PISA 2012 Field Test Questionnaires: Summary of Content and Analysis Plans

Exerpted from *Development of questionnaires for the PISA2012 Field trial: Overview of design, content, proposed analyses and outcomes of cognitive labs*, prepared by the PISA international consortium for the 30th Meeting of the PISA Governing Board Meeting, November 1-3, 2010.

Student Questionnaire

Content and design

The questions to be covered in the Student Questionnaire (StQ) together with information regarding how they fit into the questionnaire framework and whether they provide new or trend data are presented in Table 1.

Table 1. Content of Student Questionnaire for PISA2012 Field Trial

Q#	Content	Framework component	Trend/new
	Section A –the student's basic characteristics and educational career		
1	Grade level	Input – general	Trend
2	Study programme	Input – general	Trend
3	Chronological age (date of birth)	Input – general	Trend
4	Gender	Input – general	Trend
5	Whether student completed pre-primary education (ISCED 0 attendance)	Input – general	Trend
6	Starting age for primary (ISCED 1) education	Input – general	Trend
7a	Grade repeating	Outcome – general	Trend
7b	Tardiness (last month)	Outcome – general	New*
7c	Truancy (last month)	Outcome – general	New*
7d	Absenteeism (last month)	Outcome – general	New*
	Section B –the student's family context and home resources		
8	Family structure	Input – general	Trend
9a	Mother's main job 1	Input – general	Trend
9b	Mother's main job 2	Input – general	Trend
10	Mother's education (ISCED level 1-3)	Input – general	Trend
11	Mother's qualifications (ISCED level 4-6)	Input – general	Trend
12	Mother's employment status	Input – general	Trend
13a	Father's main job 1	Input – general	Trend
13b	Father's main job 2	Input – general	Trend

Q#	Content	Framework component	Trend/new
14	Father's education (ISCED level 1-3)	Input – general	Trend
15	Father's qualifications (ISCED level 4-6)	Input – general	Trend
16	Father's employment status	Input – general	Trend
17	Country of birth	Input – general	Trend
18a	If immigrant, age at time of arrival	Input – general	Trend
18b	Whether parent a national	Input – general	Trend
18c	Acculturation level 1	Input – general	New
18d	Acculturation level 2	Input – general	New
19	Home language	Input – general	Trend
20	Home resources	Input – general	Trend
21	Family wealth	Input – general	Trend
22	Books in home	Input – general	Trend
	Section C –the student's approach to learning mathematics		
23	Interest and enjoyment in mathematics	Outcome – domain-specific	Trend
23	Instrumental motivation to do mathematics	Outcome – domain-specific	New
24a	Motivation to do mathematics (situational judgment test type)	Outcome – domain-specific	New
24b	Motivation to do mathematics (situational judgment test type)	Outcome – domain-specific	New
24c	Motivation to do mathematics (situational judgment test type)	Outcome – domain-specific	New
24d	Motivation to do mathematics (situational judgment test type)	Outcome – domain-specific	New
24e	Motivation to do mathematics (situational judgment test type)	Outcome – domain-specific	New
25	Subjective norms that influence mathematics 1	Outcome – domain-specific	New
26	Subjective norms that influence mathematics 2	Outcome – domain-specific	New
27	Mathematics self-efficacy	Outcome – domain-specific	Trend
28a	Interest and enjoyment in mathematics (forced-choice)	Outcome – domain-specific	New
28b	Interest and enjoyment in mathematics (positive attitudes, more response options)	Outcome – domain-specific	New

Q#	Content	Framework component	Trend/new
28c	Interest and enjoyment in mathematics (negative attitudes, more response options)	Outcome – domain-specific	New
28d	Interest and enjoyment in mathematics (different response labels)	Outcome – domain-specific	New
29	Mathematics self-concept	Outcome – domain-specific	Trend
29	Mathematics anxiety	Outcome – domain-specific	Trend
30	Perceived control to put forth effort in mathematics	Outcome – domain-specific	New
31	Attributions of effort (failure scenario)	Outcome – domain-specific	New
32	Attributions of effort (success scenario)	Outcome – domain-specific	New
33	Mathematics work ethic	Outcome – domain-specific	New
34	Intention to put forth effort in mathematics	Outcome – domain-specific	New
35	Intention to put forth effort in mathematics (forced-choice)	Outcome – domain-specific	New
36	Mathematics behaviours	Outcome – domain-specific	New
37	Cooperative learning	Outcome – domain-specific	Trend
37	Competitive learning	Outcome – domain-specific	Trend
38	Competitive vs. competitive learning (forced-choice)	Outcome – domain-specific	New
39	Control strategies	Outcome – domain-specific	Trend
39	Elaboration strategies	Outcome – domain-specific	Trend
39	Memorisation strategies	Outcome – domain-specific	Trend
40	Control vs. elaboration vs. memorisation strategies (forced-choice)	Outcome – domain-specific	New
41	Test-taking strategies	Outcome – domain-specific	New
42a	Time spent on out-of-school-time lessons in mathematics (and other subjects)	Process – general and domain-specific	New *
42b	Type of out-of-school-time lessons (remedial or enrichment)	Process – general and domain-specific	New *
43	Hours spent on out-of-school-time (all lessons)	Process – general and domain-specific	New *
44	Hours spent on out-of-school-time (mathematics lessons)	Process – general and domain-specific	New *
45	Mark received in test language, mathematics, and science	Process – general and domain-specific	New *

Q#	Content	Framework component	Trend/new
46	Mark received in test language, mathematics, and science relative to passing grade	Process – general and domain-specific	New *
47	Opportunity to learn mathematics concepts (frequency)	Process –domain-specific	New
48	Opportunity to learn mathematics concepts (familiarity)	Process –domain-specific	New
49	Opportunity to learn mathematics concepts (problems presented and rated on experience)	Process –domain-specific	New
50	Learning time	Process –general and domain-specific	Trend
51	Opportunity to learn mathematics concepts (concepts presented and rated on experience)	Process –domain-specific	New
	Section D – the students mathematics experience		
52	Teacher support (in mathematics class)	Outcomes – domain-specific	Trend
53	Teacher support (regarding homework)	Outcomes – domain-specific	New
54	Instructional strategies of mathematics teachers	Outcomes – domain-specific	New
55	Cognitive activitation from mathematics teachers	Outcomes – domain-specific	New
56	Disciplinary climate in mathematics lessons	Outcomes – domain-specific	Trend
57	Teacher support (anchoring vignette)	Outcomes – domain-specific	New
58	Disciplinary climate in mathematics (anchoring vignette)	Outcomes – domain-specific	New
	Section E – school climate		
59	Student-teacher relations	Outcomes – general	Trend
60	Sense of belonging	Outcomes – general	Trend/New
61	Attitudes towards school 1	Outcomes – general	Trend
62	Attitudes towards school 2	Outcomes – general	New
62	Attitudes towards school 2	Outcomes – general	New
63	Subjective norms towards school	Outcomes – general	New
64	Percieved control of school environment	Outcomes – general	New
65	Intention to put forth effort in school	Outcomes – general	New
	Section F – the student's problem solving experiences		

Q#	Content	Framework component	Trend/new
66	Perserverance in solving problems	Process –domain-specific	New
67	Engagement and openness in solving problems	Process –domain-specific	New
68	Problem solving scenario (private device)	Process –domain-specific	New
69	Problem solving scenario (technology setting)	Process –domain-specific	New
70	Problem solving scenario (non-technology setting)	Process –domain-specific	New
71	Problem solving scenario (public device)	Process –domain-specific	New

Notes: * These questions are very close to those that have been used previously to obtain trend, though have significant enough changes in framing to suggest that they should be considered new.

As can be seen in Table 1, the StQ, like other instruments proposed for the PISA 2012 FT, seeks to strike a balance between obtaining trend and new data on the one hand and general and domain-specific information on the other hand while covering various aspects of inputs, processes and outcomes. Given that 2012 will be the second PISA cycle with mathematics as the major domain, domain-specific trend information that links to the information obtained in 2003 becomes of critical interest.

Coverage of constructs in the StQ has been extended from PISA 2003, to include opportunity to learn, test-taking strategies, processes associated with problem solving and a variety of new outcomes that might result from the student's experience in the mathematics classroom (*e.g.* cognitive activation).

Table 1 also highlights attempts to put forward new item formats intended to address concerns with regard to the cross-cultural comparability of indicators obtained from responses to the StQ assumed to be mainly a consequence of response styles across countries. This includes use of the situational judgment test methodology, anchoring vignettes, forced-choice, overclaiming technique and new response scales.

Analysis

The purpose of the analysis of FT data from the StQ is to gather evidence to support decisions about which scales and items to retain for the Main Survey (MS). 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 FT 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 FT data collection.

In general, the main questions to be addressed by the analyses are as follows:

a. Within countries:
   i. Do item responses behave reasonably?
      1. Is the distribution of responses across item categories reasonable?
      2. Is the mean and standard deviation as approximately expected?
   ii. Are scales suitably reliable?

1. Do scales have adequately high reliability (above $r_{xx'}$ = .80 or so)? If not, could they be made so with the addition of a few extra items (*i.e.*, is it possible to generate additional parallel items to boost reliability)?
2. Is there evidence for DIF (gender, school-type) for some items in some countries?

iii. Do scales function properly? And which of the alternative versions of scales function best?

1. Do predictor scales correlate with achievement? Which of the alternative versions (e.g, forced choice vs. Likert scale) correlates highest? (across different countries)
2. Do outcome scales correlate with other variables in expected ways? Which alternative has the most sensible pattern? (across different countries)
3. 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?

iv. Can mixed-item-type scales function adequately?

1. Do mixed-item-type scales scale properly?
2. 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

i. Do certain item types suggest greater cross-cultural consistency?

1. 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?
2. Is there measurement invariance (configural, metric, scalar) across countries?

ii. Is there any country-level DIF (*i.e.*, treating countries as groups)?

The consortium is considering several booklet designs that will enable the analyses necessary to support decisions on the design of the MS. The major issues concern whether and what to include as a common set of items across all four forms, what scales to use in an MTMM analysis, and what scales to use for a mixed-item-type analysis. These issues have been reviewed by the QEG.

In addition, several analytic methods are being considered for addressing item and scale quality issues. Mutiple Group Confirmatory Factor Analysis (MGCFA) and multilevel analyses have been used in secondary analyses of PISA 2003 questionnaire data presented at previous QEG meetings (Vieluf, Lee & Kyllonen, 2009). Item Response Theory (IRT) and Confirmatory Factor Analysis (CFA) approaches to exploring parameter invariance were compared using data from previous PISA cycles (Schulz, 2005). Differential item functioning analyses along with a comparison of the partial credit and generalised partial credit IRT models for scaling was conducted on the FT data for PISA 2009 (Glas & Jehangir, 2009; see also Walker, 2007). A latent class MTMM approach for evaluating item quality has also been shown to be effective on questionnaire data from international surveys (Oberski, Hagenaars & Saris, 2009). These are being evaluated by the consortium.

School Questionnaire

Content and design

By way of overview, the questions to be covered in the School Questionnaire (ScQ) together with information regarding how they fit into the questionnaire framework and whether they provide new or trend data are presented in Table 2.

Table 2. Content of School Questionnaire for PISA2012 Field Trial

Q#	Content	Framework component	Trend/new
	Section A – the structure and organisation of the school		
1	School type	Input – general	Trend
2	School funding source	Input – general	Trend
3	School location	Input – general	Trend
4	Competition between schools	Process – general	Trend
5	Average class size	Input – general	Trend
6	Instructional time/intended maths curriculum	Input – general and domain-specific	Trend/New
	Section B – the student and teacher body		
7	School enrolment	Input – general	Trend
8	Grade repetition	Process – general	Trend
9	% of immigrant students	Input – general	Trend
10	Composition and qualifications of teaching staff	Input – general	Trend
11	Composition and qualifications of mathematics teacher staff	Input – domain-specific	Trend
	Section C – the school's resources		
12	Computer availability to 15 year old students/ Connection to the www	Input – general	Trend
13	Access to computer hardware	Input – general	New
14	Access to the internet	Input – general	New
15	Teacher shortage / Quality of educational resources/ ICT resources/ Quality of physical resources	Input – general	Trend
	Section D – school curriculum and assessment		
16	Ability grouping in mathematics	Process – domain-specific	Trend
17	Extracurricular activities	Process – general and domain-specific	Trend
18	Curricular options for immigrants	Process – general	Trend
19	Assessment practices	Process – general	Trend
20	Use of achievement data for accountability	Process – general	Trend
21	Mathematics activities/ Mathematics extension courses	Process – domain-specific	Trend
	Section E – school climate		
22	Student (behavioural outcomes) and	Process/Outcome –general	Trend/New

Q#	Content	Framework component	Trend/new
	teacher related factors affection school climate		
23	Behavioural outcomes – drop out	Outcome – general	New
24	Parental achievement pressure	Process – general	Trend
25	Parental involvement	Process – general	New
26	Teacher morale	Process – general	Trend
27	Teacher consensus – Innovation	Process – domain-specific	Trend
28	Teacher consensus – Expectations	Process – domain-specific	Trend
29	Teacher consensus – Teaching goals	Process – domain-specific	Trend
30	Teacher evaluation	Process– domain-specific	Trend
	Section F – school policies and practices		
31	Student admission policies	Process – general and domain-specific	Trend/New
32	Educational leadership	Process – general	Trend
33	School management	Process – general	Trend/New
34	Professional development	Process – general and domain-specific	Trend
35	Responsibility for career guidance	Process – general	Trend
36	Career guidance	Process – general	Trend
37	Preparation for tertiary education	Process – general	Trend
38	Quality assurance and school improvement	Process – general	New
39	Truancy monitoring	Process – general	New
40	Truancy consequences	Process – general	New
41	School policies regarding mathematics and truancy	Process – general and domain-specific	New
42	Reasons for transfer to other schools	Process – general	Trend

As is illustrated in Table 2, the ScQ, like other instruments, seeks to balance desires regarding trend and new data on the one hand and general and domain-specific information on the other hand while covering various aspects of inputs, processes and outcomes specified in the questionnaire framework. Given that 2012 will be the second PISA cycle with mathematics as the major domain, domain-specific trend information that links to the information obtained in 2003 becomes of particular interest. In addition, coverage of outcomes in the ScQ has been extended, with a new focus on truancy as the unauthorised absence of students from school. Truancy is considered an - albeit negative - outcome of schooling and an important (negative) indicator of student's use of learning opportunities and is predictive of other types of deviant behaviour. Other new questions seek information on quality assurance and school improvement and students' access and use of the internet. This is of particular relevance given the further developments regarding computer-based testing in PISA and the rising importance of ICT in schools.

In addition, careful analyses of data from the 2009 MS have led to changes to questions and/or response scales about instructional time and school management. In some instances, for example the questions regarding the accommodation of students from different language backgrounds and teacher consensus, material was retained only after careful scrutiny of 2009 data. Still, as regards the accommodation of students from different language backgrounds, for example, changes ensued in the notes version of the questionnaire. Now countries for which this is not an issue are encouraged to drop the question as

analyses showed very little variation in many countries and a large amount of missing data in some countries.

A final point regarding the ScQ is its length. Whereas in previous cycles it took principals or their designates 30 minutes to complete this questionnaire, it is now estimated to take 40 minutes to complete. Therefore, the Questionnaire Expert Group, at its recent meeting in Budapest suggested that consideration be given to the deletion of the following questions:

a. Extracurricular activities*
b. Assessment practices*
c. Teacher morale*
d. Teacher evaluation
e. Responsibility four career guidance*
f. Preparation for tertiary education*
g. Reason for transfer to other schools

Questions marked by an asterisk (*) were those that in the break-out group discussions at the Budapest meeting of NPM which succeeded directly the QEG meeting emerged as being used the least in national reports and analyses.

Analysis

A large part of purpose of the FT is to test translations and to identify any major issues with respect to the understanding, relevance and appropriateness of question content and response scales.

The main analyses of data from the FT of the ScQ will involve checking of frequency distributions, means and plausibility of responses and missing data analysis. To check the quality of scales or constructs such as quality of educational resources, school management and school climate reliability analyses, Item Response Theory (IRT) and Confirmatory Factor Analysis (CFA) will be applied. For the purpose of these analyses, school questionnaire data from different countries will have to be combined.

In addition to these general analyses, a number of analyses with respect to new questions or items are also planned as outlined below.

Truancy. This set of questions and items attempts to link current school policy regarding truancy to how the school implements the monitoring of truancy and follows it up. In addition, the questions also try to develop a chain of events by asking whether truancy was a problem three years ago, whether it was identified as a problem and whether a policy is in place now. The analyses will be aimed at examining whether these intended aims and policies have an effect on student truancy or absenteeism. The analysis is expected to serve as a model for how PISA can study the impact of school-level policies on behavioural outcomes.

Parental involvement. With one exception, the items are identical to those that will be asked in the Parent Questionnaire in 2012. As 13 countries have indicated an interest in administering the Parent Questionnaire it is intended, for these countries, to analyse the level of correspondence between responses given by the principal and responses given by parents in the school, keeping in mind the general low response rate for the Parent Questionnaire. Indeed, one hypothesis would be that schools for which principals report higher parental involvement would have a higher response rate than other schools.

School improvement. School effectiveness research has shown that general school level policies, such as setting goals, implementing professional development, making use of external support and promoting evaluation, will impact student learning and student outcomes. Question 38 captures a range of such policies; it also includes an indicator of domain-specific (mathematics) policies.

Instructional time. To improve the data quality in the responses regarding instructional time, the items have been changed from the previous open-ended response format to a closed response format based on an analysis of PISA 2003 MS data. Careful checks of the frequency distribution across the response categories will be undertaken to examine the appropriateness of the response categories. The domain-specific question regarding instructional time in mathematics is new and again, will require careful analysis of the appropriateness of the response categories.

Teacher consensus. In 2003, when these domain-specific process questions were administered previously, the dimensional analysis methods (IRT, CFA) yielded unsatisfactory results. Only one construct, namely Teacher Consensus, was formed, based on three of the nine items. However, it is suggested that latent class analysis would be a more appropriate analytical technique to be trialled with the 2012 FT data.

Student access to the internet. Its aim is to obtain more detailed information about the type of access to computers students have at school. It covers three elements: first, the type of computer access, static or flexible; second, whether computers are also used outside class; and third, who is funding this resource in the case of one-to-one laptop access. The intention is to build an index of internet accessibility based on the seven items.

School expectation regarding student work. The main hypothesis here is that schools who expect more of their students' work to require access to the internet would be schools that provide more and more flexible access to the internet. Hence, a positive correlations with responses to the provision of computers/laptop and internet access is expected.

School management. The original items in this question were identical to the items used in TALIS. However, only two factors of the hypothesised three factors were supported by the results of a multigroup confirmatory factor analysis using PISA2009 MS data. Items that did not fit the analyses or which showed not to have sufficient cross-cultural applicability were deleted. Hence, for the analysis of the 2012 FT data, a CFA would be expected to reveal two factors, one relating more to the educational goals of the school, the other to educational problems. New items have been suggested that have been shown to measure constructs that play important mediating roles with respect to student achievement (Silins & Mulford, in press; Day, Sammons, Hopkins et. al, 2009; Leithwood & Hallinger, 2002), one regarding teacher participation in school management and principal's instructional leadership have been included. In addition, the analyses of the 2009 data revealed many empty cells, small variance and skewed distribution which gave rise to suggest new answer categories aimed at improving the spread of responses. Therefore, the analysis of FT data will focus on whether the new response scale achieves this aim.

Cognitive laboratories

All new or modified questionnaire items developed for PISA 2012 were evaluated through structured cognitive laboratory interviews prior to the FT.

A previous document (Lee, 2010) described the purpose of the cognitive laboratories (to determine item readability and usability across several languages and cultures), the anticipated participants (approximately 10 students and principals across seven languages and countries), a procedure (one-on-one interviews with standardised scripted probes), a set of issues and outcomes that would be the focus of

the cognitive laboratory studies (identification of problematic items, potential fixes), roles of cognitive laboratory supervisors, interviewers, and respondents, data recording, and a timeline (May through July 2010 data collection, and finalised items delivery by end of August 2010).

Ideally, cognitive laboratories would be conducted in every language group, for every item. This is the only way it would be possible to determine item readability and usability across languages and cultures. However, in previous PISA cycles cognitive laboratories have only been conducted in a very small number of languages, such as English, French, and German. The amount of information that can be obtained through cognitive laboratory investigations is normally quite limited, given the small sample sizes. Limiting cognitive laboratory testing to a few countries is even more limited, as questions in over 95% of the languages are not even evaluated. The assumption has been that item readability and usability actually will only be evaluated in the FT. The purpose of the cognitive laboratory as traditionally conducted in PISA is therefore limited to identifying and correcting only some of the more gross misunderstandings, misinterpretations, frustrations with what the question is asking about, and other major flaws and potential validity threats that may occur. As Norman (2010) suggested in the context of usability testing, the purpose of the cognitive laboratories "is like Beta testing of software… It is for catching bugs." Some of these may be language-specific, and some may generalise across languages and cultures. But the general presumption is that the FT is a better setting in which to capture more nuanced language- and culture-specific problems with items.

Participants

In choosing countries in which to conduct cognitive laboratories, consideration was given to various factors, ranging from ease of conducting studies, to cultural and language diversity to maximise information yield. Given these concerns, the decision was to translate questionnaire items and conduct cognitive laboratory studies in eight languages (countries). Table 3 lists each language and country, along with names and affiliations of the cognitive lab supervisors for each country.

Table 3. Countries, Languages and Cognitive Lab Supervisors

Country	Language	Contact	Affiliation
France	French	Gerben Van Lent	Educational Testing Service
Germany	German	Franzis Preckel and Julia Schembri	University of Trier
Hong Kong	Chinese	Magdalena Mok	Hong Kong Institute of Education
Jordan	Arabic	Zoubir Yazid	Educational Testing Service
Mexico	Spanish	Eduardo Backhoff	Instituto de Investigación y Desarrollo Educativo, UABC
Russia	Russian	Anastasia Lipnevich	Educational Testing Service
South Korea	Korean	Kyunghee Kim	Korea Institute of Curriculum and Evaluation
United States	English	Bobby Naemi	Educational Testing Service

Procedure

As part of the cognitive lab procedure, each contact person organised a series of interviews with at least five 15 year old students and five school administrators or principals who had experience as a parent of a 15 year old.

Efforts were made to incorporate diversity in terms of gender, ethnicity and type of school for the student samples wherever possible. No contact person reported any significant problems for either recruitment or administration of the interview sessions.

Each cognitive laboratory supervisor thus completed the following tasks:

a. Translated at least one booklet of questions into the country language for students;
b. Translated a combined school and parent questionnaire booklet for adults;
c. Recruited participants (students and adults) and interview sites;
d. Conducted cognitive interviews, which involved administering questions to participants, recording responses, and indicating suggested question revisions, and translating records back to English. (Note that each session of cognitive interviews lasted no more than two hours for both students and school administrators.)
e. Negotiated and handled payments to schools and participants.

The consortium provided the following materials to each cognitive lab supervisor:

a. Consent forms, (student participant, parent-of-student, and adult participant);
b. General probes for interviewing;
c. Recording materials (excel spreadsheet) with instructions;
d. Debriefing questionnaire
e. Compensation for cognitive laboratory supervisor.

Interview participants received a paper and pencil version of the questionnaire and filled in all questionnaire items without any interruption from the interviewer.

Immediately after the participants completed filling out the questionnaires, one-on-one interviews were carried out with the standardised scripted probes provided by the Consortium. Interviewers went through item-by-item and asked participants each of the probe questions.

Although cognitive interviews were conducted based on the standardised probes, interviewer flexibility was called upon in some situations. Although not necessary, interviewers were encouraged to use their own judgment to collect as much relevant information as possible from the interview participants.

<u>Probe Questions</u>

a. Did you understand the question? What specifically was confusing or unclear in the question?
b. What do you think the question means?
c. Did you understand the choices of answers? What specifically was confusing or unclear in the answer choices?

d. What issues did you have with the format of the question or the way the question was asked?

Answers to the probe questions, as well as any follow up questions, were coded in an item-by-item report sheet for each question.

After the interview was completed, the interviewer recorded the comments from the item-by-item reports into an Excel spreadsheet, along with a note for any recommended changes to the item.

After all interviews were completed, interviewers also completed the following debriefing questionnaire.

Debriefing Questionnaire

a. Please describe any general problems you observed with the questionnaire (*e.g.*, translation)
b. Please propose any potential solutions to these problems.
c. What are your overall comments about the questionnaires?
d. What are your overall comments about the respondents' reactions to the questionnaire items?
e. Please report any procedural issues (*e.g.*, respondents absenteeism, missing materials, equipment breakdown, respondent resistance, difficulty of using the standardised forms, problems with responses to the probes)

Results

New or modified items from the Student Questionnaire, the Parent Questionnaire, the School Questionnaire, the ICT Familiarity Questionnaire and the Educational Career Questionnaire were all subjected to cognitive laboratory interviews. Feedback from each country, including recorded student responses and overall debriefing comments from the interviewers, was combined into a master document file. Feedback was then reviewed and synthesised, resulting in modifications and recommended changes for many of the items. Feedback fell into several overarching categories:

Scaling Issues**:** These comments largely focused on problems with the scale, including dissatisfaction with the number of response categories, the labels on response categories and a mismatch or lack of agreement between the response categories and the kind of question being asked. For example, some parental respondents were dissatisfied with the lack of an option between "never" and "once a month" when asked how often they buy school supplies for their children, suggesting "once or twice a year" as an option.

Awkward Wording/Translation: These problems focused on issues where items were difficult to understand or had awkward translations. Efforts to deal with these problems largely centred on simplifying the language by removing extraneous words. However other questions simply had vague wording that could not be translated, for example asking how a child is "doing in mathematics" was confusing for both German and French respondents.

Cross Cultural Issues: These problems focused on how certain scenarios or questions were unlikely or not appropriate for a given culture or nation. For example, respondents in Russia noted that students did not have a single science course that occurred at the 15-year old grade level, and that it was possible for students to take chemistry, biology or physics at that age depending on school. German respondents noted that a teaching scenario item that mentioned a teaching arriving five minutes early to class would be unlikely, given that breaks between different subjects are usually just five minutes long, meaning that most German teachers could not possibly be in class five minutes before the lesson starts.

American respondents also noted that the likelihood of certain problem scenarios, such as driving to a wildlife park, might not be appropriate for students of various socioeconomic status levels.

There were also additional interesting contrasting cultural responses. For example, Mexican and Russian respondents reported that mathematics was not necessarily relevant to many careers and so a question referring to the importance of mathematics skills and knowledge in all careers was inappropriate, whereas Hong Kong respondents reported that most jobs required mathematics knowledge and skills and so the question was simply "asking for the sake of asking."

Conclusions

Overall, many of the cognitive laboratory interviews provided valuable information that helped serve as a form of "beta-testing" that helped "catch bugs" in the newly developed items. Feedback was incorporated into item revisions for nearly all of the newly modified or developed items. Despite issues with small budgets, timing crunches and "quick and dirty" translations for items in each of the eight countries, in the end relevant and valuable feedback was obtained in advance of the FT.

REFERENCES

Ballantine, J.A., Larres, P.M., & Oyelere, P. (2007), "Computer usage and the validity of self-assessed computer competence among first-year business students", *Computers & Education, Vol. 49*, pp. 976–990.

Birdsong, D. (Ed.) (1999), "Second language acquisition and the critical period hypothesis", Mahwah & Londong, Lawrence Erlbaum.

Brock, D.B. & Sulsky, L.M. (1994), "Attitudes toward computers: Construct validation and relation to computer use", *Journal of Organizational Behavior, Vol. 15*, pp. 17-35.

Butler, Y. G. & Hakuta, K. (2004), "Bilingualism and second language acquisition", in T. K. Bhatia & Ritchie, W. C. (eds.),*The handbook of bilingualism*, pp. 114-144, Malden/MA: Blackwell.

Cummins, J. (1981). "The role of primary language development in promoting educational success for language minority students", in California State Department of Education (ed.), *Schooling and language minority students: a theoretical framework*, pp. 3-49, Los Angeles, Evaluation, Dissemination and Assessment Center, California State University.

Cummins, J. (2000), "Language, power and pedagogy. Bilingual children in the crossfire", Clevedon, Multilingual Matters.

Cummins, J. (2003), "Bilingual Education", in J. Bourne & E. Reid (eds.), Language Education. *World Yearbook of Education 2003*, pp. 3-19, London/Sterling: Kogan Page.

Day, C., Sammons, P.,Hopkins, D., Harris, A., Leithwood, K., Gu, Q., Brown, E., Ahtaridou, E. & Kington, A. (2009). *The Impact of School Leadership on Pupil Outcomes Final Report* University of Nottingham 2009

Eurydice (2009), "Integrating Immigrant Children into Schools in Europe", Brussels, Education, Audiovisual and Culture Executive Agency (EACEA P9 Eurydice).

Greene, J. (1998), "A meta-analysis of the effectiveness of bilingual education", Claremont, CA, Thomas Rivera Policy Institute.

Grosjean, F. (2006), "Studying Bilinguals: Methodological and Conceptual Issues", in T.K. Bhatia & W.C. Ritchie (eds.), *The Handbook of Bilingualism*, pp. 32-64, Malden, MA, Blackwell.

Hakkarainen, K., Ilomäki, L., Lipponen, L., Muukkonen, H., Rahikainen, M., Tuominen, T., Lekkala, M., & Lehtinen, E. (2000), "Students skills and practices of using ICT: results of a national assessment in Finland", *Computers & Education, Vol. 34*, pp. 103–117.

Hakuta, K., Butler, Y. G. & Witt, D. (2000), "How long does it take English learners to attain proficiency?", Stanford (The University of California Linguistic Minority Research Institute).

Leithwood, K., & Hallinger, P. (eds). (2002). *Second International Handbook of Educational Leadership and Administration*. (pp 561-612) Norwell, MA: Kluwer Academic Publishers.

Limbird, C. and Stanat, P. (2006), "Sprachförderung bei Schülerinnen und Schülern mit Migrationshintergrund: Ansätze und ihre Wirksamkeit", in J. Baumert , P. Stanat and R. Watermann (eds.), *Herkunftsbedingte Disparitäten im Bildungswesen: Differenzielle Bildungsprozesse und Probleme der Verteilungsgerechtigkeit* pp. 257-308. Wiesbaden: VS Verlag für Sozialwissenschaften.

Meisel, J. (2004), "The bilingual child", in T. J. Bhatia and W. C. Richie (eds.), *The Handbook of Bilingualism*, Pp. 91-113, Malden, MA, Blackwell.

Metzger, M. J., Flanagin, A. J., & Zwarun, L. (2003), "College student Web use, perceptions of information credibility, and verification behavior", *Computers & Education, Vol. 41*, pp. 271-290.

Mouw, T. & Xie, Y. (1999), "Bilingualism and the academic achievement of first- and second-generation Asian Americans: Accomodation with or without assimilation?", American Sociological Review, Vol. 64 No. 2, pp. 232-252.

Norman, D. G. (2006). Why doing user observations first is wrong. *Interactions, VIII(4)* (Retrieved from http://interactions.acm.org/content/?p=1012, 4 June 2010).

Portes, A. & Hao, L. (2002), "The price of uniformity: language, family and personality adjustment in the immigrant second generation, *Ethnic and Racial Studies*, Vol. 25 No. 6, pp. 889-912.

Richter, T., Naumann, J. & Groeben, N. (2000), "Attitudes toward the computer: Construct validation of an instrument with scales differentiated by content", *Computers in Human Behavior, Vol. 16*, pp. 473-491.

Richter, T., Naumann, J. & Horz, H. (2010), "Eine revidierte Fassung des Inventars zur Computerbildung (INCOBI-R) [A revised version of the Computer Literacy Inventory]". *Manuscript submitted for publication*.

Schmid, C. L. (2001), "Educational achievement, language-minority students, and the new second generation", *Sociology of Education* (Extra Issue), pp. 71-87.

Silins, H. & Mulford, B. (in press), Submitted (15/02/2010) and revised (16/08/2010) on invitation to *Journal of Educational Leadership, Policy and Practice*

Slavin, R. E. & Cheung, A. (2003), "Effective reading programs for English language learners. A best-evidence synthesis", Center for Research on the Education of Students Placed at risk (CRESPAR), John Hopkins University.

Slavin, R. E. & Cheung, A. (2005), "A synthesis of research on language of reading instruction for English language learners", in Arbeitsstelle Interkulturelle Konflikte und gesellschaftliche Integration (AKI) (ed.), The effectiveness of bilingual school programs for immigrant children, pp-43-76, Berlin, Wissenschaftszentrum Berlin für Sozialforschung, Discussion Paper SP IV 2005-601.

Stanat, P. (2006), „Disparitäten im schulischen Erfolg: Analysen zur Rolle des Migrationshintergrunds", *Unterrichtswissenschaft*, Vol. 43, pp. 98-124.

Stanat, P. & Christensen, G. (2006), "Where immigrant students succeed - A comparative review of performance and engagement in PISA 2003", Paris, Organisation of Economic Co-Operation and Economic Development.

Thomas, W. P. & Collier, V. (1997), "School effectiveness for language minority students", Washington D.C., National Clearing House for Bilingual Education.

Thomas, W. P. & Collier, V. (2002), "A national study of school effectiveness for language minority students' long-term academic achievement", Santa Cruz, CA, Center for Research on Education, Diversity and Excellence, University of California-Santa Cruz.