - \rho^2(s_t, s) - \tau^2[1 - \rho^2(s, x)]]\cdot \quad \text{(A8)}$$

For details about computation of τ, see Haberman (2005).

Haberman (2005) also suggested a measure based on MSEs that is conceptually very close to the test reliability. Consider the trivial predictor $E(s)$ that predicts the true subscore of every examinee by the same number, the average subscore over all examinees. The MSE for this trivial predictor can be shown to be $\sigma^2(s_t)$. Now calculate the PRMSE for the four predictors, s, s_s, s_x, s_{sx}, and compare to the MSE for the trivial predictor. For example, for the predictor s_s, the PRMSE is given by

$$\frac{\text{MSE for the trivial predictor} - \text{MSE for } s_s}{\text{MSE for the trivial predictor}}, \qquad \text{(A9)}$$

which is equal to $\rho^2(s_t, s)$, the subscore reliability. Thus, when a true subscore is predicted by its regression on the observed subscore, the PRMSE criterion is identical to the concept of test reliability and hence the criterion should be appealing to the psychometric community as a suitable criterion. Note that smaller MSE is equivalent to larger PRMSE and hence a predictor with a larger PRMSE is preferable to one with a smaller PRMSE.

The PRMSE for the predictor s can be shown to be equal to $2 - 1/\rho^2(s_t, s)$, which can be shown to be always less than or equal to the PRMSE for the predictor s_s. Hence we will no longer consider the predictor s in this paper. The PRMSE for the predictor s_x can be shown to be equal to $\rho^2(s_t, x)$. The PRMSE for the predictor s_{sx} can be shown to be equal to $\rho^2(s_t, s) + \tau^2[1 - \rho^2(s, x)]$.

The above discussion implies that for subscores to have added value, the PRMSE has to be larger for s_s than for s_x; for example, $\rho^2(s_t, s)$ has to be larger than $\rho^2(s_t, x)$. This is justifiable from the viewpoint of correlation as well; for the subscores to have added value, it is reasonable to expect that the correlation between true subscore and observed subscore should be larger than the correlation between true subscore and observed total score.

Institutional Level Analysis

At the institutional level, the above analyses can be modified by decomposition of total scores and subscores into institutional and individual components. Thus subscore s has the decomposition $s = s_I + s_e$, where s_I, the component for the institution, is the same for each examinee in an institution and has mean $E(s)$ and variance $\sigma^2(s_I) > 0$. The component s_e

above is an examinee-specific effect that should not be confused with a typical error term in classical test theory. The score x has the decomposition $x = x_I + x_e$, where x_I, the component for the institution, is the same for each examinee in an institution and has mean $E(x)$ and variance $\sigma^2(x_I) > 0$. The residual examinee subscore $s_e = s - s_I$ within institution has mean 0, variance $\sigma^2(s_e) > 0$, and is uncorrelated with the institutional means s_I and x_I. The residual examinee total score $x_e = x - x_I$ within institution has mean 0, variance $\sigma^2(x_e) > 0$, and is uncorrelated with s_I and x_I. Denote the average observed subscore and the average observed total score for an institution as $\bar{s}$ and $\bar{x}$ respectively. We use an approach similar to that used for examinee level subscores to determine whether institutional level subscores have added value.

The predictor of institutional level true subscores based on the observed subscores is a regression of the institution's true subscore on the institution's average observed subscore and is given by

$$s_{Is} = E(s) + \rho^2(s_I, \bar{s})[\bar{s} - E(s)]\cdot \tag{A10}$$

The predictor of institutional level true subscores based on the observed total scores is a regression of the institution's true subscore on the institution's average observed total score and is given by

$$s_{Ix} = E(s) + \rho(s_I, \bar{x})[\sigma(s_I)/\sigma(\bar{x})][\bar{x} - E(x)]\cdot \tag{A11}$$

The predictor of institutional level true subscores based on the observed subscore and the observed total scores is a regression of the institution's true subscore on the institution's average observed subscore and average observed total score and is given by

$$s_{Isx} = E(s) + \beta_I[\bar{s} - E(s)] + \gamma_I[\bar{x} - E(x)], \tag{A12}$$

where β_I and γ_I are constants and can be computed as described in Haberman et al. (2006).

As with examinee level subscores, we will use the MSE criterion to compare the performance of the predictors. Haberman et al. (2006) showed that the MSE for s_{Is} is $\sigma^2(s_I)[1-\rho^2(s_I,\bar{s})]$ whereas that for s_{Ix} is given by $\sigma^2(s_I)[1-\rho^2(s_I,\bar{x})]$ and for s_{Isx} is given by $\sigma^2(s_I)[1-\rho^2(s_I,\bar{s})-\tau_I{}^2[1-\rho^2(\bar{s},\bar{x})]$ for a constant $\tau_I{}^2$.

The trivial predictor of the true subscore for any institution is a constant $E(s)$. The corresponding MSE is $\sigma^2(s_I)$. Hence, relative to the use of $E(s)$, the PRMSE for s_{Is} is the institutional subscore reliability $\rho^2(s_I, \bar{s})$ while the PRMSE for s_{Ix} is $\rho^2(s_I, \bar{x})$ and the PRMSE for s_{Isx} is $\rho^2(s_I, \bar{s}) + \tau^2{}_I[1 - \rho^2(\bar{s}, \bar{x})]$.

Hence, for the institutional level subscores to have added value, the PRMSE for s_{Is} has to be larger than that for s_{Ix}; for example, $\rho^2(s_I, \bar{s})$ has to be larger than $\rho^2(s_I, \bar{x})$. Haberman et al. (2006) discussed in detail the computation of $\rho^2(s_I, \bar{s})$ and $\rho^2(s_I, \bar{x})$, which depends on n, the institution size, using the multivariate analysis of variance technique.