ACT Research Report Series 98-5

Estimating Average Domain Scores

Mary Pommerich

W. Alan Nicewander

ACT

August 1998

For additional copies write:
ACT Research Report Series
PO Box 168
Iowa City, Iowa 52243-0168

Estimating Average Domain Scores

Mary Pommerich
W. Alan Nicewander

Abstract

A simulation study was performed to determine whether a group's average percent correct in a content domain could be accurately estimated for groups taking a single test form and not the entire domain of items. Six Item Response Theory-based domain score estimation methods were evaluated, under conditions of few items per content area per form taken, small domains, and small group sizes. The methods used item responses to a single form taken to estimate examinee or group ability; domain scores were then computed using the ability estimates and domain item characteristics. The IRT-based domain score estimates typically showed greater accuracy and greater consistency across forms taken than observed performance on the form taken. For the smallest group size and least number of items taken, the accuracy of most IRT-based estimates was questionable; however, a procedure that operates on an estimated distribution of group ability showed promise under most conditions.

Acknowledgements

We would like to thank Brad Hanson for his technical assistance and comments on the paper, Mark Reckase and Tim Miller for their many helpful suggestions on the project, and Matt Schulz and Jim Sconing for helpful comments on the paper.

Estimating Average Domain Scores

There has been some recent interest in testing circles in reporting domain-referenced scores, or scores linked to performance on a domain of items representing the skills and knowledge required for mastery of a content area. Several advantages of domain score estimates over traditional test scores are discussed in Bock, Thissen, and Zimowski (1997), who demonstrate Item Response Theory (IRT) estimation of domain scores for individuals. A primary advantage of the proposed domain scores over traditional scaled scores is the simplicity of score interpretation—a score on the domain may be expressed simply as the percentage of total possible points that are achieved. Also, the release of the domain or a representative subset of the domain would provide a more comprehensive item set from which content area strengths and weaknesses can be determined and the nature of the content areas inferred than would release of items taken alone. Thus, domain scores offer the possibility of facilitating interpretation and evaluation of performance, provided the domain has been well defined.

As part of its redesign, the National Assessment for Educational Progress (NAEP), proposes the use of "market basket" or domain score reporting based on a collection of items (the market basket) that would be released to the public (Forsyth, Hambleton, Linn, Mislevy, & Yen, 1996; National Assessment Governing Board, 1996). Different sized market baskets are proposed, including a market basket that would constitute an entire domain of items; namely, a very large pool of items operationally defining skill in the domain. Reported scores would indicate performance in the domain, although individuals would typically not take all items in the domain. Another use of domain scores that would be applicable to NAEP is discussed in Schulz, Kolen, and Nicewander (in press), who present a rationale for defining achievement levels using IRT-estimated domain scores.

The concept of a domain-referenced test stems from criterion-referenced testing, where items are associated with a well-defined behavior domain (Popham, 1978). Theoretically, a domain may consist of any clearly specified set of items (Hively, 1974). However, for a well-defined domain, both the essential attributes of the content which the student is expected to acquire and the behavior through which he or she is expected to demonstrate such acquisition should be carefully described (Baker, 1974). Nitko (1984) further clarifies that a domain is well-defined if it is clear which categories of performance or which kinds of tasks are and are not potential test items, while a domain is ill-defined if it is defined only in terms of the particular items on a single test form.

Focus on a single test form may mask the idea that items specific to that form represent larger domains of content. Observed performance in a content area on a given form may not necessarily reflect what performance would be on a broader set of items representing the content domain. Linking performance on a test to a well-defined domain of skills and knowledge shifts focus from individual items on the form taken to the content domain as a whole. This shift can provide more comprehensive information about performance, particularly if the coverage of the content area on a particular form is limited to a small number of items.

The research reported in this paper is based on a large-scale operational testing program that reports, for participating schools, an observed average percent correct on the test form taken, for items within a content area that are sampled from a content domain. The content area level is a finer level than the scale scores that are typically reported for the operational test; school performance is summarized at this level and reported with normative information to aid in the diagnosis of particular strengths and weaknesses. Because of the fine level of specification, there may be as few as five items for some content areas on a given form. In addition, the test forms are not equated or balanced at the content area level, so that the difficulty of the content areas may vary across forms, and the observed performance on a form may be dependent upon the difficulty of the form.

At the examinee level, the observed percent correct for a single form is an unbiased domain score estimate when based on a random sample of items from the domain, but is highly unreliable when the number of items on which the estimate is based is small (Hambleton, Swaminathan, Algina, & Coulson, 1978). Research suggests that computing an observed percent correct for test lengths less than 20 items could lead to unreasonable domain score estimates (Haladyna & Roid, 1983). Domain score estimates could be problematic for either examinees, or for groups, when examinee performance is aggregated at a group level. Thus, there is potential for misinterpretation at the content area level if observed examinee or group performance on a form taken is used to make inferences beyond performance on the items contained on that form. An alternative would be to supplement observed performance on the form taken with similar information about performance in the content domain from which the test items are sampled. For groups, the supplementary information would consist of the group average percent correct for each content domain.

In most testing situations, the domain performance cannot be measured outright, because the domain is typically not administered. This research was performed to determine whether schools' performance on a content domain could be accurately estimated, using item responses to a single form taken and the characteristics of the items in the domain. Our research focused only on procedures for estimating group-level (i.e., school) domain scores, specifically, the average percent correct for the group in the domain. The procedures summarized here would be applicable to NAEP, which reports only group-level scores, or any testing program providing group-level summaries. The research evaluated estimation procedures under conditions specific to the operational testing program on which the study is based; namely, very few items per content area on the form taken, small domains, and small numbers of observations per group. These are problematic conditions that might often exist in practice and that could greatly affect the accuracy of group-level domain score estimates.

An IRT-Based Domain Score Approach

A domain score for an examinee, where the domain consists only of multiple choice items, can be simply defined as the percentage of items in the domain that the examinee can answer correctly. When the domain contains open-ended items that are scored polytomously, a domain score for an examinee can be defined as the proportion of total possible domain points received by the examinee. At the group level, a domain score may be defined as the average percent correct (domain includes multiple choice items only) or the average proportion of total possible points (domain includes open-ended items) for examinees within the group.

IRT provides a convenient method for estimating domain scores using performance on the form taken and known domain item parameters. Under this approach, a domain score for an examinee (or group), where the domain consists only of multiple choice items, may be estimated as

$$\frac{1}{J}\sum_{j=1}^{J} P_j(\theta) \ , \qquad (1)$$

where θ is an IRT scale score of examinee ability estimated from item responses to the form taken (or θ is the average estimated ability at the group level), and $P_j(\theta)$ is the probability of answering domain item j correctly at ability θ. An application of IRT-based domain score estimates at the individual level was demonstrated in Bock, Thissen, & Zimowski (1997). We present a demonstration of IRT domain score estimation procedures at the group level, under conditions that could adversely affect the accuracy of the domain score estimates.[1]

One benefit of a domain score defined using IRT is that θ can be estimated from one set of items (i.e., the form taken), while the domain score can be estimated from a completely different set of items not taken by the examinees (i.e., the content domain), if the domain item parameters are known. If the IRT model holds, unbiased domain score estimates can be obtained, even if items within a form are a nonrepresentative sample of the domain, because θ is invariant across items (Hambleton & Swaminathan, 1985). IRT-based domain scores are not dependent on the difficulty of the form taken (although they are dependent on the difficulty of the domain items). Assuming the domain item parameters are known, the prevailing problem is to accurately estimate ability—a problem that is exacerbated when the number of items taken is small. Due to averaging P(θ) over examinees and/or items to obtain a group domain score estimate, it is possible that small domains and small numbers of examinees per group may also adversely affect group-level domain scores.

Because our interest is in domain scores to be reported at a group level, we could estimate each group's average ability and compute Equation (1) for each group. Mislevy (1985) proposed an EM-based solution to estimate group means for an unobserved random variable. Tate and King (1994) demonstrated the application of a group-level IRT model to estimate school ability; group-level ability estimates are provided by BILOG (Mislevy & Bock, 1990). Because of rigorous computer requirements for the EM-based solution, and restrictive assumptions for the

[1] The methods presented in this paper are discussed with respect to multiple choice items only, although the methods can be extended to open-ended items that are polytomously scored. For an application of group-level domain score estimation procedures to open-ended items see ACT (1997). Or see Bock, et al. (1997) for a discussion of the extension of the domain score estimates to open-ended items.

BILOG solution that were not met in our operational tests, it was necessary to examine alternate methods of estimating ability. We compared several IRT-based methods for estimating domain scores, some of which operated on estimates of examinee ability, and some of which operated on estimates of group ability.

Methods for Estimating Group Domain Score

Six different IRT-based methods for estimating the group domain score were evaluated. All methods computed an average percent correct (APC) for the group on the domain items. In addition, the group observed APC on the form taken was computed by averaging the observed 0,1 item responses on the form taken over all items and examinees within a group:

$$\frac{1}{N_g}\frac{1}{T}\sum_{i=1}^{N_g}\sum_{t=1}^{T} y_{it} \quad , \qquad (2)$$

where y_{it} was the response of examinee i to item t on the form taken, T was the total number of items on the form taken, and N_g was the number of examinees in group g. The group observed APC is labeled OBS. OBS was computed to provide a baseline comparison for the IRT-based estimation methods. The six IRT-based methods and OBS are summarized in Table 1.

See Table 1 at end of report.

For two of the IRT-based methods, a point estimate of ability (θ_{ig}) was computed for each examinee i within group g from item responses to the single form taken. The estimated domain score for group g was then computed as

$$\frac{1}{N_g}\frac{1}{J}\sum_{i=1}^{N_g}\sum_{j=1}^{J} P_j(\theta_{ig}) \ , \qquad (3)$$

where $P_j(\theta_{i_g})$ was the probability of examinee i in group g (N_g total) responding correctly to domain item j (J domain items total), computed from a three-parameter logistic (3PL) model (Birnbaum, 1968). The methods employed for estimating examinee ability included computing (1) the mean of the posterior distribution (*expected a posteriori* estimate; Bock & Mislevy, 1982) and (2) computing the maximum of the likelihood function (Lord, 1980). Domain scores estimated from the examinee-level *expected a posteriori* estimates are labeled EAP1. Domain scores estimated from the examinee-level maximum likelihood estimates are labeled MLE1.

For three of the IRT-based methods, a point estimate of ability (θ_g) was computed for each group g based on the single form taken. The estimated domain score for group g was then computed as

$$\frac{1}{J}\sum_{j=1}^{J} P_j(\theta_g) \text{ ,} \tag{4}$$

for J domain items, using a 3PL model. The group-level point estimate of ability was computed by (1) averaging examinee *expected a posteriori* estimates within a group, (2) averaging examinee maximum likelihood estimates within a group, and (3) directly estimating mean group ability using a latent-variable regression model. The procedure employed for estimating mean group ability using a latent-variable regression model is discussed in the Appendix. Domain scores estimated from the averaged examinee *expected a posteriori* estimates are labeled EAP2. Domain scores estimated from the averaged examinee maximum likelihood estimates are labeled MLE2. Domain scores estimated from the direct estimates of group means are labeled MU.

For the final IRT-based method, a distribution of group ability $p_g(\theta_q)$ was computed from examinee responses to items on the single form taken, as a discrete approximation to f_g, the continuous distribution of ability for group g. The estimated domain score over J domain items was then computed as

$$\frac{1}{J}\sum_{j=1}^{J}\int_{-\infty}^{+\infty} P_j(\theta)f_g(\theta)d\theta \cong \frac{1}{J}\sum_{j=1}^{J}\sum_{q=1}^{m} P_j(\theta_q)p_g(\theta_q)\ , \qquad (5)$$

where $P_j(\theta_q)$ was the probability of answering domain item j correctly given ability θ_q computed from a 3PL model, and $p_g(\theta_q)$ was group g's probability that ability $= \theta_q$ for $q = 1,2,...,m$ quadrature points. Maximum likelihood estimates of the probabilities $p_g(\theta_q)$ were computed using an EM algorithm; this estimation procedure is discussed in the Appendix. Domain scores estimated with this method are labeled EM.

The Simulation Study

Rarely in a testing program do examinees take an item set that may be classified as a content domain; hence the need for estimating domain scores. The lack of actual domain scores, however, presents a problem for evaluating the performance of domain score estimates. Equating data offers one possibility, but for most testing programs, it is unlikely to consist of examinees taking a multitude of forms. In order to obtain item responses for a fixed group of examinees on both a single form taken and a separate content domain, we performed a simulation study. The simulation allowed us to evaluate groups' estimated domain scores relative to their actual domain scores.

The simulation was performed in two stages, with conditions in each stage (i.e., number of items taken, domain size, number of examinees per group, and differences in difficulty across forms taken) chosen to reflect conditions in the existing operational testing program. Stage One systematically manipulated the conditions of the study, whereas Stage Two was modeled to match the characteristics of three actual test forms from the operational testing program. Thus, Stage One and Stage Two differ in terms of the characteristics of the distributions used to generate the item parameters, the number of items taken, and the domain size. Together the data from Stages One and Two enabled us to evaluate the potential of the estimation procedures when applied to real data.

At both stages of the simulation, examinee responses to items within a content area were generated for nine distinct test forms. Six of these test forms were used to define the content domain. The remaining three forms were each used separately as a form taken, from which the domain performance for a group was imputed.[2] The domain size was always fixed to be six times the number of items taken. The domain was chosen to be six forms because six was the number of forms in our operational program that would have been available for release. Items within content areas on the six forms, while probably not encompassing the content domain were deemed by content experts as representative of the domain as a whole, and of future items on new test forms created under the same test specifications. Thus, although the domain scores computed in the study were not indicative of true domain scores (i.e., based on all possible items in the domain), they were labeled as such because they were based on a representative sample of items from the domain.

Stage One

For each condition of number of items taken, six different forms were used to define the domain (labeled A2, B2, C2, A3, B3, and C3). Three different forms were each used separately as a form taken (labeled A1, B1, and C1), from which the domain performance for a group was

[2] Although the items on the form taken may be included in the domain, we chose to exclude them to match an operational situation where a new form is administered each year, but the domain remains constant to enable comparisons of group performance across years. The constant domain also enabled us to examine the consistency of the domain score estimates across different forms taken in the study.

imputed. The number of items taken per form was fixed at 5, 10, or 20 items with domain sizes of 30, 60, and 120 items, respectively. For each item size, domain scores were estimated for school sizes of 25, 50, and 100 examinees per school. These are typical sample sizes for schools participating in the reporting services for our operational testing program (a minimum of 25 examinees per school is required to participate). At each school size, 100 schools were simulated; the schools are treated as replications in the analyses. The items and schools were simulated to be independent of one another across the 5, 10, and 20 item conditions. Namely, separate items and schools were generated for 5, 10, and 20 items taken.

Form Taken (# items taken)		Domain
A1 (5 items)	⇒	
B1 (5 items)	⇒	Forms A2, B2, C2, A3, B3, C3
C1 (5 items)	⇒	(5 items × 6 forms = 30 domain items)
A1 (10 items)	⇒	
B1 (10 items)	⇒	Forms A2, B2, C2, A3, B3, C3
C1 (10 items)	⇒	(10 items × 6 forms = 60 domain items)
A1 (20 items)	⇒	
B1 (20 items)	⇒	Forms A2, B2, C2, A3, B3, C3
C1 (20 items)	⇒	(20 items × 6 forms = 120 domain items)

Data Generation

School ability (θ_g) was sampled from an $N(0,.4)$ distribution. Within each school, examinee ability (0_{ig}) was sampled from an $N(\theta_g,.6)$ distribution. The variances were chosen so that within-school variability would be greater than across school variability. For each test form, distinct sets of item parameters were generated for a 3PL model. The *a* parameters were sampled from a lognormal distribution with mean 1.13 and variance .36, for all item sizes on all forms. The *c* parameters were sampled from a beta distribution with $\alpha = 6$ and $\beta = 16$, for all item sizes on all forms. The *b* parameters were sampled from an *N*(.5,1) distribution on forms A1, A2, and A3; an *N*(0,1) distribution on forms B1, B2, and B3; and an *N*(-.5,1) distribution on Forms C1, C2, and C3. Manipulating the mean of the normal distribution for the *b* parameters created, for 5 and 10 items taken, differences in difficulty across forms of the magnitude observed for some content areas in the operational testing program. (For example, for 5 items taken, the observed APC was .59 for Form A1, .73 for Form B1, .71 for Form C1, and .67 for the domain, based on N=17,500 taking each form.) For 20 items taken, the form difficulty differences were larger than observed in the operational testing program, which represented a worse case scenario for the domain score estimates. For each examinee, item responses for the nine forms were coded 0 or 1 by comparing the 3PL item probability to a randomly drawn uniform deviate, where the item probability was computed using the generated examinee θ and the generated item parameters.

Evaluation of the Domain Score Estimates

For each combination of school size and item size, the APC for a school on the content domain was estimated using the EAP1, MLE1, EAP2, MLE2, MU, EM, and OBS methods. In each method, the parameters for the forms taken and the domain were assumed known, and the generated ("true") parameters were used. Three separate domain score estimates were computed for each school under each method, first using item responses to Form A1, then responses to Form B1, and finally, responses to Form C1. Because item responses were generated for the six domain forms, we were able to compute the actual domain score for each school by averaging observed responses on the domain items over all domain items and examinees within a school[3]:

$$\frac{1}{N_g}\frac{1}{J}\sum_{i=1}^{N_g}\sum_{j=1}^{J} y_{ij} \quad . \tag{6}$$

Equation 6 differs from Equation 2 (computation of OBS, the observed APC for the form taken) only in that the computation is based on the J domain items, rather than the T items on the form taken. The actual domain score was the same for all three forms taken.

Each domain score estimation method was evaluated using two criteria: the accuracy of the estimated domain scores and the consistency of the estimated domain scores across the three forms taken. The accuracy of the estimated domain scores was measured by computing, for each school on each form taken, the absolute value of the difference in the estimated domain score and the actual domain score (ABSDIF). Operationally, we would prefer ABSDIF be no greater than .05 because larger differences might lead schools to draw the wrong conclusions about their performance. The consistency of the estimated domain scores across forms taken was measured by computing for each school the standard deviation of the estimated domain scores across the three forms taken. Assessing variability over forms taken is important because schools should receive similar estimates of their domain score regardless of which form they take. If the estimated domain scores are not consistent across forms, the school could draw different conclusions depending upon which form the examinees took.

[3] The actual observed domain score was computed as the measure of the true domain score (rather than using the true school or examinee abilities in place of the estimated abilities) because with real data applications, this is the only available measure with which to evaluate the domain score estimates. Using this standard will enable us to compare results of this study to results of a future study based on real data, in which students take an actual domain of items.

Accuracy. Figures 1-3 summarize the ABSDIF for the conditions where there were 5 items taken, 10 items taken, and 20 items taken, respectively. Within these figures (and all subsequent figures), the methods are labeled at the right side of the figure according to their order at 100 examinees per school. A legend for the methods is also given below the horizontal axis. In Figures 1-3, the ABSDIF was averaged, for each school size, over the 100 schools and over the three forms taken, and thus is summarized over 300 schools at each school size. The plots indicate that both OBS and MLE2 performed very poorly on the average, while EAP1, MLE1, EAP2, MU, and EM all showed much greater accuracy. The EM method showed the greatest accuracy of all methods, on average, particularly with the 5-item test length, although the other IRT-based methods (except MLE2) approached the accuracy of the EM method as the number of items taken increased.

See Figures 1-3 at end of report.

As the number of items taken increased, the accuracy of all methods except OBS typically improved. The items and schools were simulated to be independent of one another across the 5, 10, and 20 item conditions; OBS did not improve because the difficulty of the three forms taken relative to the domain difficulty was more diverse as the number of items increased. The observed APC for the form taken can give fairly accurate estimates of the true domain score for small numbers of items if the difficulty of the form taken is very similar to that of the domain. But if the tests are not balanced at the content area level, it is unlikely that each form will be similar in difficulty to the domain.

Figure 1 shows that MLE2 was the least accurate of all methods, including OBS. For the maximum likelihood ability estimates, examinees with all correct, all incorrect responses, or unusual response patterns were assigned arbitrary values of ±5. (Mislevy and Bock (1990) state that it is necessary to set some limit, perhaps ±5 standard deviations of the latent distribution, as upper and lower bounds for θ in maximum likelihood estimation.) When the examinee MLE(θ)s were averaged, the result was either a large positive or negative estimate for the school (much more extreme than the average EAP), which, in turn, either drove the school's estimated APC for the domain up or down. As the number of items increased, the accuracy of MLE2 improved, probably because there were fewer response patterns receiving arbitrary θ values. MLE1 appeared to be less affected by the arbitrary θ values being assigned, as it performed fairly similarly to EAP1, EAP2, and MU. The different order of averaging to compute the group domain score for the MLE1 and MLE2 methods (i.e., computing θ_{ig} for examinees via MLE estimation and averaging $P(\theta_{ig})$ over examinees and items, versus averaging θ_{ig} over examinees to obtain θ_g and averaging $P(\theta_g)$ over items) yielded very different results.

The quartiles of the ABSDIF summarized across same size schools and forms (N=300 at each school size) are given in Table 2. The median value observed is labeled Q2; the 75th and 25th percentiles are labeled Q3 and Q1, respectively. Q3 and Q1 may be considered to be rough error bands for the absolute difference observed in the typical (median) school. With a few exceptions, the mean values in Figures 1-3 were higher than the median values in Table 2, suggesting that large absolute differences in some schools pulled the average above the absolute difference observed in the typical school. The quartiles show that the EM method performed comparatively well for most school sizes and items taken, particularly relative to the accuracy of the OBS method.

See Table 2 at end of report.

The domain score estimates were also evaluated across schools in terms of the proportion of times the absolute difference between the actual and estimated domain score was greater then or equal to .05 (summarized in Table 3). The proportions for OBS were quite large for all school sizes and items taken. The EM method consistently yielded the lowest proportions of the IRT-based methods (except for 50 examinees per school under 20 items taken). The other IRT-based methods showed much poorer performance than the EM method for 25 and 50 examinees per school at 5 and 10 items taken. With school sizes of 100 and items sizes of 10 and 20, the IRT-based methods (except MLE2) approached the accuracy of the EM method in terms of proportions. The robustness of the EM method to the underlying true distribution of school ability is examined in a later section.

See Table 3 at end of report.

Consistency. Figures 4-6 show the standard deviation of the estimated domain scores on the three forms taken, averaged over same size schools for 5, 10, and 20 items taken, respectively (N=100 at each school size). The plots demonstrate that on the average, OBS provided the most inconsistent estimate of all methods across forms. Clearly, if the forms taken differ in difficulty, schools will receive dissimilar domain score estimates across forms when the observed APC for the form taken is used as a domain score estimate. MLE2 also was highly inconsistent across forms. For 5 and 10 items taken, EAP1 and EAP2 provided the most consistent domain score estimates across forms, followed by the EM method. The greater consistency of the EAP methods may have been due to the inward shrinkage of the examinee ability estimates (examinees with scores at either extreme are pulled toward the group). Such inward shrinkage is typical of

Bayesian methods. As the number of items taken increased, the consistency of the EM estimates approached that of the EAP1 and EAP2 estimates.

See Figures 4-6 at end of report.

Stage Two

Data Generation

Stage Two of the simulation was performed to create item responses like those obtained by randomly equivalent groups on three actual test forms. As in Stage One, data were generated for three different content areas of varying length (5, 9, and 14 items; the lengths observed on the actual forms), for nine total forms (three forms taken and six domain forms). The domain sizes were fixed at six times the number of items taken (30, 54, and 84 items, respectively). Three school sizes were used (25, 50, and 100 examinees per school) and 100 schools were simulated at each size. Target parameters were obtained by calibrating the items from three content areas for the three actual test forms using BILOG (Mislevy & Bock, 1990). The target parameters for each form were used to generate item responses to a form taken, creating three separate forms taken. Item responses generated from these parameters yielded form difficulties that were very similar to the form difficulties observed in the real data.

For the six forms in the domain, parameters were generated to be like the target parameters. The *a* parameters for the domain items were generated from a lognormal distribution with mean and variance equal to that observed in the real parameters over the three actual forms (mean 1.22 and variance .10 for 5 items taken, mean 1.05 and variance .05 for 9 items taken, mean .94 and variance .07 for 14 items taken). The *b* parameters for the domain items were generated from a normal distribution with mean and variance equal to that observed in the real parameters over the three actual forms (mean .76 and variance .29 for 5 items taken, mean .72 and variance .41 for 9 items taken, mean -.05 and variance .43 for 14 items taken). The *c* parameters for the domain items were generated from a beta distribution with mean and variance equal to that observed in the real parameters over the three actual forms (α=5.25 and β=25.01 for 5 items taken, α=10.03 and β=68.52 for 9 items taken, α=11.74 and β=45.13 for 14 items taken).

Evaluation of the Domain Score Estimates

Accuracy. Figures 7-9 summarize the ABSDIF for the conditions where there were 5, 9, and 14 items taken, respectively (averaged over schools and forms taken). As observed in Stage One, MLE2 performed very poorly, while EM yielded the most accurate domain score estimates over all three item sizes. The greatest advantage for the EM method over the other IRT-based methods appeared to be in the 5-item case. OBS performed very well for 14 items taken because the forms were very similar in terms of difficulty, and the domain difficulty was simulated to be very similar to the form difficulties. OBS performed less well for the 5 and 9 item content areas because form difficulty and domain difficulty differences were much larger than in the 14-item case. The MLE1 estimates were more accurate in Stage Two than in Stage One, showing very similar accuracy to the EM estimates for 9 and 14 items taken. There may have been fewer unusual response patterns receiving arbitrary ability estimates in Stage Two than in Stage One, so that the MLE1 method may have performed better.

See Figures 7-9 at end of report.

The quartiles of the ABSDIF summarized across same size schools and forms are given in Table 4, while the proportion of times the ABSDIF was greater or equal to .05 over the three forms taken is summarized in Table 5. Tables 4 and 5 show the best overall performance for the

EM method, although the method showed slightly less accuracy in Stage Two than in Stage One. MLE1 showed a fairly similar performance across Stage One and Two; EAP1, EAP2, MLE2, and MU all showed less accuracy in Stage Two than in Stage One. In Stage Two, OBS performed very poorly for 5 and 9 items taken, but similarly to most of the IRT-based methods for 14 items taken. Again, this was due to the similar difficulties of the three forms taken and the domain.

See Tables 4 and 5 at end of report.

Consistency. Figures 10-12 show the standard deviation of the three domain score estimates within schools, averaged over same size schools for 5, 9, and 14 items taken, respectively. As observed in Stage One of the simulation, the EAP1 and EAP2 estimates were the most consistent across forms taken, for all item and school sizes. The EM and MLE1 methods showed similar consistency across forms for all item and school sizes. Even though the accuracy of OBS was very similar to the accuracy of the EM, EAP2, and MLE1 methods for 14 items taken, the IRT-based methods showed greater consistency in their domain score estimates across forms taken at this item size. Even where the difficulty of each form taken was very similar to the domain difficulty, the IRT-based methods appeared to provide slightly more consistent domain score estimates.

See Figures 10-12 at end of report.

Additional Analyses

The Effect of Calibrating Items on Form Taken

In the computation of domain scores, the item parameters for both domain items and items on the single forms taken were assumed known. In real applications, item parameters will have to be estimated for the test forms. If calibrating test items adds error to the domain score procedures, the accuracy of the IRT-based estimates may be affected. To address potential loss due to parameter estimation, we re-ran the 5-item condition from Stage One using estimated parameters for the form taken, rather than the "true" generated parameters. The parameters were estimated from the generated item responses using BILOG (Mislevy & Bock, 1990). The accuracy and consistency results (for three forms taken and 100 schools at each size of 25, 50, and 100 examinees) of the IRT-based methods at each school size were virtually the same, regardless of whether the estimated or true parameters were used to estimate school and examinee ability. In this case, the calibration sample size was quite large over all school sizes (N=17,500) so that the true parameters were probably fairly closely replicated in the calibration. With small sample sizes, the calibration may be less accurate and the accuracy of the domain score estimates may be affected.

The calibration was also probably aided by the fact that the item response data fit the calibration model. Provided the IRT models used fit the data (and calibration sample sizes are adequate), item calibration should pose little problem to the accuracy of the domain score estimates. If the IRT models are not appropriate, calibration may lead to some loss of accuracy in the domain score estimates. Thus, it is important that the performance of these procedures be verified with an application to real test data under calibration conditions that would exist in practice.

The Effect of Domain Size

Because the domain size always increased as the number of items taken increased, it was unclear whether improvements in accuracy and consistency of the IRT-based domain score estimates were due to the increase in the number of items taken, or the increase in the domain size. We examined this issue in two ways: (1) by comparing results for domain sizes of 60 and 120 items for 20 items taken, and (2) by comparing results for domain sizes of 30 and 60 items for 5 items taken.

First, we halved the domain size for the 20-item condition in Stage One and compared the accuracy and consistency results (for three forms taken, and 100 schools at each size of 25, 50, and 100 examinees) to those obtained for the domain size of 120 items. For the 60-item domain size, the first 60 items of the 120-item domain were treated as the domain, so that item responses were the same for items 1-60 in the 60-item and 120-item domains. For both the 60-item and 120-item domain sizes, the item responses to the forms taken remained the same. Second, we created a domain size of 60 items for the 5-item condition, and halved the domain size to compare accuracy and consistency of the estimation methods under domain sizes of 30 and 60 items. For the 30-item domain size, the first 30 items of the 60-item domain were treated as the domain, so that item responses were the same for items 1-30 in the 30-item and 60-item domains. The item responses to the forms taken remained the same for each domain size.

The accuracy and consistency results at each school size were very similar across the two domain sizes, for both 5 and 20 items taken. This indicated that the accuracy and consistency of the IRT-based domain score estimates were being affected by the number of items taken rather than the size of the domain.

The Robustness of the EM Method

Because the EM algorithm utilized a normal distribution as the starting value for each school ability estimate (see Appendix), we considered the possibility that the EM method may have performed well in the simulation because the underlying true distribution of school ability was itself normal. To assess the robustness of the EM method (and the other IRT-based methods) to the underlying distribution of school ability, we generated examinee θs within schools from a uniform distribution and assessed the performance of the methods. (The EM method still employed a normal distribution as the starting value for school ability.) As in Stage One and Two, school ability (θ_g) was drawn from an $N(0,.4)$ distribution. Within each school, examinee ability was sampled from a uniform distribution with mean θ_g and variance 3.0. Within school variability was five times greater in the uniform distribution case than for the normal distribution case (between-school variability remained constant). Domain score estimates were computed for the same conditions as Stage One (5, 10, 20 items taken, and 100 schools at each size of 25, 50, and 100 examinees), and evaluated with respect to the accuracy and consistency of results.

The results for the EM method were consistent with results noted in Stage One. Namely, the EM method did not show decreased accuracy or consistency from what it showed in Stage One. The results for the other IRT-based methods also appeared consistent with results in Stage One. As found with an underlying normal distribution, the IRT-based methods, except MLE2, provided more accurate domain score estimates than OBS. Of the IRT-based methods, the EM method showed the greatest accuracy, for all item sizes. The performance of the EM method under a uniform distribution provides some evidence for the robustness of the procedure against nonnormal distributions.

Discussion

Computing IRT estimates of ability from as few as five items is clearly a questionable practice; ability estimation may be quite poor when based on such a small number of items. With the added limitation of small group sizes, it may be unwise to consider reporting group-level domain scores estimated from performance on a form taken when certain conditions hold. Evaluating the performance of the IRT-based domain score estimates from a position of *absolute* accuracy suggests that there are some accuracy problems for some combinations of items taken and group size, although it appears less so for the EM method than the other IRT-based methods. Evaluating the performance of the IRT-based domain score estimates from a position of *relative* accuracy (relative to the observed performance on the form taken) suggests that it would be preferable to report an IRT-based domain score estimate rather than the observed performance on the form taken, even for very small numbers of items taken and small group sizes.

From a position of absolute accuracy, it appears that the EM method might be applicable with as few as 5-10 items taken and domain sizes of 30 items, if there are an adequate number of examinees per group (i.e., at least 50). As the number of items taken increases, group sizes of 25 examinees might be sufficient for applications of the EM method to estimate the group average domain score, and the other IRT-based methods might be appropriate also with larger group sizes. However, if the entire distribution of domain scores is to be summarized instead of the average domain score, the EM method may be preferable because it is the only distribution-based method. For example, in an application of the EAP1 and EM domain score estimation methods to group sizes of 2,100 (ACT, 1997), domain scores corresponding to examinees at the 90th and 10th percentiles showed some inward shrinkage for the EAP1 method, but not the EM method. For the EAP1 method, examinees with scores at either extreme were pulled toward the group, so that the estimated domain score was higher than the actual domain score at the 10th percentile and lower than the actual domain score at the 90th percentile.

Of the IRT-based methods, the EM method showed the greatest promise, particularly with 5 items taken and small group sizes, although the EAP1, EAP2, MLE1, and MU methods all performed comparatively well at times. Under conditions of few items taken and small group sizes, utilizing an entire distribution of ability in the domain score computation appears to yield more accurate domain score estimates than those based on a point estimate of ability. As the number of items taken increases, the accuracy of the point estimates of ability likely improves and the accuracy of the domain score estimates based on a point estimate approaches that of the domain score estimates based on the entire group ability distribution.

The poor performance of MLE2 relative to MLE1 suggests that the arbitrary assignment of ability levels to individuals with improbable response patterns was an ineffective method of dealing with the maximum likelihood estimation problem, when the maximum likelihood estimates were to be aggregated over examinees. Alternative procedures, such as a biweight estimator might produce more robust estimates of latent ability than maximum likelihood estimation (Mislevy & Bock, 1982). Averaging examinee EAP(θ)s to estimate group ability was much more effective than averaging the MLE(θ)s. Although averaging the examinee EAP(θ)s provided a very simplistic estimate of group ability, the EAP2 method appeared to perform as well as the MU method, which gave a more complicated point estimate of group ability. In addition, estimating ability at the examinee level and averaging probabilities over examinees and items appeared equally as effective for some conditions as using a point estimate of group ability and averaging probabilities over items.

From a position of relative accuracy, all of the IRT-based estimates except MLE2 were more consistent and accurate than OBS when the forms were dissimilar in terms of difficulty. Although the estimation of ability under the IRT methods may have been quite poor in some cases, the summary of performance over domain items and examinees within a group aided in improving the accuracy of the IRT-based domain score estimates, particularly as group size increased. OBS was much more influenced by differences in form difficulty than were the IRT-based domain score estimates. Although the OBS method gave fairly consistent and accurate domain score estimates across forms taken when differences in form and domain difficulty were small, it performed very poorly when form difficulty differences existed. OBS would also perform poorly if there were no differences in difficulty among different forms taken, but the difficulty of the forms taken differed from the difficulty of the domain. As a result, if groups use the observed APC on the form taken as a gauge, they may draw different (and incorrect) conclusions about domain performance depending upon which form the group was administered.

Our examination of real data showed that test form differences did exist at the content area level of the magnitude we simulated. When test forms are not equated or balanced at this fine of a level, form differences of this magnitude are probably likely for some content areas, and the difficulty of individual forms may differ from the difficulty of the domain. In testing programs where the differences in form difficulty are small at the level of reporting, it may be suitable to report the observed APC on the form taken as representative of the domain APC. For form differences of the magnitude observed in this study, one of the IRT-based methods would probably provide a more suitable domain score estimate.

Two separate applications of the methodology presented in this paper provide additional support for the findings of this study. First, the EAP1 and EM methods were applied in an independent simulation modeling conditions specific to the NAEP Geography and Science tests (ACT, 1997). The study used operational NAEP parameters to generate data, as opposed to simulating parameters, and the data consisted of both multiple choice and open-ended items. Second, the EAP1, MLE1, EAP2, MLE2, MU, EM, and OBS methods were applied to equating data consisting of randomly equivalent groups taking eight test forms for a large-scale operational testing program. (Note that this is a different testing program than the program we based the Stage One and Stage Two simulations on.) This application utilized item parameters calibrated on the equating group, and thus provided some check of whether the IRT model is appropriate for the real data. If the IRT model were inappropriate, we would expect the IRT-based methods to perform more poorly when applied to real data. Results from both of the applications were similar to results noted in this study (i.e., in terms of relative and/or absolute accuracy of the IRT-based methods and consistency of results across forms taken) and provide additional, independent confirmation of our conclusions about the feasibility of estimating group domain score performance and reporting school-level domain scores for our operational testing program.

Because of the nature of each study, however, we are still cautious about generalizing findings to our operational testing program. The simulation studies are somewhat artificial in that the data fit the IRT model used (although the application with equating data suggested that the IRT model was probably appropriate). In the application to equating data, no individual took the entire domain, and groups within test centers were not necessarily randomly equivalent, so that comparisons of domain score estimates based on one group's form taken to the actual performance of several different groups on the domain items may have been somewhat inaccurate. Further, the item calibration procedure for both the equating data and the simulation studies differed from how calibration would have to be performed operationally to place domain items from different forms on the same scale. Operational item calibration may add some error to the procedure in real applications.

While the results of all three studies suggest that the EM method may be a suitable method for estimating average domain scores under less than optimal conditions of small group sizes and small numbers of items taken, we intend to verify its performance with an application in which examinees are administered entire domains. ACT is in the process of conducting a study in which examinees within schools take a "domain" of items within a content area, where the domains were created from six operational test forms. Thus, we will be able to compare each school's estimated domain scores to the domain score actually observed. This information, along with the results of the studies already performed, should enable us to more fully evaluate the group-level domain score estimation procedures.

References

ACT, Inc. (1997). *ACT's NAEP redesign project: Assessment design is the key to useful and stable assessment results.* Iowa City, IA: Author.

Baker, E.L. (1974). Beyond objectives: Domain-referenced tests for evaluation and instructional improvement. In W. Hively (Ed.) *Domain-referenced testing* (Chapter 2, pp. 16-30).

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F.M. Lord & M.R. Novick, *Statistical theories of mental test scores.* Reading, MA: Addison-Wesley.

Bock, R.D. & Mislevy, R.J. (1982). Adaptive EAP estimation of ability in a microcomputer environment. *Applied Psychological Measurement, 6*, 431-444.

Bock, R.D., Thissen, D., & Zimowski, M.F. (1997). IRT estimation of domain scores. *Journal of Educational Measurement, 34*, 197-211.

Forsyth, R., Hambleton, R., Linn, R., Mislevy, R., & Yen, W. (1996). *Design/Feasibility Team Report to the National Assessment Governing Board.* Washington, DC: National Assessment Governing Board.

Haladyna, T.M., & Roid, G.H. (1983). A comparison of two approaches to criterion-referenced test construction. *Journal of Educational Measurement, 20*, 271-282.

Hambleton, R.K., Swaminathan, H. Algina, J. & Coulson, D.B. (1978). Criterion-referenced testing and measurement: A review of technical issues and developments. *Review of Educational Research, 48*, 1-47.

Hambleton, R.K. & Swaminathan, H. (1985). *Item response theory.* Boston: Kluwer-Nijhoff Publishing.

Hively, W. (1974). Introduction to domain-referenced testing. In W. Hively (Ed.) *Domain-referenced testing* (Chapter 1, pp. 5-15). Englewood Cliffs, NJ: Educational Technology Publications.

Lord, F.M. (1980). *Applications of item response theory to practical testing problems.* Hillsdale, NJ: Lawrence Erlbaum Associates.

Mislevy, R.J. (1985). Estimation of latent group effects. *Journal of the American Statistical Association, 80*, 993-997.

Mislevy, R.J., & Bock, R.D. (1982). Biweight estimates of latent ability. *Educational and Psychological Measurement, 42*, 725-737.

Mislevy, R.J., & Bock, R.D. (1990). *BILOG 3: Item analysis and test scoring with binary logistic models.* Mooresville, IN: Scientific Software, Inc.

National Assessment Governing Board (1996). *Policy statement on redesigning the National Assessment of Educational Progress.* Washington, DC: Author.

Nicewander, W.A., & Lain, M.M. (1994, June). *The correlation of IRT latent variables with one another and with observable variables.* Paper presented at the Annual Meeting of the Psychometric Society. Champaign-Urbana, IL.

Nitko, A.J. (1984). Defining "criterion-referenced test." In R.A. Berk (Ed.) *A guide to criterion-referenced test construction* (Chapter 1, pp. 8-28). Baltimore: Johns Hopkins University Press.

Popham, W.J. (1978). *Criterion-referenced measurement.* Englewood Cliffs, NJ: Prentice-Hall.

Schulz, E.M., Kolen, M.J., & Nicewander, W.A. (in press). A rationale for defining achievement levels using IRT-estimated domain scores. *Applied Psychological Measurement.*

Tate, R.L. & King, F. (1994). Factors which influence precision of school-level IRT ability estimates. *Journal of Educational Measurement, 31,* 1-15.

Woodruff, D.J., & Hanson, B.A. (1996). *Estimation of item response models using the EM algorithm for finite mixtures* (Research Report No. 96-6). Iowa City, IA: American College Testing.

Appendix

Estimating Mean Group Ability Using a Latent-Variable Regression Model (MU Method)

Group means of $\underline{\theta}$ were estimated using the cell-means regression model

$$\underline{\theta} = X\underline{\beta} + \varepsilon \ , \tag{7}$$

where X was an (N × G) design matrix indicating group (with N total observations over all G groups), $\underline{\beta}$ was a (G × 1) vector of regression weights (mean θ for each group), and θ (N × 1) was missing for all individuals. A vector of group means was obtained using least squares estimation:

$$\underline{\hat{\mu}_\theta} = \underline{\hat{\beta}} = (X'X)^{-1}X'\underline{\theta} \ . \tag{8}$$

The (G × 1) vector $X'\underline{\theta}$ was estimated by

$$X'\underline{\theta} = N D\underline{\rho} \ , \tag{9}$$

where $\underline{\rho}$ was a (G × 1) vector of correlations between each column of X and the vector $\underline{\theta}$ (subsequently labeled $\underline{\rho}(X_g,\underline{\theta})$ to signify the column of X corresponding to group g), D was a (G × G) diagonal matrix of the standard deviations of the columns of X, and N was a scalar representing the total number of observations over all G groups.

Assuming that $\theta \sim N(0,1)$, $\underline{\rho}(X_g,\underline{\theta})$ was computed as (Nicewander & Lain, 1994)

$$\underline{\rho}(X_g,\underline{\theta}) = \frac{1}{T}\sum_{t=1}^{T}\rho_{bis}(X_g,U_t)\frac{\sqrt{1+a_t^2}}{a_t} \ , \tag{10}$$

where a_t was the item discrimination parameter from a 3PL model for item t on the form taken (with T items total on the form taken). $\rho_{bis'}(X_g,U_t)$ was a corrected-for-guessing biserial between each column of X and each column of U, the (N × T) matrix of item responses, computed as

$$\rho_{bis}(X_g, U_t) = \frac{\rho_{pt.bis.}\sqrt{\pi_t(1-\pi_t)}}{(1-c_t)\phi\left[\Phi^{-1}\left(\frac{\pi_t - c_t}{(1-c_t)}\right)\right]} , \tag{11}$$

where π_t was the observed item p-value for item t on the form taken, c_t was the item guessing parameter from a 3PL model, $\rho_{\text{pt.bis.}}$ was the point-biserial correlation between item t and group g, and $\phi[\ ]$ was the height of the ordinate of the normal distribution at the point on the abscissa (i.e., the z-score, $\Phi^{-1}[\ ]$) that divided the distribution into the areas ($\pi_t - c_t$ / $1 - c_t$) and 1 - ($\pi_t - c_t$ / $1 - c_t$). Group means estimated by this method are an approximation to group means estimated by Mislevy's (1985) EM-based solution.

Estimating Distributions of Group Ability Using an EM Algorithm (EM Method)

Maximum likelihood estimates of $p_g(\theta_q)$, group g's probability that ability = θ_q, for q = 1, 2, ..., m quadrature points, were computed iteratively using an EM algorithm. The E-step at iteration s consisted of computing for each q (see Woodruff & Hanson, 1996)

$$n_{gq}^{(s)} = \sum_{i=1}^{N_g} \frac{\left[\prod_{t=1}^{T} P_t(\theta_q)^{y_{it}} Q_t(\theta_q)^{1-y_{it}}\right] p_g(\theta_q)^{(s)}}{\sum_{l=1}^{m}\left[\prod_{t=1}^{T} P_t(\theta_l)^{y_{it}} Q_t(\theta_l)^{1-y_{it}}\right] p_g(\theta_l)^{(s)}} , \tag{12}$$

where $n_{gq}^{(s)}$ was the number of the N_g examinees within group g for whom ability is equal to θ_q, y_{it} was the response of examinee i to item t on the form taken, $P_t(\theta_q)$ was the probability of answering item t correctly computed from a 3PL model, and $Q_t(\theta_q)$ = 1 – $P_t(\theta_q)$. Normalized densities from a normal distribution were used as initial starting values for the $p_g(\theta_q)^{(0)}$.

The M-step at iteration s consisted of using the $n_{gq}^{(s)}$ computed in the E-step to compute

$$p_g(\theta_q)^{(s+1)} = \frac{n_{gq}^{(s)}}{N_g} \, . \tag{13}$$

The values of $p_g(\theta_q)^{(s+1)}$ computed in the M-step at iteration s were used in the E-step at iteration $s+1$.

The E-steps and M-steps were repeated until the relative difference in the log likelihood from iteration s to iteration $s+1$ was $\leq .0001$. At each iteration, the log likelihood was computed as

$$\ln(L)^{(s)} = \sum_{i=1}^{N_g} \ln \sum_{q=1}^{m} \left[\prod_{t=1}^{T} P_t(\theta_q)^{y_{it}} Q_t(\theta_q)^{1-y_{it}} \right] p_g(\theta_q)^{(s)} \, , \tag{14}$$

while the relative difference was computed as

$$\left| \frac{\ln(L)^{(s+1)} - \ln(L)^{(s)}}{\ln(L)^{(s)}} \right| . \tag{15}$$

Several different levels of convergence were considered, ranging from .01 to .00000001. Because the results across schools did not vary much with different levels, .0001 was chosen to enable fairly quick convergence within schools. Reported results are based on 41 equally spaced quadrature points ranging from -4 to +4 (consecutive quadrature points differed by .20).

TABLE 1

Methods for Estimating Group Domain Score.

	Ability Estimation	
Method	(Based on Form Taken)	Computation of Group Domain Score
EAP1	Compute θ_{ig} for examinee *i* in group *g* via EAP estimation	$\frac{1}{N_g}\frac{1}{J}\sum_{i=1}^{N_g}\sum_{j=1}^{J}P_j(\theta_{ig})$
MLE1	Compute θ_{ig} for examinee *i* in group *g* via MLE estimation	
EAP2	Compute θ_g for group *g* as average of examinee EAP(θ)s	$\frac{1}{J}\sum_{j=1}^{J}P_j(\theta_g)$
MLE2	Compute θ_g for group *g* as average of examinee MLE(θ)s	
MU	Compute θ_g for group *g* using latent-variable regression model	
EM	Approximate $f_g(\theta)$ for group *g* using EM algorithm $[f_g(\theta) \equiv p_g(\theta_q)]$	$\frac{1}{J}\sum_{j=1}^{J}\sum_{q=1}^{m}P_j(\theta_q)p_g(\theta_q)$
OBS	—	$\frac{1}{N_g}\frac{1}{T}\sum_{i=1}^{N_g}\sum_{t=1}^{T}y_{it}$

J = number of items in domain
T = number of items on form taken
N_g = Number of examinees in group *g*
m = number of quadrature points
$P_j(\theta)$ = probability of answering domain item *j* correctly at ability θ
$p_g(\theta_q)$ = probability for group *g* that ability = θ_q at quadrature point *q*
y_{it} = response of examinee *i* to item *t* on the form taken

TABLE 2

Quartiles for the Absolute Difference in School Estimated and Observed Domain Scores, Summarized over Same Size Schools and Forms Taken (N=300), Stage One.

			Method						
Items Taken	School Size	Quartile	EAP1	MLE1	EAP2	MLE2	MU	EM	OBS
5	25	Q3	.0582	.0598	.0532	.1695	.0606	.0380	.0973
		Q2	.0309	.0332	.0284	.1066	.0328	.0200	.0680
		Q1	.0140	.0171	.0125	.0564	.0134	.0122	.0440
	50	Q3	.0533	.0522	.0483	.1668	.0469	.0300	.0904
		Q2	.0291	.0307	.0258	.1158	.0270	.0177	.0657
		Q1	.0132	.0144	.0120	.0489	.0116	.0090	.0407
	100	Q3	.0454	.0450	.0393	.1596	.0350	.0205	.0837
		Q2	.0259	.0286	.0245	.0886	.0196	.0121	.0640
		Q1	.0130	.0126	.0109	.0416	.0090	.0058	.0445
10	25	Q3	.0364	.0335	.0332	.0750	.0467	.0265	.1170
		Q2	.0223	.0216	.0210	.0503	.0259	.0155	.0613
		Q1	.0113	.0099	.0098	.0240	.0113	.0083	.0260
	50	Q3	.0335	.0280	.0307	.0712	.0311	.0186	.1184
		Q2	.0207	.0170	.0180	.0477	.0190	.0111	.0584
		Q1	.0104	.0073	.0101	.0212	.0086	.0054	.0234
	100	Q3	.0300	.0245	.0271	.0746	.0256	.0136	.1160
		Q2	.0194	.0148	.0163	.0446	.0142	.0079	.0593
		Q1	.0097	.0083	.0082	.0217	.0070	.0038	.0223
20	25	Q3	.0294	.0240	.0222	.0574	.0425	.0215	.1465
		Q2	.0181	.0149	.0132	.0351	.0250	.0128	.1137
		Q1	.0080	.0067	.0062	.0161	.0114	.0063	.0374
	50	Q3	.0238	.0197	.0177	.0520	.0302	.0145	.1484
		Q2	.0140	.0114	.0103	.0316	.0187	.0094	.1166
		Q1	.0070	.0062	.0050	.0168	.0091	.0045	.0293
	100	Q3	.0242	.0180	.0158	.0556	.0307	.0125	.1446
		Q2	.0155	.0104	.0092	.0300	.0182	.0064	.1193
		Q1	.0082	.0054	.0041	.0142	.0090	.0027	.0307

Q3 is the 75th Percentile, Q2 is the Median, and Q1 is the 25th Percentile for the Absolute Difference.

TABLE 3

Proportion of Times ABSDIF ≥ .05 over Same Size Schools and Forms Taken (N=300), Stage One.

		Method						
Items Taken	School Size	EAP1	MLE1	EAP2	MLE2	MU	EM	OBS
5	25	.3000	.3567	.2733	.7800	.3100	.1533	.6933
	50	.2667	.2733	.2333	.7467	.2333	.0700	.6667
	100	.2100	.1800	.1700	.7067	.1033	.0067	.6900
10	25	.1033	.0933	.0767	.5033	.2167	.0400	.6033
	50	.0867	.0200	.0633	.4867	.0900	.0033	.5567
	100	.0533	.0033	.0333	.4567	.0300	.0000	.5667
20	25	.0400	.0300	.0167	.3100	.1600	.0100	.7000
	50	.0167	.0000	.0033	.2833	.0667	.0033	.6767
	100	.0300	.0000	.0200	.2833	.0300	.0000	.6667

TABLE 4

Quartiles for the Absolute Difference in School Estimated and Observed Domain Scores, Summarized over Same Size Schools and Forms Taken (N=300), Stage Two.

			Method						
Items Taken	School Size	Quartile	EAP1	MLE1	EAP2	MLE2	MU	EM	OBS
5	25	Q3	.0693	.0537	.0858	.1921	.0911	.0443	.1053
		Q2	.0413	.0320	.0551	.1411	.0579	.0262	.0674
		Q1	.0183	.0156	.0339	.0862	.0293	.0123	.0293
	50	Q3	.0709	.0525	.0924	.1873	.0779	.0358	.1193
		Q2	.0424	.0316	.0562	.1283	.0458	.0189	.0710
		Q1	.0210	.0182	.0315	.0805	.0230	.0075	.0393
	100	Q3	.0557	.0425	.0799	.1879	.0712	.0252	.1165
		Q2	.0346	.0273	.0577	.1400	.0524	.0150	.0653
		Q1	.0167	.0135	.0276	.1000	.0313	.0072	.0385
9	25	Q3	.0556	.0376	.0627	.1342	.0671	.0363	.0719
		Q2	.0326	.0222	.0377	.0952	.0437	.0220	.0415
		Q1	.0141	.0115	.0183	.0559	.0208	.0105	.0215
	50	Q3	.0454	.0252	.0558	.1363	.0605	.0214	.0608
		Q2	.0276	.0139	.0395	.0997	.0383	.0131	.0437
		Q1	.0132	.0062	.0206	.0621	.0200	.0057	.0263
	100	Q3	.0363	.0190	.0526	.1396	.0582	.0168	.0622
		Q2	.0236	.0117	.0384	.1076	.0398	.0099	.0468
		Q1	.0113	.0062	.0186	.0748	.0238	.0046	.0301
14	25	Q3	.0326	.0280	.0285	.0882	.0497	.0272	.0305
		Q2	.0203	.0165	.0169	.0604	.0301	.0169	.0186
		Q1	.0104	.0075	.0078	.0318	.0160	.0074	.0105
	50	Q3	.0305	.0189	.0209	.0825	.0448	.0191	.0213
		Q2	.0167	.0112	.0112	.0571	.0277	.0102	.0125
		Q1	.0076	.0052	.0053	.0273	.0147	.0044	.0051
	100	Q3	.0287	.0146	.0163	.0791	.0401	.0125	.0162
		Q2	.0188	.0081	.0091	.0564	.0265	.0076	.0109
		Q1	.0107	.0035	.0039	.0301	.0139	.0039	.0052

Q3 is the 75th Percentile, Q2 is the Median, and Q1 is the 25th Percentile for the Absolute Difference.

TABLE 5

Proportion of Times ABSDIF ≥ .05 over Same Size Schools and Forms Taken (N=300), Stage Two.

		Method						
Items Taken	School Size	EAP1	MLE1	EAP2	MLE2	MU	EM	OBS
5	25	.4267	.2700	.5567	.8967	.5733	.2100	.6233
	50	.4433	.2767	.5500	.8767	.4633	.1067	.6667
	100	.3133	.1700	.5700	.9000	.5400	.0200	.6267
9	25	.2967	.1000	.3667	.7767	.4200	.0733	.4233
	50	.2000	.0200	.3333	.8000	.3533	.0133	.4133
	100	.1367	.0033	.2833	.8467	.3567	.0000	.4500
14	25	.0700	.0333	.0367	.5833	.2467	.0367	.0467
	50	.0167	.0000	.0000	.5567	.1933	.0000	.0067
	100	.0733	.0000	.0133	.5567	.1100	.0000	.0000

FIGURE 1. Absolute Difference in School Estimated and Actual Domain Score for 5 Items Taken, Averaged over Same Size Schools and Forms Taken (N=300), Stage One.

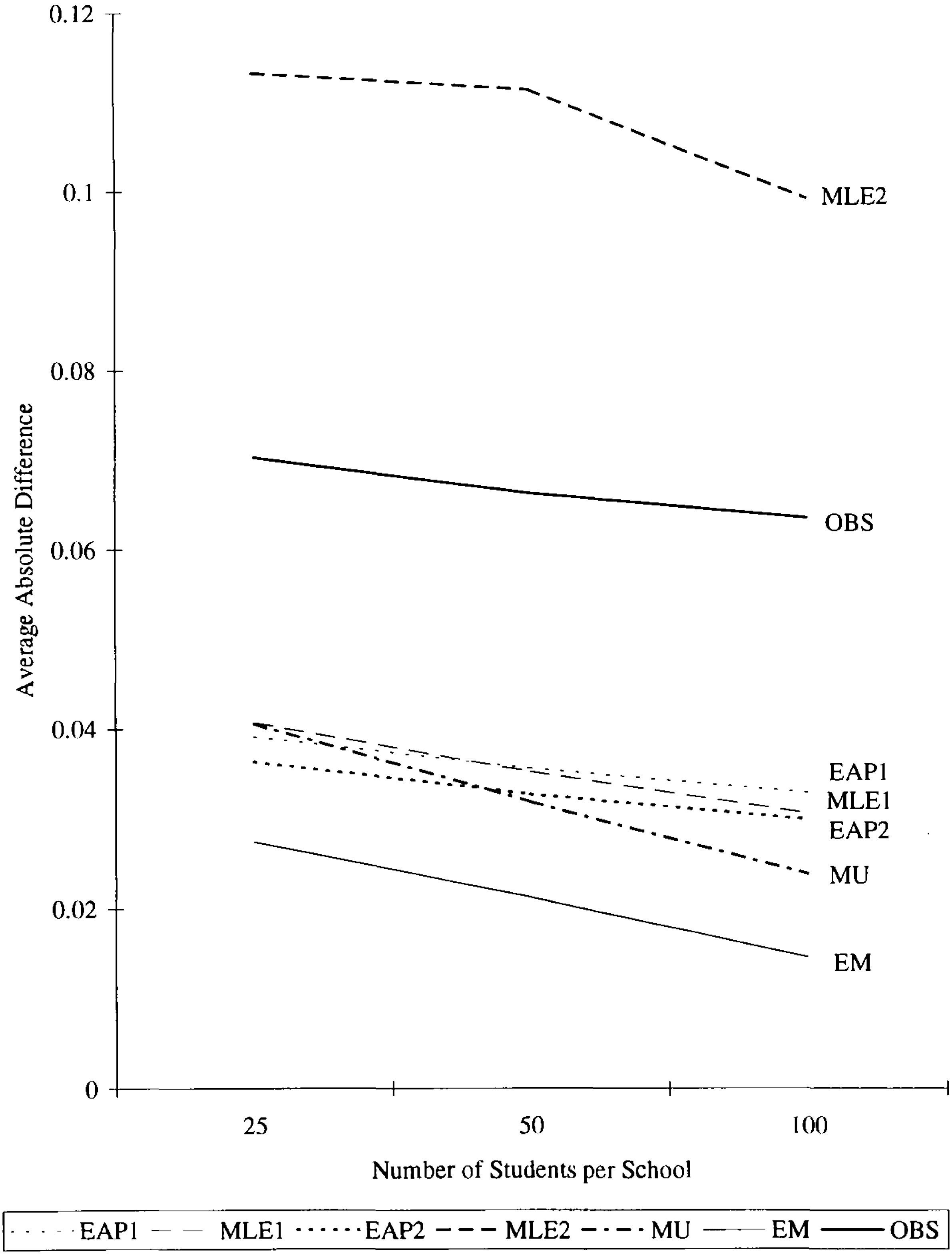


FIGURE 2. Absolute Difference in School Estimated and Actual Domain Score for 10 Items Taken, Averaged over Same Size Schools and Forms Taken (N=300), Stage One.

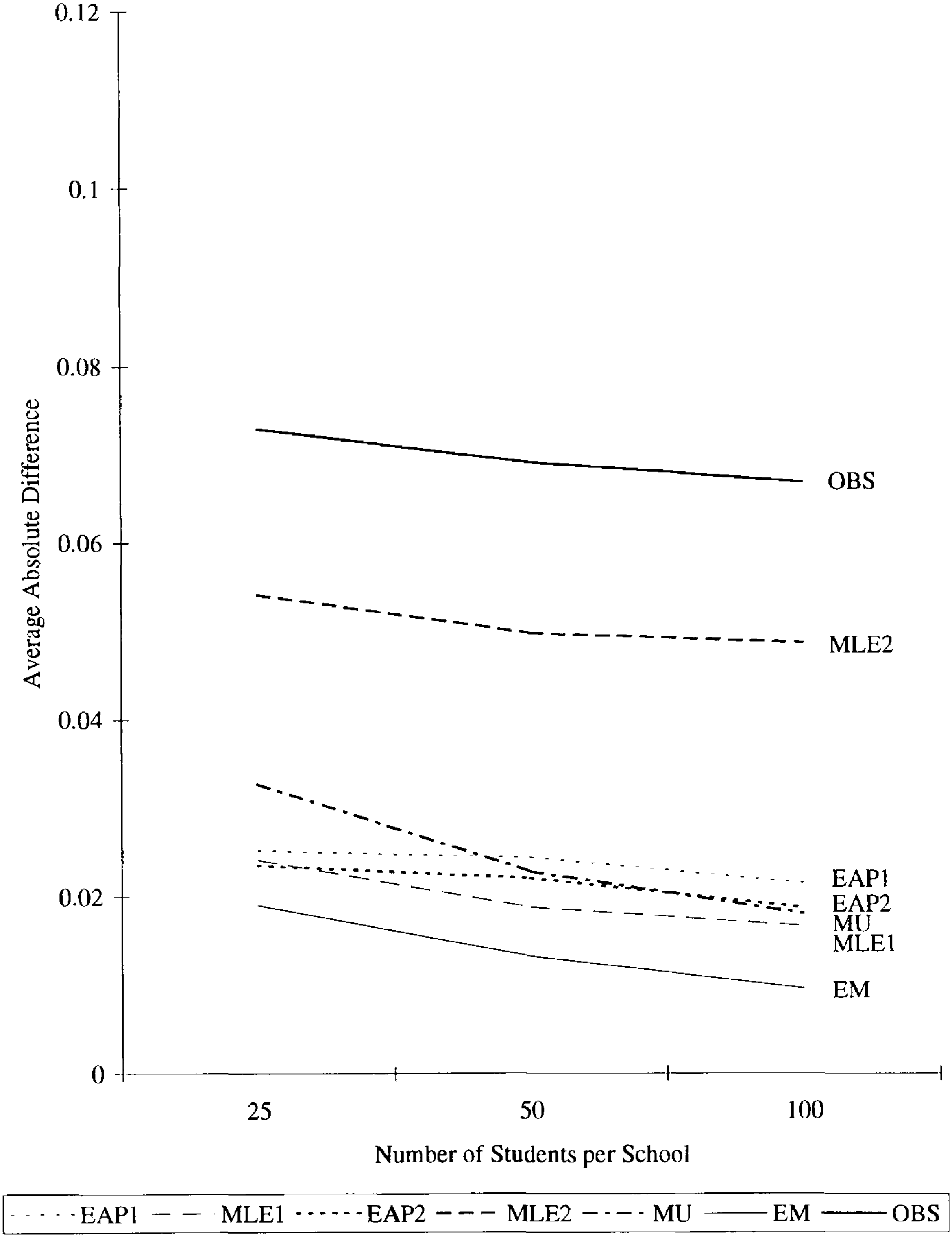


FIGURE 3. Absolute Difference in School Estimated and Actual Domain Score for 20 Items Taken, Averaged over Same Size Schools and Forms Taken (N=300), Stage One.

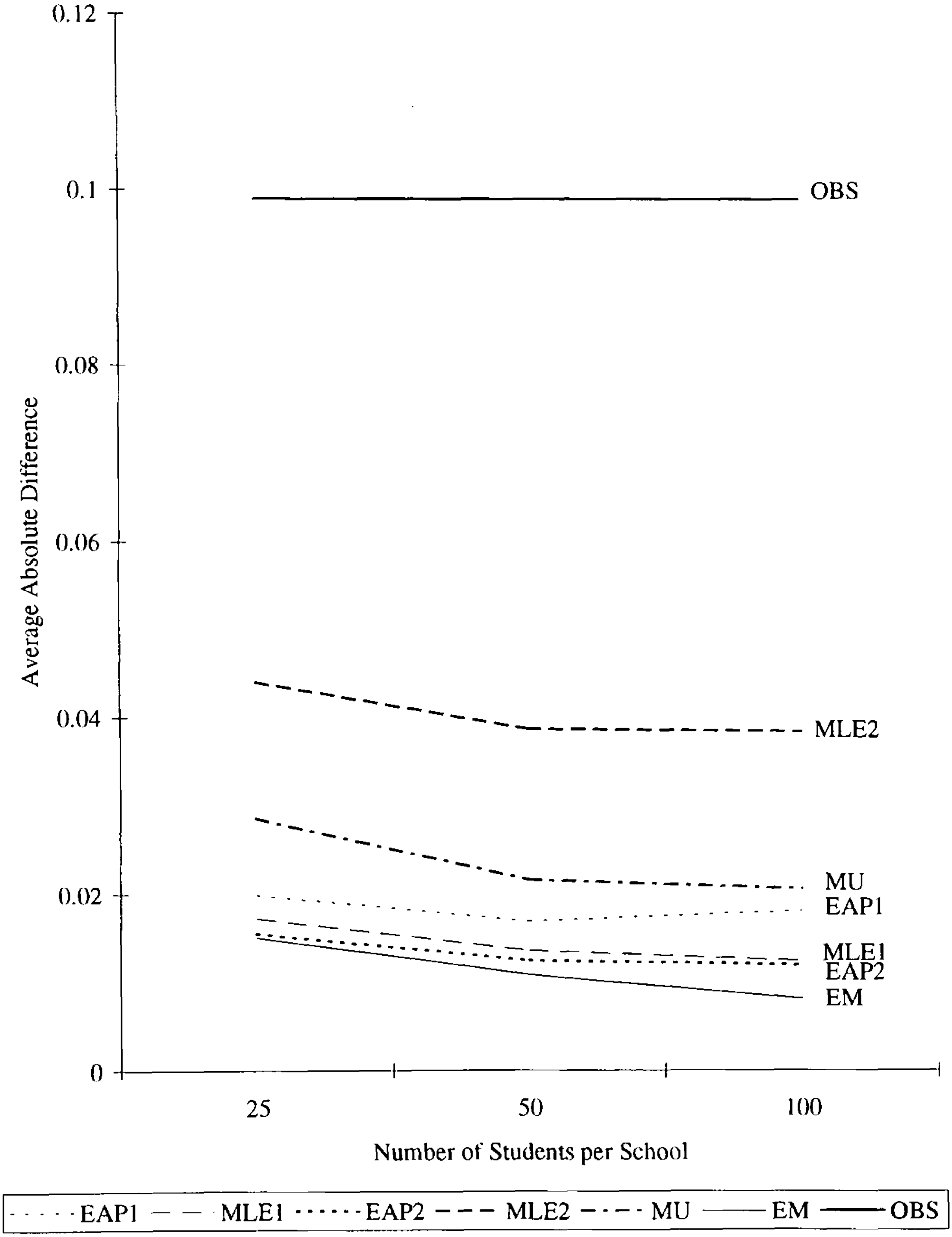


FIGURE 4. Standard Deviation of the Estimated Domain Scores over Forms Taken for 5 Items Taken, Averaged over Schools of the Same Size (N=100), Stage One.

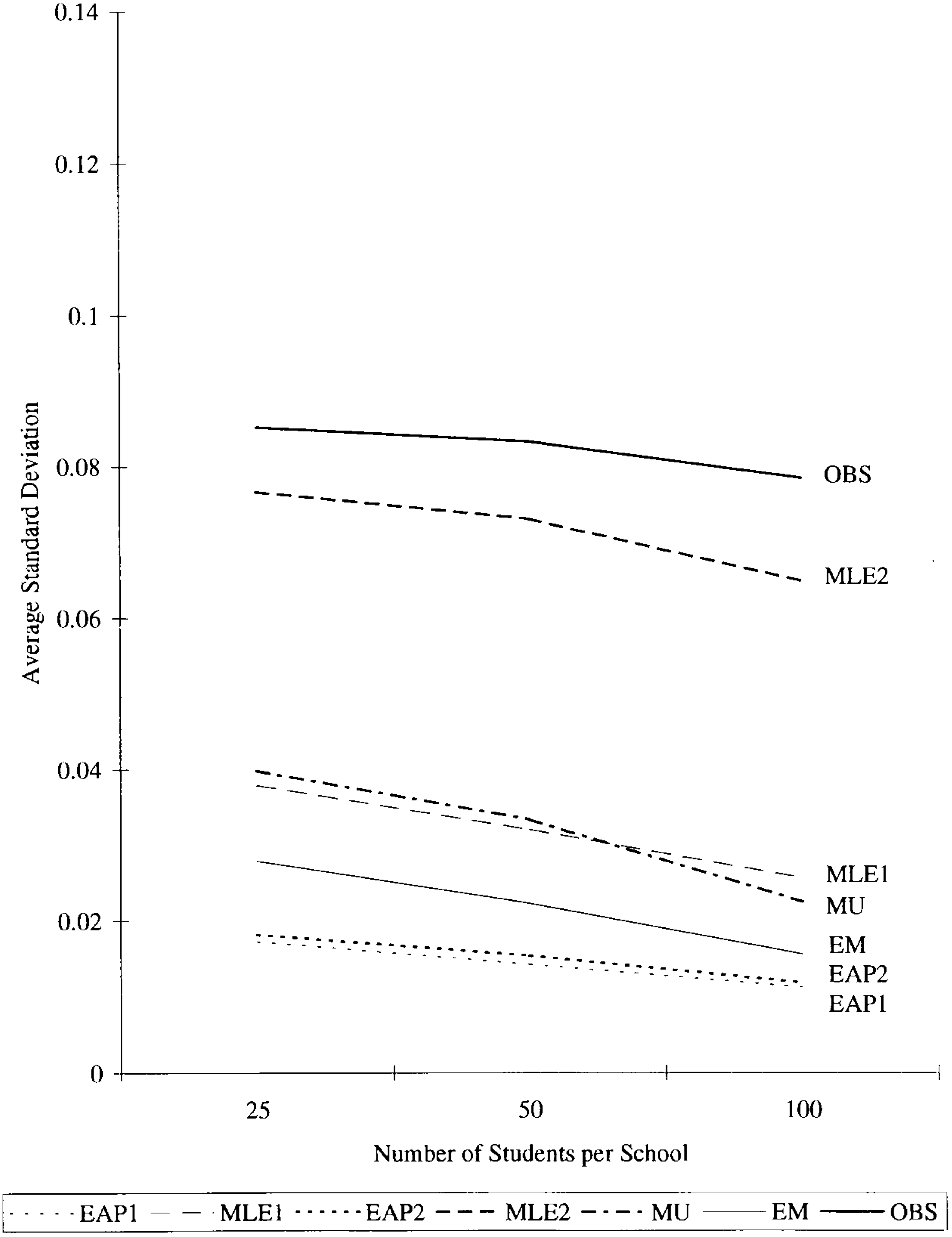


FIGURE 5. Standard Deviation of the Estimated Domain Scores over Forms Taken for 10 Items Taken, Averaged over Schools of the Same Size (N=100), Stage One.

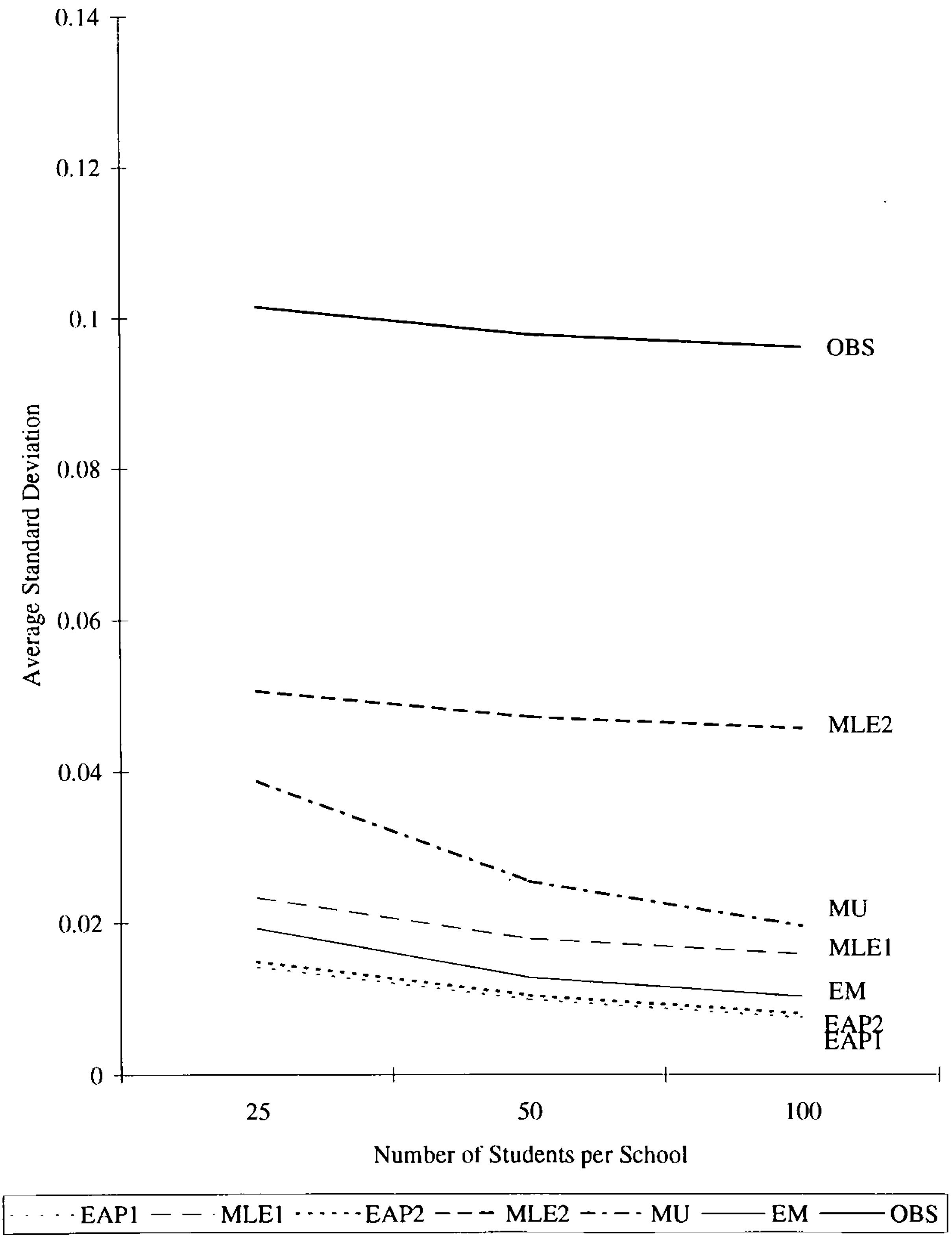


FIGURE 6. Standard Deviation of the Estimated Domain Scores over Forms Taken for 20 Items Taken, Averaged over Schools of the Same Size (N=100), Stage One.

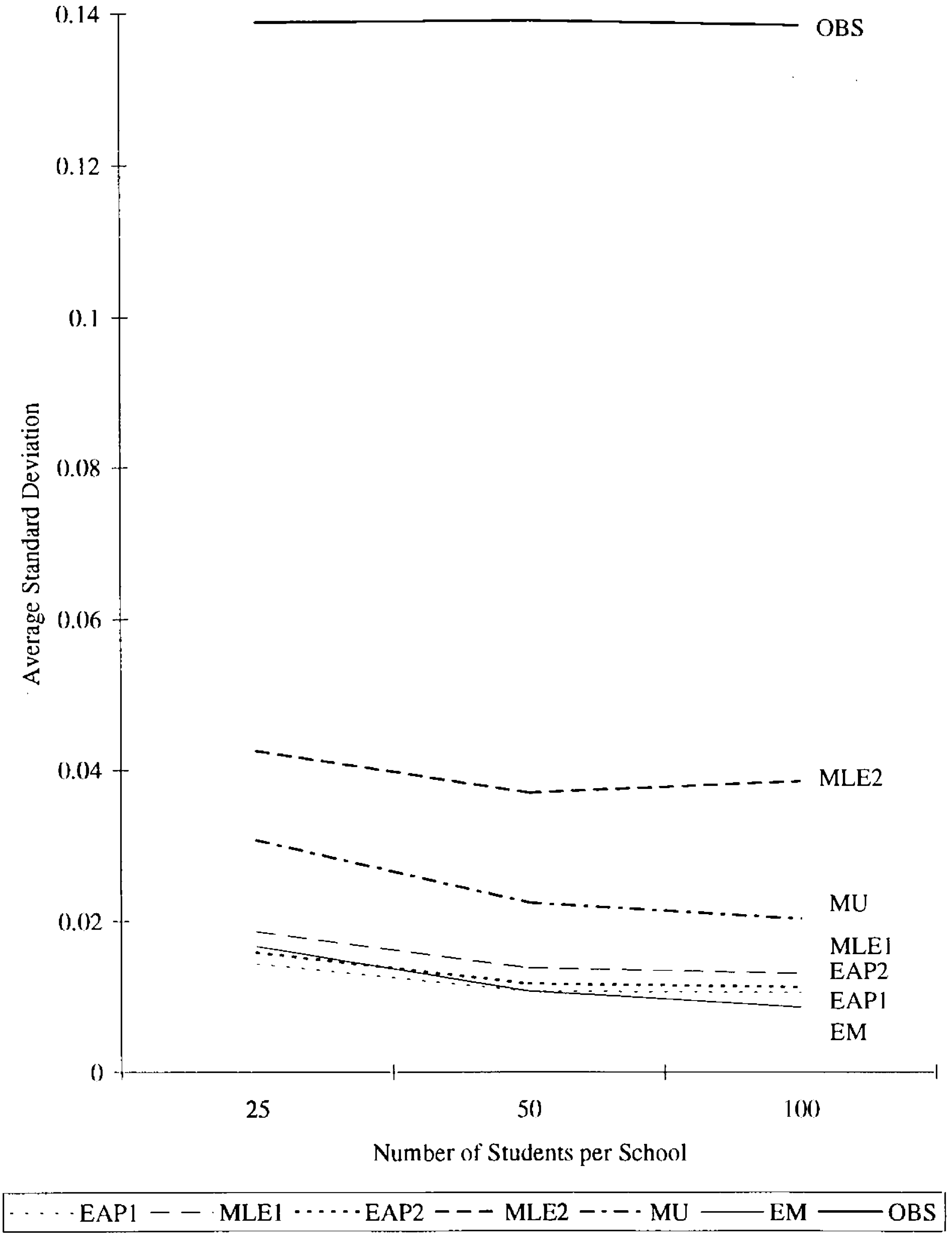


FIGURE 7. Absolute Difference in School Estimated and Actual Domain Score for 5 Items Taken, Averaged over Same Size Schools and Forms Taken (N=300), Stage Two.

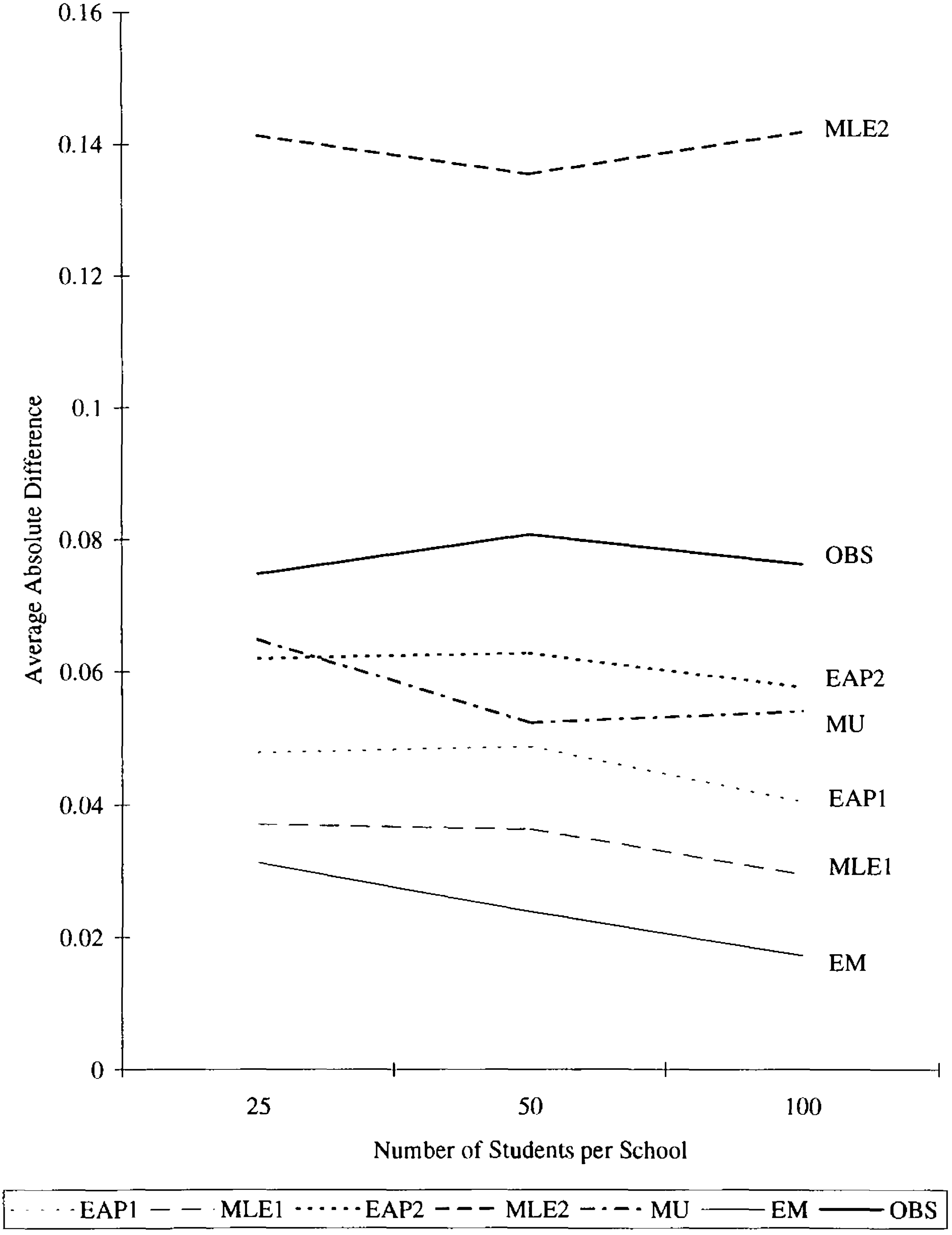


FIGURE 8. Absolute Difference in School Estimated and Actual Domain Score for 9 Items Taken, Averaged over Same Size Schools and Forms Taken (N=300), Stage Two.

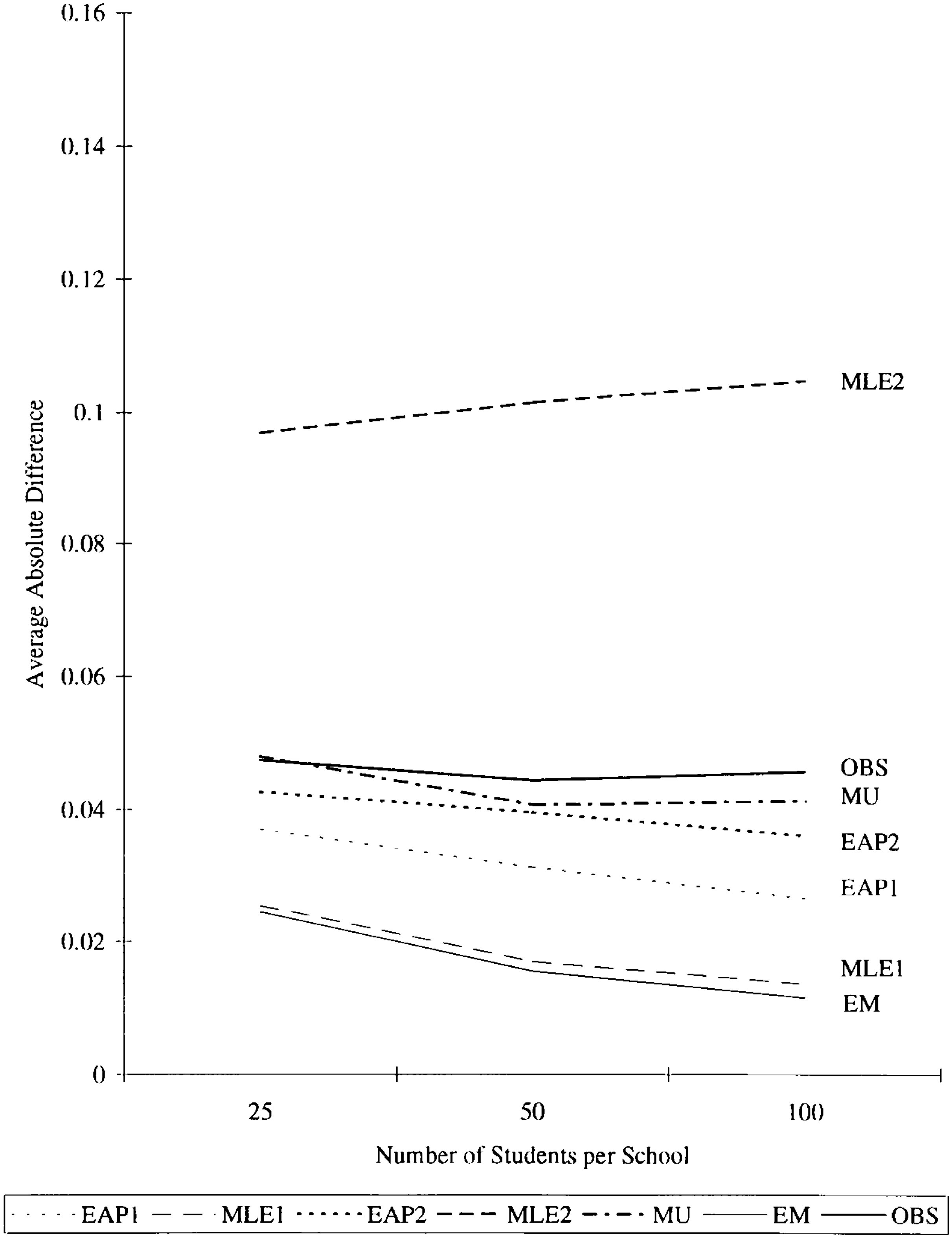


FIGURE 9. Absolute Difference in School Estimated and Actual Domain Score for 14 Items Taken, Averaged over Same Size Schools and Forms Taken (N=300), Stage Two.

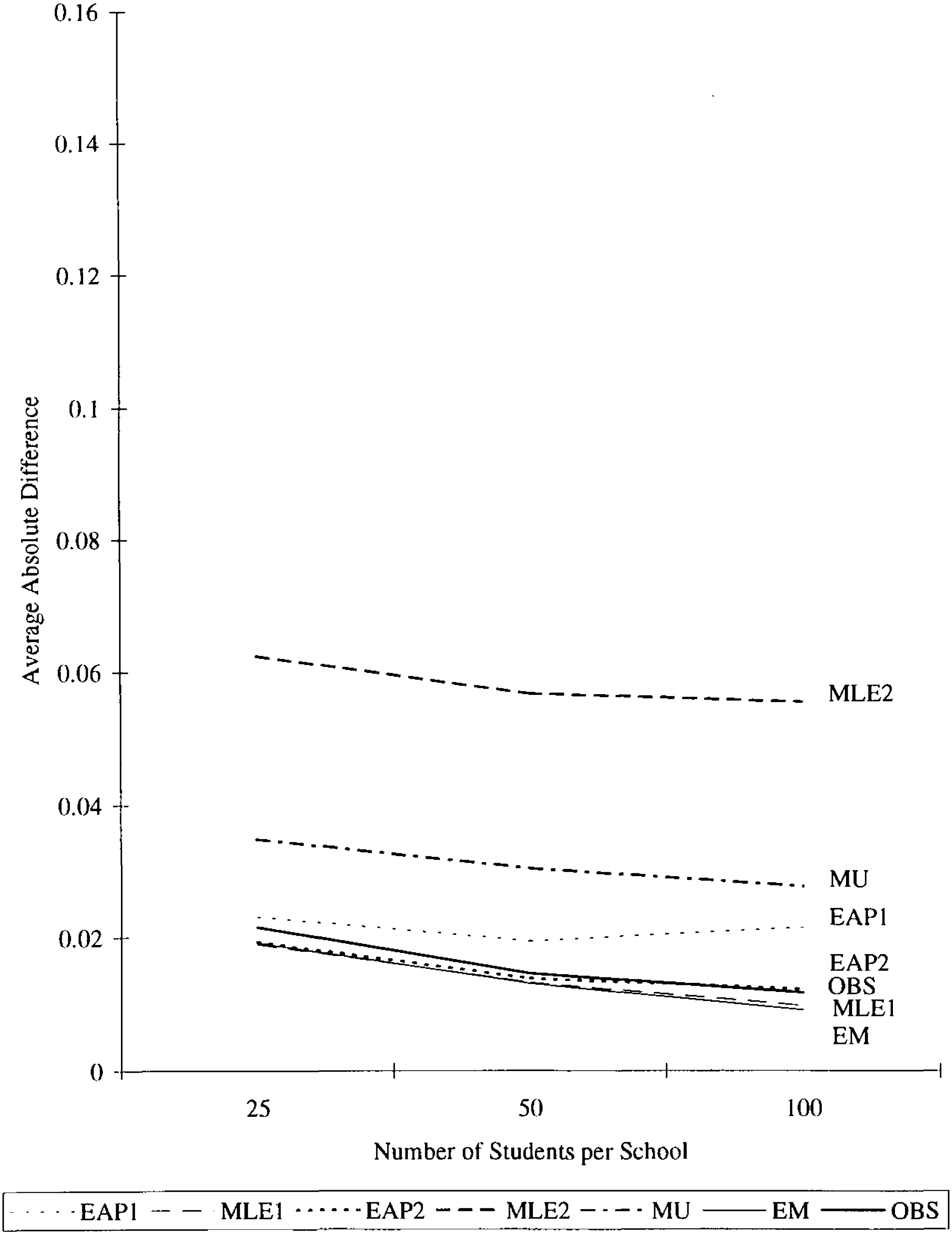


FIGURE 10. Standard Deviation of the Estimated Domain Scores over Forms Taken for 5 Items Taken, Averaged over Schools of the Same Size (N=100), Stage Two.

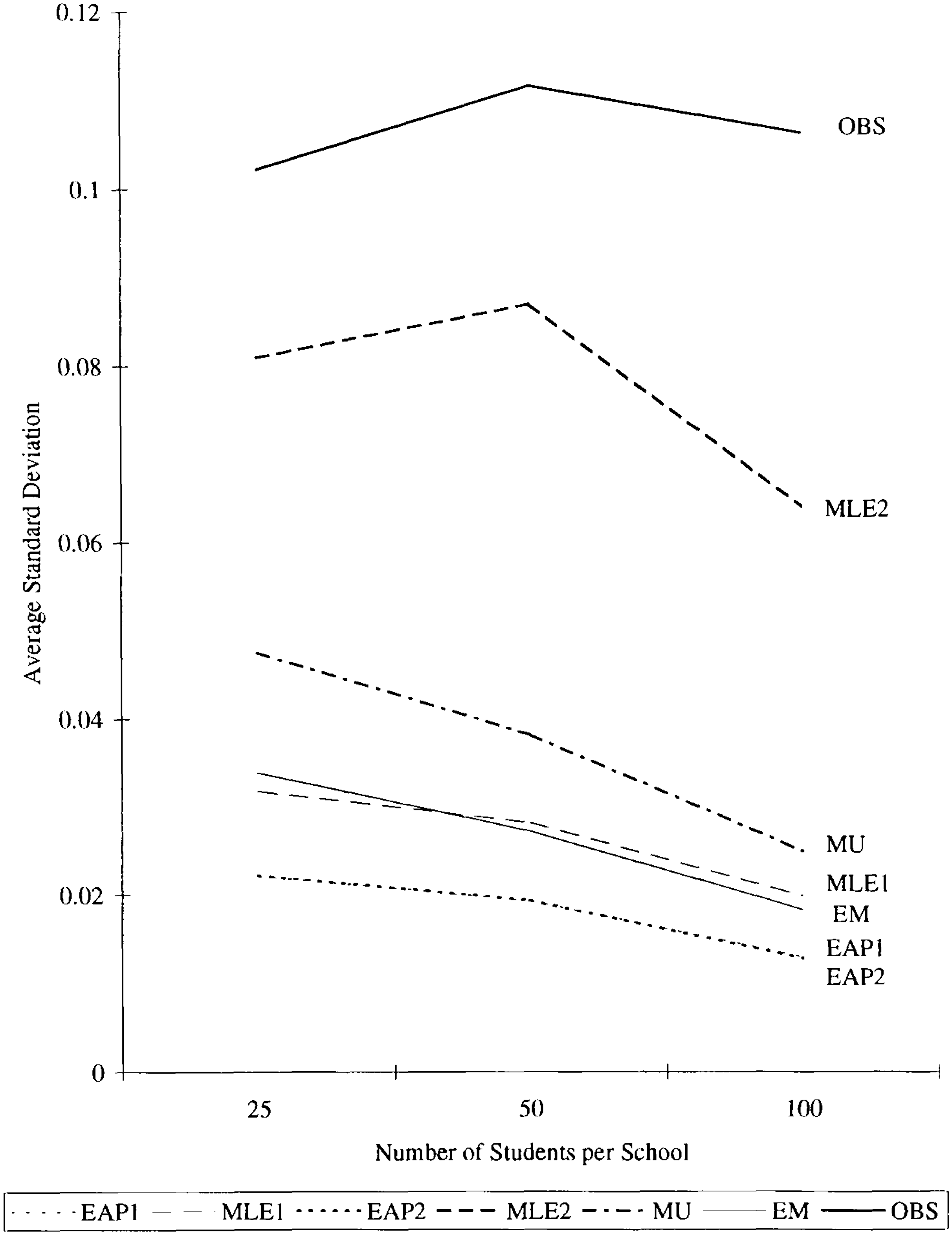


FIGURE 11. Standard Deviation of the Estimated Domain Scores over Forms Taken for 9 Items Taken, Averaged over Schools of the Same Size (N=100), Stage Two.

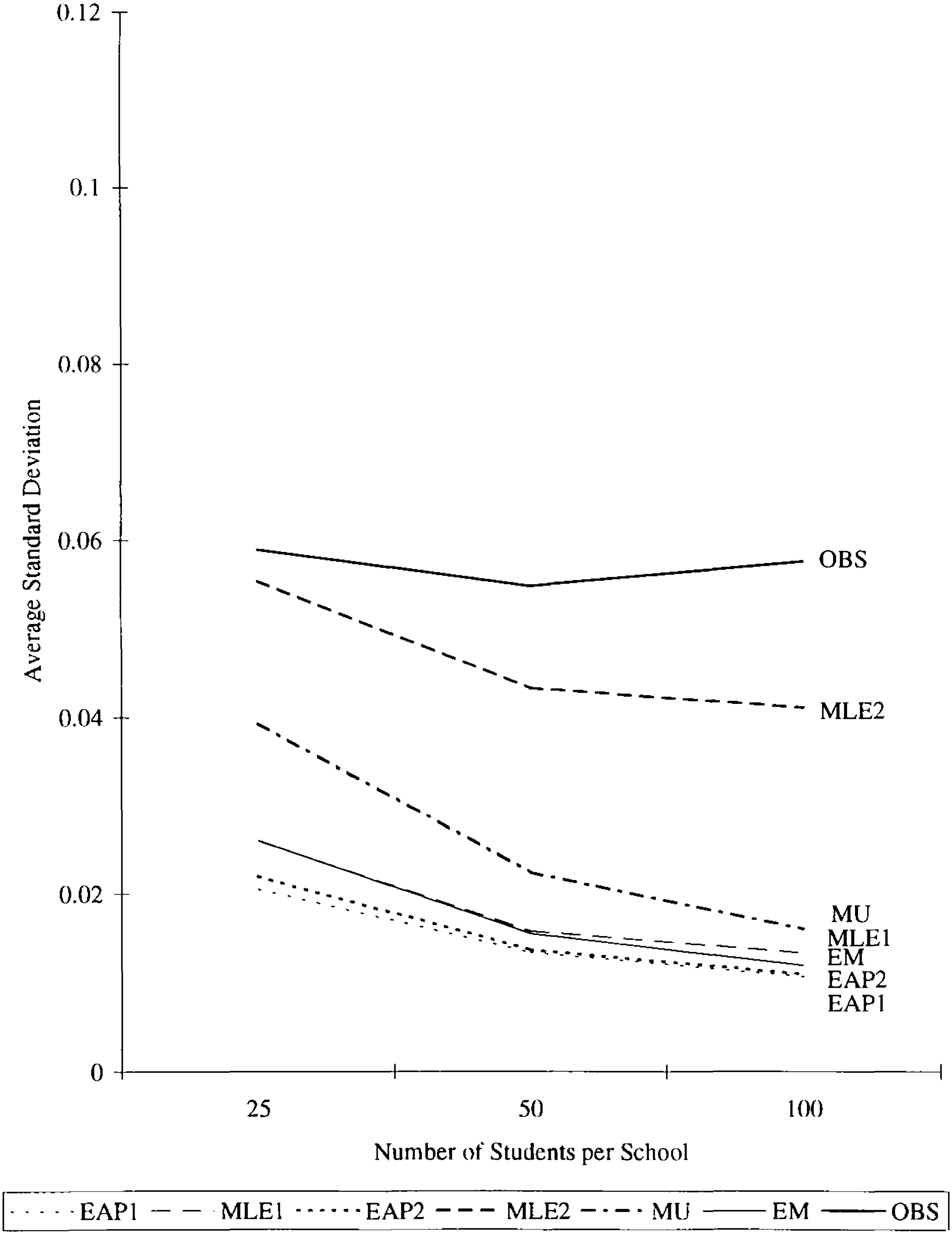
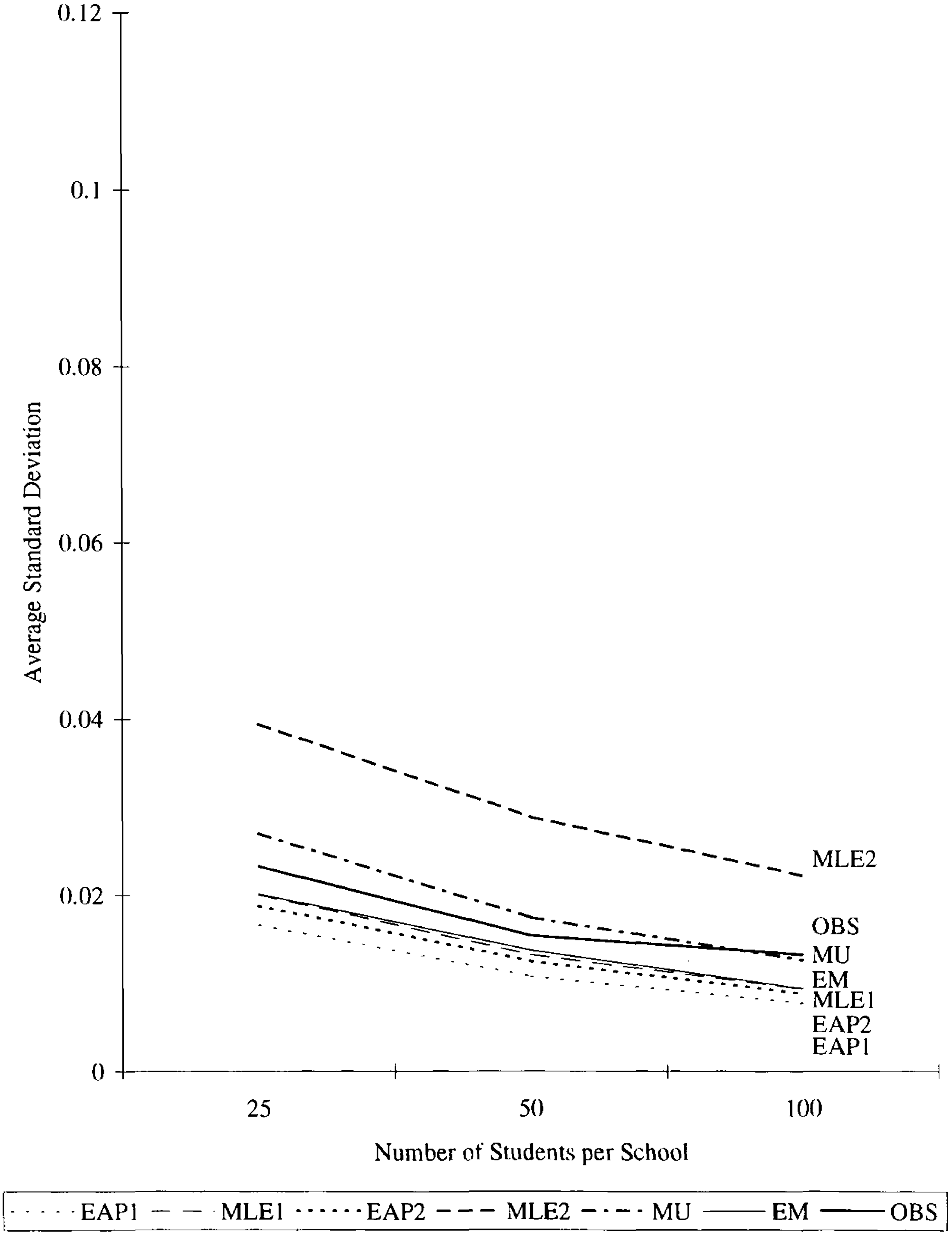


FIGURE 12. Standard Deviation of the Estimated Domain Scores over Forms Taken for 14 Items Taken, Averaged over Schools of the Same Size (N=100), Stage Two.


Average Standard Deviation
Number of Students per School
MLE2
OBS
MU
EM
MLE1
EAP2
EAP1
EAP1
MLE1
EAP2
MLE2
MU
EM
OBS