Memorandum

United States Department of Education
Institute of Education Sciences
National Center for Education Statistics

DATE: March 10, 2011

TO: Shelly Martinez, OMB

FROM: Dana Kelly, NCES

THROUGH: Kashka Kubzdela, NCES

RE: PISA 2012 Field Test and Recruitment Change Request (OMB# 1850-0755 v.11) Response to Passback 1, 2, and 3

Passback 1:

Regarding all of the proposed new content, especially on math, please provide:

1. A justification for the new content. Especially useful would be a cross walk between (a) question(s), (b) their source if from another study or what u.s. or other testing has been done, (c) the purpose.

2. The analysis plan for field test results to help narrow content, to determine question performance, etc.

Response: Thank you for the passback on PISA. Attached are three documents that provide the requested information. Attachments 6a and 6b are the international versions of the Student and School questionnaire items annotated with explanations of the purpose of each item and whether it has been administered before. Because these are the international versions, they do not include the adaptations made for the US, such as omitting some items, adapting terms, or fixing grammar so it reflects U.S. usage. We will be glad to provide those upon will request.

Attachment 7 is an excerpt from a document prepared by the international consortium to report on progress in the development of the context questionnaires to the PISA Governing Board in November. This paper summarizes the content of the FT questionnaires, describes the analysis plan for field test results to help narrow content, to determine question performance, etc., and describes the cognitive labs conducted on the draft items.

Passback 2:

We have some follow up questions about content of the student questionnaire since a justification was missing for a fair amount of the new content.

1. For some questions, such as ST35Q01-ST35Q10, there is an explanation of the item's purpose, which in this example is that "Research has shown this variable can impact behaviour, including performance on mathematics tests." For others, such as those about "situational judgment," there is no such explanation. Please provide such an explanation of purpose, the utility and the literature behind these questions.

Response: Attached in Attachment 8 please find the PISA 2012 Context Questionnaire Framework, which provides much more detail than does the annotated questionnaires on the literature behind the questionnaire items. The responses below point you to the sections of the framework that

address your specific questions. With respect to the "situational judgment" items, as described in the framework (section 3.2), PISA is measuring students' strategies, beliefs, and motivation related to mathematics literacy and the set of situational judgment items (ST30 through ST34) address motivations and intentions. As described in paragraph 148, PISA is trying out alternative item types, including situational judgment items, to address potential method bias. PISA tends to use Likert-type scales, which are susceptible to differences in response styles. In using alternative item types PISA is attempting to preclude differences in response styles as a cause of differences in mean scores and correlations across countries.

2. Some questions, such as ST49 (How often do you perform the following behaviours inside and outside of school hours?) and ST107Q01 (You are given two choices to make money: Which do you prefer?) seem to be testing a hypothesis (e.g., "These items measure student mathematics behaviours. In theory, students who often engage in these behaviours should demonstrate high mathematics achievement."). Please explain why such exploratory items are in the PISA field test rather than some small scale country-specific test. For each of these questions, please justify the practical utility of each of these questions.

   Response: ST49 provides a measure of students' opportunity-to-learn mathematics outside of the classroom. This is particularly important for PISA, because PISA purports to assess mathematical literacy learned inside and outside the classroom. The PISA 2012 Context Questionnaire Framework in Attachment 8 provides further discussion of the importance of opportunity-to-learn in the PISA theoretical framework. ST107Q01 and similar items that follow attempt to provide contextual information to help in the analysis of the new PISA financial literacy assessment. The items are intended to investigate the extent to which students' aversion (or attraction) to risk might explain students' understanding of and behavior related to money.

3. The survey includes both a former and some new self-described school climate questions (see pages 126-128, 134, and 141-3). In contrast to several other measures which also seem to be about school climate (ST88Q01-ST88Q04), these have no justification or reference to existing literature. Please provide evidence of the performance of the former questions, the evidence base for the new questions, and a clear justification for their inclusion in the PISA field test.

   Response: ST77 (page 126) is a repeat of a question used in PISA 2003 to measure teacher support in the classroom and ST78 (page 127-8) is a new question designed to measure teacher support via homework. The theoretical role of teacher support in student learning is developed in Attachment 8, the PISA 2012 Context Questionnaire Framework. Section 2.4 discusses classroom inputs and processes. Section 3.3 provides more detail on expanded notions of opportunity-to-learn and how it relates to quality of instruction and learning environments. ST81 (page 134) is repeated from PISA 2009, but modified to the mathematics context (PISA 2009 focused on reading). The item asks about discipline in the classroom. Again, it is associated with opportunity-to-learn and how it relates to learning environments that provide opportunities for students to engage productively with content in the classroom. ST86 and ST87 (pages 141-3) are repeated from PISA 2003, when mathematics was last the PISA focal subject. ST87 addresses teacher support, discussed above. ST88 asks about students' sense of well-being. The PISA 2012 Context Questionnaire Framework (Attachment 8) discusses the role of students' sense of well-being in section 3.2 on students' strategies, beliefs, and

motivation. The theoretical link between student beliefs and performance rests on Bandura's work on self-efficacy.
PISA is an international study in which the United States is one of many participants. While the United States has had opportunities to comment on the context questionnaire framework and items at multiple points, the development of the instruments is a collaborative effort among the international consortium, questionnaire expert group, and participating countries, and there are compromises and trade-offs. The field test will provide an opportunity to look at how the proposed items function and items that do not work will not be included in the main study. NCES will review the field test data carefully and make recommendations to the international consortium for items to include or exclude from the main study. Ultimately, the PISA Governing Board, on which the United States is a voting member, will approve the final instruments.

4. Are questions ST82Q01-ST82Q03 supposed to measure school climate as well? Please explain purpose and justify these as well.

   Response: ST82Q01-ST82Q03 are vignettes that will be used to anchor students' responses to attitudinal questions that follow. PISA is attempting to address concerns about cross-cultural validity of attitudinal questions through the use of techniques such as anchoring vignettes, which allow the use of statistical models to adjust student response to self-assessment questions based on their "anchoring" responses to hypothetical situations.

5. The analysis plans seems to assume that there is already demonstrated utility of each item on the questionnaire. Particularly for the items we call out above, we do not know what the utility is and see no criteria in the analysis plan that would weed out such questions. Please clarify how the concept of utility is operationalized in the selection of PISA items.

   Response: The process for the developing PISA questionnaire items includes empanelling an international group of experts, development by the experts and the international contractor of a theoretical framework (attached as Attachment 8), review and approval of the framework by the board of participating countries, development of items that address the priorities outlined in the framework for the field trial, review of the proposed items by the expert group and all participating countries, and approval of the field test instruments by a board of participating countries. A similar review process will be used for item selection for the main study, based on analysis of the field trial results. The field trial analyses will examine how well items perform in terms of the variation in responses, the extent to which responses fit theoretical expectations (how well groups of items scale, the extent to which items expected to correlate with each other correlate, etc.), as well as cross-national differences in the performance of items. Items that do not work will not be included in the main study instruments. Throughout the field test analysis and review process NCES will carefully review the field test data and make recommendations for items to include or exclude from the main study. Based on these analyses and taking into consideration countries' recommendations and the total time allocated for background questions in the main study, the international contractor and questionnaire expert group will recommend a final set of main study items. The recommended final instruments will be presented to the board of participating countries for approval.

Passback 3

In passback 3 OMB provided the following feedback on our responses to questions in passback 2: “Answers to questions 2 through 5 are not fully responsive. Please reread them carefully. In #2 we are looking for whether there is a strong literature or whether this really is testing an untested hypothesis. Item 3 asks for several things, such as performance evidence of past items, not just a restatement that they are past items, etc. For item 4, we'd like a clear "no" if that in fact is the answer. For item 5, we are looking for NCES's detailed analysis plan. Please do not send us to "Attachment 8" for answers to these follow up questions.”

We have revised our responses to questions 2-5, below, to be more responsive to the questions OMB has asked.

2. Some questions, such as ST49 (How often do you perform the following behaviours inside and outside of school hours?) and ST107Q01 (You are given two choices to make money: Which do you prefer?) seem to be testing a hypothesis (e.g., "These items measure student mathematics behaviours. In theory, students who often engage in these behaviours should demonstrate high mathematics achievement."). Please explain why such exploratory items are in the PISA field test rather than some small scale country-specific test. For each of these questions, please justify the practical utility of each of these questions.

   Response: ST49 has been included to provide a measure of student engagement in mathematics, based on their reported behaviors in and out of school hours. Previous studies, including PISA, have found that engagement in subject matter is related to performance and that engagement (for example, reading habits and interest or participation in science-related activities) is also related to gender and socio-economic status (Campbell, Voelkl and Donahue, 1997; OECD, 2007; OECD, 2010). Item ST49 also provides a measure of students’ opportunity-to-learn mathematics through students’ participation in various activities in and out of school hours which is particularly important for PISA because PISA is intended to assess mathematical literacy learned inside and outside the classroom. The notion of opportunity to learn was introduced by John Carroll in the early 1960s, and was initially meant to indicate whether students had sufficient time and received adequate instruction to learn (Carroll, 1963; cf. Abedi *et al*., 2006). It has since been an important concept in international student assessments (Husén, 1967; Schmidt & McKnight, 1995; Schmidt *et al*., 2001), and has been shown to be strongly related to student performance, especially in cross-country comparisons (Schmidt & Maier, 2009). The meaning of the construct broadened over time to include students’ learning opportunities and activities outside of the classroom. While Item ST49 is different in its content focus than items used in previous administrations of PISA which have focused on reading and science, it is similar in spirit to items that appear in PISA 2006 and 2009 (e.g., Q19 from the 2006 student questionnaire) and thus is not considered by NCES to be an exploratory item. The field test will provide information about whether students’ behaviours related to mathematics do correlate with performance in mathematics literacy, as they have in other subjects, and whether that is consistent across countries.

   ST107Q01 and similar items that follow (ST108Q1 through ST114Q1) are designed to provide contextual information to help in the analysis of the new PISA financial literacy assessment. Attitudes are considered important aspects of financial literacy (OECD 2010b). Moreover, individual preferences are important determinants of financial behaviour and can interact with financial literacy. Research from general behavioural psychology suggests that one’s risk tolerance (willingness to accept the possibility of a loss in order to achieve greater gain) might explain students’ financial literacy (Barsky, Juster, Kimball, & Shapiro, 1997; Holt & Laury, 2002). As the first major cross-national assessment of financial literacy, PISA is in a position to examine whether this is the case in multiple countries in the field test.

3. The survey includes both a former and some new self-described school climate questions (see pages 126-128, 134, and 141-3). In contrast to several other measures which also seem to be about school climate (ST88Q01-ST88Q04), these have no justification or reference to existing literature. Please provide evidence of the performance of the former questions, the evidence base for the new questions, and a clear justification for their inclusion in the PISA field test.

Response: The items on the pages cited above (ST77, ST78, ST81, ST86, and ST87) are related to four aspects of school climate: teacher support, discipline in the classroom, student-teacher relations, and students' sense of belonging. Below we provide additional justification for the inclusion of these items, drawn from the PISA questionnaire framework and other sources. For items that have been used before, we provide prior reliability coefficients and how they have been used in reports. ST88 is not intended to measure school climate: it is a measure of students' general attitudes toward school and so further information is not provided about this item.

Teacher Support. Research has shown that student learning is generally supported by a positive and respectful atmosphere that is relatively free of disruption and focused on student performance (Creemers & Kyriakides, 2008; Harris & Chrispeels, 2006; Hopkins, 2005; Scheerens and Bosker, 1997). The major facets of a positive classroom climate are: Supportive teacher-student interactions, good student-student relationships, achievement orientation, and an orderly learning atmosphere with clear disciplinary rules. ST77 (page 126) is a repeat of a question used in PISA 2003 to measure teacher support in the mathematics classroom and ST78 (page 127-8) is a new question designed to measure teacher support via homework. The latter question has been added to strengthen the construct by addressing how students perceive their teachers to support them in a significant element of the learning experience. Together, these two items will provide information about students' perceptions of the degree of support they receive from their teachers with respect to learning mathematics.

Information about the prior performance of ST77:

Reliability: In 2003 (the last time mathematics was the focal subject and therefore the last time the scale was used), the median reliability (Cronbach's alpha) across OECD countries for this scale was .83 (from PISA 2003 technical report: OECD 2005).

Utility: The scale was featured in the OECD's report releasing the PISA 2003 results (OECD 2004, pp. 212-214): The report compared countries' average scores on the scale and on the scale's individual items and described within-country variation, including variation across schools and by student gender within countries.

Disciplinary Climate of the Classroom. Classrooms and schools with more disciplinary problems are less conducive to learning since teachers have to spend more time creating an orderly environment before instruction can begin (Gamoran and Nystrand, 1992). ST81 (page 134) measures the disciplinary climate of the mathematics classroom. This item was administered in 2000 and 2009 but has been modified to the mathematics context (PISA 2000 and 2009 focused on reading).

Information about the prior performance of ST81:

Reliability: The 2009 technical report is not yet available to provide the reliability coefficient but a similar scale was used in 2000 and had a reported median reliability of .81 across countries (from PISA 2000 technical report).

Utility: The scale, focused on reading, was reported in the OECD report releasing the PISA 2009 results (OECD 2010, pp. 90-92): The report compared countries' average scores on the scale and on the scale's individual items and described within-country variation, including the extent to which within-country variation on the scale was between schools or within schools.

Student-teacher Relations. Positive student-teacher relations have been found to be related to the establishment of environments conducive to learning. Research finds that students, particularly disadvantaged students, learn more and have fewer disciplinary problems when they feel that their teachers are devoted to their academic success (Gamoran, 1993) and when they have good working relations with their teachers (Crosnoe, Johnson and Elder, 2004). ST86, administered in PISA 2000, 2003, and 2009, forms a student-teacher relations scale.

Information about the prior performance of ST86:

Reliability: In 2003 the median reliability the median reliability (Cronbach's alpha) across OECD countries for this scale was .76 (reported in the PISA 2003 technical report: OECD 2005).

Utility: ST86 was reported in the OECD's report releasing the PISA 2009 results (OECD 2010, pp. 88-90): The report compared countries' average scores on the scale, the scale's individual items, and the extent of within-country variation on the scale.

Student Sense of Belonging. ST87 (pages 141-43) measures students' sense of belonging (a-f) and well-being and life satisfaction (g-i). At the school level, relatively strong effects have been found for the Sense of Belonging scale as process predictors of mathematics performance as well as mathematics interest (see PISA Context Questionnaire Framework). ST87 was administered in 2003 and has been modified somewhat.

Information about the prior performance of ST87:

Reliability: In 2003 the median reliability (Cronbach's alpha) across OECD countries for this scale was .74 (reported in the 2003 technical report: OECD 2005).

Utility: ST87 was included in the OECD's report releasing the PISA 2003 results (OECD 2004, pp. 128-132): The report compared countries' average scores on the scale, the scale's individual items, and the extent of within-country variation.

The above items were included in the field test because of the evidence above and through consultation among the many countries participating in PISA. While the United States has had opportunities to comment on the context questionnaire framework and items at multiple points, NCES is beholden to the international schedule. Moreover, the development of the instruments is a collaborative effort among the international consortium, questionnaire expert group, and participating countries, and there are compromises and trade-offs. The field test will provide an opportunity to look at how the proposed items function and items that do not work will not be included in the main study. NCES will review the field test data carefully and make recommendations to the international consortium for items to include or exclude from the main study. Ultimately, the PISA Governing Board, on which the United States is a voting member, will approve the final instruments.

4. Are questions ST82Q01-ST82Q03 supposed to measure school climate as well? Please explain purpose and justify these as well.

   Response: No, these items are not intended to measure school climate. ST82Q01-ST82Q03 are vignettes that will be used to anchor students' responses to attitudinal questions that follow. PISA is attempting to address concerns about cross-cultural validity of attitudinal questions through the use of anchoring vignettes, which allow the use of statistical models to adjust student response to self-assessment questions based on their "anchoring" responses to hypothetical situations (King *et al.,* 2004; King & Wand, 2007; Hopkins & King, 2009).

5. The analysis plans seems to assume that there is already demonstrated utility of each item on the questionnaire. Particularly for the items we call out above, we do not know what the utility is and see no criteria in the analysis plan that would weed out such questions. Please clarify how the concept of utility is operationalized in the selection of PISA items.

   Response: In our prior response to this question, we summarized the PISA item selection process involving the international group of experts, the international contractor, PISA Governing Board of country representatives, and the developmental, approval, and testing steps taken. However, we failed to address specifically how utility is operationalized.

   The utility of the items is determined by countries' representatives to the PISA Governing Board and based on countries' information priorities and the framework drafted by the international group of experts and the international contractor, including the framework's references to prior research. The PISA Governing Board sets priorities for what is to be measured and the international contractor and group of experts is tasked with developing items and testing their validity and reliability. As you note, the field trial analysis does not directly assess utility; utility has already been established.

   Our response on the utility of specific items is contained in the response to questions 2 and 3. The United States questioned the utility of some of the items you are questioning, as well as other items and topical areas that were removed from the field test or not pursued for item development, because of the objections of the United States and other countries. These items remained because they had sufficient support from the countries for inclusion in the field test. If OMB continues to question the utility of some of the items after reviewing our response, we would be happy to discuss the items and consider objecting to the items inclusion in the main study.

   Pasted below, please find additional detail on the analysis plan for the field trial (adapted from "PISA Field Test Questionnaires Summary and Analysis Plans.doc" provided in response to passback 1):

   The purpose of the analysis of field trial data is to gather evidence to support decisions about which scales and items to retain for the main study. In some cases, the issue is comparing alternative methods for measuring certain scales. In other cases the issue is simply whether a newly introduced scale behaves well psychometrically. In either case, it is useful to anticipate the kind of data that will be helpful in making decisions about keeping and deleting of questions and items, and for designing the field trial study to ensure the collection of such data. In particular, it is important to design booklets which will allow the most useful data analyses following field trial data collection.

   The main questions to be addressed:

   A. Within countries: Do item responses behave reasonably? Is the distribution of responses across item categories reasonable? Is the mean and standard deviation as approximately expected? Is there evidence for DIF (gender, school-type) for some items in some countries? Are scales

suitably reliable? Do scales function properly? And which of the alternative versions of scales function best? Do predictor scales correlate with achievement? Which of the alternative versions (e.g., forced choice vs. Likert scale) correlates highest? (across different countries) Do outcome scales correlate with other variables in expected ways? Which alternative has the most sensible pattern (across different countries)? Do scales (and items) (both predictor and outcome) behave appropriately from the context of a multi-trait-multi-method (MTMM) design? That is, do constructs measured in different ways still measure the same underlying trait? Can mixed-item-type scales function adequately? How do mixed-item-type scales compare to same item-type scales in their predictive validity with achievement, and in their correlations with other variables?

B. Across countries: Do certain item types suggest greater cross-cultural consistency? Particularly for scales in which we have observed positive ecological correlations and negative within-country student-level correlations (e.g., mathematics interest, instrumental motivation), are there scale versions that "show"/"have" or maybe "scale versions with" greater consistency of correlations at the country and student level? Is there measurement invariance (configural, metric, scalar) across countries? Is there any country-level DIF (i.e., treating countries as groups)?

References

Abedi, J., M. Courtney, S. Leon, J. Kao, and T. Azzam (2006), *English Language Learners and Math Achievement: A Study of Opportunity to Learn and Language Accommodation* (CSE Report 702, 2006), University of California, Center for the Study of Evaluation/National Center for Research on Evaluation, Standards, and Student Testing, Los Angeles, USA.

Adams, R.J. & Wu, M.L. (2002). *PISA 2000 Technical Report*. OECD, Paris.

Barsky, R. B., Juster, F. T., Kimball, M. S., & Shapiro, M. D. (1997). Preference parameters and behavioural heterogeneity: An experimental approach in the health and retirement study. *Quarterly Journal of Economics, 11*, 537-539.

Campbell, J.R., Voelkl, K.E., & Donahue, P.L. (1997). *NAEP 1996 trends in academic progress* (NCES Publication No. 97985r). Washington, DC: U.S. Department of Education.

Carroll, J.B. (1963), "A Model of School Learning", *Teachers College Record*, Vol. 64, pp. 723-733.

Creemers, B.P.M. and L. Kyriakides (2008), *The Dynamics of Educational Effectiveness: A Contribution to Policy, Practice, and Theory in Contemporary Schools*, Routledge, London.

Crosnoe, R., M. Johnson and G. Elder (2004), "Intergenerational Bonding in School: The Behavioral and Contextual Correlates of Student-Teacher Relationships", *Sociology of Education*, Vol. 77, No. 1, pp. 60-81.

Gamoran, A. (1993), "Alternative Uses of Ability Grouping in Secondary Schools: Can We Bring High-Quality Instruction to Low-Ability Classes?", *American Journal of Education,* Vol. 102, No. 1, pp. 1-12.

Gamoran, A. and M. Nystrand (1992), "Taking Students Seriously", in F. Newman (ed.), *Student Engagement and Achievement in American Secondary Schools*, Teachers College Press, New York City, New York.

Harris, A. and J.H. Chrispeels (eds.) (2006), *Improving Schools and Educational Systems: International Perspectives*, Routledge, London.

Holt, C., & Laury, S. (2002). Risk aversion and incentive effects. *American Economic Review, 95*, 1644-1655.

Hopkins, D. (ed.) (2005), *The Practice and Theory of School Improvement: International Handbook of Educational Change*, Springer, Dordrecht.

Hopkins, Daniel, and Gary King. "Improving Anchoring Vignettes: Designing Surveys to Correct Interpersonal Incomparability." *Public Opinion Quarterly* (2009

Husén, T. (1967), *International Study of Achievement in Mathematics*, Vol. 2, Wiley, New York, USA.

King, Gary, Christopher J. L. Murray, Joshua A. Salomon, and Ajay Tandon. "Enhancing the Validity and Cross-cultural Comparability of Measurement in Survey Research." *American Political Science Review* 98 (2004): 191-207.

King, Gary, and Jonathan Wand. "Comparing Incomparable Survey Responses: New Tools for Anchoring Vignettes." *Political Analysis* 15 (2007): 46-66.

Organisation for Economic Cooperation and Development (ed.) (2004), *PISA 2003. Learning for Tomorrow's World: First Results from PISA 2003.* OECD, Paris.

Organisation for Economic Cooperation and Development (ed.) (2005), *PISA 2003 Technical Report*, OECD, Paris.

Organisation for Economic Cooperation and Development (ed.) (2007), *PISA 2006. Science Competencies for Tomorrow's World,* OECD, Paris.

Organisation for Economic Cooperation and Development (ed.) (2010), *PISA 2009 Results: What Makes a School Successful? Resources Policies and Practices (Volume IV).* OECD, Paris.

Scheerens, J. and R.J. Bosker (1997), *The Foundations of Educational Effectiveness*, Pergamon, Oxford.

Schmidt, W.H., C.C. McKnight, R.T. Houang, H.C. Wang, D.E. Wiley, L.S. Cogan and R.G. Wolfe (2001), *Why Schools Matter: A Cross-National Comparison of Curriculum and Learning*, Jossey Bass, San Francisco, California, USA.

Schmidt, W.H. and A. Maier (2009), „Opportunity to Learn", in G. Sykes, B. Schneider and D.N. Plank (eds.), *Handbook of Education Policy Research,* Routledge, New York, pp. 541-559.

Schmidt, W.H. and C. McKnight (1995), "Surveying Educational Opportunity in Mathematics and Science: An International Perspective", *Educational Evaluation and Policy Analysis*, Vol. 17, No. 3, pp. 337-353.