Review of the Programme for International Student Assessment (PISA) Test Design: Recommendations for Fostering Stability in Assessment Results.

John Mazzeo & Matthias von Davier

Executive Summary

With the completion of the 2006 Programme for International Student Assessment (PISA), the third assessment cycle conducted by the program, full frameworks have been established for each of the PISA assessment areas – Reading, Mathematics, and Science. *Literacy scales*, the main reporting scales for PISA, which summarize results over item sets that reflect the full range of cognitive outcomes delineated in the frameworks, have been established for Reading (in 2000), Math (in 2003) and Science (in 2006). With these frameworks and scales in place, the PISA Governing Board (PGB) has indicated that "the establishment of reliable trends should become the overriding priority. With this in mind, the Organization for Economic Cooperation and Development (OECD) PISA Secretariat has asked us to carry out an external review of the current test design and to provide design recommendations relevant to maintaining reliable trend results for forthcoming PISA assessments (2009 and beyond). In particular, we have been asked to provide guidance on: (1) Criteria indicating sufficient stability/precision in the establishment of trends, (2) Suggestions on how to improve the stability of the link, and (3) Recommendations regarding the PISA test design and the number of link items for each of the assessment domains. This report presents the results of our review.

It should be understood at the outset that , in our view, the degree to which PISA trends are more or less stable than other comparable assessment programs and the extent to which such instability is a function of test design issues seems to us to be a very much an open question. In addition, it should be noted that factors beyond test design also impact the degree to which the conditions for stable trend measurement can be established. Such factors include accuracy and consistency over time in the translation of assessment instruments, consistent adherence to standardized administration procedures, consistent establishment of sampling frames, adherence to prescribed school and student sampling procedures, and consistency of scoring of open-ended items across time in each country and language. Consistent with our understanding of the task, with one exception, we have chosen not to focus our efforts in these latter areas. However, it should be understood that these factors may play as great or a greater role in achieving stable patterns of trend results as do test issues of test design.

The report contains three main sections. The first section discusses the issue of criteria indicating sufficient stability/precision. The second section provides our thoughts on the two design issues on which we were explicitly asked to comment – the balance between major and minor domains and the number of link items for each of the PISA assessment domains. The third section discusses some other general test design and analysis issues for the PISA program to consider regarding procedures and designs for future assessments. A fourth and final section of the report provides an overall summary and concluding comment.

Criteria indicating sufficient stability/precision in the establishment of trends

One of the issues on which the PGB is seeking clarification is that of criteria indicating sufficient stability/precision in the establishment of trends. We have interpreted the PGB request as asking for guidance on how to evaluate whether the trend results being obtained under the current design, or results obtained in the future from similar or modified designs, are sufficiently stable and precise for their intended purpose. With respect to precision, we feel the most fruitful course of action is to examine the magnitude, in effect size terms, of the trend differences currently being declared as statistically significant. If it is

judged that there are too many *prima facie* large differences that fail to reach statistical significance, then changes to design parameters – most likely, the number of schools and students sampled for the assessment – will need to be made. As a point of reference, we provided information regarding the magnitude of trend differences observed in the U.S.'s National Assessment of Educational Progress (NAEP) and the extent to which such differences are declared as statistically significant. We compared the magnitude of trend differences in PISA declared as statistically significant to those found in NAEP and found PISA results to be quite similar to those found for NAEP.

We feel similarly about the issue of stability. We know of no credible evaluative criteria to appeal to other than the experience of other assessment enterprises that share similar goals and features as PISA. To that end, we provided as a point of reference data from main NAEP and Long Term Trend (LTT) NAEP regarding the magnitude of cycle-to-cycle changes in U.S. national and state-by-state changes. Again, we compared the magnitude of country-level PISA changes in reading and mathematics to those found in NAEP and found similarity between the two assessments with respect to the mathematics results. For the reading results, however, it did appear that large cycle-to-cycle changes were somewhat more common in PISA than in NAEP. The analyses presented were meant to be illustrative and could, in principle, be extended to other kinds of estimands (e.g., male/female differences) or to results from other international assessments judged by the PISA program to provide useful comparative criteria.

Recommendations regarding the PISA test design

Whether it is PISA, TIMSS, or NAEP, it is typically the case that multiple content domains are assessed in each cycle and the different assessments (PISA, NAEP, and TIMSS) have taken somewhat different design approaches to addressing the challenges of assessing multiple subjects in each cycle. PISA and TIMSS have consistently made use of "mixed" designs, in which each assessment booklet contains clusters from multiple content domains. In contrast, the main NAEP assessments, and LTT Trend assessments beginning in 2003, have made use of "focused" designs in which each assessment booklet contains blocks from only a single content domain.

To be sure, mixed designs offer a number of attractive features but they are not without their challenges. We believe the most serious challenge mixed designs are facing from a trend perspective is the potential impact of context effects on assessment results, both within an assessment cycle and across time. Context effects, which are discussed again in Section 3 of this report, occur when the psychometric functioning of items or clusters of items differs depending on factors such as the item position within a cluster, the position of the cluster within an assessment booklet, or the other material that an item or cluster is paired with. In our experience, contexts effects or their absence can only rarely be predicted, can sometimes be detected after the fact if the right data has been collected, and may not be presumed to be the same across jurisdictions (like states in NAEP, or countries in international assessments). Context effects, if present, are problematic because the standard Item Response Theory (IRT) models that are currently used to analyze large-scale survey assessments assume some degree of invariance of psychometric properties – across booklets within a particular assessment cycle, or across assessment cycles – in order for results to be reported on common and comparable scales. If the survey data violate such assumptions, analysis methods must be modified to account for such violations in assumptions and may only be partially successful in doing so.

Context effects may be hard to control in mixed designs, particularly in the situation like that of PISA where the emphasis placed on a particular content area changes across assessment cycles. There is consistent evidence from within PISA that such context effects impact the psychometric functioning of items in general, and the link items, in particular. The PISA Consortium is well aware of these "booklet" effects, which have been documented in the PISA technical reports, as well as in the document TAG(0505)4, and they have implemented innovative statistical methods that attempt to estimate and mitigate the impact of these booklet effects on PISA results. While these efforts provide further evidence of the care

and technical sophistication of the analysis work done in support of PISA, there are limits to the degree to which such adjustments can fully control for these booklet effects.

Simultaneously maintaining a mixed design and the major/minor domain distinction may not be desirable going forward in a system whose principal goal is stable trend measurement. We would recommend the use of focused designs and believe a very good example of such a design was suggested by Hambleton and associates (Hambleton *et al*, 2005). As with all designs, there are trade-offs that need to be considered. For a fixed total sample size (e.g., the 4,500 students as is currently typical in PISA), there would clearly be fewer students contributing to results for each content under the Hambleton focused design than is the case under the current design, which fully exploits the possibility of multivariate scaling and correlations between performance in the various domains cannot be estimated. Offsetting these negatives are, in our view, some key positives. We believe that the likelihood of booklet effects and the need to adjust for them would be greatly reduced. In NAEP, which has used focused designs for many years, we have seen no evidence of such booklet effects. Perhaps the biggest challenge in implementing any such design change for PISA will be how to effect the transition without disrupting trends. The PISA program would likely need to plan for some form of bridge study in which the old and new designs of the assessments are administered to randomly equivalent samples.

On the number of link items for each of the assessment domains

Another issue on which we have been asked to comment is that of the number of link items for each of the assessment domains. It is generally acknowledged that educational constructs are multifaceted and the unidimensional summaries typically employed to analyze such surveys capture, at best, the overall average performance across the different facets of the construct. Consequently, all survey assessments, whether they are single-language national assessments like NAEP, or multi-language international assessments like PISA, can expect to observe variability in assessment results, trend results in particular, if one looks within the overall collection of items used to measure trend. In our view, this is not an indication of "unstable trend measurement" per se. We would argue that stable trend measurement is observed if country-by-country trend results are relatively invariant over multiple **collections of trend items**, each of which was considered *a priori*, an appropriate measure of the intended construct. Such considerations highlight the importance of making the overall set of items on which trends are based sufficiently large and sufficiently representative of the full content domain framework to ensure that the assessment results serve what we see as their intended purpose – to report reliably on the overall trend with respect to the full construct defined by that assessment's framework. The PISA Consortium appear quite conscious of this issue and have been investigating its impact on overall trend measurement and its implications for the selection of trend items (Gebhart and Adams, 2007; TAG(0505)4).

Clearly, one fair question to ask is whether current plans vis-à-vis the number and nature of the trend items seem reasonable for 2009 or as a basis for future trend assessments. To our way of thinking, there are two aspects to this question – Is the nature of the trend item set appropriate and is the number of items appropriate. Regarding the first question, it is reported that the trend items for both the mathematics and science assessments were selected with the expressed intent that, as a collection, they span salient subdimensions of the overall construct. The degree that this is in fact the case is a matter for others with subject matter expertise to judge and is beyond the scope of our current activity. The content make-up of the Reading trend link may be somewhat more open to question.

As to the question of number of trend items, we would approach this issue, as we did the question of trend stability criteria, from a pragmatic standpoint, by comparing current PISA practice with our experience in NAEP. NAEP practice has been to base trend results on considerably larger samples of items than has been done in PISA to date and to ensure a larger degree of overlap between adjacent assessments than is the case in PISA, though the more recent PISA plans with respect to Mathematics and Science represent some attempt to expand on the size of the common item pool. The reasons for the differences in approach

between NAEP and PISA are the result of myriad factors and it may not be practical within the constraints of PISA to approach the levels of item reuse that undergird NAEP trends. However, our general recommendation would be to consider prioritizing an increase in the amount of linking items included in future assessments in the interest of ensuring the stability of trends. Such an increase should, of course, maintain the recent practice in Mathematics and Science of attempting to ensure representative coverage of the full content domain by the linking items.

Other suggestions on how to improve the stability of the link

The importance of controlling context with respect to trend measurement is well illustrated by NAEP's experience in the 1980's with its Reading Anomoly (Beaton, 1988; Beaton and Zwick, 1990). Based on this experience, the NAEP program has adopted a conservative stance with respect to keeping context as consistent as possible from one assessment cycle to the next. Basically, we have assumed, absent evidence to the contrary, that all changes potentially matter and should be avoided where possible. We would encourage the PISA program that, to the extent possible, a similarly conservative stance be adopted in the future. NAEP relies entirely on repeating intact clusters of items as its means of linking results from one assessment cycle to previous assessment cycles. We believe that the degree to which such policies can be emulated in PISA, the risks of encountering context effects – that potentially add instability to trend results – can be reduced. When changes to test designs that can impact context are considered in NAEP such changes are accompanied by "bridge studies" that are designed to estimate and appropriately adjust trend results for the potential impact of said changes.

The order of presentation may have a significant impact on the response behavior of examinees. The possibility that item position may affect item functioning leads us recommend that PISA maintain a high level of consistency across assessments in regards to cluster timing and mode of presentation; cluster position; and cluster composition.

For all large scale assessment – not just PISA, the passage or common-stimulus design carries some threat to the precision of the link due to the fact that questions using the same common passage as a reference may be statistically dependent in ways that go beyond what a single student variable may be able to explain. Suggestions for all large scale assessments, PISA included, that are aimed at reducing passage effects include the reduction of the number of questions per prompt or passage, while, at the same time to limiting the length of passages, so that fewer dependent observations, but more independent passages can be fit into the assessment timeframe. Such test construction practices also have the potential to reduce the linking error due to the passage structure of the assessment. We must acknowledge, however, that implementing changes in test construction to reduce passage effects carries with it its own threats to trend stability. Making such changes, in particular limiting passage lengths in reading assessments, carries with it its challenges to trend in that, by doing so, one could be to some degree changing the nature of the construct being measured. Thus, such changes are best introduced at a point in time when new frameworks, and new trend lines based on instruments from those frameworks, are being established.

Questions with a multiple-choice response format can be automatically scored as correct or wrong, while questions that allow for an open (or constructed) response may need to be scored by human scorers. Scorers who evaluate constructed responses and provide scores for these responses need to be trained to ensure that they adhere to comparable scoring rules and these rules should be applied consistently over multiple assessment cycles, and across participating countries.

The need for comparable measures over time and across participating countries suggests that the task material and the response format should facilitate reliable rating processes. This suggests to us that extended responses should be used sparingly, and short constructed responses should be preferred, since these can be rated and categorized into a limited number of categories more reliably. We know that PISA currently invests considerable effort in monitoring the equivalency of constructed-response scoring across countries and languages within an assessment cycle and strongly encourage the Consortium to

maintain their standards in this regard. However, we would also encourage PISA to carefully consider their current procedures for monitoring comparability of scoring across time points in each country where necessary. We feel that, with the increase in the number of participating countries seen in each cycle, PISA should consider investing further resources to strengthen their processes and procedures devoted to scoring equivalency, particularly equivalency of scoring at multiple time points, as needed.

The statistical methodologies use to derive measures of student proficiency from a series of responses to questions administered in PISA are based on the assumption of a systematic relationship between the likelihood of a correct response and an underlying proficiency variable. This relationship is assumed to be a parametric mathematical function. In the PISA assessment, this is the logistic function with one location parameter (often referred to as the Rasch model). This location parameter describes the difficulty of a question, i.e., it is a measure of how likely a correct response is given a certain level of a student's proficiency. In international assessments, the difficulty of a question is assumed to be the same across translations and participating countries, while the distribution of proficiencies across countries can vary freely.

The choice of a measurement model is an important decision for any assessment program, and is also one possible source of lack of model-data fit for some subset of questions within as well as across countries and assessment cycles. Our suggestion for the PISA consortium is to devote some resources for explorations of more general modeling approaches to study the effect of differential item functioning across assessment cycles and countries under various item response models. We recognize that the Rasch model chosen for PISA has unique mathematical properties, and there are good reasons to use a model that involves fewer rather than more parameters. However, we feel that there is some justification for the decision by NAEP (and other assessments) to go with a more general IRT model (2PL/Generalized Partial Credit model, and 3PL) in the face of an assessment that is designed to provide a broad coverage of the domain using multiple item formats and test versions. In our experience, these more general IRT models do accommodate the functioning of items in diverse populations better than the Rasch model, which assumes that all items contribute the same amount of information to the measurement of student proficiencies. We assume that using a more general IRT model may also help reduce some of the country-by-item interactions observed in PISA, since the adoption of a more general measurement model improves model-data-fit considerably in our experience.

Review of the Programme for International Student Assessment (PISA) Test Design: Recommendations for Fostering Stability in Assessment Results.

John Mazzeo & Matthias von Davier

September 9, 2008

With the completion of the 2006 Programme for International Student Assessment (PISA), the third assessment cycle conducted by the program, full frameworks have been established for each of the PISA assessment areas – Reading, Mathematics, and Science. *Literacy scales*, the main reporting scales for PISA, which summarize results over item sets that reflect the full range of cognitive outcomes delineated in the frameworks, have been established for Reading (in 2000), Math (in 2003) and Science (in 2006). With these frameworks and scales in place, the PISA Governing Board (PGB) has indicated that "the establishment of reliable trends should become the overriding priority. With this in mind, the Organization for Economic Cooperation and Development (OECD) PISA Secretariat has asked us to carry out an external review of the current test design and to provide design recommendations relevant to maintaining reliable trend results for forthcoming PISA assessments (2009 and beyond). In particular, we have been asked to provide guidance on: (1) Criteria indicating sufficient stability/precision in the establishment of trends, (2) Suggestions on how to improve the stability of the link, and (3) Recommendations regarding the PISA test design and the number of link items for each of the assessment domains. This report presents the results of our review.

In preparing this report, we reviewed several documents provided to us by the OECD PISA Secretariat: (1) Proposal for Securing Trends in PISA 2009 [EDU/PISA/GB(2008)1]; (2) The Draft Technical Report for PISA 2006, (3) The paper – *The computation of equating errors in International surveys in education* – by Monseur and Berezner (2007), (4) PISA 2003 Follow-Up Analyses and Discussion Issues [TAG(0505)4]; (5) Technical Review of PISA (Draft Version, October 4, 2005) by Hambleton, Gonzalez, Plake and Ponocny; and (6) The PISA Consortium's response to the Hambleton report. In addition to these reports, we also reviewed pertinent chapters of the PISA 2000 and 2003 Technical Reports, as well as sections of the 2006 PISA report – *Sciences Competencies for Tomorrows World: Volume 1 – Analysis*, and, Draft Technical Note on Comparisons over Time on the PISA Scales [EDU/PISA/GB(2007)42], all of which are available on the PISA website, www.pisa.oecd.org. In addition to these PISA-related documents, we also reviewed the article – *The Influence of Equating Methodology on Reported Trends in PISA* by Gebhart and Adams (2007) .

In reviewing this documentation, one cannot help but be impressed by how complex an endeavor PISA is and the degree of sophistication reflected in all aspects of its operation, and, in particular its test design and statistical analysis procedures – the principal focus of this review. From our experience as contractors for the U.S. National Assessment of Educational Progress (NAEP), we are cognizant of the fact that test designs and analysis procedures for all assessments are developed by trying to strike an appropriate balance among competing forces – the policy and informational goals of the assessments sponsors and participants, the practical and fiscal realities associated with actually carrying out the assessment, and the psychometric realities of what kinds of results can be reliably produced from data collected under a particular test design. It is evident to us from our review that the current PISA test design and analysis procedures have been capably developed by the current Consortium to be responsive to the values of the PISA program with respect to these competing forces. Without

reconsideration of current constraints (both practical and fiscal) and values, test design improvements may be difficult to achieve.

We should state clearly at the outset that, our discussion of stability criteria presented in the ensuing section notwithstanding, the degree to which PISA trends are more or less stable than other comparable assessment programs and the extent to which any such instability is a function of test design remains, in our view, an open question. We do note however that there has been some recent research around the topic of trend stability and related issues. These studies have been carried out by staff of member institutions within the PISA consortium (Gebhardt and Adams, 2007), as well as members of the PISA technical advisory group (Monseur and Berezner, 2007) and other researchers concerned with the science of this and other large scale survey assessments (for example: Monseur *et al.*, 2008: Park and Bolt, 2008; Xu and von Davier, 2008). These studies indicate that models used in the analysis, the test design (based on item clusters, or otherwise), the interactions between test items and country- and language-specific factors, as well as the selection and composition of the link sets used in the trend may have a non-trivial and therefore non-negligible effect on the reported trends. Hence some discussion of the current PISA test design and consideration of changes for future assessments seems warranted if the focus of future assessments will be on trend results.

Making specific practical suggestions as to how to improve the current and future test designs with respect to the stable measurement of trend, in light of the above discussion and absent detailed information about program goals, values, and constraints, is a challenging endeavor. Despite our efforts to review the extant documentation on PISA, there is no way that, as external reviewers, we can understand the goals and constraints (practical, political, and financial) within which PISA operates at a level of depth comparable to that of the current Consortium members. In making such suggestions, there is always a strong possibility that they are simply not logistically, technically, or politically feasible. Consequently, we have not proposed a specific alternative test design, but rather have made suggestions about more general aspects of what a test design that prioritizes stable trend results might entail. As was evident to us in reviewing the background material provided to us, many, if not all, of these suggestions have been considered in one way or another by the current Consortium or in the previous review commissioned by the OECD (Hambleton et al., 2005).

That all being said, in working extensively on design and analysis activities for many years on assessments similar in nature to PISA, in particular NAEP, we have been forced to address many of the same challenges currently facing PISA. We felt that the most valuable way for us to be of assistance to the Secretariat and the PGB is to share our thoughts on the current PISA test design from our own perspective and experiences in addressing the question of test designs for stable trend measurement in NAEP. We are certainly conscious of the fact that international assessments like PISA face unique challenges, and most likely resource and practical constraints, not faced by national assessments like NAEP. We are hopeful, nonetheless, that at least some of our design choices and lessons learned can be of assistance to the PGB.

Establishing stable country-by-country trend lines in the context of an international assessment like PISA requires maintaining control over many aspects of the design and conduct of the entire study, including:

a) Measuring the same constructs in all assessment cycles depicted jointly in the trend line

b) Ensuring, over time, that the instrument measures the same construct in all participating countries, jurisdictions, and other subgroups for which trends are reported

c) Ensuring that the relationship between questions and underlying proficiency stays the same across cycles for those questions that are common over cycles

d) Ensuring that questions translated into the language of instruction are functioning the same across multiple official languages used in the assessment in the same country

e) Ensuring that the set of questions that make up the assessment are presented under comparable, standardized conditions across countries and over time

f) Ensuring for each country that, over time, the sample of students was drawn from the same population of students (15-year-olds in the common schooling system as defined in PISA), with the exception that the students assessed in each cycle come from different birth cohorts

This non-exhaustive list of assumptions provides a collection of potential sources of error that interfere with the comparability of results across assessment cycles and across countries within an assessment cycle. If any of these assumptions is violated, the trends reported may have a larger margin of error than the associated measure of precision (i.e., the standard error) reported in conjunction with the trend measure.

Many of the considerations, particularly (a) through (c) listed above are impacted directly or indirectly by the kinds of test design factors we have been asked for input on. Many of PISA's current policies and practices reflect a strong awareness of the importance of these considerations. In particular, concerns regarding (a) are obviously reflected in PISA's trend reporting practices in Mathematics and Science. Using interim trend scales, which have been discontinued when full frameworks are established (e.g., in Science), or in restricting trend reporting to only those subdomains with adequate item coverage in prior assessment cycles seem to us to be very wise policies indeed and, other things being equal, are certainly courses of action that we would endorse or recommend were we operating under current program constraints.

We hasten to add, though, that other factors beyond test design also impact the degree to which the conditions for stable trend measurement can be established. Such factors include accuracy and consistency over time in the translation of assessment instruments, consistent adherence to standardized administration procedures, consistent establishment of sampling frames, and adherence to prescribed school and student sampling procedures. The challenges associated with ensuring consistent translation, administration, and sampling in an international context are areas that the PGB and its contractors have considerable experience with. Consistent with our understanding of the task we have been assigned and our own personal areas of expertise, we have chosen not to focus our efforts in these areas. However, it should be understood that these factors may play as great or a greater role in achieving stable patterns of trend results as do issues of test design. The issues we raise below, and the test design changes we recommend that PISA consider in the short or longer term, if implemented, may or may not substantially affect the stability of PISA results. We believe our suggestions reflect sound design principles that would minimize the potential impact of test design factors on the stability of trends.

The next three sections below will address the test design issues the PGB asked for guidance on. As test design issues typically impact analysis issues as well, we have, where appropriate, discussed issues related to these as well. The first section discusses the issue of criteria indicating sufficient stability/precision. The second section provides our thoughts on the two design issues on which we were explicitly asked to comment – the balance between major and minor domains and the number of link items for each of the PISA assessment domains. The third section discusses some other general test design and analysis issues for the PISA program to consider regarding procedures and designs for future assessments. The fourth and final section of the report provides an overall summary and concluding comments.

Section 1 – Criteria Indicating Sufficient Stability/Precision in the Establishment of Trends

One of the issues on which the PGB is seeking clarification is that of criteria indicating sufficient stability/precision in the establishment of trends. It is clear from the background material we were provided and from the fact that this review has been commissioned that some concerns have been expressed regarding the stability/precision of the limited trend information available to date in PISA. We have interpreted the PGB request as asking for guidance on how to evaluate whether the trend results being obtained under the current design or results obtained in the future from similar or modified designs are sufficiently stable and precise for their intended purpose.

First, we would like to clarify that, in our view, precision and stability are distinct, though related, concepts. To us, the concept of precision is directly tied to statistical ideas around variance of estimates, which, in turn, leads to issues of sampling and replication. Assessments like PISA and NAEP necessarily involve sampling – sampling schools and students to be assessed in each of the participating countries, creating collections of items that implicitly define the educational construct being measured, carrying out the assessment under prescribed administrations procedures, and conducting the analysis. One can conceive of repeatedly replicating the assessment exactly as conducted but, using differing samples of schools of students selected in identical fashion, creating multiple collections of items, each of which is representative of the content domain to be assessed[1], and carrying out and analyzing the assessment in accordance with the prescribed procedures. The degree of variability in assessment results over these "exact hypothetical replications" of the assessment, which is typically estimated by the standard errors associated with any particular assessments result, is what **we** think of as precision. Specifically, in the context of measuring trends, precision would refer to the standard error of the difference between two sets of assessment results which is a direct function of the precision associated with the separate results from each of the two years.

Stability, in our view, refers to the observed pattern and magnitude of changes that one observes in a time series of assessment results. Typically, how large are changes in average scores, or percentages of students exceeding cut-scores of interest, from one assessment cycle to the next? How similar are trend results based on sub-areas of the construct (e.g., in geometry or algebra) compared to results for math overall? How similar in magnitude are the differences between relevant subgroups within a population (e.g., males and females) across assessment cycles? Does a time series of average-score results for a particular country tend to move consistently up or down for substantial periods or is this change irregular (up sharply in one assessment cycle and down the next). To be sure, one factor that contributes to stability is precision since one source of fluctuation from one time point to the next is the sampling uncertainty inherent to the assessment results. However, as discussed above and in other sections of this report, stability of results will depend on other factors as well – the degree to which collections of items in each of the assessments define similar educational constructs, the degree to which country-by-subdomain-by-assessment cycle interactions are present in the data, and the degree to which sample-selection, instrumentation and administration procedures can be held constant over time. With this perspective, we discuss criteria for precision and stability in separate sections below.

Section 1.1 – Precision

In theory, statistical methods could be used to establish precision criteria for trends in PISA or any other survey assessment. One approach might involve some variation of the

[1] Alternatively, one could also consider the set of items being used for the assessment as fixed over replications, though this implies a more constrained definition of the construct being measured.

following: (1) First, policy decisions are made concerning the minimum effect sizes[2] (in this case, trend differences) one wants to be able to detect (i.e., to declare as statistically significant); (2) Second, policy decisions are made with respect to what power (i.e., the probability of finding a statistically significant difference in the sample-based assessment results if the results for the entire country have in fact changed by that minimum amount) and "type-I" error control (the probability that of declaring a difference on the basis of the assessment results when results for the entire country have not changed); (3) Based on one and two, sampling designs, including school and student sample sizes, as well as instrument designs, are developed that will achieve the criteria delineated in steps one in two.

In practice, the "textbook approach" described above is exceedingly complicated to implement for complex assessment programs like PISA or NAEP. Precision in survey assessments is determined largely by the number of schools and students in the sample, the total number of items in the assessment, and the number of items administered to each student in each of the domains to be reported on. Among other considerations, the resulting sampling plans, implied school and student sample sizes, and assessment designs required to meet precision criteria defined *a priori* may not be feasible or affordable. In our experience, it is far more common that surveys in general, and survey assessment programs in particular, optimize precision for their assessment given the financial and practical constraints within which they operate. Our suggestion would be to take a pragmatic approach to establishing precision criteria. One way to evaluate whether the resulting precision levels for trends are reasonable is to examine the size of country-level trend differences that are and are not declared statistically significant. If differences in results judged *prima facie* to be of substantial size frequently fail to meet statistical significance criteria, then precision levels for the assessment may not be adequate and efforts would be required to address the logistical, financial, or design constraints that are limiting current precision levels.

To provide a point of reference, it may be helpful to take a look at a set of typical results from Main NAEP[3]. Though certainly not identical in all respects, NAEP and PISA share much in the way of measurement goals, design, and analysis. Both are cross-sectional survey assessments intended as measures of group-level educational achievement, as opposed to the achievement of individual students. Both NAEP and PISA make use of matrix sample designs, in which samples of students from participating jurisdictions respond to a sample of test questions from a much larger collection of questions. Both assessments use modern data analysis approaches based on item-response theory (IRT) to summarize assessment results in terms of scale scores.

For illustrative purposes, we will use changes in average NAEP scale scores between the two most recent grade 8 Main NAEP Mathematics assessments (2005 and 2007). Fifty-three states and other jurisdictions within the U.S. participated in both assessments, with typical sample sizes of about 100 schools and 2,500 students per state/jurisdiction. Each student is tested for 50-minutes in one and only one subject area – in this case, Mathematics. Overall results for each state/jurisdiction are reported on a composite scale which, at grade 8, has a standard deviation of 36 points for the United States as a whole. With respect to these state-level results, trend differences of two scale score points or less (an effect size of about .06) were rarely statistically significant. Differences of three or four points, (effect sizes between .08 and .11) were statistically significant about two-thirds of the time, and differences in excess of four points (i.e., effect sizes of .14 or greater) were uniformly declared to be statistically significant.

The obvious point of comparison here would be to overall country-level results in PISA. To that end, we used the country-level trend data in Tables 6.2c and 6.3b from the PISA 2006 Volume 2, Data (OECD, 2007) to calculate similar effect size estimates for the

[2] We use the general term "effect size" to refer to a difference in average scores divided by a standard deviation . Effect sizes have the convenient property of allowing one to make comparisons across assessments that use different reporting scales.

[3] NAEP conducts two ongoing time-series, Main and Long Term Trend NAEP. The distinctions between these two assessments are described in more detail below.

most recent PISA mathematics assessments (2006 and 2003). Specifically, we divided the differences between 2006 and 2003 average scores (for OECD and partner countries as given in Table 6.3b), by the OECD average within-country standard deviation (as given in Table 6.2c) to convert the scale score differences to comparable effect size measures[4]. Comparison of PISA to NAEP on this effect-size/statistical significance criteria shows quite similar results for overall average score trends. Like NAEP, effect sizes below .06 were almost never significant (1 out of 22), half or the differences between .07 and .11 (4 of 8) were declared statistically significant, and all of differences exceeding .11 (8 of 8) were declared statistically significant.

The comparison presented above is not precise in that different conventions for determining statistical significance, in particular with respect to controlling type-1 error over multiple-comparisons may be used by NAEP and PISA. Further, and more detailed comparative analysis, may or may not confirm the analyses presented here. In addition, we have, for illustrative purposes, restricted ourselves to examining overall average scores. Other estimands, such as changes over time in subgroup differences or percentiles could in principle be examined in similar fashion. That being said, given the data readily available to us, PISA precision levels for country-level average scores, at least with respect to the evaluative criteria advanced in this report, appear comparable to those for state-level results in Main NAEP. Whether such levels represent adequate precision for PISA's purposes is a judgment that only the PGB, the Consortium, and the participating countries can determine.

Section 1.2 – Stability

We would recommend taking a similarly pragmatic approach to the issue of stability criteria. What are reasonable expectations for the amount of change one should expect to see across adjacent time points? How likely are reversals – i.e., large swings up and down. We believe that many educational researchers, if queried, would expect that change in large populations over small time periods is likely to be modest. But we do not know of a truly credible way to determine the answer to such questions. The size of the populations being assessed, the effectiveness of potential educational interventions, and myriad other factors must certainly affect such matters.

As with the question of precision, one way to approach the issue of stability is to consider the degree to which assessment results, like average scale scores for participating countries, tend to vary from one assessment cycle to the next and to use extant data from other large scale assessments to develop reasonable expectations about the degree of volatility one might expect to see. To this end, we can share with the PGB the degree of stability we have observed in our years working on NAEP. NAEP has as its **principal goal** the measurement of trends in educational achievement in a large number of subject areas and, as discussed in more detail below, NAEP designs reflect that assessment's emphasis on stable trend measurement. The three NAEP subject areas of most relevance to PISA are Reading, Mathematics, and Science, though admittedly the NAEP content frameworks on which its assessments are based differ from the corresponding PISA frameworks of the same name.

Since its inception, NAEP has conducted two separate on-going time series in each of the three PISA-related subject areas, Reading Mathematics, and Science. The two NAEP assessments[5] in each of the subject areas are based on somewhat different content frameworks

[4] It could be argued that the optimal way to create PISA effect sizes comparable to the NAEP effect sizes reported would be to divide PISA results by the U.S. standard deviation as was done in obtaining NAEP effect sizes. However, as discussed later in the report, effect sizes were also calculated for the two most recent PISA reading assessents and U.S. results are not available in that subject. In order to ensure consistency across PISA subjects, we chose the average within-country OECD standard deviation for the denominator the effect size. In mathematics, the two standard deviations (U.S. and OECD average) were similar (90 and 92, respectively).

[5] The reason there are two distinct NAEP assessments in each of these subject areas is pertinent to this paper . Briefly, the creation of two separate NAEP's reflects how the

and assessment instruments. The Long-Term Trend (LTT) NAEP assessments were initiated in the late 1960s (for Reading) and early 1970s (for Mathematics and Science) and scale score results (back to 1971 for Reading and back to 1978 for Mathematics) are currently reported (for the U.S. as whole, for regions, for males and females, and for key demographic and educational subgroups) for three age-cohorts – 9-year-olds (typically in their fourth year of primary education), 13-year-olds (typically in their eighth and final year of primary education), and 17-year-olds (typically in the next-to-last year of secondary education). The main NAEP assessments, whose precisions were referenced above, are based on more up-to-date content frameworks. These assessments were initiated in the 1990s (1990 for Mathematics, 1992 for Reading, and 1996 for Science) and, like LTT NAEP, results are reported for the U.S. as a whole and for key subgroups of interests. Main NAEP assessment results are also available for individual states and, in recent years, for large urban school districts (e.g., Los Angeles and New York). Main NAEP reporting is grade-based, rather than age-based and results are reported for 4th, 8th and 12th grade students, with state-level results available only for the first two grades.

The NAEP LTT results between the years 1984 and 1999 provide one admittedly conservative benchmark for establishing expectations about the stability of trend results. The LTT assessments during those time periods were carried out by **administering the exact same assessment booklets in successive assessment cycles using exactly the same sampling, administration, scoring, analysis, and quality monitoring procedures.** The time between LTT assessments, two or four years, is similar to that between successive PISA assessments. In Reading, seven LTT assessments were conducted between 1984 and 1999 (1984, 1988, 1990, 1992, 1994 and 1996). The Age-13 assessment consisted of 107 questions based on 43 reading passages while the Age-17 assessment contained 95 questions based on 36 passages. The LTT Mathematics and Science assessments were carried out six times during the same time period (1986, 1990, 1992, 1994, and 1996). The Age-13 Mathematics assessment contained 127 questions and the Age-17 contained 132. The Age-13 and Age-17 Science assessments contained 83 and 82 questions, respectively.

LTT Reading Results are shown in Table 1. Changes in the U.S. average results from adjacent assessments are shown in terms of both scale scores and effect sizes[6]. Changes between successive assessments ranged from zero to three points in absolute magnitude on the NAEP scale. In effect size terms, the changes were between .00 and .08. Results for the Mathematics assessments are shown in Table 2 and for Science in Table 3. The differences in scale score units are similar, though perhaps slightly larger than those seen in reading. In one instance, an absolute effect size differences as large as .15 was found. In large part, however, differences rarely exceed .10.

National Center for Education Statistics and its contractors chose to deal with the ongoing tension between accommodating change and the requirement for stable trends. Main NAEP was changed to reflect up-to-date content frameworks and assessment methodologies while LTT NAEP was maintained to provide stable trend results relative to the original NAEP frameworks and assessment technologies.

[6] For each age and subject area presented, effect sizes were calculated by dividing the differences in scales scores by the standard deviation for the U.S. national population obtained from the most recent LTT assessment.

Table 1– Age 13 and 17 NAEP Long-term Trend Reading Changes in Average Scale Score (1984 – 2004)

	Age 13			Age 17		
Year	Average Scale Score	Scale score change	Change as effect size	Average Scale Score	Scale score change	Change as effect size
1984	257	-	-	289	-	-
1988	257	0	0.01	290	+1	0.03
1990	257	-1	-0.02	290	0	0.00
1992	260	+3	0.08	290	0	-0.01
1994	258	-2	-0.05	288	-2	-0.04
1996	258	0	0.00	288	-1	-0.01
1999	259	+1	0.04	288	0	0.00
2004	259	-1	-0.02	285	-3	-0.07

Source: National Assessment for Educational Progress, 1984 – 2004

Table 2 – Age 13 and 17 NAEP Long-term Trend Mathematics Changes in Average Scale Score (1986 – 2004)

	Age 13			Age 17		
Year	Average Scale Score	Scale score change	Change as effect size	Average Scale Score	Scale score change	Change as effect size
1986	269	-	-	302	-	-
1990	270	+1	0.04	305	+3	0.09
1992	273	+3	0.08	307	+2	0.07
1994	274	+1	0.04	306	-1	-0.02
1996	274	0	0.00	307	+1	0.03
1999	276	+2	0.05	308	+1	0.03
2004	281	+5	0.15	307	-1	-0.05

Source: National Assessment for Educational Progress, 1986 – 2004

Table 3 – Age 13 and 17 NAEP Long-term Trend Science Changes in Average Scale Score (1986 – 1999)

	Age 13			Age 17		
Year	Average Scale Score	Scale score change	Change as effect size	Average Scale Score	Scale score change	Change as effect size
1986	251	-	-	289	-	-
1990	255	+4	0.10	290	+2	0.04
1992	258	+3	0.08	294	+4	0.08
1994	257	-1	-0.03	294	0	0.00
1996	256	-1	-0.02	296	+2	0.04
1999	256	0	-0.01	295	0	-0.01

Source: National Assessment for Educational Progress, 1986 – 1999

As noted, NAEP LTT results for the United States as a whole probably provides a somewhat conservative baseline against which to evaluate the stability of PISA assessment results. The United States is a large country and it may be reasonable to expect greater stability in its results than will be the case for countries of smaller size. Education policy in the U.S. is in large part the responsibility of individual states and therefore its overall national results might not be particularly sensitive to educational reform efforts controlled and implemented at the state level. It could be argued that results for countries where the education system is more centrally-controlled and reform efforts can be more universally implemented might in fact be expected to see more dramatic results between assessment cycles than is evident in U.S. NAEP LTT results. Moreover, the fact that the frameworks for NAEP LTT are not necessarily aligned with up-to-date curriculum and instructional practice may also make them less sensitive to current efforts focused on improving educational achievement.

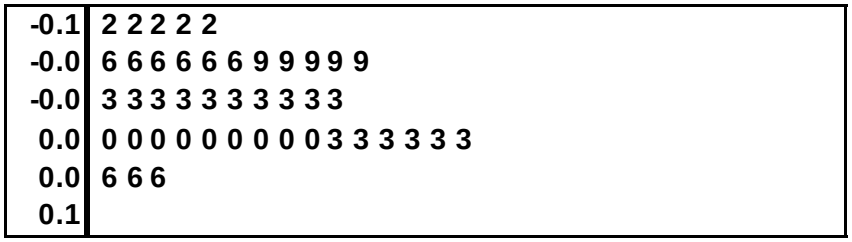
Perhaps a somewhat more realistic benchmark can be provided by the Grade-8 U.S. state-by-state results obtained with the main NAEP assessments. Like the NAEP LTT, main NAEP assessments retain tight control and comparability over sampling, administration, scoring, and analysis procedures. Unlike NAEP LTT, main NAEP does allow some changes in the make-up of its assessment booklets. Specifically, after each NAEP assessment, a relatively small percentage (about one quarter) of the assessment items, are released to the public. These items are replaced in subsequent assessment cycles. Therefore, successive main NAEP assessments share about three-quarters of their questions.

U.S. state-by-state reading assessment results for main NAEP are available for five assessment cycles (1998, 2002, 2003, 2005, 2007), though not all states participated in each of the assessments. Figures 1 through 3 present stem-and-leaf plots (Emersen and Hoaglin, 1983) of state-by-state differences, expressed as effect sizes[7], in Grade 8 average scale scores between successive assessment cycles for the four most recent main NAEP assessments. The period covered corresponds to the passage of legislation in the U.S. that effectively mandated state-by-state participation in NAEP and focused considerable policy interest in improving educational achievement. In each figure, the "stems" (i.e., the digits to the left of the bold

[7] For each grade and subject, effect sizes were calculated by dividing differences in the average scores for states in successive assessments by the most recent estimate of the standard deviation for the U.S. public school population.

vertical line) represent the effect size digit immediately to the right of the decimal point while the "leaves" represent the effect size digit for the second decimal place. Each leaf corresponds to a result for a single NAEP jurisdiction. The duration between assessment cycles ranged between one and four years. As is evident from the figures, the absolute magnitude of the differences, in effect size terms, are typically less than .10 and do not exceed .12.

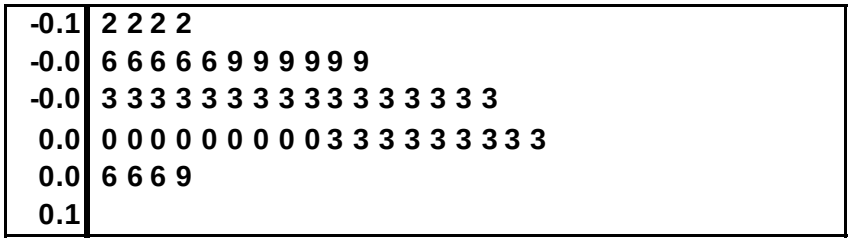
Figure 1 – Histogram of Changes in Average Scales Scores, Expressed as Effect Sizes Grade 8 State NAEP Reading 2003-2002



-0.1	2 2 2 2 2
-0.0	6 6 6 6 6 6 9 9 9 9 9
-0.0	3 3 3 3 3 3 3 3 3 3
0.0	0 0 0 0 0 0 0 0 0 3 3 3 3 3 3
0.0	6 6 6
0.1	

Source: National Assessment for Educational Progress, 2002 - 2003

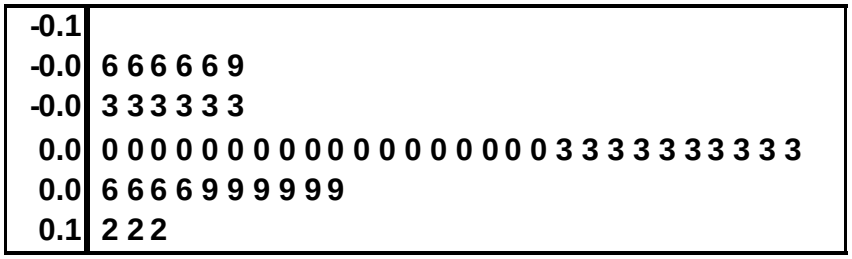
Figure 2 – Histogram of Changes in Average Scales Scores, Expressed as Effect Sizes Grade 8 State NAEP Reading 2005-2003



-0.1	2 2 2 2
-0.0	6 6 6 6 6 9 9 9 9 9 9
-0.0	3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3
0.0	0 0 0 0 0 0 0 0 0 3 3 3 3 3 3 3 3 3
0.0	6 6 6 9
0.1	

Source: National Assessment for Educational Progress, 2003 - 2005

Figure 3 – Histogram of Changes in Average Scales Scores, Expressed as Effect Sizes Grade 8 State NAEP Reading 2007-2005



-0.1	
-0.0	6 6 6 6 6 9
-0.0	3 3 3 3 3 3
0.0	0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 3 3 3 3 3 3 3 3 3 3
0.0	6 6 6 6 9 9 9 9 9 9
0.1	2 2 2

Source: National Assessment for Educational Progress, 2005 - 2007

Figures 4 and 5 present comparable stem-and-leaf displays of differences, expressed as effect sizes, for the two most recent PISA reading assessments. Figure 4 is restricted to only OECD countries, while Figure 5 includes the results for the partner countries. The effects sizes are based on Tables 6.1.c and Table 6.3a from the PISA 2006 Initial Report, Volume 2 (OECD/OCDE, 2007) and were calculated in analogous fashion to the PISA mathematics effect sizes described above. The figures indicate that changes in Reading scores larger than those experienced in LTT and Main NAEP. For example, from Figure 4, 5 of the 28 effect sizes for OECD countries are larger than the largest effect size observed in the NAEP reading assessment. If the partner results are included (Figure 5), there are 8 of 38 effect sizes larger than anything yet observed in NAEP. Determining whether it is reasonable to expect larger changes in reading performance country-by-country in an international context than state-by-state within a country like the U.S. or whether the large change results for particular countries make sense requires an understanding of educational policies and issues that the authors of this report do not claim to have. What can be said, however, is that it does appear that assessment-to-assessment changes in reading results larger than those observed for the U.S. NAEP occur in PISA with some frequency.

Figure 4 – Histogram of Changes in Average Scales Scores, Expressed as Effect Sizes PISA Reading 2006-2003 (OECD Only)

-0.2	0
-0.1	5
-0.1	3 3
-0.0	5 6 6 6 7 7 7 8
-0.0	1 1 3
0.0	0 0 0 0 1 2 2 3 4
0.0	6
0.1	1 1
0.1	
0.2	2
0.2	

Source: PISA 2006: Vol. 2 Data

Figure 5 – Histogram of Changes in Average Scales Scores, Expressed as Effect Sizes PISA Reading 2006-2003

-0.2	0 2
-0.1	5 5
-0.1	0 1 3 3
-0.0	5 5 6 6 6 7 7 7 8
-0.0	1 1 2 3 3
0.0	0 0 0 0 1 2 2 3 4
0.0	6 6
0.1	1 1 1
0.1	
0.2	2
0.2	7

Source: PISA 2006: Vol. 2 Data

Figures 6 through 8 present stem-and-leaf displays of the state-by-state differences in average NAEP Mathematics results, expressed as effect sizes, between successive assessment cycles for the same time periods as were shown for NAEP Reading. Effect size differences greater than 1 are far more frequent than in NAEP reading. Effect sizes in excess of .20 are not uncommon and in one instance an effect size as large as .36 was observed. There are two additional aspects of the NAEP data worth commenting on. First, the largest differences are associated with the 2003/2000 comparison. Second, the vast majority of the changes are positive. In interpreting this pattern, it may be relevant to note that the 2003 was the first NAEP mathematics assessment after the passage of U.S. No Child Left Behind (NCLB) Act which mandated state-level accountability testing, established economic consequences for states based on improvements in state test scores and required state participation in NAEP. These facts and data underscore, in our view, some of the challenges involved in interpreting whether trends results are "sufficiently" stable. In our view, such judgements require substantial knowledge of relevant demographic and educational-policy trends for the jurisdictions in question and cannot be made solely on the basis of statistical or quantitative criteria[8].

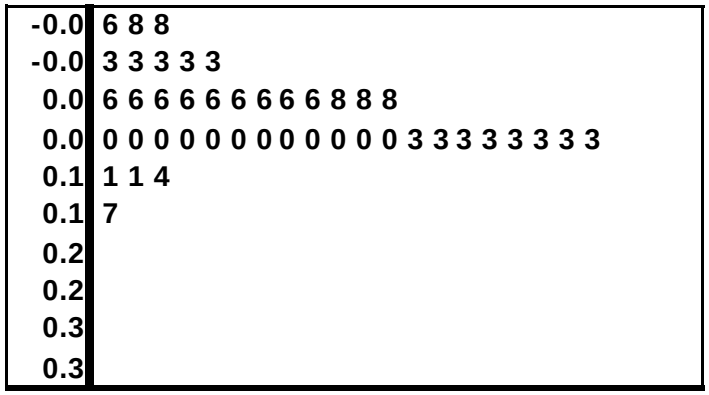
Figure 6 – Histogram of Changes in Average Scales Scores, Expressed as Effect Sizes Grade 8 State NAEP Mathematics 2003-2000

-0.0	
-0.0	3 6
0.0	6 6 8 8 8 8 8 8
0.0	0 3 3 3 3 3
0.1	1 1 1 1 1 1 1 1 4 4 4
0.1	7 9 9 9 9 9
0.2	2 2 2 2 2 2 2 2
0.2	
0.3	
0.3	6

Source: National Assessment for Educational Progress, 2000 - 2003

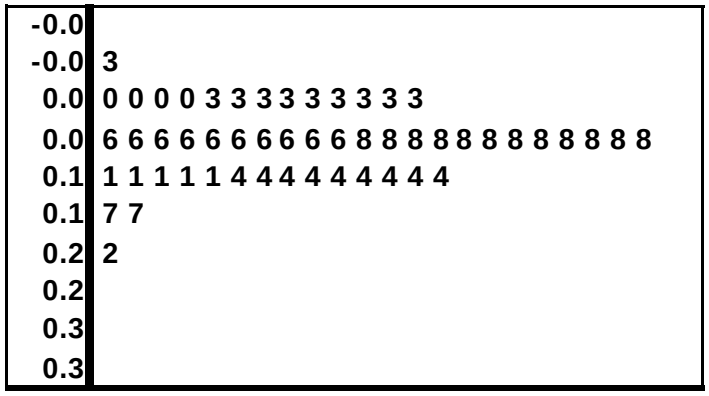
[8] State-by-state science assessment results for main NAEP are available for only two assessment cycles (2005 and 2003) and are not included here. However, results are quite in line with those of Reading and Math.

Figure 7 – Histogram of Changes in Average Scales Scores, Expressed as Effect Sizes
Grade 8 State NAEP Mathematics 2005-2003



Source: National Assessment for Educational Progress, 2003 - 2005

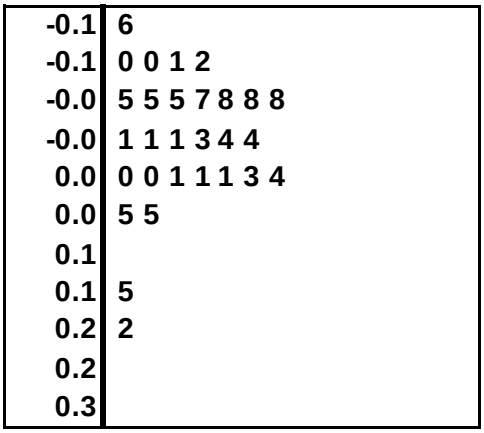
Figure 8 – Histogram of Changes in Average Scales Scores, Expressed as Effect Sizes
Grade 8 State NAEP Mathematics 2007-2005



Source: National Assessment for Educational Progress, 2005 - 2007

Changes in PISA average Mathematics scores between 2006 and 2003, expressed as effect sizes, are shown in Figures 9 (OECD countries only) and 10 (with partner data included). The absolute value of changes in the PISA mathematics results look comparable to those encountered in NAEP, particularly for the post NCLB (2003 – 2007) period. Unlike the NAEP results, the PISA results reflect a more even balance between positive and negative changes.

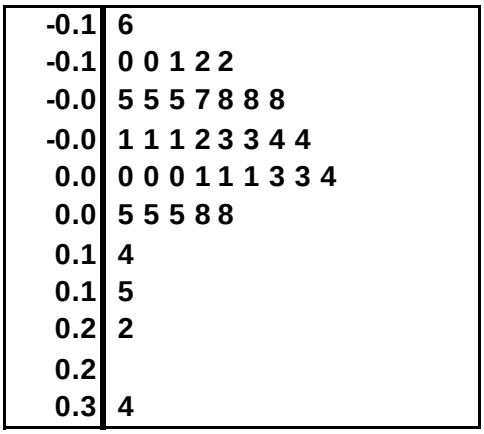
Figure 9 – Histogram of Changes in Average Scales Scores, Expressed as Effect Sizes
PISA Mathematics 2006-2003 (OECD Only)



-0.1	6
-0.1	0 0 1 2
-0.0	5 5 5 7 8 8 8
-0.0	1 1 1 3 4 4
0.0	0 0 1 1 1 3 4
0.0	5 5
0.1	
0.1	5
0.2	2
0.2	
0.3	

Source: PISA 2006: Vol. 2 Data

Figure 10 – Histogram of Changes in Average Scales Scores, Expressed as Effect Sizes
PISA Mathematics 2006-2003



-0.1	6
-0.1	0 0 1 2 2
-0.0	5 5 5 7 8 8 8
-0.0	1 1 1 2 3 3 4 4
0.0	0 0 0 1 1 1 3 3 4
0.0	5 5 5 8 8
0.1	4
0.1	5
0.2	2
0.2	
0.3	4

Source: PISA 2006: Vol. 2 Data

Section 1.3 – Summary

So, to summarize, we recommend taking a pragmatic approach to evaluating precision and stability. With respect to precision, we feel the most fruitful course of action is to examine the magnitude, in effect-size terms, of the trend differences currently being declared as statistically significant. If it is judged that there are too many *prima facie* large differences that fail to reach statistical significance, then changes to design parameters – most likely, the number of schools and students sampled for the assessment – will need to be made. As a point of reference, we provided information regarding the magnitude of trend differences observed in NAEP and the extent to which such differences are declared as statistically significant. We carried out a similar analysis for the two most recent PISA assessments and found similar results to those obtained in NAEP.

We feel similarly about the issue of stability. We know of no credible evaluative criteria to appeal to other than the experience of other assessment enterprises that share

similar goals and features as PISA. To that end, we provided as a point of reference, data from main NAEP and LTT NAEP regarding the magnitude of cycle-to-cycle changes in U.S. national and state-by-state changes. We conducted a similar analysis based on the PISA data from the two most recent PISA assessments. For the mathematics assessments we found that cycle-to-cycle changes in PISA country-level average scores are similar in magnitude to those encountered for state-level results in Main NAEP. For the reading assessments, we found that large cycle-to-cycle changes in PISA average scores are more common than what we have experienced in NAEP.

The analyses presented here were meant to be illustrative. We focused on changes in overall average scores. Similar analyses could be carried out to examine things like the stability of gender differences, of changes in percentile locations, of changes in relationships between background variables and performance, etc. Furthermore, we used NAEP as our point of reference because we are familiar with its results and methods, and because the stable measurement of trend has been a key priority of the program. International multi-language assessments like PISA face challenges over and above those faced by single-language national assessments like NAEP. So, the NAEP results most likely present a conservative benchmark against which to evaluate current precision and stability levels in PISA. Results from other international assessments may provide additional helpful comparative criteria in this regard.

Section 2 – Thoughts on the PISA Test Design and the Number of Link Items for Each of the Assessment Domains

Section 2.1 – On the Current PISA Design

PISA, like other large-scale survey assessments such as NAEP and TIMSS, makes use of matrix sample designs in which samples of students from participating jurisdictions respond to a sample of test questions from a much larger collection of questions. There is much similarity across such assessments in how these matrix samples are implemented. Typically, the pool of items that make up the assessment for a given domain are assembled into a set of non-overlapping collections designed to be administered as a separately timed unit. In PISA, these units are referred to as clusters; in NAEP they are referred to as blocks. From these clusters, a collection of assessment booklets are constructed, each of which consists of multiple units[9]. In PISA, each booklet contains four clusters of assessment questions. In NAEP, depending on the assessment, each booklet consists of two or three blocks. With rare exception, each assessment booklet shares units in common with one or more of the other assessment booklets – a design feature that facilitates analysis using IRT methods and the combining of results based on the different assessment booklets.

Whether it be PISA, TIMSS, or NAEP, it is typically the case that multiple content domains are assessed in each cycle. PISA has assessed Reading, Mathematics, and Science in each of its three assessment cycles, along with, other special areas of focus such as problem solving. TIMSS assesses Mathematics and Science in each assessment cycle. NAEP conducts assessments annually, with different combinations of subjects being assessed in any given year. Main NAEP Reading and Mathematics assessments are conducted every other year, Science and Writing assessments every fourth year, and other subjects, such as Geography and History, far less frequently. So, for example, in 2005, main NAEP assessments were conducted in Reading, Mathematics and Science. In 2007, main NAEP assessments were conducted in Reading, Mathematics and Writing.

The different assessments (PISA, NAEP and TIMSS) have taken somewhat different design approaches to addressing the challenges of assessing multiple subjects in each cycle. PISA and TIMSS have consistently made use of “mixed” designs, in which each assessment

[9] Some survey assessments also include separately timed sections of background or attitudinal questions in assessment booklets, in addition to the sections consisting of items measuring academic achievement . For simplicity of exposition, this distinction is ignored in the current section .

booklet contains clusters from multiple content domains. In PISA, each assessment booklet contains clusters from at least two of the domains being assessed. In TIMSS, each assessment booklet contains clusters of both Mathematics and Science items. From the 1980s on through to 2003, NAEP LTT assessments similarly made use of "mixed" designs, with students being assessed either with booklets that contained both Mathematics and Science blocks or booklets made up of Reading and Writing blocks. In contrast, main NAEP assessments, and LT Trend assessments, beginning in 2003, have made use "focused" designs in which each assessment booklet contains blocks from only a single content domain.

To be sure, mixed designs offer a number of attractive features. In situations like TIMSS, where each student is assessed in both content domains, assessment results for each of the domains can be based on the full sample of schools and students. If, in a jurisdiction of interest, the assessed sample consists of 100 schools and 3 000 students, assessment results for both domains are based on that 100-school/3 000-student sample. For more ambitious designs, like PISA, where students are tested in two of three domains and all pairings of domains occur in the design, the multivariate scaling approaches used in analysis can, in essence, produce a similar result. This is accomplished by estimating results for the missing content domain using item responses from the other correlated content domains and exploiting the relationships between student background characteristics and domain proficiencies estimated from the other portions of the sample. In focused designs, it is typically the case that school and student samples near the size of of those used in the mixed design[10] must be obtained for **each** of the subject areas assessed in order to maintain identical precision levels. In cluster samples like those used in NAEP and PISA, this can usually be achieved in a cost-efficient manner by expanding the student sample sizes within each school. However, practical limitations on the available numbers of students typically dictates some modest expansion of school sample sizes as well. Mixed designs offer other attractive features as well. Sample sizes for school-level analyses are larger in mixed designs than focused designs, making such analyses potentially more tractable and informative. Furthermore, the relationships between achievement in the different domains can be examined in mixed designs – a situation that is obviously not possible with focused designs.

But mixed designs are not without their challenges. Including items from multiple content domains in each booklet means that each student is assessed per unit of testing time with fewer clusters of items in that domain than under a focused design. If students can be tested for relatively long periods of time, as is the case in PISA, this feature of mixed designs may be somewhat less serious. In situations where student testing time is limited, as it is in NAEP, problems can arise if, for example, there is interest in reporting results by subdomains (e.g., algebra, geometry, and statistics within the general domain of mathematics), monitoring trends in the subdomains, or estimating the relationships in performance among the subdomains. With limited student testing time, mixed designs may require rather large numbers of test booklets to be produced and spiralled to achieve stable reporting at the subdomain level. Within this context, the option of focusing the design and increasing school and student sample sizes may afford a more attractive and less complex alternative. The preference in NAEP for using focused designs is in part the result of reconciling the constraint of limited student testing time with the desire for more disaggregated measurement and reporting at the level of subdomains.

The more serious challenge mixed designs are facing is the potential impact of context effects on assessment results, both within an assessment cycle and across time. Context effects, which are discussed again in Section 3 of this report, occur when the

[10] This is the case, strictly speaking, only in situations where multivariate proficiency estimation is used and where plausible values are generated for students in all subject areas assessed, regardless of whether they have been administered items in that subject. Even when NAEP employed mixed booklet designs for data collection purposes in the LTT assessment, proficiency estimation for a given subject area was carried out separately based only on the sample of students administered items in that subject area. Therefore, for NAEP, the move from mixed to focused designs had little impact on the sample sizes required to achieve comparable levels of precision.

psychometric functioning of items or clusters of items differs depending on factors such as the item position within a cluster, the position of the cluster within an assessment booklet, or the other material that an item or cluster is paired with. Depending on which questions were given prior to the current item in question, the response may be more difficult or easier to give for all or some of the students. Intelligence tests use these context effects systematically, in that similar questions of higher difficulty are given later in the test, since students get used to the problem type and are able to respond to more difficult questions as they move along. The standard progressive matrices test (SPM, Raven) is a good example for such a test. However, large-scale assessments like PISA and NAEP operate under other circumstances than intelligence tests in that their booklets are covering a broader content domain with multiple types of questions and response formats. Apart from the position of local context surrounding an item within a cluster, the context in which **a cluster** appears can have an impact on how its questions function. This is particularly true for assessment designs that use booklets with mixed content coverage and which vary the depth of coverage for any given content area from assessment cycle to assessment cycle as is done with the rotating major/minor domain distinction in PISA.

Under the different contexts, students may respond to questions in systematically different ways. This is an effect that can be viewed as an additional source of error rather than a source of systematic variance between participants of the study. For one thing, context effects may not be affecting all students taking the test. Context effects may be observed when assessments combine clusters from different content domains, and use different combinations across booklet and cycles. In addition, in our experience, contexts effects – or their absence – can only rarely be predicted or designed on purpose. They can sometimes be detected after the fact if the right data has been collected. Moreover, in our experience, we see little evidence to suggests that context effects can be presumed to be the same across jurisdictions (like states in NAEP, or countries in international assessments), or even the same across subpopulations within a country.

As an example, some students may be motivated by a science block following a reading block in the first block position, while other students with reading difficulties may be affected negatively by having this arrangement. While some effects of increase in difficulty and omission rates through position cannot be avoided, but can at least be assumed to affect all examinees to some extent, context effects are differential effects; that is, different groups of students taking the test may be affected in very different ways. In essence, like cluster order effects, contexts effects may affect the way students in different countries respond to questions in the assessment. As outlined below, the lack of control over context effects and the strong possibility that such effects, if present, will be differential by country, suggests to us that a prudent approach is to choose a test design strategy that minimizes their likely impact, rather than to use statistical methods to account for these effects after the fact.

Context effects, if present, are problematic because the standard IRT models that are used to analyze all current large-scale survey assessments assume some degree of invariance of psychometric properties, across booklets within a particular assessment cycle, or across assessment cycles, in order for results to be reported on common and comparable scales. If the survey data violate such assumptions, analysis methods must be modified to account for such violations in assumptions and may only be partially successful in doing so.

Context effects may be harder to control in mixed designs than in focused designs, even within an assessment cycle. To illustrate why, let us consider a simple design in which there are only 2 clusters per booklet. Let us denote the Reading clusters as R1, R2, … etc., and Mathematics clusters as M1, M2, …etc. In a focused design, students get two blocks of Reading items. Consider the situation where reading blocks are slightly harder when they appear in the second position than in the first position. When a single IRT model is fit that ignores position within the test booklet, the item parameters represent the average of the item difficulty across the two positions. As a consequence, student reading proficiencies are overestimated, relative to this average, for test takers that get the cluster in position 1 and underestimated for students that get the cluster in position 2. In a focused design, all test takers get two reading blocks, one in the first position, and one in the second position. So, for

each test taker, there is the potential for over- and under-estimation effects to effectively cancel out to some degree. In the corresponding mixed design, each test taker would be administered one and only one reading block and, depending on position, proficiencies would be over or under estimated with no concomitant opportunity for position effects to cancel out. Such context effects, if differential by country or subgroup within country introduce potentially distortions to cross-group and cross-country comparisons.

Across assessment cycles, additional complications can arise with respect to the measurement of trends. Trend measurement is accomplished in survey assessment by ensuring items, or preferably clusters of items, are repeated across assessment cycles. Successful analysis depends on being able to assume that the psychometric functioning – in particular, the average difficulty of the clusters and the relative ordering of item difficulty within the clusters – is the same across assessment cycles. It is typically possible to monitor to some degree whether constancy of item function across assessment cycles holds at an aggregate level – e.g., in PISA at the level of the international scaling sample used to estimate item parameters – and adjust analysis procedures appropriately if the data suggest such assumptions do not hold. However, it is usually far more challenging to determine whether such assumptions of constancy hold when the data is disaggregated – e.g., to the country level – and far less obvious whether and how analysis procedures should be changed if the requisite degree of constancy in not evident.

In light of this, controlling context seems to us an important safeguard to maximize the probability that such assumptions hold and one potentially important aspect of context is the content domain of the other clusters of items. Under focused designs, by definition, Reading blocks always appear paired only with other Reading blocks, Mathematic blocks with Mathematics blocks, etc. This is not the case in mixed designs unless such a feature is explicitly made a constraint of the design, as it was during the period in which the NAEP LTT Trend was administered as a mixed design.

The domains assessed in PISA are Reading, Mathematics, and Science, and each domain serves as the major assessment domain every third cycle, while assessed as a minor domain with less coverage and fewer questions in the other two cycles. Reading was a major domain in 2000, and will be in 2009 again, while Mathematics was the major domain in 2003, as was Science in 2006. This design deviates from other assessments that we are familiar with which tend to provide equal coverage of the construct domains in every cycle. As an example, the National Assessment of Educational Progress (NAEP) assesses Reading and Mathematics every two years, while other subjects are assessed using larger intervals and a different timetable. However, every Mathematics and Reading assessment conducted in NAEP uses item pools and sets of booklets of similar or identical size and covers the full content domain defined in the framework to the same extent. One could say mathematics (or any NAEP subject for that matter) is a major domain each time math is assessed.

The major/minor domain design distinction has served PISA well to date, allowing assessment in Mathematics and Science to be conducted beginning in 2000 and prior to the establishment of fully developed assessment frameworks in these content domains. However, per the proceeding discussion on context effects, we do believe that the major/minor domain design distinction in PISA poses potential complications in that it removes from consideration one of the key tools available for controlling potential context effects in mixed designs – i.e., the option of keeping context constant for trend items. Consequently, simultaneously maintaining a mixed design and the major/minor domain distinction may not be desirable going forward in a system whose principle goal is stable trend measurement.

As an example, there are two clusters of items (28 items in total) that have appeared in all three PISA Reading Assessments. In 2000, when Reading was the major domain, each of these clusters appeared in three assessment booklets as the first, second, or third cluster of material presented to the student. When appearing in positions two or three, all prior clusters administered to the students were Reading clusters. In 2003, the same clusters appeared in four assessment booklets – once in each of the four possible positions within a booklet. When appearing in the second position, one of the two common blocks was preceded by another Reading block (as was the case in 2000) while the other was preceded by a Science block.

When appearing in position four, the Reading blocks were preceded by a sequence of Mathematics blocks.

There is consistent evidence from within PISA that such context effects impact the psychometric functioning of items in general, and the link items, in particular. Given that PISA spirals tests booklets within administration sessions, one would expect that, all things being equal, the average PISA results for students taking each booklet should be more or less the same and differ only within the bounds of sampling variability. Such a pattern of results would be expected within each country, as well as at the aggregate level (i.e., in a combined data set including results from all PISA participating countries). Contrary to expectation, results in each of the three assessment cycles reveal that average scores vary substantially across booklets. In virtually all countries, the variability in booklet means far exceeds what would be expected based solely on sampling variability.

The PISA Consortium is well aware of these "booklet" effects, which have been documented in the PISA technical reports, as well the document TAG(0505)4 and they have implemented innovative statistical methods that attempt to estimate and mitigate the impact of these booklet effects on PISA results. In effect, adjustments are calculated which equalize the average PISA scores for each of the booklets. As we understand it, separate adjustments for each booklet and content area are calculated within each assessment cycle based on an international sample of data. These international adjustments are then used in the subsequent generation of each country's results. While these efforts provide further evidence of the care and technical sophistication of the analysis work done in support of PISA, there are limits to the degree to which such adjustments can fully control for these booklet effects.

It is clear from the document TAG(0505)4, that there is considerable country-to-country variability in the magnitude of these booklet effects. For a number of reasons, conceptual and perhaps technical as well, a single set of international adjustments are used (see pages 13 – 14 from Chapter 9 of the draft 2006 PISA Technical Report). Therefore, within an assessment cycle, the within-country booklet differences are only partially accounted for by the single international set of adjustments. Moreover, there is reason to suspect that the effectiveness of the international adjustments for a particular country may vary from assessment cycle to assessment cycle, and hence, constitute an additional noise component potentially affecting trend results.

We want to be clear here that our point is not to question or criticize the adjustment procedures currently being used, but rather to point out that, in the presence of such kinds of effects there will be limits to what can be accomplished by statistical adjustments. It should be acknowledged that while such effects can in principle have an impact on trend measurement, the actual impact of such effects vis-à-vis the stable measurement of trend is not known. Considerable study using experimental designs would probably be required to disentangle whether changes in the subject matter make up of test booklets, or simply the serial position of the block within a booklet, or both are inducing the kinds of booklet effects that are seen in PISA. Despite these caveats, our strong preference and recommendation would be to make changes to the design of PISA that minimize or remove the possibility of such an effect. We believe such changes would be far easier to accomplish in a design that either no longer maintained the major/minor domain distinction and rotation across cycles or that moved from a mixed design to a focused design.

We believe a very good example of a design that preserves the major/minor distinction but moves away from the mixed design was suggested by Hambleton *et al* (2005). In their design, Reading, Mathematics, and Science are all assessed with separate sets of booklets. Reading booklets contain only Reading clusters, Mathematics booklets contain only Mathematics clusters, etc. What distinguishes major and minor domains in such a design is the number clusters (seven for major domains, four for minor domains) and number of booklets (again, seven and four, respectively). Hambleton *et al*. also included examples of schemes for rotating content across assessments to preserve the ability to measure trend. Furthermore, the Hambleton *et al*. design recommends shortening student testing time from two hours to 90-minutes. We believe this is also a good suggestion in that it may mitigate the potential for factors such as fatigue effects to impact results in unpredictable ways.

As with all designs, there are trade-offs that need to be considered. As noted earlier, for a fixed total sample size (e.g., the 4 500 students as is currently typical in PISA), there would clearly be fewer students contributing to results for each content under the Hambleton *et al*. focused design than is the case under the current design, which fully exploits the possibility of multivariate scaling. Similarly, sample sizes by subgroups within each country are reduced. Moreover under the Hambleton *et al*. proposed alternative design, correlations between performance in the various domains cannot be estimated. Offsetting these negatives are, in our view, some key positives. We believe that the likelihood of booklet effects and the need to adjust for them would be greatly reduced. In NAEP, which has used focused designs for many years, we have seen no evidence of such booklet effects.

Furthermore, presuming that booklets for all domains – major and minor – can be spiralled within each assessment session, we do not believe the reduction in precision will be as severe as might be presumed from a comparison of domain specific student sample sizes. Under simple random sampling, reducing the student sample size by half would result in about a 42 percent increase in the size of standard errors. However, for multistage cluster samples like those used in NAEP and PISA, both the number of schools and number of students affects the precision of the results. If only the latter is reduced under designs like those proposed by Hambleton *et al*. design, reductions in precision will be considerably less, depending on design effects (i.e., the ratio of standard errors for the actual design being used to those obtainable through simple random sampling of students). In NAEP, for example, where design effects for state-level results tend to be between 3 and 4, reductions in precision would be on the order of 20 to 25 percent (Rust, 2008, personal communication). The losses in sampling precision will be further offset, particularly for minor domains, by small increase in the easurement precision resulting from obtaining three clusters of domain-specific item responses for each sampled student. Lastly, any reductions in precision could be further offset by considering some modest increases in school- and student-sample sizes.

Perhaps the biggest challenge in implementing any such design change for PISA will be effecting the transition without disrupting trends. As should be evident by now from the discussion above, simply embedding common items or clusters from previous PISA assessments into a new design is not likely to be adequate due to the potential impact of context effects. The PISA program would likely need to plan for some form of bridge study in which randomly equivalent samples are administered the assessments under the old and new designs to allow for trend results to be adjusted for the impact of the change. One of the key challenges of such a bridge study is that it is often only practical, financially or logistically, to conduct such a study and to carry out such an adjustment at the aggregate level. For example, the vast majority of students in each country might be administered the assessment under the new design, while a small portion is administered the assessment under the old design. Data obtained from the old design, aggregated over participating countries, would provide a sufficient sample to estimate an overall design effect adjustment. However, as is the case with the current booklet adjustments, if the effect of the design change varies significantly country-by-country, no overall adjustment will account for its impact.

Section 2.2 – On The Number of Link Items and The Balance between Major and Minor Domains

Another issue on which we have been asked to comment is that of the number of link items for each of the assessment domains. The number and nature of link items is a very appropriate topic for the PGB to be concerned about. In an ideal world, one would measure trends in a given content domain by repeatedly administering a single pool of items that comprehensively measures the framework associated with that domain. Practical realities such as the need to release items, and in some cases, concerns about the security of items, make this ideal impossible. In PISA, additional challenges are faced in that the amount of time available for testing a given content domain varies from cycle to cycle under the current major/minor rotation scheme. So, in practice, trends are measured on a subset of the item pool.

If the assumptions of the IRT models that are used to analyze survey assessments were true (i.e., the educational constructs defined by the item pools used in the assessments were truly unidimensional), then the size and make-up of the set of trend items would matter little. The rank order of countries in any given assessment cycle, or the changes in average scores for a given country would depend little, if at all, on which subset of the total item pool were selected for use in subsequent assessments. In point of fact, it is generally acknowledged that educational constructs are multifaceted and the unidimensional summaries typically employed to analyze such surveys capture, at best, the overall average performance across the different facets of the construct. Writing about NAEP scaling procedures, Mislevy (1990) acknowledged this point:

> "We hasten to point out that any procedure of aggregation, from a simple average to a complex multidimensional scaling model, highlights certain patterns at the expense of other potentially interesting patterns that may reside within the data. ….The choice of an aggregation procedure must be driven by a conception of just which patterns are salient for a particular purpose" (p. 230)

Consequently, all survey assessments, whether they are single-language national assessments like NAEP, or multi-language international assessments like PISA, can expect to observe variability in assessment results, trend results in particular, if one looks within the overall collection of items used to measure trend. States and countries differ in terms of curricula emphases and instructional practices, both at a given point in time as well as over time. Add to that the additional challenges associated with translation and ensuring comparable scoring for constructed-response items, and it would be surprising indeed **not** to find considerable item-by-country interaction. In our view, this is not an indication of "unstable trend measurement" *per se*. We would argue that stable trend measurement is observed if country-by-country trend results are relatively invariant over multiple **collections of trend items**, each of which was considered *a priori*, an appropriate measure of the intended construct. When dissected, each such collection might show considerable country-by-country variability in results. However, so long as results in the aggregate are stable, we would argue that stable trend measurement *vis-à-vis* the overall construct has been achieved.

Such considerations highlight the importance of making the overall set of items on which trends are based, sufficiently large and sufficiently representative of the full content domain framework to ensure that the assessment results serve what we see as their intended purpose – to report reliably on the overall trend with respect to the full construct defined by that assessment's framework. The PISA Consortium appear quite conscious of this issue and have been investigating its impact on overall trend measurement and its implications for the selection of trend items (Gebhart and Adams, 2007; TAG(0505)4). One can see the impact of these investigations, at least partly, in their *Proposal for Securing Trends in PISA 2009* [EDU/PISA/GB(2008)1], in the way the selection of trend items has evolved over time, and in the careful way the program had approached the reporting of trend results.

Mathematics was a minor domain in 2000 and the major domain for the first time in 2003. As a major domain, the total assessment for Mathematics consisted of 85 test items grouped into seven distinct clusters, a total of 3 ½ hours of testing time. The 2000 and 2003 assessments in Mathematics shared only 20 items. Trend results between these two assessments were reported, but only for the two subscales (Space and Shape; Change and Relationships) within the domain with "adequate" item representation in both the 2000 and 2003 item pools. In 2006, 48 of the 85 mathematics items (56 percent) were repackaged into four 30-minute clusters and readministered. The trend between 2006 and 2003 is therefore based on this subset of the original 2003 Mathematics item pool. According to document EDU/PISA/GB(2008)1, three of these same four clusters (36 items) will be administered in 2009 when Mathematics is again a minor domain. Therefore the 2009 trend point is based on 42 percent of the original 2000 item pool, though not in their original cluster configuration while 2009 and 2003 will share 75 percent of their items.

Science was a minor domain in 2000 and 2003. According to the PISA Technical Reports, both assessments consisted of 35 items organized into 30-minute clusters, and, according to Gebhart and Adams (2007), 25 items were common to the two assessments. Trend results were reported between 2000 and 2003 on an interim Science scale. Science was a major domain for the first time in 2006, and the total assessment consisted of 108 items grouped in seven clusters (3 ½ hours) total testing time. Included among the 108 items are 22 items from the 2003 assessment, which permit some degree of trend reporting between 2006 and 2003 with respect to the limited content reflected in these items. However, no trends between 2006 and 2000 are reported given the evolution in content for the science assessment. The original interim (2003 - 2000) science scale and the 2006 - 2003 trend scales are based on different items and are not comparable. According to EDU/PISA/GB(2008)1, plans for 2009 are to use 53 items (49 percent of the 2006 item pool) when Science is again planned as a minor domain.

Reading was the major domain in 2000 and the total assessment that year consisted of 141 items, grouped into nine distinct clusters. This represented a total of 4 ½ hours of testing time. In subsequent assessment cycles (2003 and 2006), 28[11] items of these 141 items (two intact clusters) have made up the entirety of the Reading assessment when it has been a minor domain. According to document EDU/PISA/GB(2008)1, these same 28 items will be administered in 2009 when Reading is again the major domain. Therefore, the trends for Reading are based on each country's performance over time on this 28-item subset (approximately 20 percent) of the original pool of Reading items.

Clearly, one fair question to ask is whether current plans *vis-à-vis* the number and nature of the trend items seem reasonable for 2009, or as a basis for future trend assessments. To our way of thinking, there are two aspects to this question: Is the nature of the trend item set appropriate, and is the number of items appropriate? We turn first to the question of appropriateness: Do the collection of trend items appear to have been selected to be representative of the full domain constructs?

Based on the documents that we reviewed, it is reported that the trend items for both the mathematics and science assessments were selected with the expressed intent that, as a collection, they span salient subdimensions of the overall construct. The degree that this is in fact the case is a matter for others with subject matter expertise to judge and is beyond the scope of our current activity. The content make-up of the Reading trend link maybe somewhat more open to question. The document Draft Technical Note on Comparisons Over Time on the PISA Scales (EDU/PISA/GB(2007), page 8), indicates that linking item clusters which appear in 2003, 2006, and are planned for 2009 may contain a larger proportion of items that measure the reflection and evaluation aspect of the framework than did the 2000 item pool. This may or may not be problematic from the point of view from interpreting the meaning of the trend results. But, in practical terms, it may be adding little in the way of instability to actual trend results. The Reading trend clusters effectively make up the entire 2003 and 2006 Reading assessments, and therefore this differential emphasis does not impact comparisons between these two assessments. Moreover, it appears that plans for 2009 Trend Reporting entail using only data from these trend items to generate the 2009 data point for comparison to earlier years. In future assessments, 2012 and beyond, we presume that the trend collections will be selected, to the extent possible, to be representative of literacy constructs reflected in the 2009 scales.

As to the question of number of trend items, we would approach this issue, as we did the question of trend stability criteria, from a pragmatic standpoint, by comparing current PISA practice with our experience in NAEP. In making comparisons it must be acknowledged that the difference in assessment design – in particular the major/minor domain distinction that exists in PISA but not in NAEP is a complicating factor. This complication

[11] Note that the documentation we were sent is inconsistent is this regard. The 2003 PISA Technical Report and Gebhart and Adams (2007) states that the 2003 Reading Assessment consisted of 28 items, two intact clusters from the 2000 assessment. The 2006 PISA Technical Report states that there were 31 Reading items in 2003.

not withstanding, we will describe current main NAEP practice using the most recent Grade-8 assessments in Reading, Mathematics and Science, to provide at least a frame of reference.

As noted earlier, in main NAEP, each content area is effectively a major domain each time it is assessed. The item pools used for these assessments each contain clusters unique to that year and clusters common to the previous assessment. Both for the item pool as a whole, and for the clusters common to previous assessment, care is taken to assure that the item sets span the key content and process subdimensions specified in the NAEP frameworks. The two most recent Reading and Mathematics assessments occurred in 2005 and 2007. For each of these the two Reading assessments, the total item pool was made up of thirteen 25-minute clusters (about 140 Items) – 325 minutes, or roughly 5 ½ hours, of testing time. The trend link between these two assessments consisted of ten clusters – essentially 77 percent of the total. This was also true for each of the two most recent Mathematics assessments, which occurred in 2005 and 2007. In each of these two assessments, the total item pool was made up of ten 25-minute clusters (about 180 Items) – 250 minutes, or a little over four hours, of testing time. The trend link between these two assessments consisted of seven clusters – essentially 70 percent of the total. The two most recent Science assessments were conducted in 2000 and 2006. The former consisted of 14 clusters (199 items,) the latter, 16 clusters (224 items). Eleven clusters (69 percent of the 2000 pool and 79 percent of the 2005 pool) were common across assessments.

As a second point of reference, it may be useful to look at NAEP practice within its LTT component during the 1980s and 1990s. For LTT, identical instruments were administered at two to four year intervals for an extended period of time. That is to say, there were no unique items. All items were common across assessments. This, in some sense, mirrors the PISA situation in Reading for the 2003, 2006, and 2009 assessments in which trend results are based on responses to identical clusters of items. Using the Age-13 LTT assessments as examples, the Reading assessment consisted of 107 questions based on 43 reading passages, the Mathematics assessment contained 127 questions, and the Science assessment contained 83 questions.

As is evident from the above, NAEP practice has been to base trend results on considerably larger samples of items than has been done in PISA to date. Moreover, NAEP has tended to ensure a larger degree of overlap between adjacent assessments than is the case in PISA, though the more recent plans with respect to Mathematics and Science represent some attempt to expand on the size of the common item pool. No doubt, the reasons for the differences in approach between NAEP and PISA are the result of myriad factors, which may include practical and financial considerations, as well as the aforementioned design differences. It may simply not be practical within the constraints of PISA to approach anything like the levels of item reuse that undergird NAEP trends. Furthermore, to our knowledge, there is no compelling empirical research as to what levels of reuse and overlap are necessary, in a given context, to achieve some predescribed level of stability of trends. Clearly, the PISA consortium's own work in this regard (TAG(0505)4) is relevant and has already influenced design decisions for 2009.

That all being said, our general recommendation would be to consider prioritizing an increase in the amount of linking items included in future assessments in the interest of ensuring the stability of trends. Such an increase should, of course, maintain the recent practice in Mathematics and Science of attempting to ensure representative coverage of the full content domain by the linking items. Even absent specific empirical evidence as to relationship between the size of trend set and we think there is good logical reason to suspect that the stability of overall trends cannot help but be improved by the more robust coverage of the full construct that a larger item pool could provide. Moreover, if there is any interest in reporting trends on subaspects of the domain, these would benefit immensely from the larger sample of trend items.

In order to increase the number of trend items, modifications to current practice in PISA would need to be considered with respect to policies around the public release of items as well as to the balance between major and minor domains. Maintaining a larger number of items as common across assessments will require the PISA program to commit to releasing

fewer items so as to maximize the number and nature of items available to form the link set. Absent other changes, this reduces item development needs and therefore could result in considerable cost savings to the program. In minor domain subjects, however, increasing the number of trend items will almost certainly increase the number of clusters that define the assessment for that year and blur, to some degree, the distinction between major and minor domains. For any design, but particularly for the current design, the only way to accommodate such an increase in the size of the minor domain will be to increase the total number of assessment booklets – which will likely add cost and complexity over current procedures.

Section 3 – Some Other Suggestions on How to Improve the Stability of the Link

Section 3.1 – Further Thoughts on the Importance of Context Effects

The importance of controlling context with respect to trend measurement is well illustrated by NAEP's experience in the 1980s with its Reading Anomaly (Beaton, 1988; Beaton and Zwick, 1990). ETS had conducted its first NAEP assessments in Reading and Writing in 1984, introducing IRT scaling and the use of marginal estimation and plausible values – methodologies quite similar to those used in PISA – and re-expressing results from NAEP assessments back to 1971 on these newly derived IRT scales. In 1986, NAEP Reading assessments were planned as well, along with assessments in Mathematics and Science and the introduction of IRT scaling methods to these subject areas. Several changes were made to the design of the NAEP Reading assessment, in particular to the booklet design, between 1984 and 1986:

1) Passages/items from 1984 were repeated in 1986 but not as part of intact clusters. Therefore, local context was different and item positions within cluster were different
2) Other subjects-areas within the booklet were different. In 1984 Reading clusters were paired with Writing while in 1986 Reading was paired with Math, Science and (for students at age 17, Computer Competence, History, Literature)
3) Timing of clusters was different – one to two minutes longer in 1986 than in 1984, and the number of questions per block increased in 1986 given the extra time that was allocated
4) Booklet formats differed with respect to ink color, line length of reading passages, and response formal (i.e., circle the correct option in 1984 versus fill in the oval 1986)

The original analysis of the 1986 reading trend data showed anomalous results. Average scores dropped precipitously for two of the three age groups assessed and magnitude of the change over this two year period far exceeded anything observed since the inception of the NAEP assessment program in 1971.

The 1986 Reading results were not published at that time. Instead, an experiment was planned and conducted in 1988 to disentangle the aggregate effect of the Reading Assessment design changes made between 1984 and 1986. Two randomly equivalent samples at each age level were selected. One sample was assessed with the identical instruments and procedures used in 1984. The second group was assessed with the 1986 instruments/procedures. Data from both samples were separately analyzed and compared to each other, the differences providing an estimate of the impact of the collective set of changes described above. Results differed by age group, ranging from 2 to 8, NAEP scale score points. In effect size terms, this amounted to changes between .05 and .22. The larger differences (.16 and .22 in effect size terms for ages 13 and age 17, respectively) exceeded in magnitude any of the reported changes in NAEP Reading results between 1971 and 1984. Moreover, as is evident from Table 1, such changes are bigger than any changes in LTT Reading results observed since that time.

As a result of our experiences with the Reading anomaly, the NAEP program has adopted a conservative stance with respect to keeping context as consistent as possible from one assessment cycle to the next. Basically, we have assumed, absent evidence to the contrary, that all changes potentially matter and should be avoided where possible. We would encourage the PISA program that, to the extent possible, a similarly conservative stance be adopted in the future. NAEP relies entirely on repeating intact clusters of items as its means of linking results from one assessment cycle to previous assessment cycles. For mixed designs – like those used in LTT during the 1980s and 1990s, context was held constant by repeatedly administering the exact same assessment booklets from one cycle to the next. We believe that the degree to which such policies can be emulated in PISA, the risks of encountering context effects – that potentially add instability to trend results – can be reduced. When changes to test designs that can impact context are considered in NAEP – e.g., the move from mixed designs to focused designs in LTT NAEP that occurred in 2003 – such changes are accompanied by "bridge studies" modeled after the Reading Anomaly experiment that are designed to estimate and appropriately adjust trend results for the potential impact of said changes.

Section 3.2 – Position and Passage effects

The order of presentation may have a significant impact on the response behavior of examinees. Questions that appear in clusters early in the test booklet may seem easier, while the same question given in the last cluster of a booklet may seem more difficult, or more prone to response omission, or both, since students tend to run out of time at the end of the booklet and may rush to responses or do not respond at all. This may be due to the fact that examinee fatigue may increase, or motivation may decrease, or examinees may run out of time, so that responses are either more prone to errors, or are omitted completely.

The effect of this change in omission rates or increased difficulty is that questions presented at different block positions may be functioning differently. This may be true homogeneously for all questions in a cluster, in which case it may be possible to correct for this effect using a cluster order parameter similar to the facets model as has been done in past PISA analyses. However, if the effects differ for different item types, or are not homogeneous across questions, or translated versions of question, such analysis-based approaches may only partly correct for such effects. The need to correct cluster order effects by means of block-position facet parameter may also constrain the design for future assessments, since certain design features have to be applied in the same ways so that the same cluster order parameters can be assumed to hold for linking across different assessment cycles.

The possibility that item position may affect item functioning leads us recommend that PISA maintain a high level of consistency across assessments in three areas: (1) cluster timing and mode of presentation; (2) cluster position; and (3) cluster composition. The first recommendation is meant to ensure that questions are presented under the same time constraints and the same instructions to test takers across assessment cycles, so that omission rates and functioning will not be affected by these factors. The second recommendation of maintaining cluster position is meant to make sure that in the presence of position effects, these are held constant by presenting the same questions in the same cluster positions at the same rates. Finally, keeping the clusters intact will ensure that within the cluster, a question will appear in the same context and relative position in that cluster.

Apart from being an economic method for balancing item positions, the balanced incomplete block design used in PISA and other large scale assessments provide another advantage for test construction: Tests that involve a common stimulus such as a reading passage can be constructed much easier. These passage or common-stimulus arrangements are used to pose a set of questions, all related to the same passage. Clusters are a 'natural' design feature for this type of assessment, since passages and related questions obviously have to go together.

However, for all large-scale assessment – not just PISA – the passage or common-stimulus design carries some threat to the precision of the link due to the fact that questions

using the same common passage as a reference may be statistically dependent in ways that go beyond what a single student variable may be able to explain. In statistical terms, this means that one of the basic assumptions, local independence, used in the measurement models for these assessments is potentially violated through this arrangement of several questions under one stimulus. Student responses to questions based on passages of text carry information beyond their ability to read; every passage has a topic, and students may vary with regard to their familiarity with the topic. In addition, when students have difficulty with one part of the passage, where, for example, they may encounter a word they are not familiar with, students with such a difficulty understanding parts of the stimulus may be disadvantaged on more than one of the subsequent questions.

As mentioned above, clusters that involve passages or common stimuli required for a series of questions may not be adequately modeled by methods that rely on the assumption of local independence given unidimensional student proficiency variables. Responses to questions based on the same stimulus may still be statistically dependent, even if the students overall proficiency is known. This is due to the fact that variables other than the overall student ability may affect the responses to several questions at once, if these questions refer to one common stimulus. These local dependencies can be adequately modeled using so-called testlet models or multidimensional extensions of IRT models that allow specification of an overall factor (the ability of interest) and specific factors that reflect what is unique in terms of systematic variance of questions within one common stimulus group. Wainer *et al*. presented the testlet model, which is a constrained version of a well known bi-factor model design known both from structural equation modeling as well as multidimensional IRT (Thissen and Steinberg, 2008; Xu and von Davier, 2006; Brandt, 2008). The effect of passage dependencies on measurement are usually seen in a reduction of reliability of the assessment, since the number of independent items is effectively reduced if clustered under a common stimulus. More severe effects can be seen in linking or trend designs involving passage based assessments, since the unique sources of variances can lead to undesirable distortions of the direction the trend takes (Monseur and Berezner, 2007). Verhelst (1997) has taken a simple, but effective approach to eliminate the effect of local dependencies: The sum of responses per passage is used as a polytomous item per passage instead of using the questions individually. This can be shown to be a permissible data reduction under the Rasch model. The sum score of the passages follows the partial credit model if the collection of the items per passage follow a Rasch model (Verhelst, 1997).

We would recommend that all large-scale assessments, PISA included, take some steps aimed at reducing passage effects. These are: (1) reducing the number per prompt or passage, while, at the same time, (2) limiting the length of passages. In addition to reducing the number of dependent observations in the tests, this allows the program to administer more independent passages within the same assessment timeframe. Such test construction practices also have the potential to reduce the linking error due to the passage structure of the assessment. As Monseur and Berezner (2007) point out, the release and drop of passages from the linking set of two consecutive assessments may have substantial effects on the stability of the trend.

We must acknowledge, however, that implementing changes in test construction to reduce passage effects carries with it its own threats to trend stability. Making such changes, in particular limiting passage lengths in reading assessments, carries with it its challenges to trend in that, by doing so, one could be to some degree changing the nature of the construct being measured. Thus, such changes are best introduced at a point in time when new frameworks, and new trend lines based on instruments from those frameworks, are being established.

If the arrangement of questions under a common stimulus in an assessment leads to local non-ignorable dependencies between these questions, Verhelst's suggestion to use sum scores and polytomous IRT measurement models for these is a practical and appropriate remedy. Alternatively, explorations with Rasch or more general multidimensional IRT models that allow for within-item-multidimensionality may be carried out to assess the effect of the specific, passage related factors on the results in more detail.

Section 3.3 – Controlling comparability of open response scoring:

Questions with a multiple-choice response format can be automatically scored as correct or wrong, while questions that allow for an open (or constructed) response may need to be scored by human scorers. These scorers evaluate the response and classify it in one out of several response categories. These response categories are often also dichotomous, so that correct versus incorrect is distinguished, or partially correct responses are considered in addition, by allowing for a score with multiple, ordered categories that reflect the level of correctness. Constructed response questions used in large scale survey assessments may allow for three or even four ordered levels of 'correctness' where the highest level indicates a complete and correct answer, and levels below the highest indicate partially correct responses.

Scorers who evaluate constructed responses and provide scores for these responses need to be trained to ensure that they adhere to comparable scoring rules. These rules should be applied consistently over the course of scoring one question across multiple students within one assessment cycle, as well as over multiple assessment cycles, and across participating countries. The consistency of the application of the scoring rules, and the quality of these scoring rules is crucial, so that every answer can be classified uniquely into one of the rating categories, independently of who rated the response. If scorer dependent variables have an adverse impact on this process, the reliability of the questions with human-rated response decreases. This potential decrease is more severe for open-ended questions that require complex productions as responses and holistic evaluation of these than for questions that require relatively short phrases or constructions, or written entries of results. Short open-ended responses can be rated generally more reliably than extended constructions, which undergo more of an integrative process on the side of scorers, who have to find a single adequate category for these responses.

This leads to the question of how to test whether the scoring process adhered to the same rules across scorers, assessment cycles, and countries. One way to answer these questions is to design a set of studies that examine whether scoring is consistent in all the levels mentioned above. Scoring equivalency studies use multiple scores provided by multiple scorers as the basis to test whether scoring rules are applied consistently by different scorers. Responses from previous assessments are rescored by scorers working on current assessments in order to check whether 'old' responses are scored in the same way as 'new' responses from the current cycle. Scorers from different countries may receive responses to be (re-)scored in a language often spoken across countries as the first foreign language. As an example, responses in English collected in countries where this is the official language may be used in English-speaking countries as well as in countries where other languages are spoken, but scorers who are also proficient in English can be found in sufficient numbers.

Apart from these off-line or asynchronous rescoring studies, online or real-time scoring monitoring may also be advisable if resources permit. In these real-time studies, the scoring process is monitored, and scorers are given feedback about their performance compared to other scorers as they go along. This way, drift while scoring can be avoided, or monitored, and rescoring of subsets of the responses may be carried out to keep the scoring process on track.

The statistical methodologies used to analyse off-line scoring equivalency studies involve measures of agreement for categorical ratings. Of these, Cohen's kappa (Cohen, 1960) is one that is simple to compute and maybe the most commonly used statistic, while there are alternative approaches that are mathematically more involved, but allow more thorough and in-depth analysis across scorers (Klauer and Batchelder, 1996), and allow one to test specific hypotheses about the cognitive process involved in the scoring of constructed responses.

The need for comparable measures over time and across participating countries suggests that the task material and the response format should facilitate reliable rating processes. This is, extended responses should be used sparingly, and short constructed responses should be preferred, since these can be rated and categorized into a limited number

of categories more reliably. In addition, extended constructed response questions take a lot of time, and provide comparably less reliable information, therefore, the use of such item types in trend or anchor sets is less desirable than the use of question types with short responses. The number of response categories used in scoring open ended responses should also be considered carefully. These categories have to be anchored by example solutions and verbal descriptions; finding unique examples, and translating these into languages used for the assessment, or not translating example responses and asking scorers to make this leap of transfer is not trivial. Short constructed response questions ratings can be trained much more efficient, since responses to these questions are less ambiguous. As noted above in the discussion of passage effects, changes in the mix of item types within the assessment are probably best implemented at those points in time where new frameworks or trend lines are to be implemented.

We know that PISA currently invests considerable effort in monitoring the equivalency of constructed-response scoring across countries and languages within an assessment cycle and strongly encourage the Consortium to maintain their standards in this regard. However, we would also encourage PISA to carefully consider current procedures for monitoring comparability of scoring across time points in each country where necessary. We feel that with the increase in the number of participating countries seen in each cycle, PISA should consider investing further resources to strengthen processes and procedures devoted to scoring equivalency, particularly equivalency of scoring at multiple time points, as needed.

Section 3.4 – Country-by-item interactions, and the effect of the model chosen for measurement:

The statistical methodologies use to derive measures of student proficiency from a series of responses to questions administered in PISA are based on the assumption of a systematic relationship between the likelihood of a correct response and an underlying proficiency variable. This relationship is assumed to be a parametric mathematical function. In the PISA assessment, this is the logistic function with one location parameter (often referred to as the Rasch model). This location parameter describes the difficulty of a question, i.e., it is a measure of how likely a correct response is given a certain level of a student's proficiency. In international assessments, the difficulty of a question is assumed to be the same across translations and participating countries, while the distribution of proficiencies across countries can vary freely.

Students with a high level of proficiency are assumed to respond correctly with high probability, and students with a low proficiency will most likely respond incorrectly. However, for two questions with the same location (difficulty) parameter, the probability of a correct response at a given level of proficiency will be exactly the same under the Rasch model used in PISA. Questions from the same content domain are, in terms of the prediction of student responses, completely determined by their location relative to the proficiency measured.

The invariance of this systematic relationship between response and proficiency, for the same question posed at different times, or presented in different translated versions across countries, is one of the fundamental assumptions for constructing a comparable proficiency scale across assessment cycles and countries. The invariance assumption may be violated with regard to some link items, that is, relative to the other link items. In that case, the items in question can not be described by the same location across all assessment cycles. This is usually detected during statistical analyses with so-called item-fit diagnostics; these are statistical checking methods that show whether the relationship between student's responses to the questions and their proficiency is stable over time within some limits of tolerance.

There are multiple reasons for violations of this invariance assumption. Some are related to changes in the student population assessed, and some are related to the choice of the model used to derive proficiency information from student response data. Others are more trivial reasons that nevertheless in consequence may violate the basis for maintaining the status of a link item for trends reported over time.

The choice of a measurement model is an important decision for any assessment program, and is also one possible source of lack of model-data fit for some subset of questions within as well as across countries and assessment cycles. The model used in PISA and other large-scale assessments can be described as a multiple-scale, simple-structure item response model, where each question is assumed to belong to exactly one sub-domain of the content area of interest. In mathematics, NAEP used to distinguish five content sub-domains, while in recent assessment frameworks, this number was reduced to four subdimensions. In Reading, dependent on the grade level, there are two or three subdomains. While every question is assumed to belong to one subdomain only, these domains are not assumed to be independent of each other. Quite to the contrary, the modeling and reporting assumes and estimates correlations between sub-domains, and these domains turn out to be substantially correlated in PISA, NAEP and other assessments.

The assumption of a single domain per question is restrictive and may be violated in cases where for example, passage effects are not negligible. In that case, specific factors may be introduced in order to model the dependency of responses on an additional passage or unique factor. Models of this type can be specified in multidimensional IRT using a bifactor structure, or in so-called testlet models, which are constrained versions of the bifactor model as discussed in section 3.1.3.

Another, maybe more relevant source of modeling error is the assumption that items of different type contribute the same amount of information to the measurement of student proficiencies. While this assumption is a very useful one to maintain parsimony in tests where very similar items are presented, it may be less so in cases where mixed open (short and extended constructed responses) and closed (multiple-choice) item forms are presented in the same instrument.

In the mixed-format case, some items may contribute more reliable information about student proficiencies than other item types. Models that can accommodate and estimate these differences are using discrimination parameters that are estimated (for example, in the case of the 2PL or generalized partial credit model), or imputed (for example the OPLM – the one parameter logistic model). These models can be shown to have statistical advantages in terms of model-data fit in most real data from assessments involving mixed item types (see, for example, Haberman, (2005a, 2005b), and von Davier *et al*, (in press)).

Apart from considering more general IRT models to accommodate differences between items within the same and different language versions, there have been empirical studies on the effect of country-by-item interactions. Gebhardt and Adams (2007) have studied the effects of variation of average item difficulties for the unique items between country specific calculation and the international parameters used in PISA. This study gives some indication of how much impact can be expected in the face of some level of difficulty variation across translations. The studies conducted by Monseur and Berezner (2007) and Monseur, Sibberns and Hastedt (2008) may also be interpreted as examinations of the effect on linking when there is variation in how questions are responded to that is unrelated to the targeted proficiency variable. In the Monseur *et al*. studies the effect of dropping questions from the linking were examined. The effects seen in these studies can be attributed to model misspecifications for the questions omitted. One reason of these model errors can be attributed to questions that function differently in different subgroups, or that may measure somewhat differentially across groups.

A study by Park and Bolt (2008) addresses a model-based approach to quantifying the differences between question functioning across different translations, or across groups presented with different versions of an assessment. These authors utilize extensions of item response models that include a random variance component in addition to the overall item difficulty estimated across groups. The model-family used can be described as Random-Item models (DeBoeck and Wilson, 2004; Jannsen, 2002), which take into account that difficulties of questions may exhibit small variations across modes of assessments, or across assessment cycles or samples (or translations) assessed. While explorations with these models are useful to learn about these variations, they cannot provide an alternative to operational standards and procedures aimed at minimizing effects of translations and country specific adaptations of

questions on the statistical properties of the items presented in different countries or languages.

There are models that can be used to check whether the collection of threats is indeed an issue for the assessment at hand, but rather than modeling deviations from local independence using passage-specific factors, or using facet parameters to reduce block-order effects, assessment designers strive for avoiding these adverse effects by optimizing the test design. However, there are issues that cannot be completely avoided, and for these, tools like multidimensional IRT and random item models are useful to assess the extent to which the assumptions of the simpler, operational model are not met.

Our suggestion for the PISA consortium is to devote some resources for explorations of more general modelling approaches to study the effect of differential item functioning across assessment cycles and countries under various item response models. We recognize that the Rasch model chosen for PISA has unique mathematical properties, and there are good reason to use a model that involves less rather than more parameters. However, we feel that there is some justification for decision by NAEP (and other assessments) to go with a more general IRT model (2PL/Generalized Partial Credit model, and 3PL) in the face of an assessment that is designed to provide a broad coverage of the domain using multiple item formats and test versions. In our experience, these more general IRT models do accommodate the functioning of items in diverse populations better than the Rasch model, which assumes that all items contribute the same amount of information to the measurement of student proficiencies. We assume that using a more general IRT model may also help reducing some of the country–by–item interactions observed in PISA, since the adoption of a more general measurement model improves model-data-fit considerably in our experience.

Conclusion

In the initial stages of assessments like PISA and state-by-state NAEP many questions of educational policy interest are addressed by examining the relative rankings of countries (in PISA) or states (in NAEP) and attempting to correlate such rankings and levels of performance with educational organizational or policy variables. There tends to be a considerable degree of variability in the average performance levels across the jurisdictions, and, as a result, relative rankings of countries within an assessment, and even across assessment cycles, are often shown to be fairly robust across different subsets of the item pool or to the use of different data analysis models. We would guess that relative rankings are also quite robust to issues associated with test design.

When the focus becomes measuring trends in performance over time for countries and states, where it generally assumed that the magnitude of changes is typically much smaller, program procedures and design issues may have a greater impact. As Al Beaton (1990) wrote in reflecting on the so-called NAEP Reading Anomaly

> It should be noted that the inferences of IRT are valid given the truth of the assumptions, but the assumptions may not be true; they are assumptions about the state of nature, not natural laws,….changes in format and context that may be considered negligible when comparing individuals may not be negligible when comparing differences in subpopulations over time. In the particular case of NAEP, the effects of changes in measurement were apparently larger than the trend effects being measured. Thus, maintaining identical instruments is critical when looking for small differences. (p. 11)

The PISA program represents an impressive and ambitious endeavour. Its current policies, procedures, test designs and analytic approaches are clearly state-of–the-art for international assessments and have served the program well for its first nine years. Certainly any review of the technical documentation indicates that, within current program goals and constraints, much care has been taken in design and analysis to minimize the impact of methodological artifacts on the program results. However, if accurate trend reporting is now

to be the focus and it is judged by important constituencies that current levels of trend stability are not adequate , we believe the program will need to invest greater resources in ensuring that sample sizes, quality assurance procedures, test designs and accompanying analysis procedures are attuned to that more ambitious focus. We have tried above to lay out some general directions for change that the program might consider.

References

Beaton, A. E. (1988), *The NAEP 1985-86 Reading Anomaly: A Technical Report*, ETS, Princeton, New Jersey, USA.

Beaton, A. E. and R. Zwick, (1990), *Disentangling the NAEP 1985-86 Reading Anomaly*, Princeton, New Jersey, USA.

Brandt, S. (2008), "Estimation of a Rasch Model Including Subdimensions", in M. von Davier and D. Hastedt (eds.), *Issues and Methodologies in Large-Scale Assessments*, IEA-ETS Research Institute, Hamburg, Vol. 1. p. 53-71.

Cohen, J. (1960), "A Coefficient of Agreement for Nominal Scales", *Educational and Psychological Measurement*, Vol. 20, No.1.

von Davier, M., X. Xu and C. H. Carstensen (in press), *Measuring Learning and Change with the General Diagnostic Model*, ETS, Princeton, New Jersey, USA.

DeBoeck, P and M. Wilson, (2004), *Explanatory Item Response Models*, Springer, New York.

Emerson, J. D. and D. C. Hoaglin (1983) *Stem-and-Leaf Displays*, in D.C. Hoaglin, F. Mosteller and J.W. Tukey (eds.), *Understanding Robust and Exploratory Data Analysis*, John Wiley & Sons, New York.

Gebhardt, E. and R. J. Adams, (2007), "The Influence of Equating Methodology on Reported Trends in PISA", *Journal of Applied Measurement*, Vol. 8, No. 3.

Glas, C.A. and W. van der Linden, (2001), *Modeling Variability in Item Parameters in Item Response Models*, retrieved 05/08/08 from http://eric.ed.gov/ERICWebPortal/custom/portlets/recordDetails/detailmini.jsp?_nfpb=true&_&ERICExtSearch_SearchValue_0=ED467377&ERICExtSearch_SearchType_0=no&accno=ED467377

Haberman, S. J. (2005a), *Identifiability of Parameters in Item Response Models with Unconstrained Ability Distributions*, ETS, Princeton, New Jersey, USA.

Haberman, S. J. (2005b) *Latent Class Item Response Models*, ETS, Princeton, New Jersey, USA.

Hambleton, R.K., E. Gonzalez, B.S. Plake, and I. Ponocny, (2005), *Technical Review of PISA*, unpublished manuscript.

Janssen, R. (2002), "Estimating a Random Effects Version of the Linear Logistic Test Model Using SAS", paper presented at the annual meeting of the American Educational Research Association, New Orleans, Louisiana, USA, April.

Klauer, K. C. and W. H. Batchelder, (1996), "Structural Analysis of Subjective Categorical Data", *Psychometrika*, Vol. 61, pp. 199-240.

Mislevy, R. (1990), "Scaling Procedures", in E.G. Johnson and R. Zwick (eds.), *Focusing the New Design: The NAEP 1988 Technical Report*, ETS, Princeton, New Jersey, USA.

Monseur, C. and A. Berezner, A, (2007), "The Computation of Equating Errors in International Surveys in Education", *Journal of Applied Measurement*, Vol. 8, No. 3.

Monseur, C., H Sibberns, and D. Hastedt, (2008), "Linking Errors in Trend Estimation in International Surveys in Education" in M. von Davier and D. Hastedt (eds.), *Issues and Methodologies in Large-Scale Assessments*, IEA-ETS Research Institute, Hamburg, Vol. 1. p. 113-122.

OECD (2005). *PISA 2003 Technical Report*, OECD, Paris.

OECD (2007), *PISA 2006 Science Competencies for Tomorrow's World*, OECD, Paris, Vol. 1.

OECD (2007). *PISA 2006 Vol. 2: Data*, OECD, Paris.

OECD (in press), *PISA 2006 Technical Report*, OECD, Paris.

Park, C. and D. Bolt, (2008), "Application of Multilevel IRT to Investigate Cross-National Skill Profiles on TIMSS 2003", in M. von Davier and D. Hastedt (eds.), *Issues and Methodologies in Large-Scale Assessments*, IEA-ETS Research Institute, Hamburg, Vol. 1. p. 73-98.

Rijmen, F. and P. De Boeck, (2002), "The Random Weights Linear Logistic Test model", *Applied Psychological Measurement*, Vol. 26, pp. 269-283.

Thissen, D. and L. Steinberg, (2008), "Using Item Response Theory to Disentangle Constructs at Different Levels of Generality", in S. Embretson and J. Roberts (eds.), *New Directions in Psychological Measurement with Model-Based Approaches*.

Verhelst, N. (1997), *Modeling Sums of Binary Responses by the Partial Credit Model*, Cito, Arnhem, The Netherlands.

Xu, Xueli and M. von Davier (2008), "Linking with the General Diagnostic Model", in: M. von Davier and D. Hastedt (eds.), *Issues and Methodologies in Large-Scale Assessments*, IEA-ETS Research Institute, Hamburg, Vol. 1. p. 99-113.

ANNEX: Response to comments by the PISA consortium. Extract from the consortium comments, with replies embedded in Arial font:

Consortium Comments on *Mazzeo and von Davier*: Review of the Programme for International Student Assessment (PISA) Test Design: Recommendations for Fostering Stability in Assessment Results.

The Consortium welcomes the review undertaken by Mazzeo and von Davier and is strongly supportive of the vast majority of the content. It is our suggestion that this review be shared with the Technical Advisory Group, members of which can discuss in detail the recommendations and guide the Consortium in the construction of an action plan.

Before providing an initial commentary on the report we provide a summary of the key recommendations and observations made in the review.

Observations

1. The degree to which PISA trends are more or less stable than other comparable assessment programs and the extent to which such instability is a function of test design issues seems to us to be a very much an open question…We know of no credible evaluative criteria to appeal to other than the experience of other assessment enterprises that share similar goals and features as PISA. Further, the reviewers state that an assessment of the adequacy of the PISA trends is a judgment call outside the range of their expertise.
2. The current PISA test design and analysis procedures have been capably developed by the current Consortium to be responsive to the values of the PISA program with respect to these competing forces. Without reconsideration of current constraints (both practical and fiscal) and values, test design improvements may be difficult to achieve.
3. Using interim trend scales, which have been discontinued when full frameworks are established (e.g., in Science), or in restricting trend reporting to only those subdomains with adequate item coverage in prior assessment cycles seem to us to be very wise policies indeed and, other things being equal, are certainly courses of action that we would endorse or recommend were we operating under current program constraints.
4. Perhaps the biggest challenge in implementing any design change for PISA will be effecting the transition without disrupting trends….The PISA programme would likely need to plan for some form of bridge study in which randomly equivalent samples are administered the assessments under the old and new designs to allow for trend results to be adjusted for the impact of the change.

Recommendations

1. There is consistent evidence from within PISA that such context effects impact the psychometric functioning of items in general, and the link items, in particular. It is therefore recommended that
    a. PISA maintains a high level of consistency across assessments in three areas: (1) cluster timing and mode of presentation; (2) cluster position; and (3) cluster composition. In fact the use of intact clusters is preferred.
    b. a focused design (rather than mixed design) be used and believe a very good example of such a design was suggested by Hambleton and associates.
    c. the removal of context effects would be far easier to accomplish in a design that no longer maintained the major/minor domain distinction.
2. Priority should be given to increasing the number of linking items included in future assessments in the interest of ensuring the stability of trends.

3. The *passage* or *common-stimulus* design that is used by PISA carries some threat to the precision of the link due to the fact that questions using the same common passage as a reference may be statistically dependent in ways that go beyond what a single student variable may be able to explain. It is therefore recommended that the number of questions per prompt or passage is reduced and length of passages be limited so that fewer dependent observations, but more independent passages can be fit into the assessment timeframe.
4. Alternatives to the Rasch models used in PISA could be considered. In particular:
    a. Verhelst's suggestion to use sum scores and polytomous IRT measurement models for item sets with dependencies should be explored. Models of this type can be specified in multidimensional IRT using a bifactor structure, or in so-called testlet models, which are constrained versions of the bifactor model.
    b. the PISA consortium should devote some resources to the exploration of more general modelling approaches to study the effect of differential item functioning across assessment cycles and countries under various item response models.
5. It is recommended that PISA should carefully consider the current procedures for monitoring comparability of scoring across time points in each country where necessary. With the increase in the number of participating countries seen in each cycle, PISA should consider investing further resources to strengthen their processes and procedures devoted to scoring equivalency, particularly equivalency of scoring at multiple time points, as needed.

Response to the Observations and Recommendations

The Consortium's main concern with this document is that it proposes a set of solutions without having identified whether there is a problem. The reviewers themselves state that "The degree to which PISA trends are more or less stable than other comparable assessment programs and the extent to which such instability is a function of test design issues seems to us to be a very much an open question[12]"

In the document no evidence on this matter is presented. On pages 13 to 22 the reviewers discuss the precision and stability of NAEP yet they do not discuss comparable properties of PISA.

While the review does not include a presentation of its terms of reference the reviewers indicate that the review has been motivated by a concern about lack of stability or precision.

It is clear from the background material we were provided and from the fact that this review has been commissioned that some concerns have been expressed regarding the stability/precision of the limited trend information available to date in PISA. As will become evident below, we believe this is a judgment call outside the range of our expertise. We have interpreted the PGB request as asking for guidance on how to evaluate whether the trend results being obtained under the current design or results obtained in the future from similar or modified designs are sufficiently stable and precise for their intended purpose. (p12)

Reply: Our understanding of the task as presented by OECD was not that there are imminent problems with the trend as reported in PISA. However, as the programme goes into the next cycles, 2009 and beyond, we understand that OECD is concerned with how to maintain and improve the quality of the programme, given that current research (cited in the main body of the revised report) shows that the measures reported do depend in non-trivial ways on a multitude of factors. In that context we were asked to provide guidance on: (1) Criteria indicating sufficient stability/precision in the establishment of trends,

[12] This is the last sentence of the second paragraph of their Executive Summary

(2) Suggestions on how to improve the stability of the link, and (3) Recommendations regarding the PISA test design and the number of link items for each of the assessment domains. We tried to respond to this request by discussing the challenges facing every large scale assessment in terms of requirements on test design and analysis, and by comparing PISA to a large scale national survey assessment as a benchmark that operates under less complex conditions than international assessments. Regarding the our recommendations for changes, we tried to meet the spirit of the task given to us by putting forth suggestions based on the view that reducing potential sources of instability inherent in the current design, if practically and fiscally feasible, could beneft the program in the future even if the present test design is not a significance source of instability.

Note that the reviewers have declared their inability to judge the adequacy of the precision/stability of PISA. They indicate that they instead will provide guidance on how to evaluate whether the results being obtained are sufficiently stable and precise.

Reply: PISA as an international assessment involve a large number of diverse participating countries. For any given country, we cannot judge whether the many components required to operate hand-in-hand have worked in comparable ways over time. In addition, policy changes (such as the NCLB act in the USA), as well as massive changes in population composition (e.g., influx of refugees from neighboring countries in times of conflict and unrest) may provide plausible potential causes for systematic changes in trend results. Since we do not claim to have deep understanding about these specific processes in the vast array of countries that are involved in PISA, our ability to judge specific concerns about trends in participating countries is limited.
This is why we stuck to a discussion of criteria, general suggestions, and recommendations in the area of test design, and models assumed, since in that area, the experience accumulated is independent of the specific subject domain or group of countries/states participating in the assessment.

Given this stated purpose it is disappointing that the reviewers present some results for NAEP, but do not compare those results to PISA. Further they fail to consider the results of other comparable studies such as TIMSS.

OECD have posited that PISA lacks precision; we suspect that the real complaint, justified or not, is that it lacks stability.

Interestingly, on the matter of precision the reviewers show that for NAEP state results:

- differences of 0.06 effect size or less were rarely significant;
- those between 0.08 and 0.11 were significant about 2/3 of the time;
- differences of 0.14 or greater were significant always.

Comparable figures from PISA 2006 (tables 6.3a and 6.3b) give:

- differences of 0.06 or less are significant 1 time in 45 (and that one is Canada);
- differences of between 0.08 and 0.11 are significant 6 times out of 11;

– differences of 0.14 or greater are always significant.

On the matter of stability the reviewers comment on NAEP but make no comparison to PISA. A comparison of stability that we have undertaken is reported in the Annex to this response. The annex, which is extracted from a letter from the Consortium to the Secretariat, demonstrates the relativity of stability of PISA and NAEP and TIMSS.

Perhaps the failure of the review to compare PISA with studies other than NAEP is its key flaw. NAEP is administered in one country and the resources that it is allocated far exceed those allocated to PISA. To illustrate this point, the 2009 NAEP grade 12 national assessment of reading, mathematics, and science has a total of 408 books and a sample size of about 80,000 students. If PISA can apply that scope of design we imagine that the measurement of trend can be improved.

Reply: The comparison to PISA to NAEP has been addressed in the revised text. The choice of NAEP as the means of comparison was based on the fact that we are well aware of the added complexity PISA is facing, and wanted to provide a comparative benchmark of an assessment that lacks some of the complexities, such as multiple languages, PISA is facing. A comparison to trend measures reported in TIMSS did not seem advisable to us, since such a comparison would have multiplied the number of potential causes of discrepancies to be discussed between the assessment frameworks, the sampling, the models used, and the management and processes applied within countries and in the international coordination of the project. In contrast, by focusing on NAEP, which is administered in one language within the US, with homogeneous administrations under direct supervision by a single contractor, we were able to provide data that can be viewed as a benchmark of what can be expected under somewhat more controlled conditions than the ones an international assessment is facing.
In order to provide a benchmark that is useful, however, we based the vast majority of the PISA/NAEP comparisons in the text on state-level trend measures (and state-level sample sizes, not the 80,000 cited above as the national sample size across all three domains in grade 12). Therefore, sample sizes are more similar between country-level comparisons in PISA, and state results in NAEP reported in the text.

We now make some more specific initial comments about what we have taken as the reviewers' recommendations.

Recommendation 1a: Use intact clusters
The Consortium is in complete support of the recommendations made about context affects and the need, wherever possible, to use intact clusters. This support is reflected in *EDU/PISA/GB(2008)1 - Proposal For Securing Trends In PISA 2009*, which is the document we prepared for the April 2008 Governing Board Meeting.

Our proposal, as given in *EDU/PISA/GB(2008)1* is to use intact clusters for linking mathematics and reading in PISA 2009 to previous PISA cycles. In the case of science the use of intact clusters is not possible because every PISA 2006 science cluster contains either released or embedded attitude items.

In forming science link clusters we have payed considerable attention to ensuring that the link items are a random and representative selection from PISA 2006. Further, we have ensured that they will be presented in the same context and position as they were in 2006.

Given that all three domains have now been fully developed we see no impediment to maintaining intact link clusters for the foreseeable future.

Reply: The consortium's and our assessment with respect to the advantages of using intact clusters seem to agree to a large extent.

Recommendation 1b: Use a focussed design and not a mixed design
As discussed above it is possible to use intact clusters with a mixed design and despite a considerable amount of research the Consortium has no evidence that a mixed design limits the potential stability of PISA results from cycle to cycle. That being said it is reasonable to conjecture that the fact that students respond to mixed-domain booklets could influence their performance.

The reviewers suggest that the focussed design proposed by the Hambleton review should be reconsidered. The TAG reviewed this proposal when it was proposed by Hambleton *et al.* and rejected the design.

On the positive side the TAG concluded that: *If there are substantial context effects, as yet undemonstrated, from mixing domains, these can be eliminated; and the testing time per student can be shortened, without substantially increasing the number of booklets.*

On the negative side the TAG concluded that: *either the sample size (of students per country) must be virtually doubled; or standard errors will be increased very substantially and opportunities for effective multilevel modelling of the data will be severely curtailed. Further, there will be no opportunity for cross-domain analyses.*

In the view of the TAG the need to double the size of PISA and the loss of the capacity to do multi-level modelling outweighed the potential benefit of removing an as yet undemonstrated context effect.

Note that any major change to the design would also require a very expensive bridging study and trends would be put at great risk by the design change.

Reply: Our suggestion to use a focused design rather than a mixed design, at least within the context of the major/minor subject distinction, stems primarily from our concern over potential context effects and the ongoing evidence from PISA itself regarding the presence of booklet effects. The consortium notes that they have no direct evidence for context effects in the current design due to the change of test subject from major to minor across assessment cycles. As we indicated in several places throughout the report, we tend to adopt a conservative viewpoint and assume that things that can matter will matter and, other things being equal, prefer designs that remove the possibility of such effects. Moreover, based on experiences with the NAEP assessment, focused designs have not given rise to the kinds of booklet effects seen in PISA and such designs have proven within NAEP to provide advantages in terms of assessment design requiring linkages across adjacent cycles based on intact blocks or item clusters. The same ends, i.e., elimination of booklet effects may be achievable within mixed designs, particularly, if shorter testing times are required and the major/minor subject distinction is somehow eliminated. In our view, any design modifications that achieve these ends are potentially beneficial. Note that, we also indicate in the report that, as contractors, we fully understand that all designs have advantages and disadvantages and, ultimately, design choice involves judgment and compromise. We recognize in the main text that implementing design changes in general, and a move from a mixed to a focus design in particular would require a (potentially very expensive) bridge study. We understand that whether and when to implement such changes requires judgments to be made by the different entities conducting the PISA programme based on their values of the

relative merits of the designs and their understanding of the constraints the programme is operating under.

Recommendation 1c: Do not maintain the current major/minor distinction
The Consortium has no preferred position on this matter.

-/-

Recommendation 2: Increase the number of link items
The proposed number of link items between PISA 2006 and PISA 2009 is 28 items for reading, 36 items for mathematics, and 53 items for science. The number of reading link items between 2006 and 2009 cannot be increased beyond 28 because the set of 28 reading link items is *all* of the items used in PISA 2006 (and also in 2003). Note that 2009 will also include approximately 12 items selected from 2000 that were not used in 2003 or 2006. This brings the total potential reading link between 2000 and 2009 to 40 items.

All other things being equal it is difficult to disagree with a recommendation to increase the number of the link items; just as increasing the sample size will reduce sampling errors, increasing the number of link items will decrease measurement (link) error.

The important questions to answer before increasing the number of link items are:

(i) What are reasonable criteria for stability of PISA trends?

(ii) Are those criteria being met?

(iii) If they are not met will increasing the number of link items assist in meeting those criteria?

As discussed above we do not believe that either (i) or (ii) have been established. Further, the review does not discuss the expected gains that would be obtained by increasing the number of link items, nor does it discuss the other changes that would be necessary to enable such an increase.

Reply: As stated in the consortiums response above, it is difficult to disagree with this recommendation, while practical considerations may limit the achievable number of link items heavily. However, recent research cited in the text has indicated that the number and selection of link items is important (see, for example, Monseur et. al. 2007, 2008). In that sense, some indication is given on (i) and (ii) above, namely that having no unique, but only link items covering the domain to be tested thoroughly, and being administered in the same way over time, is the theoretically most stable way to assess trend. At the same time setting a real and achievable criterion for stability is much harder, since this depends not only on the assessment design and the target of inference, but also on realities of the educational system, as well as political and fiscal constraints rather than only on measurement targets to be met.

Recommendation 3: Increasing the number of independent passages and decreasing the number of items per passage
PISA items are arranged in units – groups of independently-scored items (questions) based on a common stimulus. Many different types of stimulus are used including passages of text, photographs, tables, graphs, and diagrams, often in combination. This unit structure enables the employment of contexts that are as realistic as possible and that reflect the complexity of life situations, while making efficient use of testing time. The use of authentic and often complex situations is central to the items being consistent with PISA domain definitions. Using situations about which several questions can be posed, rather than asking separate questions about a larger number of different situations, reduces the overall time required for a student to become familiar with the material relating to each question.

A disadvantage of this approach is that it reduces the number of different assessment contexts – hence it is important to ensure that there is an adequate range of contexts so that bias due to the choice of contexts is minimised. While every effort is made to ensure that the items within a unit should be independent of each other it is inevitable that the unit structure of PISA causes dependencies between PISA items that share a stimulus.

The Consortium has sponsored research on the degree of dependency and the implication for any such dependency on inter-country comparisons and trends. The review cites some of this research. The review fails to acknowledge that the effect of such dependencies has been included in the estimation of PISA 2006 equating standard errors. Further, the review does not discuss the potential consequences of dependency for the PISA assessment. To the extent that the degree of dependency is invariant across countries and PISA cycles, dependency does not bias PISA results. It does however result in an overestimation of the reliability.

In summary, before adopting a recommendation of this nature it would be wise to be clear about (a) the extent and impact of the dependency and (b) the impact on such a change on the construct validity and stability of PISA.

Reply: This recommendation addresses a general concern that local dependencies exist in passage-based items common in, for example, reading assessments. Even a very careful design of questions arranged under a common reading passage cannot completely eliminate the fact that all questions address the same item stem, namely the reading passage. As an implication, the extent to which local dependency can be observed will vary to some extent from passage to passage, depending on the nature of the text presented as common stem. We hasten to acknowledge the consortiums efforts to study and incorporate this effect in the report of linking errors for PISA 2006. The same rule stated above applies here, the ideal of observation of item responses that only depend on the student variable and no other variable is a goal that cannot be reached, but can be viewed as a gold standard. In practice, all real large scale assessments struggle with some form of dependencies of student responses on other factors such as block order positions, common passages as item stems, and context effects. Our recommendation was targeted at minimizing the effect of common passages by suggesting that fewer items per common stem, and more (but shorter) stems, approximate the ideal of collecting (locally) independent observations of student performance better than fewer, longer passages with more items per passage.

Recommendation 4: Alternatives to the Rasch models used in PISA could be considered.
The Consortium and its TAG regularly discuss the PISA scaling models and the merits of considering alternative (and typically more general) models. In fact the analysis plans for the PISA 2009 Field Trial[13] take up TAG recommendations to explore alternative scaling exactly as proposed by Mazzeo and von Davier.

The Consortium has already undertaken work on the degree of dependency between items with units and the impact that this has on the contribution that item sampling makes to the uncertainty in trend results. Further work on this is proposed on the basis of 2009 Field Trial data. It is worth noting however that the explicit suggestion by the reviewers of using a polytomous model to analyse item bundles (or testlets) does not *solve* the problem of dependency between items that share a common stimulus. The use of such an approach would both expose and *recognise* the dependency but it would not change trend estimates. The use of such models would better recognise the uncertainty associated with the trends and take the dependency into account in the estimation of measurement errors.

[13] These plans have been submitted to the Secretariat and will be agenda items at the coming TAG and NPM meetings

The Field Trial analysis plans also describe the Consortium's intention to explore the use more general item response models and to examine cross-participant invariance of the item parameter estimates.

Reply: We welcome the explorations of the consortium targeted at using more general psychometric models. Our recommendation of using models that involve more flexible parametric item functions agrees with the direction the consortium is taking on this issue.

Recommendation 5: *Increase the resources devoted to ensuring cross-country and cycle equivalency in coding.*

PISA includes a greater proportion of open items that require coding by trained professionals than do the majority of comparable studies. As with the unit structure the use of such items has been central to a valid assessment of the PISA constructs.

The Consortium is in strong agreement with the suggestion that even minor systematic variation between counties in their approach to the coding of such items can influence the comparability of the assessment. As such, we too support the recommendation that increased resources be allocated to both ensuring and monitoring cross country and cross cycle equivalence in coding. Unfortunately recent budget cuts have required the scaling down of this area of activity.

Reply: We welcome the consortium's strong agreement with our recommendation.