CENTER FOR LAW AND EDUCATION

99 Chauncy Street, Suite 402
Boston, MA 02111
(617) 451-0855
FAX: (671) 451-0857

Reply to:
1875 Connecticut Ave., NW, Suite 510
Washington, DC 20009
(202) 986-3000
FAX: (202) 986-6648

COMMENTS ON PROPOSED RULE FOR TITLE I

34 CFR Part 200
73 Federal Register 22020
April 23, 2008

The Center for Law and Education, a national organization with offices in Boston and Washington, strives to ensure high quality education for all students, particularly those in low-income communities. We have spent much of the last 30 years on improving education in Title I schools.

These proposed regulations address issues that are very important for effective implementation of Title I in ways that best serve children. In one area – the statutory requirement for multiple assessment measures – we believe the NPRM is right to draw attention to the law, but that the regulatory approach proposed is wholly inadequate and will actually have a very negative impact on implementation of the law. In other areas – minimum group sizes for disaggregation, supplemental educational services and choice, and adequate yearly progress – many of the proposed changes are significant positive steps in the right direction, although other fundamental changes are also needed. The same is true of our view of many of the proposals on graduation rates, but we believe that the proposals in two sub-areas (on goals and on alternative definition of standard number of years to graduate) are quite problematic. Further, as will be seen, the issues in addressing multiple measures and the issues in addressing minimum group size (along with confidence intervals) are each illuminated by looking at th connections between the two.

§200.2 State Responsibilities for Assessment – Multiple Measures

We are glad that the Department is concerned about correcting the misperception that Title I requires a single test when it not only permits but actually requires multiple measures. We need to correct this misunderstanding and, even more important, to secure implementation of that requirement. But the regulatory language that is proposed to address this area – which says nothing more than that multiple measures "[m]ay include– (i) [s]ingle or multiple question formats that range in cognitive complexity within a single assessment; and (ii) multiple assessments within a subject area"[1] – will not secure meaningful implementation at all. Indeed, the NPRM itself states that the clarifications in the proposed rule "do not impose new

[1]Proposed §200.2(b)(7), at 73 Fed.Reg. 22040.

requirements *or require States to change their current assessment systems.*"[2] This is, most unfortunately, a quite accurate assessment of the impact of the proposed language, under which pretty much anything counts as multiple measures. Single-format questions within a single assessment, multiple-format questions within a single assessment, or multiple assessments. We believe that quite different language, discussed below, should be substituted that will provide meaningful guidance and direction for how to implement this provision of law. Short of that, however, we believe that providing *no regulatory language at all,* and instead having the Department both widely disseminate a simple clarification that the statute indeed not only permits but actually requires multiple measures, together with additional non-regulatory efforts to assist in grappling with the issues involved in full implementation of the requirement, would be far preferable to promulgating the language in the proposed rule. In other words, the proposed rule – by making the requirement largely meaningless – is worse than no rule at all, letting the statute speak for itself (but at least making sure people hear what it says).

Addressing the Purposes of Multiple Measures. Part of the problem here comes from not providing adequate attention to the purpose of using multiple measures in the context of Title I assessment, and their connection to ensuring that the assessments result in valid and reliable decisions under Title I. The NPRM attempts to identify the purpose of including the measure: "The proposed language would clarify what is meant by this concept, which is included in the law to ensure that a State's assessment system measure the full range of cognitive complexity in the State's academic content standards."[3]

Even if the above statement from the NPRM adequately captured the underlying purpose of the requirement, the proposed regulation does not actually effectuate that purpose, for three reasons. First, because of the structure of the proposed provision, and the use of only "may" without any "shall," the language actually does not ensure or require cognitive complexity at all– it only says that (among other possibilities) multiple measures may include single or multiple questions formats that range in cognitive complexity within a single assessment . Second, the language on including a range of complexity applies only to reliance on a single assessment and does not on its face apply to the other permissible example, multiple assessments. Third, if this were rewritten to clarify that multiple measures, in whatever form, *must* provide for a range in cognitive complexity, it would still not meet the purported purpose identified by the NPRM "to ensure that a State's assessment measure the *full range of cognitive complexity in the State's academic content standards*" (emphasis added) – it is quite easy to meet the standard of providing some range in the complexity, without including the full range found in the State's standards. It is difficult to imagine a State assessment that could not be reasonably argued to contain a range of cognitive complexity, regardless of how woefully it fails to capture the full range of complexity identified in the State's standards.

---

[2] 73 Fed. Reg. 22021. Emphasis added. The problem lies not in the first part of this sentence – regulatory language that fostered sound implementation of the multiple measures requirement that has been in the law since 1994 would not be imposing a new requirement. It would secure compliance with that requirement, which would indeed require changes in many current assessment systems.

[3] *Id.*

More importantly, the identification of the purpose of the multiple measures requirement in the NPRM is itself inadequate. First, neither the legislative history of the provision nor the statute itself would seem to support the conclusion that ensuring measurement of the full range of cognitive complexity in the State standards (let alone a range of cognitive complexity of some kind) is the sole significant purpose of multiple measures, or that a showing that this particular purpose was met would obviate the need for multiple assessments. This provision was included in Title I in the 1994 reauthorization, and the legislative history accompanying it is scant. The only text we could find that went beyond restating the provisions in the bills was in the Senate Committee Report, which stated, "The committee intends that States develop a set or system of assessments, rather than a single test. The set should include multiple measures, and may include assessments developed at the local level as one of the measures if a State chooses."[4] The statute itself does not provide a basis for subsuming the multiple measures requirement under the requirement to assess higher-order skills and understanding. It is a leap of logic to go from the language of the Act itself – "involve multiple up-to-date measures of student academic achievement, including measures that assess higher-order thinking skills and understanding" to the conclusion that assessing higher-order thinking skills and understanding" is the sole purpose of the multiple measures requirement. A requirement that says "Do X [use multiple measures], including Y [measure higher order skills]" cannot logically be reduced to "Do Y." The meaning of "X" cannot be assumed to be derived solely from "Y." [5]

---

[4]103rd Congress, 2nd Session; Senate Rept. 103-292 (Part 1 of 5) (June 24, 1994). The Report also states, "The bill encourages States to move toward using new forms of assessments, such as performance based measures, for the purpose of determining adequate yearly progress and for assessing the performance of children served under title I Part A." *Id.* Seven years after the enactment of the multiple measures requirement, during the 2001 reauthorization, Senator Kennedy made a statement that does highlight the role of multiple measures in assessing "critical thinking and true problem solving," but this statement similarly states that the assessment must be done through "multiple test items, varying formats, and multiple tests." Senator Kennedy made a statement that does seem to link the multiple measures requirement with assessment of more complex skills, but even this falls far short of indicating that the sole significant purpose is measuring more complex skills or that a single test that addressed more complex skills would achieve the purposes of the multiple measures requirement: "There must be multiple measures within the test multiple test items, varying formats, and multiple tests to assess the higher order of understanding and thinking; not just memorizing, but critical thinking and true problem solving." 147 Cong Rec S 13322 at 13348 (December 17, 2001).

[5]Reinforcing the need to consider the independent meaning and purpose of the multiple measures requirement, without assuming that they are necessarily solely derived from its role in assessing higher-order skills, is that, unlike the Senate version, the House bill included a requirement for multiple measures without any reference to measures of higher order skills. It is also instructive that while the Senate version (and the version ultimately adopted) contained the language on including measures that assess higher-order thinking skills and understanding, the House version required multiple measures but not the including higher order skills clause, further that it has independent meaning and purpose that are not necessarily derived solely from its role in assessing higher order skills. See the conference report, House Report 103-761, Part 4, at

In giving independent meaning, as we must, to the requirement for "multiple measures," it is also important to recognize that the requirement to assess skills with a varying range of cognitive complexity, including higher order skills, already exists under the Act without reference to multiple measures. Under the Act, the State's standards themselves must contain such a range – the State's academic content standards must be "challenging," "contain coherent and rigorous content," and "encourage the teaching of advanced skills," and the State's academic achievement standards must describe basic, proficient, and advanced levels of achievement.[6] The assessment in turn must be aligned with those standards and capable of determining the basic, proficient, and advanced levels of performance. Saying that multiple measures means nothing more than any assessment that has a range of cognitive complexity, which is already covered by these other requirements, would thus essentially write the language requiring multiple measures out of the Act because it would have no independent meaning and add nothing.[7]

In the broadest and most comprehensive sense, the overarching purpose of using multiple measures is to enhance and help ensure the validity and reliability of decisions based on assessment. This is clear from the leading professional standards for educational assessment,[8] with which the assessments under Title I must conform.[9] The relevant decision here is whether students are proficient in the pertinent subject areas, as defined by mastery of the skills and knowledge that the state has determined that all children should know and be able to do. In the terms of the professional standards, multiple measures contribute to the validity and reliability of the determinations of proficiency by bringing to bear multiple sources of information about that proficiency. In other words, having identified the relevant construct – proficiency in the pertinent subject area as defined by the skills and knowledge articulated for all students in the State's standards – the Act then calls for multiple measures of that construct.

Thus, a prime reason for having multiple measures has been ignored in the proposed regulations and accompanying commentary – to ensure validity and reliability of the judgments about proficiency, as required by the Act, by providing *multiple ways* for students to demonstrate proficiency in the *same* skills and knowledge.

---

note 98.

[6]Sec. 1111(b)(1)(D).

[7]Similarly, consider a single test that provided subscores for reading comprehension and grammatical usage (both of which are addressed in the State's standards), for example. If this constitutes multiple measures, then the language that requires it likewise has no independent meaning and adds nothing – since anything less than providing for assessment of both of these important areas of the State standards would constitute clear failure to meet the most basic requirement of alignment with the standards, without reference to multiple measures.

[8][Joint Standards etc. – !! Complete the references.]

[9]Sec. 1111(b)(3)(C)(iii).

Moreover, the utility and need for providing multiple measures of the same proficiencies is not merely an abstract possibility, it is a concrete reality under Title I. As discussed at greater length in the next section (on disaggregation of data), in many schools, the assessment results are throw out, rather than used, for certain groups of students, not because they are too small a sample from which to draw valid and reliable conclusions about the proficiencies of the larger group they represent (given that all students, not a sample, are assessed under Title I), but rather because the assessments (purportedly) do not actually provide a valid and reliable measure of the proficiencies of the actual students being assessed – i.e., in light of the volatility of scores, we cannot be sufficiently sure that a student's score actually reflects his/her proficiency and so with groups of this size, the determination of what portion of the students assessed are proficient is of questionable accuracy. Similarly, in other schools, the results are counted and used, but instead are treated with confidence intervals that result in deeming groups of students as having reached a requisite rate of proficiency when in fact the odds are high that many of these groups do not in fact have that rate of proficiency, again because of limited confidence as to how well the scores reflect the actual proficiencies of the actual students being assessed.[10] In both cases, this frustrates the overarching goal of No Child Left Behind in identifying and addressing the needs of students who are not on an adequate path to proficiency in the skills and knowledge all students are expected to master. And in both cases, the problem stems from the acknowledged inability to determine proficiency of these groups of students with the single assessment instrument the State is using, even when that instrument is among the better ones in assessing more complex skills.[11]

Further, if the results of a single assessment are too unstable and unreliable to make accurate and valid judgments about the proficiencies of the students in a group when the number in the group drops below a certain level, then of course they are unable to do so when the group size drops to one – i.e., determining the proficiency levels of a single, individual student. Yet the Act requires that the assessment results be shared and used for that purpose as well, to allow parents, teachers, and principals to understand and address the academic needs of students, and it

---

[10]This happens as a result of (a) using a confidence interval create a range around the AYP target rate of proficiency for the group, to reflect the sense that within that range we cannot be sufficiently confident whether the students' scores on the test mean that their "actual" rates of proficiency, if measured more accurately, would be above or below the target rate of proficiency; and (b) then using the bottom of that range as the cut-off for determining AYP. This is done to minimize the number of false negatives, in which despite the scores being below the AYP rate, the actual proportion of proficient students is above it. The result is to maximize the number of false positives, in which the group is deemed to have reached the requisite rate of proficiency when in reality they have not. As with the application of minimum group size, this application of confidence intervals is not because of uncertainty in whether the results of a sample can be generalized to a larger group they represent, but because of uncertainty about how well the scores reflect the proficiency of the actual students assessed.

[11]For one thing, for smaller groups, the fact that a few students who score near the cut-score for proficiency can be pushed over or under the line by answers to one or two questions that may not accurately separate those who are and those who are not actually proficient, can in turn push the group over or under the AYP target rate of proficiency.

also requires that the assessments be valid and reliable for each of the purposes for which it is used. So concluding that we don't have valid and reliable data on whether individual students have attained proficiency because we have chosen to constitute our assessment system as a single assessment that is not adequate for that purpose, and letting it go at that, is simply unacceptable under the Act.

If the state's system of assessments does not allow us to accurately determine whether students are proficient, the solution – as mandated by NCLB as well as good practice and our obligation to children and their families – must be not to throw out the data and not count the kids, but rather to revise the system of assessment so that it can tell us when students are proficient. This is central to the core requirements for assessment under NCLB, under which the assessments must be valid and reliable for each of the purposes for which they are used. Note also that beyond the purpose of making determinations of proficiency of subgroups, another required use is for individual student reports allow parents, teachers, and principals to understand and address the academic needs of students. If the system of assessments in the state are not allowing valid and reliable judgments about the proficiencies of smaller groups of students, they are certainly doing so, as required, for individual students.

Both the problem (not having sufficient information to make a valid and reliable decision about which students are actually proficient) and a key part of the solution (using multiple measures of the same proficiencies in order to bring more evidence to bear on the question of whether these students are proficient and thus to increase the accuracy of those decisions) are staring us in the face. The Building this conception of multiple measures into the definition and criteria for use of multiple measures serves a critical purpose of the assessment provisions of No Child Left Behind.

In other words, faced with certain facts :

* the requirement to generate and use, as part of the state assessment system, valid and reliable data about the proficiencies of individual students in order to understand and address their academic needs;

* the fact that relying on a single assessment will not produce sufficiently valid and reliable data about the proficiencies of groups below a certain size, let alone individual students;

* the requirement that the state assessment system use multiple measures of achievement of th e skills and knowledge identified in the state standards for what all students should know and be able to do;

* the requirement that the state assessment system be consistent with relevant nationally recognized professional and technical standards, which in turn require that the validity and reliability of the assessment system in relation to the purpose it is being used be built by identifying each of the inferences being made and then ensuring that the evidence supports sufficient justification for those inferences;

* the overarching goal of the Act of leaving no child behind in relation to achievement of

the skills and knowledge that all children are expected to master;

it cannot be permissible to allow reliance on a single assessment, thereby resulting in a decision not to count and use the results of certain groups of students under minimum group size rules (or alternatively, a decision to deem certain groups as having made adequate progress toward proficiency when in fact they have not, through selection of the bottom of a confidence interval as the cutoff) rather than to use the multiple measures requirement to avoid these undesirable and unfortunate results by instead using a set of multiple measures that increase the validity and reliability of the judgments that can be made about the proficiencies of the students being assessed. The former will be the continued result of using the characterization of multiple measures found in the proposed regulation. The latter is the remedy that will be achieved by instead using the approach we delineate here.

Addressing the requirement for multiple measures in terms of providing multiple ways of measuring the same skills and knowledge identified in the state standards, thereby bringing to bear additional measures of the construct of proficiency, will also address other issues related to relying on a single test which also jeopardize the validity of the decisions being made. The use of a single test for assessing student proficiency increases the likelihood of teaching to the test, rather than to the broader set of skills and knowledge that the test is designed to assess. Not only does teaching to the test raise concerns about the quality and appropriateness of curriculum and instruction, it also directly threatens the validity of the use of the test itself. That validity depends on the test being a good representation of that broader set of knowledge and skills. When a student is taught to the test rather than to the broader and deeper set of knowledge and skills, that connection starts to break down, and successful performance on the test no longer represents the same mastery of those underlying skills and knowledge or of the ability to use them in different contexts. Measuring those proficiencies in more than one way, thereby reducing the felt pressure and tendency to teach to a single test, thus contributes to the validity of the decisions made and inferences drawn in this sense as well.

Similarly, while it is widely recognized that the use of multiple measures can contribute to ensuring validity of interpretation of performance for diverse populations, this benefit too depends upon the measures providing different but equivalent ways of demonstrating proficiency in the same knowledge and skills. Otherwise, we are left with only one way to demonstrate proficiency in certain areas of skills knowledge and skills and a different, but again sole, way to demonstrate proficiency in regard to other skills and knowledge; and these sole methods may not provide accurate interpretation for certain populations.[12]

Additional considerations regarding reliance on a single assessment. Even if the Department were to conclude that either our analysis of the importance of providing multiple ways to demonstrate proficiency in the same skills and knowledge is completely misguided or that that

[12]None of this is intended to challenge the notion that multiple measures can indeed also enhance the assessment system's ability to adequately assess more complex skills, as one of the purposes of multiple measures. But that is, as demonstrated above, a far cry from (a) declaring it to be *the* purpose of multiple measures and (b) concluding that anything with a range of cognitive complexity constitutes multiple measures, without reference to any other criteria.

goal could be met by providing different measures within the same assessment, it would still be a mistake to frame the regulation as permitting reliance on a single assessment. First, it invites a variety of problems. For example,

> It increases the possibility that the results will be skewed by some spurious event around the time of, or affecting the administration of, the assessment. While we generally think of this as affecting an individual student ("having a bad day"), which is more of a factor in assessments for high-stakes individual decisions than in assessments of groups of students for purposes of school improvement and accountability, some of these events s can also affect groups of students.
>
> Conjoining two ways of measuring the same set of skills and knowledge into a single event can negatively impact student performance. A student having a lot of difficulty with one measure may have a hard time switching gears as s/he continues to be preoccupied by the difficulty or experiences an overall sense of frustration. And depending on how it is administered, s/he may simply spend too much time on a particular measure and thus not enough time to demonstrate proficiency on another that would more accurately assess the student' proficiency.
>
> The effort to do this in a manageable amount of time can easily push the state in two opposite directions, both to the detriment of the validity and reliability of the assessment and decision process. On the one hand, the effort to do more than one measure well in the same time period may create an assessment that produces fatigue part-way through, to the detriment of students' performance and to an accurate picture of their skills and knowledge. On the other hand, the effort to avoid that fatigue by cutting back on the length of each may seriously compromise the capacity of each to accurately measure the requisite range of knowledge and skills.

More broadly, even if the Department concludes that the regulations should not flatly rule out the use of a single assessment incorporating multiple measures, should not affirmatively include among the limited number of provisions that get attention because of promulgation in the regulations a statement that multiple measures can be met through a single assessment (and indeed a single format). Such promulgation will have the effect of actively encouraging reliance on a single assessment, with the attendant problems noted above, and the conclusion that there may be a set of conditions under which a single assessment could fulfill the purposes of multiple measures, the provision as written will encourage inadequate approaches to multiple measures which do not effectively accomplish those purposes. It sends a message that a very diluted and minimal approach to multiple measures is welcome. In practice, the scope of multiple measures will shrink to fit the size of the container deemed acceptable (with accompanying rationales as necessary), rather than be based on a careful analysis of what kinds of measures are needed to accomplish the purposes of the Act.

Combining the multiple measures to make decisions about student proficiency. The NPRM is silent on this complex task, even though it is central to the question of using multiple measures to ensure and enhance the validity of the decisions. Once appropriate multiple measures have been selected, how much they contribute to valid and reliable decision depends on the decisions of how to combine the The goal should be to combine them in a way that builds a body of

reliable and accurate information about the students' mastery of the requisite skills and knowledge and that addresses conflicts in the data in a way that contributes to the validity and reliability of the overall judgments – rather than, for example, combining them in a way that deems students proficient for school improvement and accountability purposes simply because they are proficient by one measure if other measures cast doubt on that conclusion. In any event, states (and or districts) involved in these decisions should be under the same obligation, under the professional Joint Standards [!! Revise citation], that applies to other aspects of validation – first to identify each of the specific assumptions or inferences underlying the approach for combining the measures and then to collect and carefully analyze the evidence that supports or challenges each of these inferences.

In summary, we strongly recommend that:

1. The regulations clarify that in order to achieve the overall purposes of ensuring validity and reliability of the proficiency determinations made under the Act, multiple measures must include different ways of measuring the same proficiencies of students in the knowledge and skills identified in the state's standards.

2. Regardless of how the final regulations respond to recommendation number 1. above, the final regulations should strike proposed paragraphs (7)(i) and (ii), which define multiple measures in a way that will undermine achievement of the Act's purposes.

3. The regulations provide guidance and require and enable thoughtful attention to the question of how the multiple measures are to be combined in order to make valid and reliable determinations of students' proficiencies.

§200.7 Disaggregation of Data– Minimum Subgroup Size

(a)(2)(B) Maximum inclusion of all student subgroups

1. We support the language here requiring that the determination of the minimum subgroup size be determined in a way sufficient to ensure, to the maximum extent practicable, inclusion of all student subgroups in accountability determinations, particularly at the school level. Making sure that the process of determining minimum subgroup size is conducted with maximum fealty to the purpose of the disaggregation requirements and to the underlying principles of leaving no child behind is important. In this regard, we also commend the commentary's recognition (on pp.. 22022-23) of various developments that facilitate the capacity to generate reliable data, thus reducing the need to exclude certain subgroups because of lack of sufficient reliability – the implications of which also affect the discussion below.

2. At the same time, the additional requirement raises questions about how this new language will be interpreted and implemented. To give it meaning, and avoid its becoming meaningless out of a sense that, of course we already do that, further direction, through regulation and guidance, would be helpful.

3. Part of a state's ensuring maximum inclusion of all student subgroups lies in connecting this provision with the multiple measures requirement.

In this regard, the commentary in the notice is instructive.

First, as the commentary in the notice correctly notes, the usual reason for needing a minimum group size – in order to ensure that data drawn from a sample can reliably be used to draw conclusions about a larger population which that sample represents – simply does not apply here because all students, not a sample, are assessed. (By analogy, we cannot reliably draw conclusions about prevalence of strep throat in a neighborhood by testing one 4-member family. If, however, we are seeking only to draw conclusions about who in that family has strep, this issue disappears – the size of the group being tested, namely 4, is just right for the purpose. Similarly, if we're trying to determine whether *these* students are on a path to proficiency in order to see whether more needs to be done, then for that purpose size doesn't matter.)

Next, the commentary says that the use of minimum subgroup size, if not needed to address the normal purpose of reliable sampling, is a protection that minimizes the potential instability of scores in small populations and reduces the likelihood that a single extremely high or low score will skew the overall score for the group. However, this latter concern is also not relevant to the use of assessments for NCLB purposes, because they do not rely on a average score but rather the proportions of students achieving basic or advanced levels, so the extent to which an individual student scores way above or way below the requisite score is irrelevant.[13] What that leaves is the notion that minimum subgroup sizes are needed because a single assessment is not reliable and accurate enough to determine whether the students being assessed are indeed proficient. (In this sense, the problem is not an individual with an extremely high or low score. To the contrary the problem is most clear for a score that is right near the cut-off point for proficiency.)

In one sense, the need for a bright-line determination – proficient or not – makes the problem inevitable. A good assessment development process can reduce the arbitrariness of deciding that one particular minimum score constitutes proficiency and that anything just below it does not but it cannot eliminate the arbitrariness altogether, and the problem gets sufficiently diluted only when you have a large number of students' scores. *But* this problem, along with the larger related problem of the instability of scores, is not inherent in assessment itself, but only in an assessment determination that relies on a single measure.

In other words, what is really being said here is that, given the instability of the test scores on the state assessment, we don't actually know whether these students are proficient. The solution to that problem provided by the minimum subgroup size is to not count (thus not to attend to) that group of students. That solution, however, is typically neither necessary, educationally sound, nor consistent with other, overriding requirements of No Child Left Behind.

---

[13]Similarly, a growth model for AYP, relying on the proportions of students making sufficient annual gains toward proficiency, also does not rely on averaging scores and will not be affected by a particularly low or high score.

If the state's system of assessments does not allow us to accurately determine whether students are proficient, the solution – as mandated by NCLB as well as good practice and our obligation to children and their families – must be not to throw out the data and not count the kids, but rather to revise the system of assessment so that it can tell us when students are proficient. This is central to the core requirements for assessment under NCLB, under which the assessments must be valid and reliable for each of the purposes for which they are used.

Note also that beyond the purpose of making determinations of proficiency of subgroups, another required use is for individual student reports allow parents, teachers, and principals to understand and address the academic needs of students. If the system of assessments in the state are not allowing valid and reliable judgments about the proficiencies of smaller groups of students, they are certainly not meeting the requirement to do so about the still smaller group of one individual. (And that requirement has no escape clause.) And, as with all uses of assessment under the Act, the assessments used must be valid and reliable for that purpose. Stated conversely, if the systems of assessments provides, as under the law it must, valid and reliable information for determining the proficiencies of the smallest group of all, namely one student, then it should also be able to provide sufficiently valid and reliable information about the rate of proficiency among a group of students.

Having located the problem not in using a sample to make judgments about a larger population but rather in having an assessment that does not validly and reliably allow judgments about the proficiencies of those being assessed, two things should be evident. First, the response needs to be to improve the states' systems of assessments so that they meet the statutory requirements for validity and reliability, rather than to throw out the results for certain groups because they do not. Second, critical to that constructive, and mandated, response is meaningful implementation of the related statutory requirement for multiple measures of achievement of the state's identified learning goals. Because a score on a single test may be too volatile and unreliable to make valid judgments about the proficiencies of subgroup or individual, there is – as the Act recognizes – a need for multiple measures of achievement.[14] We should, through a valid and reliable system of assessments, using multiple measures of achievement, be able to maximize the extent to which all subgroups are included in the accountability system.

Stated differently, when faced with a choice between relying on a single measure of proficiency and then having to throw out a subgroup's scores because the results are too volatile to be a valid and reliable measure of the group's proficiency rate versus developing the multiple measures of proficiency that the law requires and which enhance the reliability and validity of the information, the choice should be clear in relation to both the specific mandates and the overarching purposes of the Act. States that have not developed the multiple measures needed to ensure that the various uses of assessment under the Act (including information about the proficiencies of individual students), as required by the

---

[14]As our comments to the provisions on multiple measures discuss, the ability of multiple measures to ensure and enhance the validity and reliability of judgments about students' proficiencies depends upon a proper understanding of the nature of multiple measures, consistent with the Act – in terms of providing additional information useful in validly determining proficiency.

Act, should not be permitted to exclude groups of students from the assessment system on that basis or to claim that they have ensured that all subgroups are included to the maximum extent feasible. The regulations and guidance should incorporate that recognition.

(a)(2)(ii) Revision of the Consolidated State Application Accountability Workbook

(B) Other components impacting inclusion of subgroups.

1. We support the requirements for revision of the Workbook, and in particular the requirement to explain how other components of the State's definition of AYP interact to affect the statistical reliability of the data and to ensure maximum inclusion of all students and subgroups. This is a positive and important complement to paragraph (A) in seeking to ensure maximum inclusion.

2. Specific attention to the use of confidence intervals as one of the other components is needed. For the reasons discussed at greater length in our comments on multiple measures, the issues with confidence intervals have a lot in common with those of minimum group size, arising from a similar lack of confidence in the validity and reliability of judgments about the very students whose proficiency is actually being assessed (rather than judgments about a larger group based on a sample), and resulting in a similar failure to identify and address a significant numbers of students and student groups who in reality are not on a proper path to proficiency.[15]
And similarly, it should be made clear that **the requirement to enhance the validity and reliability of judgments about students' proficiency through proper use of multiple measures should be explicated and implemented before using confidence intervals to respond to the uncertainties resulting from reliance on a single measure.**

3. Specific attention to uniform averaging is also warranted. In keeping with treatment of the other issues, it should be made clear that while states may choose whether or not to include multiple year of data, **states may not opt not to include multiple years of available proficiency data for purposes of determining and then turn around and use that decision to declare that they don't have enough data to meet a minimum group size or to adopt a confidence interval that is larger than it would be if they used the available data, thereby increasing the numbers of false positives in terms of student subgroups deemed to be making adequate progress when they are not.** This is contrary to the principle of making decisions in a way that maximizes proper inclusion in the accountability system.

[15] As also discussed earlier, in the case of confidence intervals, this choice is compounded by the decision to use the bottom of the interval as the target for determining adequate yearly progress.

§200.19 Other Academic Indicators – Graduation Rates

(a)(1)(i)(A)(*2*) Removing a student from the cohort.

We support these provision for confirmation, including the documentation needed to confirm enrollment in another program that culminates in award of a regular diploma.

(a)(1)(i) and (a)(1)(i)(B) Focus on four-year graduation rates

We support the focus on four-year graduation rates because for the vast majority of students who do not graduate in the standard number of years the delay reflects lost credit through not doing well and progressing, not students who are thriving and progressing in a well-designed program based on alternate time assumptions related to their needs.

At the same time, provided that all the other elements were in place (including the change in approach to goals and AMOs discussed below), we would be open to a system which also took some account of the fact that students have graduated even if they did not do so in the standard number of years. This balance could be struck by first counting students only in the denominator of the graduation rate when they do not graduate with their cohort, but then counting them again, in both numerator and denominator of the class with which they later graduate. This would keep the primary focus on four-year rates while providing some incentive to continue to try to engage and assist students who've dropped out or been left behind.

(a)(1)(i)(C)(*2*) Alternate definition of "standard number of years" for certain students

The provision for an alternate number of years for "limited categories of students who, under certain conditions, may take longer to graduate" is **far too amorphous, open-ended, overly discretionary, and overbroad , particularly for students with disabilities**, who are identified as one target group from which to identify students. While a high-school restructuring initiative *for all students* (regardless of race, income, disability, etc.), such as the early college high schools, might acceptably be seen to change the norm and redefine "standard number of years," basing a redefinition on disability will promote discrimination and lowered expectations, except with the tightest controls that ensure its use only where it has been clearly demonstrated that the student, even with the best instruction, assistance, and supports, would under no circumstances be able to progress at a rate sufficient to graduate within the same time expected for others. In that narrow case, the group would need to be targeted to students with the most severe cognitive disabilities who, based on meeting the above criterion, are instead performing *at a proficient level* on a State's alternate assessment based on alternate achievement standards, with the number of such students counted for this purpose capped at no more 1% of all student assessed.

(d)(1) Graduation rate goals and continuous and substantial improvement measures

1. Allowing a state to select any target rate it wants, and requiring undefined substantial and continuous progress toward it, will not address the problems ED well describes – for example a state selecting a goal of 50% of its students graduating on time. This is akin to

where we were on academic performance before NCLB, with minimal gains toward low goals being acceptable. Consistent with the rest of *No Child Left Behind*, adopt a goal of 100% on-time graduation with ambitious AMOs. (For the latter, we would recommend AMOs based on gap closing in 4 years, with even increments over the timeline applicable to each subgroup.)

On the achievement side, in NCLB, we set the goal as 100% proficiency because we want all children to acquire the skills and knowledge the state has said all children should master, and we attempt to set annual targets toward that goal in a way that ensures that students get attention if they are not an expeditious path to meet that goal. Similarly, as the commentary for this section begins, "There is an urgent need to improve America's high schools and ensure that all students graduate rom high school ready for postsecondary instruction or the workforce." Setting a goal that is less than that or setting progress markers at a level at which accept that no attention need be paid in this and upcoming years for students who are not getting there, is inconsistent with the goal and with the premises of No Child Left Behind. As with setting an ambitious achievement goal that is what we want for all children, in order to attend to any child who is not getting there, its viability depends on a non-punitive continuous improvement approach – in which expectations of attention to children, not sanctions, are triggered when those children are not on the necessary path to the goal.

2. **However, the single best thing ED could do for students not on a path to graduating in the standard number of years is to acknowledge the obvious – students who are no longer in the grade being assessed because they've been left back or dropped out haven' demonstrated proficiency – and to thus count *all students* in the cohort, *including those*, in determining the proportion of students who are proficient.** This will (a) provide a much more accurate picture of who has attained proficiency, (b) eliminate all incentive to let struggling students drop out, (c) create positive incentives to keep kids in and on grade level. Unlike anything that can be done with the graduation rate measure alone, it will truly mean leaving no child behind. This approach will not necessarily cause a dramatic increase in the number of identified schools – on average the true proficiency rates will be lowered by including these kids, but so will the annual targets.

§200.20 Making Adequate Yearly Progress – Growth Models

We support retaining the goal of all students attaining proficient or higher levels, by the same 2013-2014 target year, is the critical touchstone for ensuring that growth does not become a diluted substitute for attaining proficiency in the knowledge and skills the state has said all students should master.

§200.37 Notice of Identification for Improvement, Corrective Action, or Restructuring

In (b)(4)(iv), 14 calendar days is far too little notice for many parents to be assured of having enough time to understand, explore, and weigh the choices for their children's school. This would be true at any time of year, but the problem is compounded by the inability at this time of

year to see schools and classes in session, the likelihood of key personnel in schools and in supporting non-profits being away; the family itself being away when the fourteen day notice arrives, etc.

§200.48 Funding for Choice-Related Transportation and Supplemental Educational Services

We support the criteria in (d)(1) for demonstrating success in reaching and adequately informing parents of their options before reallocating unused funds. This is a well-designed set of criteria (and indeed be useful in thinking about adequate outreach and information in relation to other forms of parent involvement).