ACT Research Report Series 2000-1

Issues in Creating and Reporting Concordance Results Based on Equipercentile Methods

Mary Pommerich

Bradley A. Hanson

Deborah J. Harris

James A. Sconing

ACT

January 2000

Issues in Creating and Reporting Concordance Results Based on Equipercentile Methods

Mary Pommerich
Bradley A. Hanson
Deborah J. Harris
James A. Sconing

Abstract

Educational measurement practitioners are often asked to link scores on tests that are built to different content specifications. Standard equating procedures, such as equipercentile methods, can be applied to link distinct tests, but the resulting scores are not interchangeable. If the linked tests are not written to the same specifications, the linkage is considered a concordance, rather than an equating. The same procedures are used to establish both types of linkage, but concordance is a much weaker form of linkage than equating. This paper focuses on the use of equipercentile methods to link distinct tests, and discusses issues in creating and reporting concordance results that may require different treatment than equating results. Examples are presented from a concordance between ACT and SAT I scores.

Acknowledgements

We thank Richard Sawyer for interesting and helpful discussions on this topic, and Dean Colton and Ron Cope for thorough reviews and helpful comments on the report.

Issues in Creating and Reporting Concordance Results Based on Equipercentile Methods

Practitioners appear to be increasingly faced with the difficult task of linking scores across distinct tests that were never intended to be linked. The tests typically have different specifications, different populations, different score scales and distributions, and varying degrees of relationship between scores to be linked, among other differences. Examples include linking performance on a state assessment to performance on the National Assessment of Educational Progress tests, linkages between ACT and SAT I scores for use in college admissions decisions, or linkages between placement exams targeted for community college students and entrance exams targeted for traditional college-bound students.

Various types of linkages have been defined and discussed in the literature (e.g., Linn, 1993; Mislevy, 1992), of which equating is considered to be the most statistically rigorous. The rigor of equating comes not from the statistical procedures applied, but from the way the tests are constructed, namely to the same specifications (Mislevy, 1992), and from careful design of equating studies. As Linn (1993) warns, there is nothing to prevent the use of statistical equating procedures with tests that do not meet the assumptions of equating. A recent example is given in Dorans, Lyu, Pommerich, and Houston (1997), in which equipercentile "equating" procedures are applied to develop concordance tables between ACT and SAT I scores. Although the ACT and the SAT I are both college entrance exams, they are unique tests developed for different purposes with different contents. Results of equipercentile procedures applied to ACT and SAT I scores are considered to be concordant, rather than equated; concordance is a much weaker form of linkage. Even though the same procedures are used to establish both types of linkages, concordant scores should not be considered interchangeable as equated scores are, because the tests being linked are

not constructed to the same specifications. Equating procedures adjust for differences in difficulty, not for differences in content (Kolen & Brennan, 1995).

This paper focuses on methodological issues in applying equipercentile methods to link distinct pairs of tests that are not constructed to the same specifications. This situation will be referred to as a concordance situation, as opposed to an equating situation, and the end result is a concordance table that gives "comparable" scores between the tests. (It is the situation and not the methodology that lends the distinction between an outcome that is a concordance or an equating.) Equipercentile methods yield scores that have the same percentile rank within the linking groups (for both an equating and a concordance). Concordant scores, therefore, are comparable in terms of the proportion selected by either test within the linking sample. Interpretation of results is more difficult for a concordance than an equating, because of weaker ties between the tests, so that inappropriate use of results may be a greater concern. In addition, concordance study designs are typically governed by convenience rather than rigorous methodological considerations. Whereas equatings are typically planned in advance and conducted under carefully controlled circumstances, concordances are typically conducted after the fact with little possibility of attention to design. This substitution of convenience for rigor in design may lead to problems in creating and reporting concordance results that require different treatment than equating results.

For example, as samples of convenience, concordance samples are likely to be smaller than equating samples. What do we do about small sample problems such as gaps in frequency distributions, zero frequencies at the tails, or sparse data in general? Questions may also arise concerning procedures that have been well researched for equating situations, but which may need to be considered anew in a concordance situation. When and how should concordance results be smoothed? Are traditional equating standard errors appropriate for a concordance situation?

What are appropriate sample sizes for computing concordances? Further, because concordance results may be less stable and more easily misunderstood than equating results, it is important to consider whether there are score points for which concordances should not be reported.

Examples are presented of specific issues that arose in applying equipercentile methods to create concordance tables between ACT Composite and SAT I Verbal + Math (V+M) scores. ACT and SAT I scores were obtained from 14 institutions and two states, and 17 sets of concordance tables were developed: one for each of the 14 institutions and two states, and one for the entire pool of examinees (the pooled group). Examples from four institutions and the pooled group are presented. The institutions and states provided ACT and SAT I scores for examinees taking both tests, as part of a collaboration between ACT, The College Board, and The Educational Testing Service to develop concordance tables between ACT and SAT I scores. The data and sample are described in Dorans, et al. (1997). Initially, we applied procedures that were traditionally used in equatings of the ACT Assessment. As we began to look more closely at the data and results, particularly at the institution level, we began to rethink the traditional procedures and to make adjustments more appropriate for the case at hand. The paper presents examples of problems encountered in creating and reporting the concordances and highlights issues that possibly require different treatment in a concordance situation than in an equating situation.

Data Collection and Cleaning Issues

Unlike equating, where special designs are implemented to collect appropriate data, when distinct tests are to be linked, a concordance sample is likely to be a sample of convenience, derived from readily available data on examinees taking each test of interest. If the same examinees take both tests (comparable to a single group design in an equating), consideration must be given to screening for time between testing, for order of testing, and for students with repeat

scores on one or both tests. These factors are of concern for a single group design in an equating situation also, but often can be controlled in advance by a carefully maintained design. These factors are perhaps more worrisome in a concordance situation because they typically cannot be systematically controlled in the data collection process, and there may be some temptation to analyze the data without screening, which could yield different interpretations of results than those based on screened data. Decisions regarding a final sample may, in part, be determined by the intended use of the concordance table.

If too much time is allowed between testing in a concordance sample, examinees may show better performance on the second test because they learned more of the test content between tests. In addition, examinees may perform better on the exam taken second because of practice effects. Thus, it may be necessary to screen the concordance sample so that the order of testing is counter-balanced to some degree. For a linkage between the ACT and the SAT I, too much time between testing could bias the concordance results if the order of testing is not balanced. The large-scale ACT-SAT I linkage reported in Dorans et al. (1997) excluded examinees that took the tests more than 217 days apart. Research supporting that decision is also reported in Marco and Abdel-Fattah (1991). On average, students in Dorans et al. (1997) took the SAT I 15 days prior to the ACT. A concordance based on a restricted sample, so that average time between testing was 0 days, yielded virtually the same results. Thus, while eliminating examinees with a lot of time between testing and counter-balancing the order of testing may be important, the degree to which the sample actually needs to be restricted should be carefully examined.

Data cleaning issues will be specific to each concordance situation. There are likely to be tradeoffs between strict screenings to obtain counter-balancing and minimal time between testing, and restricted sample sizes that can create problems in interpreting equipercentile results.

Decisions about which scores to keep for repeat testers could also influence counter-balancing and time between testing. Practitioners may also want to consider how schools treat multiple scores per examinee when making data screening decisions. Because equipercentile relationships can be affected by group characteristics, care should be taken to make concordance samples as similar as possible to the population to whom the results will be applied.

The Equipercentile Method

Concordances between ACT and SAT I scores were created using an equipercentile method, often referred to as "equipercentile equating" in the equating literature. Equipercentile equating procedures may technically be applied to any set of test scores. Although the label commonly assigned to the procedure is that of "equating," the results are considered to be equated only if the assumptions of equating are met in the data. In a concordance situation, results cannot be considered to be equated because the tests being linked are not written to the same specifications. The equipercentile method is discussed in detail in Kolen and Brennan (1995), along with implementation issues.

The equipercentile function applied to the ACT-SAT data utilizes the percentile ranks of scores A and B (where A represents ACT scores and B represents SAT scores), and is given by:

$$\frac{\Pr(A < i) + .5 \times \Pr(A = i) - \Pr(B < u^*(i))}{\Pr(B = u^*(i))} + u^*(i) - .5 \quad , \qquad (1)$$

where $u^*(i)$ is the smallest score such that $Pr(A < i) + .5\ Pr(A = i) < Pr(B \leq u^*(i))$, for scale score i. Equation 1 results in concordant scores on test B corresponding to score points on test A. Although A and B are discrete scores, the equipercentile function is based on continuous approximations of A and B by treating the percentile rank of the scores as the percentage of examinees scoring below the midpoint of the interval containing that score. The equipercentile

function essentially spreads each discrete score point to a range about the score point. A rationale for such a continuization process is presented in Holland & Thayer (1989).

Accepted definitions of percentile rank vary from percent *at or below* a given score (e.g. Hays, 1988), percent *below* a given score (e.g. Thorndike & Hagen, 1986), and *one-half percent at plus percent below* a given score (e.g. Angoff, 1971). The equipercentile function in Equation 1 continuizes the discrete score distribution using a uniform kernel that spreads the density at each score point uniformly in an interval one-half point below and one-half point above the score point (Holland & Thayer, 1989). This is equivalent to the one-half percent at plus percent below definition of percentile rank, and is the continuization procedure that is applied in equipercentile equatings of the ACT. Equipercentile functions for the other definitions of percentile rank can be derived by using different kernels to continuize the distribution. The continuization corresponding to the percent below definition of percentile rank uses a uniform kernel that spreads the density at a point in a unit interval above the point. The continuization corresponding to the percent at or below definition of percentile rank uses a uniform kernel that spreads the density at a point in a unit interval below the point. Changing the definition of percentile rank in the equipercentile function changes the resulting concordances.

Table 1 contains unrounded and rounded concordance results between ACT Composite and SAT I V+M scores for an application of the equipercentile function to examinees within a large institution (N=12,280).[1] In the equipercentile function, percentile rank was defined in three different ways: percent below, the traditional definition of one-half at plus percent below, and percent at or below. ACT Composite scores range from a minimum of 1 to a maximum of 36, by

[1]Results are reported only for ACT to SAT concordances in all tables and figures, and not for SAT to ACT concordances, because of the smaller number of score points to be reported (36 for ACT to SAT versus 121 for SAT to ACT).

an increment of 1. Concordances are not reported for ACT scores below 11 because those scores typically represent chance level on the ACT, or for ACT scores of 36 because nobody in the sample received a 36. SAT I V+M scores range from a minimum of 400 to a maximum of 1600, by increments of 10. The rounded concordant score values in Table 1 differ typically by 30 to 40 SAT score points overall as the definition of percentile rank changes.

See Table 1 at end of report.

Presumably, most applications of the equipercentile function will use a fixed definition of percentile rank, so that concordance results themselves will not be influenced merely by a varying definition. However, if users of a concordance table compare data summaries (i.e., frequency distributions) at concordant score points for their own local group of ACT- and SAT-tested examinees, and use a different definition of percentile rank than that used in creating the concordance table, problems may arise. For example, suppose a high school counselor wanted to determine whether students within the school performed better on the ACT or the SAT I (where many students had taken both tests), so as to advise his or her students to take one or the other. Suppose also that the counselor used the data for all examinees within the school that had taken both the ACT and SAT I and computed the observed score frequency distributions for both tests in terms of the percentile rank at each concordant score point. Further suppose that the counselor defined percentile rank as the percent below in computing the frequency distributions, rather than the less intuitive one-half at plus percent below definition that was used in the equipercentile function to create the concordance table.

Table 2 gives an example of a misapplication of the rounded concordances for the half at plus percent below definition of percentile rank given in Table 1. The "Percent Below ACT Score" and "Percent Below Concordant SAT Score" give the percentage of the 12,280 examinees who scored below the given ACT score, and the percentage who scored below the concordant SAT score, respectively. The difference in percentile rank is computed as the percent below the concordant SAT score minus the percent below the ACT score.

See Table 2 at end of report.

Because of the definition of equipercentile (i.e., two concordant scores have the same percentile rank), if we were unaware of the different definitions of percentile rank used, we would expect that the percentile rank would be the same for both tests at each score point. If differences in percentile rank were observed in a local group, they would therefore be attributed to differential performance for the group on the two tests. Table 2 indicates that the percentage of examinees below is consistently greater at each concordant SAT I score point than at the corresponding ACT score point. The difference in percentile rank appears to favor the ACT because a greater percentage is consistently scoring at or above the concordant score points on the ACT than the SAT. However the differences in percentile rank are spurious, caused by the fact that one definition of percentile rank was used to create the concordance table, while a different definition of percentile rank was used to evaluate the group at the concordant score points.

Figure 1 gives a graphical depiction of the difference in observed percentile rank reported in Table 2, along with the difference in observed percentile rank based on the two alternate definitions of percentile rank. The line labeled "Below" represents the difference in observed

percentile rank (SAT – ACT) as given in Table 2, where percentile rank was defined as percent below. The lines labeled "Half At + Below" and "At Or Below" represent the other definitions of percentile rank. The appropriate comparison to make between observed percentile ranks for ACT and SAT I concordant scores in this case is to define percentile rank in the same manner as in the equipercentile function ("Half At + Below"), which will result in differences around zero. Differences are not exactly zero because of rounding of the concordance results.

See Figure 1 at end of report.

Defining observed percentile rank as percent below appears to favor examinees taking the ACT, while defining observed percentile rank as percent at plus percent below appears to favor examinees taking the SAT. In both cases, the alleged favoritism is merely an artifact of using a different definition of percentile rank in comparing performance on each test at the concordant score points than was used in the equipercentile function to create the concordance table. Because users may unknowingly apply a varying definition of percentile rank in evaluating their local group relative to the concordance table, there is a possibility for misinterpretation of comparisons of performance on each test.

In an equating situation, it is unlikely that anyone would want to make this sort of comparison of results. In a concordance situation, it is entirely possible that someone (such as high school counselors or admissions personnel) might compare the performance of a local group on the two tests via percentile rank. Using the percent below or percent at or below definition of percentile rank in the equipercentile function can also be problematic when comparing group performance, if a different definition of percentile rank is used to make comparisons based on the

concordance results[2]. Because usage of results may differ across equating and concordance situations, careful attention must be paid to the ramifications of extending standard equating procedures to a concordance situation.

To Smooth or Not to Smooth

Equipercentile methods are often employed jointly with smoothing methods in order to reduce the effect of sampling error on the results (Kolen & Brennan, 1995). Sampling error is typically referred to as the "standard error of equating" in an equating context. Smoothing can be applied at either of two times with the equipercentile method. Presmoothing procedures involve smoothing each test score distribution prior to applying the equipercentile function. Postsmoothing procedures involve smoothing the outcome, i.e., the concordant score points. Both presmoothing and postsmoothing methods have been shown to improve the estimation of the equipercentile function by reducing sampling error (e.g., see Hanson, Zeng, & Colton, 1994). Smoothing can also introduce systematic error (i.e., bias) that may result in greater total error (sampling error + systematic error) than no smoothing at all, although that outcome is unlikely.

Table 3 presents smoothed and unsmoothed concordances (rounded) for the equipercentile function applied to ACT Composite and SAT I V+M scores for a large pooled group taking both tests (N=103,525), along with frequencies at each ACT score point. Percentile rank in the equipercentile function was defined as one-half at plus percent below. (Note: All future concordances presented in the paper are based on this definition of percentile rank and all concordance results are rounded unless specified otherwise.) The results in the column labeled "Unsmooth" were presented in Dorans et al. (1997). The results in the column labeled

[2] The occurrence noted in Figure 1 is heightened by a large difference in the number of scale points between the ACT Composite and the SAT I V+M scores. The different definitions of percentile rank tend to be important only for those pairs of tests with differing numbers of scale points.

"Presmooth" are based on a presmoothing of the scale score distributions using a polynomial log-linear model of degree six. The polynomial log-linear model used for smoothing is presented in Holland and Thayer (1987) and Kolen (1991). The results in the columns labeled Postsmooth (.10), Postsmooth (.25), and Postsmooth (.50) are based on a postsmoothing method that applies a cubic spline function to the concordance results, with increasing degrees of a smoothing parameter (.10, .25, or .50). Values of a smoothing parameter between 0 and 1 have commonly been applied in practice. The method is discussed in detail in Kolen (1984) and Kolen and Brennan (1995).

See Table 3 at end of report.

It is important to note that when postsmoothing was used, the cubic spline function was applied to a restricted range of score points, excluding scores with percentile ranks below 0.5 and above 99.5 (which corresponds to ACT scores of 35 or greater and 12 or less in Table 3). A linear interpolation procedure was then used to obtain smoothed score points outside the range of the spline function. There is no restriction against applying the cubic spline function to the entire range of score points; however, smoothing will generally be poor at score points where very few people score. Applying the cubic spline to a restricted range, followed by linear interpolation to smooth the remaining score points is recommended by Kolen (1984) for applications of the postsmoothing procedure. The cubic spline function with linear interpolation is the procedure that is employed in smoothing equating results for the ACT Assessment.

Unshaded score points in Table 3 indicate that the concordant scores are the same across the unsmoothed and four smoothed results. The shaded score points indicate that at least one concordant score differed from the others at that ACT score point. With the exception of some

ten-point differences (equal to one SAT I scale score point) at ACT scores of 32, 15, and 13, results for all methods are essentially the same between ACT scores of 13 and 34. Marked differences in the unsmoothed, presmoothed, and postsmoothed results occur at the extreme score points (above 34 and below 13), corresponding to the range in which the results of the spline function were interpolated. Over all score points, the presmoothing results closely match the unsmoothed results. With such a large sample size (N=103,525) this is not surprising. The postsmoothing results also closely match the unsmoothed results, except at the tails. The postsmoothing method produces different results in the tails because of the linear interpolation that occurs in that region. These results suggest that smoothing is not necessary for a concordance between the ACT and SAT I based on a sample of this size.

This raises the question of what sample sizes in general have standard errors of "equating" such that concordances would require smoothing. Bootstrap standard errors of unsmoothed concordances between ACT Composite and SAT I V+M scores are presented in Figure 2 for sample sizes of 1000, 3000, 5000, 8000, and 103,525. One thousand bootstrap replications were used at each sample size to compute the standard errors. In evaluating the magnitude of the standard errors, we chose to create a standard comparable to the typical standard error of equating for the ACT. Because equating for the ACT is very carefully maintained and monitored each year, this was viewed as an acceptable amount of error to expect. The average standard error of equating for the ACT is about one-third the standard error of measurement for the ACT, and one-third of the standard error of measurement for the SAT I V+M is about 14. Figure 2 shows a horizontal line drawn at a standard error of 14. Only for the sample size of 103,525 did all score points show bootstrap standard errors less than 14. Thus, very large sample sizes are typically needed to concord without smoothing.

See Figure 2 at end of report.

As sample size decreases, the standard errors increase, most notably in the tails, where the magnitude surpasses the value of 14. The large portion of problematic score points occur in the lower tail of the ACT score distribution, at score points where very few examinees typically score. In the sample with 103,525 observations, only 0.14 percent of examinees received a score of 11 or below. For a sample size of 8000, the standard errors surpass 14 only at ACT score points of 11 or lower. Thus, if those score points were not reported in a concordance table, it might not be necessary to smooth results based on samples of this size. (Concordances for scores below 11 were not reported in Dorans et al. (1997), because those are typically chance-level scores on the ACT.) As sample size decreases from 8000 to 1000, many more score points within the range of 11-36 show standard errors greater than 14, suggesting the need to employ smoothing at the smaller sizes.

Computing Concordance Standard Errors

A computer program used to equate the ACT was used in this study to compute ACT-SAT concordant score points. In addition to providing equipercentile equivalents, the program outputs standard errors of equating. The standard errors are estimated based on Lord's (1982) analytic derivation of the standard error of equating for the equipercentile method applied to discrete score points under a random (independent) groups design. Unfortunately, naive application of the software results in an inappropriate measure of the standard error because the sample for the ACT-SAT data is not independent, so that the true standard errors would actually be less than the magnitude indicated by the output standard errors. Lord (1982) offers an alternative analytic

derivation for the equipercentile method applied to discrete score points under a single group design, which would be appropriate in our concordance situation.

Bootstrap methods also provide an alternative way of suitably estimating the standard error for any data collection design. The bootstrap method utilizes resampling procedures, and calculates the standard deviation of the statistic of interest (in this case, the concordant scores) over the samples drawn. Both bootstrap methods and analytic procedures for computing standard errors are discussed in Kolen and Brennan (1995). One advantage of bootstrap methods over analytic procedures is that the summary statistics may be computed for either rounded or unrounded concordance results, whereas the analytic derivations reported in Lord (1982) are based on unrounded concordance results only. It is important to look at standard errors for rounded results, since only the rounded results are reported. Also, groups to be linked may be less clearly defined in a concordance situation than in an equating situation. In applications of equipercentile procedures to link statewide assessments to NAEP (Ercikan, 1997; Linn & Kiplinger, 1995), the groups are probably not independent, and probably not completely overlapping either. Using independent groups or single group analytic standard errors in this situation could give misleading impressions about the magnitude of the standard errors. Bootstrap procedures should provide more accurate estimates of standard errors than analytic procedures when concordance groups are ill defined.

Figure 3 shows analytic and bootstrap standard errors for unsmoothed concordances between ACT Composite and SAT I V+M scores. The analytic standard errors are computed from the observed data (N=103,525), while the bootstrap standard errors are computed from 1000 repeated computations of concordances for sample sizes of 103,525 drawn with replacement from the original pooled sample. Note that the analytic standard errors (which come from our

equipercentile computer program) are based on the assumption of two independent groups, and are thus incorrectly defined for our sample. Also, the analytic standard errors are summarized only for the unrounded concordance results, whereas the bootstrap standard errors are summarized for both the unrounded and rounded concordance results.

See Figure 3 at end of report.

A comparison of the analytic and bootstrap standard errors for the unrounded concordances, shows that the analytic standard errors (unrounded) are consistently greater than the bootstrap standard errors (unrounded) for ACT score points of 10-36. This suggests that if analytic methods are to be applied for the computation of standard errors, the appropriate derivation for the group(s) at hand should be employed. A comparison of the bootstrap standard errors for the unrounded and rounded concordances shows less consistent trends for the rounded concordances: in some ranges, the standard errors of the rounded concordances are substantially greater than the standard errors for the unrounded concordances, at other times, they are zero. Clearly the standard errors based on unrounded and rounded concordances can give a different impression of the magnitude of the sampling error. Bootstrap standard errors are perhaps more flexible than analytic standard errors in a concordance situation, because they offer the opportunity to evaluate standard errors for both unrounded and rounded concordance results and are applicable to any group design.

Trouble-Shooting for Applications with Smaller Samples: Reporting Decisions for Institutions

For the ACT-SAT concordances based on the pooled sample, we had a very large sample (N=103,525), which eliminated a lot of sparse data problems that would typically need to be

addressed in a concordance situation. Creation of concordance tables for individual institutions participating in the study (based on much smaller sample sizes) invoked problems related to zero frequencies at the tails of the score distributions, gaps within the score distributions, and sparse data in general. As sample sizes decreased, the problems became more prevalent. These problems can also occur with raw scores in an equating situation, but may be exacerbated in a concordance situation because the scale score distributions differ across the two tests. Such data problems can result in individual score points for which resulting concordances are quite unstable.

Because users may be inclined to treat any reported concordant scores as interchangeable, even when cautioned against employing such an interpretation, practitioners may choose to restrict the information given to users so that misuse of concordance results is at a minimum. The effect of data problems on concordance results should be carefully examined, so that informed decisions can be made about what to report to users. Reporting decisions may, in part, be determined by the intended use of the concordance table.

This section of the paper shifts in focus from creating concordances and their accompanying statistics to reporting concordance results. We present four examples of institutional-level concordances, discuss the reporting decisions we made for each institution, and examine how those decisions were driven by the underlying data. It should be kept in mind throughout the discussion that concordance tables provide comparable decision points at a *group* level, and are not intended to predict student performance at an *individual* level. If the data for an institution do not support reporting concordances at certain score ranges, it is likely because of low frequencies in those score ranges, which suggests that the institution should not need to establish group-level comparable scores in those ranges.

Sample Size and Reporting Decisions. Sample sizes for the 14 institutions and two states that contributed data to the recent large-scale ACT-SAT concordance study (Dorans et al., 1997) are, in order of size from smallest to largest: 49; 868; 1,179; 1,724; 1,774; 2,276; 2,385; 3,473; 4,555; 5,536; 5,849; 8,076; 8,354; 12,280; 21,592; and 23,555. Separate concordances were computed for all groups (including the pooled group), except the institution with 49 observations. Only the pooled group concordances are presented in Dorans et al. (1997). Kolen and Brennan (1995) suggest that sample sizes of about 1,500 will result in acceptable standard errors for equipercentile procedures applied to random groups in an equating situation. We chose to apply equipercentile procedures for some institutions with sample sizes less than 1,500.

Although some sample sizes were smaller than recommended for equating, we compensated by imposing strict restrictions on what score points would be reported to an institution. Namely, decisions to report ACT to SAT concordances for individual score points were made by evaluating bootstrap standard errors (based on 1000 bootstrap replications) for each score point. All score points were included in the application of the equipercentile function, but ACT to SAT concordances for a given score point were reported only if the standard error was less than or equal to one-third the standard error of measurement for the test being linked. As discussed earlier, this created a standard for evaluation for the standard errors comparable to what would be expected in a carefully maintained equating situation (i.e., the ACT).[3]

The pooled ACT-SAT concordances may not be appropriate for use by institutions that differ greatly from the pooled sample, either in terms of score distributions or demographically. For that reason, we might consider performing a unique concordance based on a sample size as

[3] Decisions to report SAT to ACT concordances (not presented here) were made based on the ACT score points reported in the ACT to SAT concordances. Namely, every SAT score point corresponding to an ACT score point reported in the ACT to SAT concordance was reported in the SAT to ACT concordance. Because there are many more SAT I V+M score points than ACT score points, multiple SAT points may correspond to the same ACT score point in the SAT to ACT concordance.

small as 500, if the institution differed greatly from the pooled sample. If the pooled concordances adequately represent the relationship between scores that exists for an institution, the institution could use the pooled concordances to determine comparable score points. If not, a unique concordance may be more suitable for that institution, even if based on a fairly small sample, as long as restrictions block the reporting of unstable score points. The restrictions may mean that only a portion of possible score points be reported for that institution, but the score points to be reported would be driven by the data. Namely, score points would be reported only where the data support them. If too few score points are reported to be useful, more data should be collected prior to running the concordance.

Example 1. Use of the standard error requirement to make reporting decisions is demonstrated in Table 4, which shows the unsmoothed concordances between ACT Composite and SAT I V+M scores for an institution with a sample size of 868. The sample for this institution was much more academically able than the pooled ACT-SAT sample. The average ACT Composite score was 29.2 versus 23.2 for the pooled sample, while the average SAT I V+M score was 1320.4 versus 1071.4 for the pooled sample. Table 4 also gives examinee frequencies, bootstrap standard errors for rounded concordances (labeled "SE"), and reporting decisions at each ACT score point between 10 and 36. The shading highlights scores of 10-11, both of which show acceptable standard errors, but frequencies of zero. Unsmoothed results are reported to facilitate comparison with the pooled results, which are also unsmoothed. When reporting results for an institution, we would not report scores below 11 (i.e., scores at chance level on the ACT) to match what was reported in the pooled ACT-SAT concordance (Dorans et al., 1997); results for an ACT Composite of 10 are included for demonstration. By our standard error requirement alone (≤ 14), we would conclude that concordances for ACT scores of 12-20 should not be reported, while

scores of 10-11 and 21-36 should be reported. However, nobody in the sample received scores of 10 or 11, thus the acceptable standard errors at those score points are probably misleading.

See Table 4 at end of report.

Because standard errors can be deceiving at score points where zero or few examinees scored (usually at the tails), we adopted a stronger standard that allowed reporting only for *contiguous* score points that met the standard error requirement. Beginning with the first score point not meeting the standard error requirement, all other score points above (if the score is in the high tail) or below (if the score is in the low tail) would not be reported, even if an individual score above (or below) that point meets the standard error requirement. Applying this criterion to the institution in Table 4, scores of 10, 11, and 21-36 all meet the standard error requirement. Of those points, only 21-36 are contiguous. Thus, by the contiguous standard error requirement, we would only report scores between 21 and 36.

The concordance results of Table 4 are plotted relative to the pooled sample concordances (unsmoothed also) in Figure 4. The solid plotted line represents the pooled concordances, while the institution's results are plotted at each ACT score point with error bars representing one standard error about the observed institution concordance.[4] The vertical line through the plot splits the ACT score points into sections of points to report and points not to report, corresponding to the reporting decision given in Table 4. The section to report is labeled "Report." Clearly the concordance estimates at the lower tail of the ACT score scale shown are much less stable than

[4] The institution's result ± one standard error intersects the pooled group results only if the vertical portion of the error bars intersects the pooled line, *not* the horizontal portion.

score points in the middle and upper tail, leading to our decision not to report ACT-SAT concordances at those score points for this institution.

See Figure 4 at end of report.

Unfortunately, the lower tail is also the point where the concordance results appear to be most different from the pooled sample concordances, which if the differences were legitimate, is where a unique institution concordance would be most helpful. But the results are largely unstable because there were not enough examinees scoring at each point less than 21 (see individual cell frequencies in Table 4). Because so few people scored at those points, the data do not support reporting results at those score points. If the distribution of ACT scores in this sample is representative of the distribution for the institution's applicant pool, the institution is likely to be making admissions decisions at a higher level than at the score points we chose not to report, where more of the applicants fall. If the institution would like to make decisions in a range for which no score points are reported, it would be necessary to collect more data so that the lower score points are better represented.

Example 2. Table 5 presents unsmoothed concordances for an institution with 1,179 observations. The table gives examinee frequencies, unsmoothed concordances, unsmoothed concordances with linear interpolation at the extremes, postsmoothed concordances with .10 smoothing parameter (with linear interpolation), bootstrap standard errors computed for rounded, unsmoothed concordances, and reporting decisions at each ACT score point. Linear interpolation was applied to unsmoothed results only for demonstration; the shading indicates score points at which linear interpolation occurs (i.e., score points where the percentile rank was greater than 99.5 or less than 0.5). Normally, we would consider smoothing for small sample sizes (say less than

10,000), and would investigate different smoothing and non-smoothing alternatives before choosing an appropriate solution. However, Table 5 shows that careful attention must be paid to the results of a smoothing, particularly if they are to be compared to the results based on the pooled sample.

See Table 5 at end of report.

Concordance results in the "Unsmooth" and "Unsmooth (with interpolation)" columns are the same in the unshaded rows, because neither results were smoothed or interpolated at those points. In the shaded rows where the interpolation was applied, however, the results for the "Unsmooth" and "Unsmooth (with interpolation)" columns are quite different. The similarity of results in the shaded rows for postsmoothing and no smoothing with linear interpolation indicate that the differences we see between the unsmoothed (with no interpolation) and postsmoothed concordances in the shaded rows are largely due to the interpolation, not the smoothing. A similar effect between unsmoothed and postsmoothed results was demonstrated in Table 3 for the pooled sample.

If individual institution results are to be compared to the pooled sample results, applying postsmoothing procedures to the institutional concordances could lead an institution to mistakenly conclude that they were quite different from the pooled group at the tails. However, these differences could occur largely because of the application of linear interpolation to smoothed results for the institution, but no linear interpolation for unsmoothed results for the pooled sample. The intended use of results should be taken into consideration when choosing procedures to apply in practice. If results from one application are to be compared to results from another application (i.e., across years, or from one group to another), it might be practical to follow the same

smoothing procedures, so that observed differences in any reported score points can be interpreted as real and not due to procedural differences. This may mean choosing to ignore certain procedures that would normally be viewed as appropriate. If results are to be stand-alone, then the practitioner is free to choose the procedures that are most appropriate for the problem at hand.

The unsmoothed concordance results of Table 5 are plotted relative to the pooled sample concordances (unsmoothed also) in Figure 5. The solid plotted line represents the pooled concordances, while the institution's results are plotted at each ACT score point with error bars representing one standard error about the observed institution concordance. The vertical lines through the plot split the ACT score points into sections of points to report and points not to report, corresponding to the reporting decisions applied in Table 5. Concordances for this institution are fairly similar to the pooled concordances, beginning to diverge in the right tail at about an ACT score of 27. By our contiguous standard error requirement, we would not report concordances for ACT scores below 14 or above 28. If we didn't require contiguity, we might be led to believe we could report results for ACT scores of 33 and 34 because the standard errors are 0.00. Examination of frequencies at each score point show that these standard errors are deceptive because of the zero frequencies that occur at scores of 33-35.

See Figure 5 at end of report.

A comparison with the shaded rows in Table 5 indicates that our standard error requirement for reporting acts in nearly the same score region as the linear interpolation, and may be considered a proxy of sorts for interpolating, albeit slightly more conservative. Because smoothings that work well where the vast majority of the data are do not necessarily extrapolate well outside the bounds of the majority of the data (Dorans et al., 1997), we chose not to report

extreme score points rather than perform linear interpolation there. The standard error requirement restricts the information that is reported, but the restriction is driven by what is observed in the data. The N-counts in Table 5 show that very few people scored at the points that are not reported.

Example 3. The occurrence of small frequencies at extreme score points is typically less of a problem as the concordance sample size increases. Increasing the concordance sample size will in general reduce standard errors, but will not necessarily result in all score points being reported. Figure 6 presents the unsmoothed concordance results for a high-performing institution with 5,849 observations. The average ACT Composite for this institution was 28.2 versus 23.2 for the pooled sample, while the average SAT I V+M was 1263.7 versus 1071.4 for the pooled sample. Only two examinees scored below a 16 in this sample. As a general rule, standard errors are smaller than observed in Figures 4 and 5 (except in the lower tail), particularly in the middle and upper scores. Although this example used a much larger sample size than Examples 1 and 2, our standard error requirement would still have us report concordances only for scores of 18 or higher.

See Figure 6 at end of report.

Example 4. Even for seemingly adequate sample sizes, careful attention still needs to be paid to the data and the concordance results. Larger sample sizes may not necessarily enable the reporting of all score points. Figure 7 presents the unsmoothed concordance results for an institution with 12,280 observations. The concordances for this institution are given in Table 1 (Half At + Percent Below definition). According to the plot, concordances would not be reported for ACT scores of 12 or less, or for a score of 36. The "Don't Report" decision for scores less

than 13 is based on the standard error requirement. The "Don't Report" decision for a score of 36 is based on the fact that no examinee received a score of 36 in this sample.

See Figure 7 at end of report.

Although the standard error is acceptable at 36, we prefer not to report concordances for score points higher (or lower) than the highest (or lowest) observed score. Assuming the contiguous standard error requirement was met, we would report concordances only for the highest and lowest ACT scores observed in a sample, along with all scores in between, even if some in-between score points have zero frequencies. The equipercentile procedure as implemented in our computer program has procedures in place to handle cases of zero frequencies at any score point. Whatever the underlying score distributions, whether skewed or sparse in general, the computer program can churn out answers that on the surface look acceptable. Closer examination of the data, however, may yield evidence of unstable results that might be misleading if reported.

Discussion

This paper was intended to be instructional, walking the reader through issues that were considered in the process of creating concordance tables between ACT and SAT I scores. With regard to the issues considered, it covers much of the ground already covered in extensive research on equating procedures, but from a much different perspective. Because concordance situations are typically not as controlled or well-designed as equating situations, it is important that these procedures be researched under conditions relevant to concordance situations. The examples convey that we cannot take for granted that procedures or reporting decisions appropriate for an equating situation are also appropriate for a concordance situation. Also, each pair of tests to be

linked must be considered to be a unique situation, for which these issues must be considered anew[5].

In the creation of the concordance tables, the reporting decisions we made were very conservative. Reporting decisions were in part driven by the fact that concordances between ACT and SAT I scores are increasingly used by institutions to make high stakes decisions. Users without full knowledge of the procedures used and assumptions made in creating the concordance tables may blindly use what they are given, even if given guidelines for interpreting results. Rather than take the chance of misuse or misunderstanding of results, we chose to report only score points for which results were acceptably stable. Some users may still be inclined to treat the ACT and SAT I scores as interchangeable, despite our remonstrances against such a practice. By eliminating unstable results from the reported tables, we feel more comfortable in making the information available to users. The different issues demonstrated in each example lead us to conclude that we need to closely examine the data and results for each concordance in order to determine which procedures to use and what results to report.

Although the examples presented here were limited to an ACT-SAT linkage, the issues raised are relevant to any concordance based on equipercentile methods, when two distinct tests are to be related. It is the responsibility of practitioners to make appropriate choices for screening data, procedures to apply, and results to report in concordance situations. The choices will largely be driven by the specific pair of tests to be linked, particularly the characteristics of the data, and the intended uses of the results. Clearly, the issues discussed here may not be relevant to all concordance situations, and there are issues not considered here that might arise in concording

[5] The ACT Composite and SAT I V+M concordance is itself a unique situation, because of the strong correlation between the two tests (~.96 after correcting for attenuation). Attempting to concord less highly related tests might cause additional problems than those presented here. This paper focuses on the act of creating concordances, and makes no distinction about when other classes of linking might be more appropriate. For discussions along that nature, refer to Dorans and Lyu (1999) and Hanson, Harris, Pommerich, Sconing, and Yi (1999).

other pairs of tests. The discussion is provided merely to raise the level of awareness of the types of decisions that need to be made for each pair of tests that are to be related.

References

Angoff, W.A. (1971). Scales, norms, and equivalent scores. In R.L. Thorndike (Ed.), *Educational measurement* (2nd ed., pp. 508-600). Washington, DC: American Council on Education.

Dorans, N.J, & Lyu, C.F. (April, 1999). *Distinctions among classes of linkages*. Paper presented at the annual meeting of the National Council on Measurement in Education, Montreal, Quebec, Canada.

Dorans, N.J., Lyu, C.F., Pommerich, M., & Houston, W.M. (1997). Concordance between ACT Assessment and recentered SAT I sum scores. *College & University*, *73*, 24-34.

Ercikan, K. (1997). Linking statewide tests to the National Assessment of Educational Progress: Accuracy of combining test results across states. *Applied Measurement in Education*, *10*, 145-159.

Hanson, B.A., Harris, D.J., Pommerich, M., Sconing, J., & Yi, Q. (April, 1999). *Suggestions for the evaluation and use of concordance results*. Paper presented at the annual meeting of the National Council on Measurement in Education, Montreal, Quebec, Canada.

Hanson, B.A., Zeng, L., & Colton, D. (1994). *A comparison of presmoothing and postsmoothing methods in equipercentile equating* (Research Report No. 94-4). Iowa City, IA: American College Testing.

Hays, W.L. (1988). *Statistics*. New York: Harcourt Brace Jovanovich College Publishers.

Holland, P.W., & Thayer, D.T. (1987). *Notes on the use of log-linear models for fitting discrete probability distributions* (Research Report No. 87-31). Princeton, NJ: Educational Testing Service.

Holland, P.W., & Thayer, D.T. (1989). *The kernel method of equating score distributions* (Research Report No. 89-7). Princeton, NJ: Educational Testing Service.

Kolen, M.J. (1984). Effectiveness of analytic smoothing in equipercentile equating. *Journal of Educational Statistics*, *9*, 25-44.

Kolen, M.J. (1991). Smoothing methods for estimating test score distributions. *Journal of Educational Measurement*, *28*, 257-282.

Kolen, M.J., & Brennan, R.L. (1995). *Test equating: Methods and practices*. New York: Springer.

Linn, R.L. (1993). Linking results of distinct assessments. *Applied Measurement in Education*, *6*, 83-102.

Linn, R.L., & Kiplinger, V.L. (1995). Linking statewide tests to the National Assessment of Educational Progress: Stability of results. *Applied Measurement in Education*, *8*, 135-155.

Lord, F.M. (1982). The standard error of equipercentile equating. *Journal of Educational Statistics*, *7*, 165-174.

Marco, G.L., & Abdel-Fattah, A.A. (1991). Developing concordance tables for scores on the Enhanced ACT Assessment and the SAT. *College & University*, *66*, 187-194.

Mislevy, R.J. (1992). Linking educational assessments: Concepts, issues, methods, and Prospects. Princeton, NJ: ETS Policy Information Center.

Thorndike, R.L., & Hagen, E.P. (1986). *Measurement and evaluation in psychology and education* (4th ed.). New York: Macmillan.

TABLE 1

Unrounded and Rounded Concordances Between ACT Composite and SAT I V+M Scores for an Institution with N=12,280, Based on Three Different Definitions of Percentile Rank.

	Unrounded Concordances			Rounded Concordances		
ACT Composite	Percent Below	Half At + Percent Below	Percent At Or Below	Percent Below	Half At + Percent Below	Percent At Or Below
35	1580.00	1590.00	1600.00	1580	1590	1600
34	1509.44	1530.63	1570.00	1510	1530	1570
33	1464.00	1479.17	1499.44	1460	1480	1500
32	1420.24	1435.71	1454.00	1420	1440	1450
31	1378.43	1392.38	1410.24	1380	1390	1410
30	1334.46	1349.71	1368.43	1330	1350	1370
29	1297.87	1310.87	1324.46	1300	1310	1320
28	1261.43	1273.18	1287.87	1260	1270	1290
27	1222.65	1236.07	1251.43	1220	1240	1250
26	1185.22	1199.17	1212.65	1190	1200	1210
25	1150.00	1162.79	1175.22	1150	1160	1180
24	1109.07	1124.23	1140.00	1110	1120	1140
23	1073.49	1086.83	1099.07	1070	1090	1100
22	1032.54	1048.13	1063.49	1030	1050	1060
21	993.45	1008.49	1022.54	990	1010	1020
20	950.86	967.32	983.45	950	970	980
19	907.75	926.69	940.86	910	930	940
18	865.51	881.83	897.75	870	880	900
17	818.88	838.94	855.51	820	840	860
16	773.27	793.29	808.88	770	790	810
15	714.00	743.59	763.27	710	740	760
14	653.33	684.84	704.00	650	680	700
13	596.00	626.58	643.33	600	630	640
12	510.00	561.25	586.00	510	560	590
11	480.00	496.67	500.00	480	500	500

TABLE 2

Difference in Percentile Rank (Defined as Percent Below SAT – Percent Below ACT) for Concordant ACT Composite and SAT I V+M Score Points.

ACT Composite	Percent Below ACT Score	Concordant SAT I V+M	Percent Below Concordant SAT Score	Difference in Percentile Rank (SAT-ACT)
35	99.9	1590	99.9	0.0
34	99.5	1530	99.7	0.2
33	98.7	1480	99.1	0.4
32	97.4	1440	98.0	0.6
31	95.0	1390	95.7	0.7
30	91.5	1350	92.9	1.4
29	87.3	1310	88.8	1.5
28	82.3	1270	83.6	1.3
27	75.8	1240	78.9	3.1
26	68.1	1200	71.1	3.0
25	60.6	1160	62.7	2.1
24	52.4	1120	54.6	2.2
23	44.5	1090	48.0	3.5
22	35.7	1050	39.5	3.8
21	28.1	1010	31.3	3.2
20	21.2	970	24.3	3.1
19	15.4	930	18.1	2.7
18	10.6	880	12.1	1.5
17	6.7	840	8.3	1.6
16	4.0	790	4.9	0.9
15	2.1	740	2.8	0.7
14	0.8	680	1.2	0.4
13	0.3	630	0.5	0.2
12	0.1	560	0.2	0.1
11	0.0	500	0.0	0.0

TABLE 3

Smoothed and Unsmoothed Concordances (Rounded) Between ACT Composite and SAT I V+M Scores for a Large Pooled Group (N=103,525).

ACT Composite	N	Unsmooth	Presmooth	Postsmooth (.10)	Postsmooth (.25)	Postsmooth (.50)
36	24	1600	1600	1590	1590	1590
35	187	1580	1570	1550	1550	1550
34	611	1520	1520	1520	1520	1520
33	1345	1470	1470	1470	1470	1470
32	2126	1420	1430	1420	1430	1430
31	3059	1380	1380	1380	1380	1380
30	4081	1340	1340	1340	1340	1340
29	4662	1300	1300	1300	1300	1300
28	5342	1260	1260	1260	1260	1260
27	6109	1220	1220	1220	1220	1220
26	6709	1180	1180	1180	1180	1180
25	6862	1140	1140	1140	1140	1140
24	7346	1110	1110	1110	1110	1110
23	7491	1070	1070	1070	1070	1070
22	7558	1030	1030	1030	1030	1030
21	7269	990	990	990	990	990
20	6980	950	950	950	950	950
19	6359	910	910	910	910	910
18	5544	870	870	870	870	870
17	4699	830	830	830	830	830
16	3429	780	780	780	780	780
15	2579	740	740	740	740	730
14	1676	680	680	680	680	680
13	953	620	630	620	620	630
12	375	560	560	590	590	590
11	107	500	500	570	570	580
10	34	450	440	560	560	560
9	3	410	400	540	540	540
8	5	410	400	520	520	520
7	0	400	400	510	510	510
6	1	400	400	490	490	490

TABLE 4

Unsmoothed Concordances Between ACT Composite and SAT I V+M Scores for an Institution with N=868.

ACT Composite	N	Concordant SAT I V+M	SE	Reporting Decision
36	1	1600	2.06	Report
35	17	1580	7.14	Report
34	44	1520	5.83	Report
33	104	1480	5.30	Report
32	107	1420	5.59	Report
31	85	1380	5.91	Report
30	100	1340	4.98	Report
29	88	1310	5.56	Report
28	84	1270	5.36	Report
27	60	1230	7.42	Report
26	42	1200	6.83	Report
25	44	1160	10.04	Report
24	31	1110	8.53	Report
23	23	1080	11.18	Report
22	13	1040	9.94	Report
21	8	1020	13.01	Report
20	3	970	26.51	Don't Report
19	3	930	29.30	Don't Report
18	6	900	29.20	Don't Report
17	1	800	45.51	Don't Report
16	1	790	36.65	Don't Report
15	1	780	28.24	Don't Report
14	1	740	30.16	Don't Report
13	0	730	30.22	Don't Report
12	1	720	24.15	Don't Report
11	0	500	12.06	Don't Report
10	0	490	0.00	Don't Report

TABLE 5

Unsmoothed and Smoothed Concordances Between ACT Composite and SAT I V+M Scores for an Institution with N=1,179.

ACT Composite	N	Unsmooth	Unsmooth (with interpolation)	Postsmooth (.10)	SE	Reporting Decision
36	1	1520	1580	1580	33.60	Don't Report
35	0	1520	1540	1540	14.40	Don't Report
34	0	1520	1490	1490	0.00	Don't Report
33	0	1520	1440	1440	0.00	Don't Report
32	1	1480	1400	1400	14.47	Don't Report
31	4	1330	1350	1350	17.36	Don't Report
30	6	1310	1310	1300	15.81	Don't Report
29	6	1270	1270	1270	16.22	Don't Report
28	21	1230	1230	1230	11.86	Report
27	18	1190	1190	1200	7.43	Report
26	29	1170	1170	1170	6.93	Report
25	49	1140	1140	1130	5.03	Report
24	69	1100	1100	1100	7.02	Report
23	74	1060	1060	1060	5.62	Report
22	114	1030	1030	1020	6.31	Report
21	137	980	980	980	4.82	Report
20	133	940	940	940	4.54	Report
19	125	900	900	900	5.14	Report
18	126	860	860	870	4.66	Report
17	103	830	830	830	4.68	Report
16	80	780	780	780	6.72	Report
15	51	740	740	740	8.06	Report
14	19	680	680	680	13.52	Report
13	7	610	610	620	24.23	Don't Report
12	3	550	600	600	35.03	Don't Report
11	3	530	580	580	36.94	Don't Report

FIGURE 1. Difference in Observed Percentile Rank Between ACT Composite and Concordant SAT I V+M Score Points for Three Definitions of Percentile Rank; Percentile Rank in Equipercentile Function is Defined as One-Half Percent at Plus Percent Below.

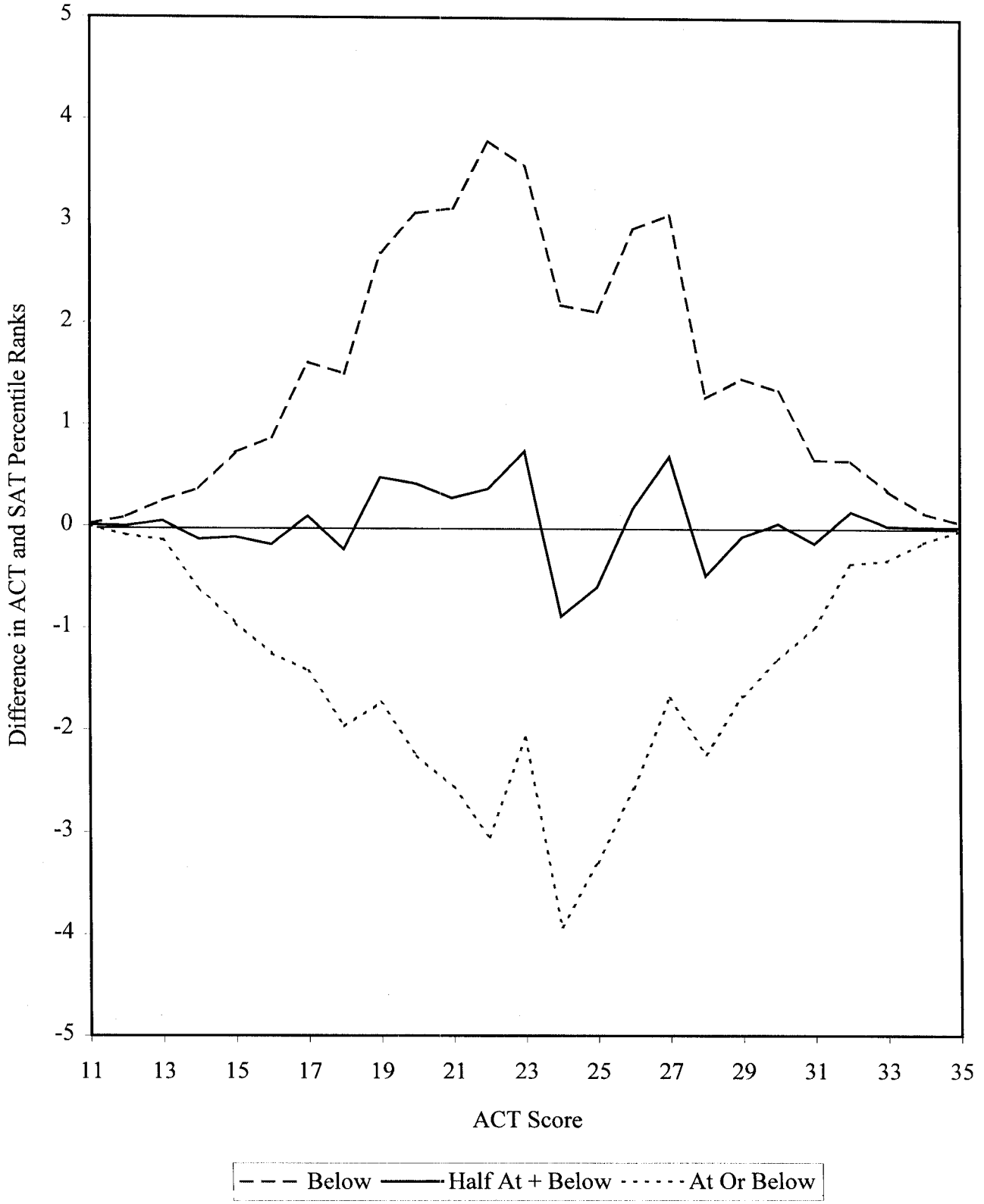


FIGURE 2. Bootstrap Standard Errors for Unsmoothed Concordances Between ACT Composite and SAT I V+M Scores, by Sample Size.

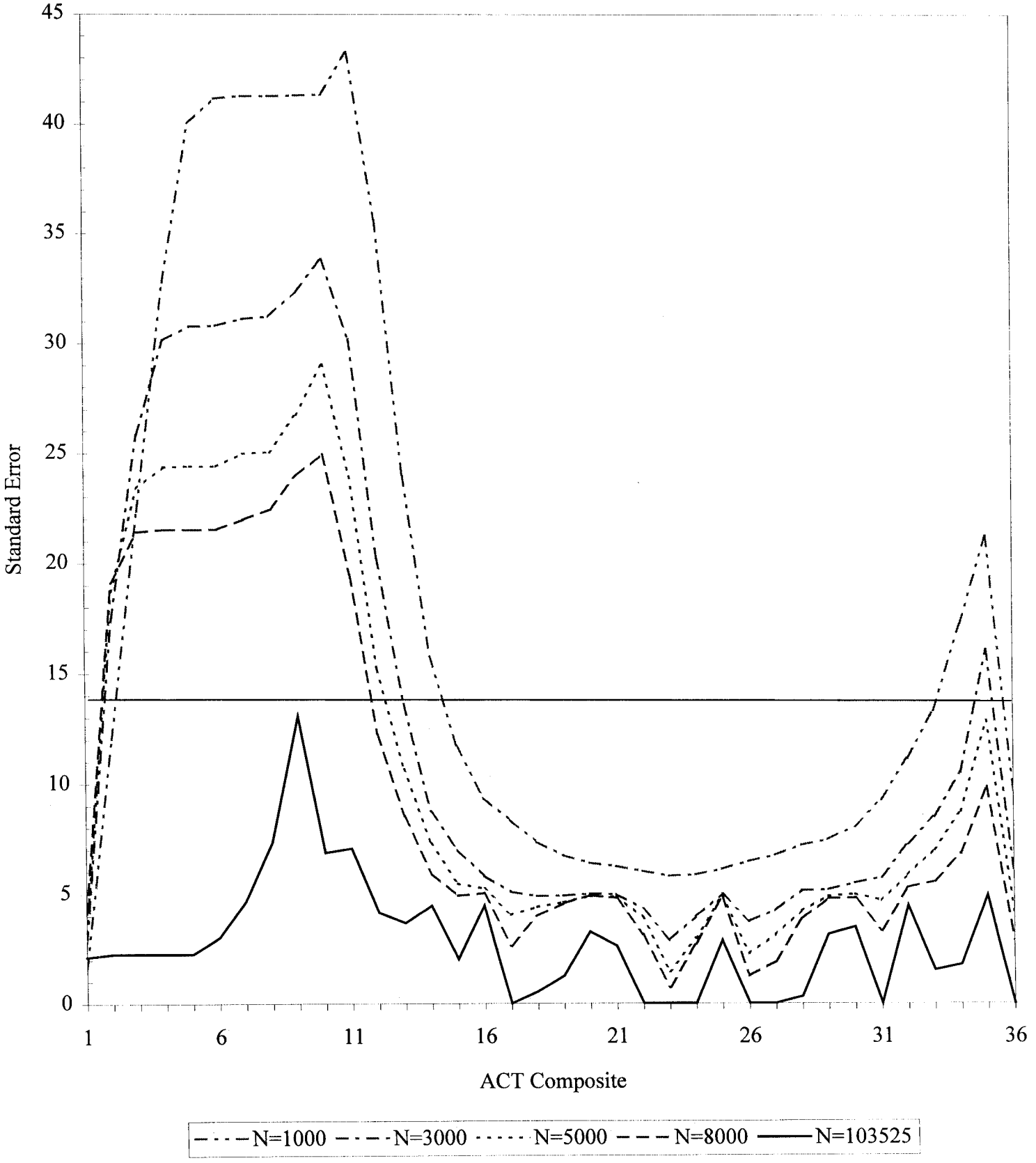
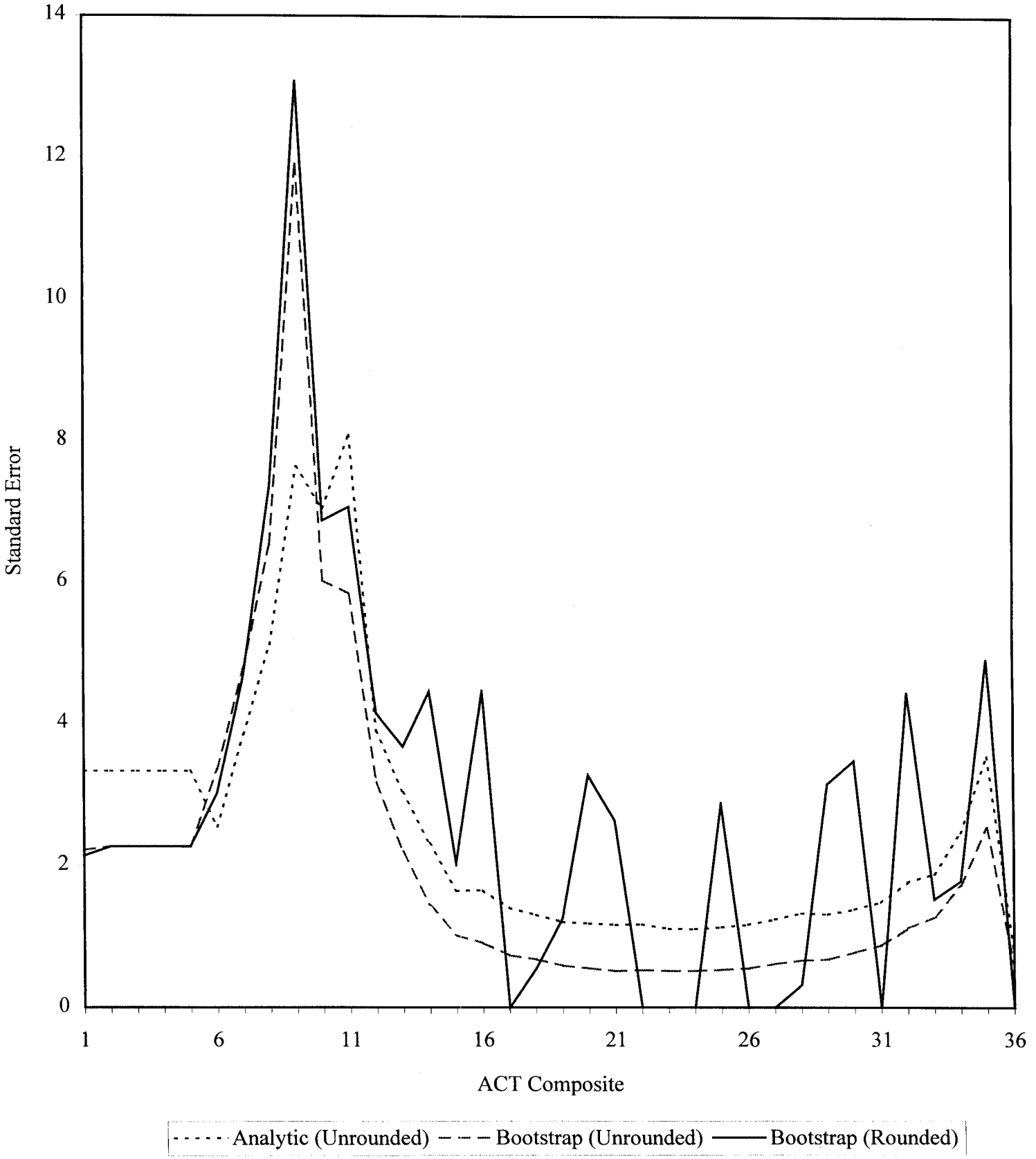


FIGURE 3. Analytic and Bootstrap Standard Errors for Unsmoothed ACT Composite to SAT I V+M Concordance.*



* The analytic standard errors assume independent groups, and are incorrect.

FIGURE 4. Unsmoothed Concordances Between ACT Composite and SAT I V+M Scores for an Institution with N=868.

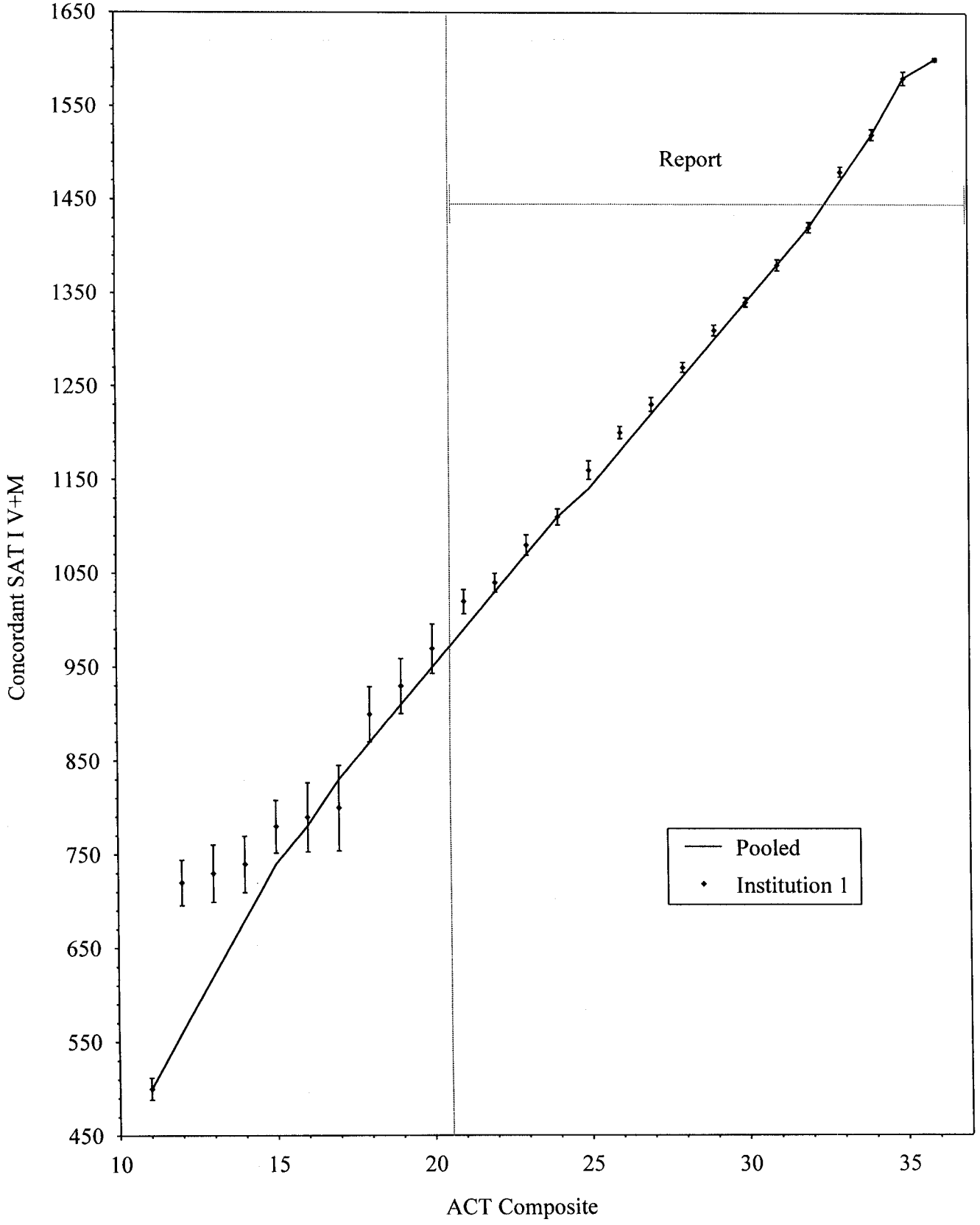


FIGURE 5. Unsmoothed Concordances Between ACT Composite and SAT I V+M Scores for an Institution with N=1,179.

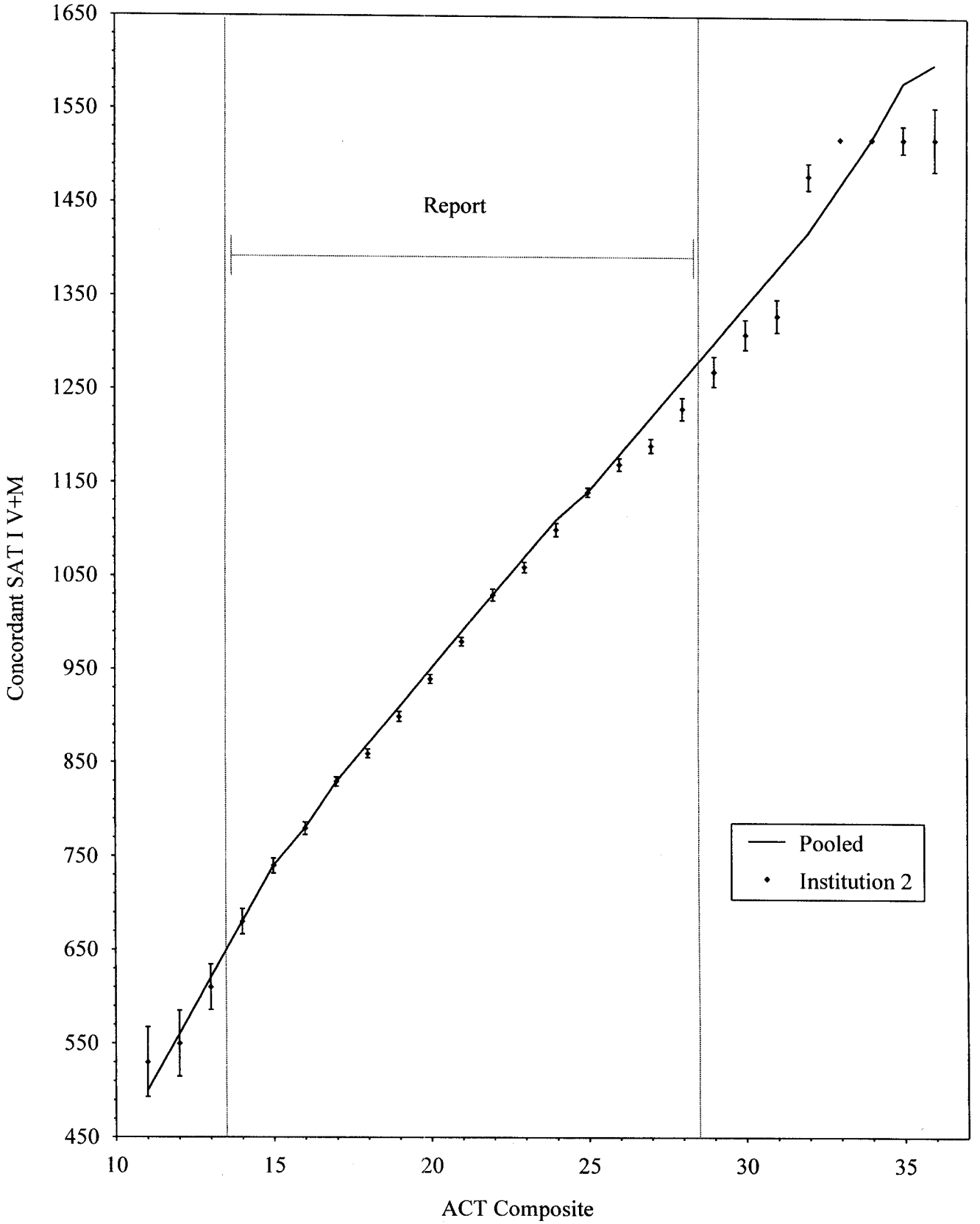


FIGURE 6. Unsmoothed Concordances Between ACT Composite and SAT I V+M Scores for an Institution with N=5,849.

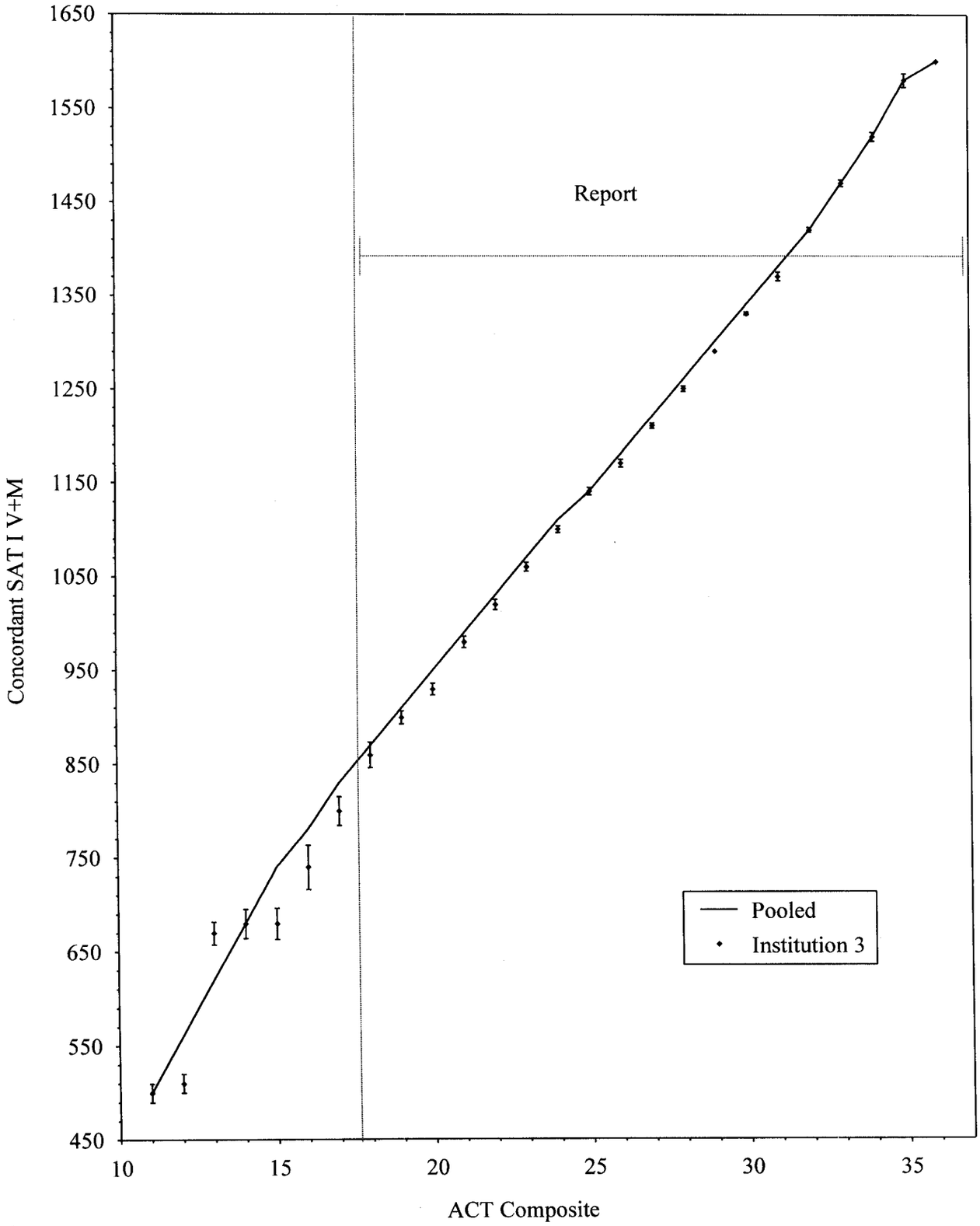
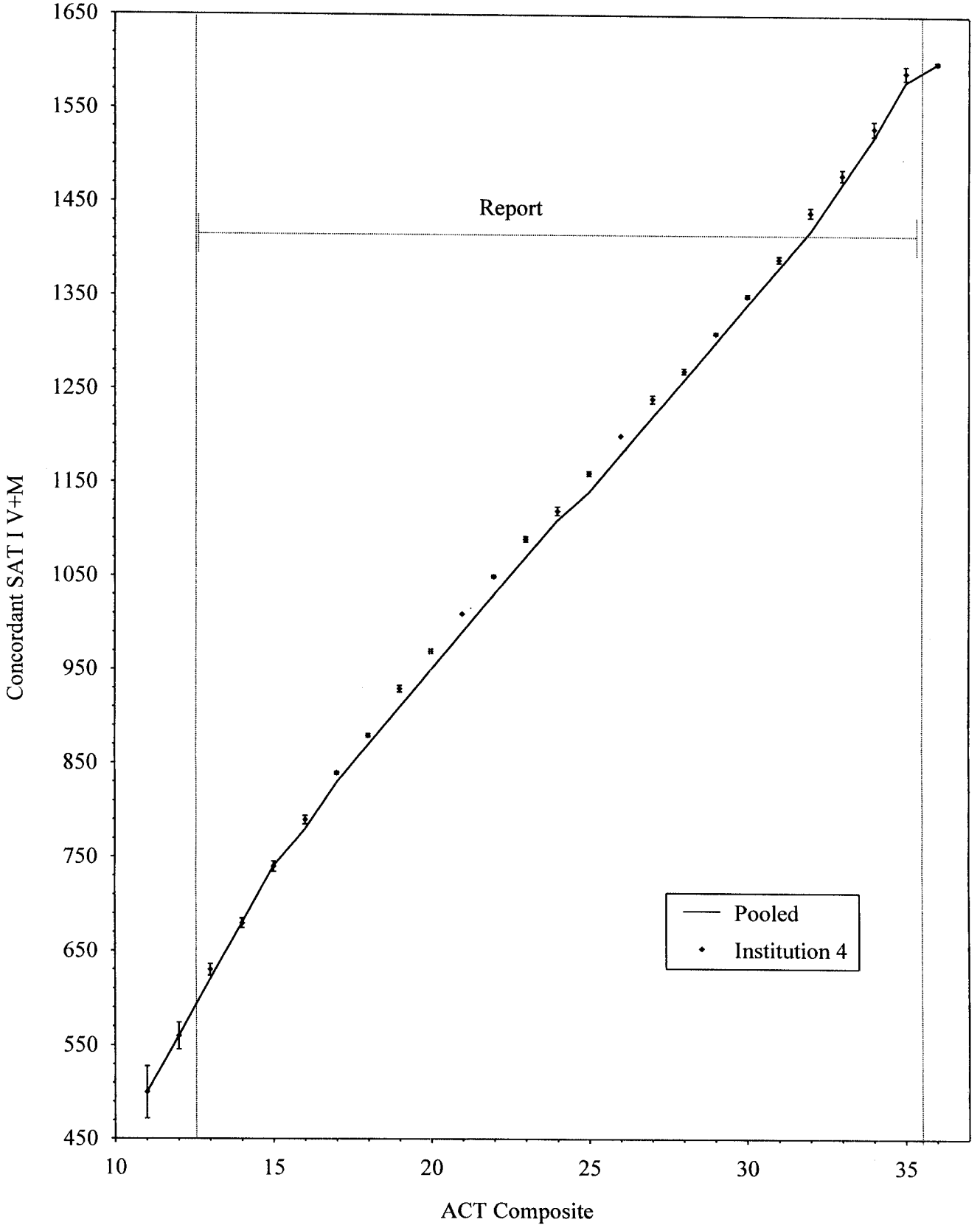



FIGURE 7. Unsmoothed Concordances Between ACT Composite and SAT I V+M Scores for an Institution with N=12,280.