ANNEX A6: Anchoring Vignettes in the PISA 2012 Student Questionnaire

Since PISA 2000, self-report items based on a Likert-type scale – where respondents are asked to report on the four-point category scale "strongly agree", "agree", "disagree" and "strongly disagree" – have been used to measure many of the contextual factors captured in the student questionnaires. A robust finding across all previous PISA assessments is that the directionality of relationships between some background constructs measured with Likert scales (e.g. mathematics interest) and achievement outcomes on the individual student level is not consistent with those at the aggregated country level. While such inconsistencies might stem from real differences in how relationships play out at the individual and country levels and might be due to omitted variable bias, it is not possible to rule out that they might be the result of systematic differences among countries in how students interpret the agreement response scale or in response styles (e.g. Buckley, 2009; Cheung and Rensvold, 2000).

In order to address this problem, several new survey methods were introduced in the PISA 2012 Student Questionnaire to enhance the validity of questionnaire indexes, especially for cross-country comparisons (Kyllonen and Bertling, 2013). One of the new methods introduced is an alternative scoring of Likert-type items based on so-called anchoring vignettes (King and Wand, 2007; Hopkins and King, 2010). The anchoring vignettes approach has been used for cross-country comparisons in various fields of research (Kapteyn, Smith and Van Soest, 2007; Salomon, Tandon and Murray, 2004; Kristensen and Johansson, 2008), but PISA 2012 is the first large-scale education assessment to use this approach.

Two sets of anchoring vignettes (see Tables A6.1 and A6.2) were included in the PISA 2012 Student Questionnaire, as presented in Tables A6.1 and A6.2. These allow for alternative scoring of self-reported items based on students' defined standards when using the 4-point agreement scale (strongly agree, agree, disagree, strongly disagree). Each of these vignettes describes behaviours of a hypothetical mathematics teacher that are indicative of lower or higher levels of classroom management or teacher support, respectively. Each vignette combines several behavioural aspects. Students read the vignettes and were asked to indicate their level of agreement with a statement about the hypothetical teachers described in the vignettes. Differences in these ratings can be attributed to differences in the interpretation of the rating scale and general differences in preferred response behaviours, as the underlying levels in the hypothetical teachers were held constant across countries.

When items are scored based on vignettes, numerical values for student responses are not assigned based on the concrete response option chosen (e.g. the value 4 for "strongly agree" and 3 for "agree") but based on the self-reported answer *relative to* the personal standard captured by the rating of three vignettes. The extension of the nonparametric scoring procedure (e.g. King and Wand, 2007) is described step by step below; more details are given in the *PISA 2012 Technical Report* (OECD, forthcoming). Clear interpretation of the vignettes in terms of the relative ordering of low, medium and high levels is one requirement for the use of the vignettes.

Table A6.1: Anchoring vignettes based on classroom management behaviours

Low level	The students in Mr. <name's> class frequently interrupt his lessons. As a result, he often arrives five minutes late to class.	ST84Q03
Medium level	The students in Ms. <name's> class frequently interrupt her lessons. She always arrives five minutes early to class.	ST84Q01
High level	The students in Ms. <name's> class are calm and orderly. She always arrives on time to class.	ST84Q02

Note. For each vignette, students were asked to indicate how much they agree with the statement "Mr./Ms. <name> is in control of his/her classroom."

Table A6.2: Anchoring vignettes based on teacher support behaviours

Low level	Ms. <name> sets mathematics homework once a week. She never gets the answers back to students before examinations.	ST82Q03
Medium level	Mr. <name> sets mathematics homework once a week. He always gets the answers back to students before examinations.	ST82Q02
High level	Ms. <name> sets mathematics homework every other day. She always gets the answers back to students before examinations.	ST82Q01

Note. For each vignette, students were asked to indicate how much they agree with the statement "Mr./Ms. <name> is concerned about his students' learning."

The original anchoring vignette method was extended so that multiple items within one index as well as multiple indices can be anchored based on the same set of anchors (Bertling, Kyllonen, Roberts and Blew, forthcoming). Results from analysis of field trials and main survey data showed that the vignettes capturing classroom management behaviours produced clearer results (e.g. regarding the correct rank order of low, medium and high vignettes) and were better suited as anchors for students' self-reported answers. Therefore, students' responses to the anchoring vignettes capturing classroom management behaviours were used for all adjusted indices listed in Table A6.3. Details regarding the comparison of the two sets of vignettes are provided in the *PISA 2012 Technical Report* (OECD, forthcoming).

Table A6.3 summarises 12 adjusted indices included in the PISA 2012 international database.

Table A6.3: Adjusted indices in the PISA 2012 international database

Variable name	Variable label
ANCATSCHL	Attitude towards School: Learning Outcomes (Anchored)
ANCATTLNACT	Attitude towards School: Learning Activities (Anchored)
ANCBELONG	Sense of Belonging at School (Anchored)
ANCCLSMAN	Mathematics Teacher's Classroom Management (Anchored)
ANCCOGACT	Cognitive Activation in Mathematics Lessons (Anchored)
ANCINSTMOT	Instrumental Motivation for Mathematics (Anchored)
ANCINTMAT	Mathematics Interest (Anchored)
ANCMATWKETH	Mathematics Work Ethic (Anchored)
ANCMTSUP	Mathematics Teacher's Support (Anchored)
ANCSCMAT	Mathematics Self-Concept (Anchored)
ANCSTUDREL	Teacher-Student Relations (Anchored)
ANCSUBNORM	Subjective Norms in Mathematics (Anchored)

Description of the alternative scoring based on vignettes for increased validity of international comparisons

The three vignettes used in the anchoring procedure shown in Table A6.1 capture three different levels of classroom management that can be described as low, medium and high. Students were asked to read the vignettes and indicate their level of agreement with the statement that the described teacher is in control of his or her classroom using the same 4-point agreement scale that is also used for most questionnaire indices in the student questionnaire. Depending on their rating standards and their interpretation of the four levels of the agreement scale, students might place the three vignettes on different agreement categories. For instance, one student might "agree" that a teacher described in the first vignette is in control of his/her classroom while another student might "strongly agree" or "disagree" with this statement. Since the actual levels of teachers' classroom management presented in the vignettes are invariant over respondents, differences in students' response to the vignettes signal that students differ with regard to how they interpret the response scale, and that any comparisons based on raw item responses might have validity problems.

The alternative scoring based on the vignettes proposed by Bertling et al. (forthcoming) addresses this problem. Regardless of where on the 4-point agreement scale a student places the vignettes, a student's self-report can be scored relative to his/her rating of low, medium and high for the vignettes. Based on

this approach, in PISA 2012, students' responses on the original 4-point agreement scale were re-scaled into a 7-point scale representing all possible relative rank comparisons of students' self-reports and their rating of the vignettes. On this 7-point scale, the value one represents a rating lower than the low vignette, the value two represents a rating at the level of the low vignette, the value three represents a rating higher than the low but lower than the medium vignette, and so forth. The maximum score, seven, is assigned when a student's self-reported response is higher than the rating of the high vignette. In other words, low values are assigned when a self-report rating is relatively low compared to the evaluation of the vignettes, and high values are assigned when a self-report rating is relatively high compared to the evaluation of the vignettes. In this way, the three vignettes are used to anchor student judgements, providing context for the ratings on other questions sharing the same response scale. Scoring is applied on the individual student level using each student's responses to the vignettes as an anchor for this student's self-reported responses to various Likert-type questions. Table A6.4 illustrates the differences in possible values assigned to original and anchored item responses.

Table A6.4: Possible values for original and anchored item responses

Responses to question as presented in questionnaire	Strongly disagree	Disagree	Agree	Strongly agree			
	1	2	3	4			
Anchored responses	Lower than low vignette	Same as low vignette	In between low and medium vignette	Same as medium vignette	In between medium and high vignette	Same as high vignette	Higher than high vignette
	1	2	3	4	5	6	7

Two special cases are given when there are ties in the responses to the anchoring vignettes (i.e. a student chooses the same agreement category for two or all three vignettes) or when responses to the anchoring vignettes violate the expected order of vignettes (i.e. a student chooses a higher agreement category for a vignette representing a low value on the underlying construct than for a vignette representing a high value on the underlying construct). The scoring method used in PISA 2012 addresses these two cases in the following way.

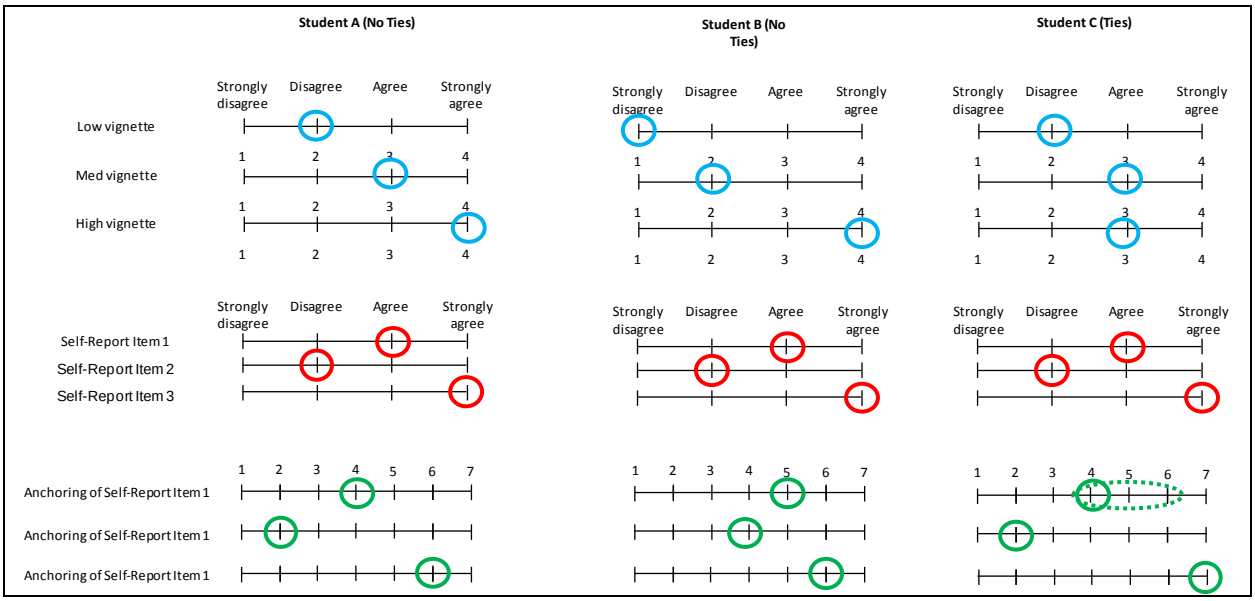
If ties in the vignette ratings are present, students' self-reported responses are scored based on "lower bound scores". This means that the lowest possible score among the range of possible scores is assigned. This score reflects the value on a latent continuum that clearly pertains to the respondent (i.e. the minimum) rather than a higher value that just might pertain to the respondent.

If order violations in the vignette ratings are present, order violations are re-classified into ties. That is, if a student rates the highest vignette lower than the medium vignette, responses for this student would be rescaled in a way that the ratings for the medium and high vignette are tied. For instance, the rank order

"low, high, med" would be rescaled into "low, {med, high}", with the brackets indicating that the same rank is assigned to the medium and high vignette. Note that, in most cases, order violations are rescaled into complete ties of all vignettes (i.e. "{low, medium, high}"). Ties are created at the highest response category chosen by the student. For instance, in the example used above ("low, {med, high}") the tie is created at the value the respondent assigned to the high vignette. Ties are then analysed as described above.

A graphical illustration of the scoring procedure based on vignettes for three examples with and without ties is given in Figure A6.1. The three hypothetical students in this example provided exactly the same responses to the three self-reported items shown, but differ in their responses to the vignettes. As a result, scores on the anchored items also differ between the three students. Further detail about the scoring approach is provided in Bertling et al. (forthcoming) and in the *PISA 2012 Technical Report* (OECD, forthcoming).

Figure A6.1: Illustration of scoring based on vignettes for three hypothetical students



Source: Bertling and Kyllonen (2013)

Assumptions and Cautions

The alternative scoring approach for Likert-type items based on vignettes makes the frame of reference for scoring of questionnaires items more transparent and can thereby help in interpreting students' responses across different countries and education systems. There are, however, several assumptions that underlie the use of anchoring vignettes in the context of PISA; therefore, caution is advised when interpreting adjusted indices using anchoring vignettes.

First, the scoring approach is based on two main identifying assumptions: "vignette equivalence" and "response consistency" (see e.g. Kapteyn et al., 2011). The vignette equivalence assumption posits that different respondents interpret the vignette scenario in the same way. In other words, there is an assumption that all differences in the ratings of the vignettes should be attributable to the differences in how respondents interpret and use the agreement scale, but not to the differences in how respondents

interpret the vignette scenario themselves. The response consistency assumption posits that respondents use the same standards both in evaluating themselves and in providing an evaluation of the vignette scenario.

Second, the original anchoring vignette method was developed to anchor stand-alone questions only. By contrast, in the context of the PISA 2012 Student Questionnaire, the anchoring vignette method was extended so that the same vignette scenario is applied to a large set of different items. This extension is possible because of an assumption that an individual's rating standards are invariant across different context whenever the same rating scale is used. This means that students are expected to use a four-point Likert scale with the categories "Strongly disagree" to "Strongly agree" in a reasonably comparable way for the different questions included in the Student Questionnaire, whether these refer to items such as "I learn mathematics quickly" or items such as "My teacher helps students with their learning".

The scoring process anchoring student responses using vignette scenarios depends on the particular vignette scenarios (i.e. where on the continuum of the underlying construct the vignettes are located) and the number of vignettes used. While PISA 2012 data suggests reasonable consistency of results across the two sets of vignettes, further research is needed to fully understand the effects of different vignette context and how the validity of results depends on the number of vignettes and number of scale points used. For instance, gains in validity might be larger for questions that capture similar constructs as the constructs described in the vignettes.

The order of vignettes and self-reports in the questionnaire may have an influence on the results. As Hopkins and King (2010) showed, administering vignettes first might have a priming effect that reduces inter-individual differences in interpretation of the response scale. In the PISA 2012 Student Questionnaire some self-reported questions using the four-point Likert scale are presented before the vignettes and others are asked after the vignettes.

Finally, in order to use data from all students, including students with tied anchor evaluations (e.g. students who give the same ratings for two vignettes classified as low and medium) or "order violations" (e.g. students who give lower ratings to a vignette classified as high as to a vignette classified as medium or low), additional assumptions are needed, as described in the previous sections. Future research is needed to fully understand students' response processes.

It is recommended that adjusted indices using anchoring vignettes should be interpreted in addition to classical indices, not as a replacement. Both values on classical questionnaire indices and on adjusted indices can be influenced by students' systematic or unsystematic response behaviors. Examining both of these indices provide a basis for a more general picture of relationships and effects that is less tied to a single survey method only.

References

Bertling, J. P. and P.C. Kyllonen (2013), "Using anchoring vignettes to detect and correct for response styles in PISA questionnaires", in: Prenzel, M. (Chair), *The Attitudes-Achievement-Paradox:How to Interpret Correlational Patterns in Cross-Cultural Studies*, Invited Symposium at the EARLI 2013, Munich, Germany.

Bertling, J. P. and P. C. Kyllonen, R. D. Roberts and E. Blew (forthcoming), *Anchoring Vignettes for Improved Cross-Cultural Comparability of Survey Responses in International Educational Large-Scale Assessments.*

Buckley, J. (2009), *Cross-National Response Styles in International Educational Assessments: Evidence from PISA 2006.* Last accessed 10/02/2012 https://edsurveys.rti.org/PISA/documents/Buckley_PISAresponsestyle.pdf

Cheung, G. W. and R. B. Rensvold (2000), "Assessing Extreme and Acquiescence Response Sets in Cross-Cultural Research Using Structural Equations modelling", *Journal of Cross-Cultural Psychology*, Vol. 31, pp. 187-212.

Hopkins, D. J. and G. King (2010), "Improving Anchoring Vignettes: Designing Surveys to Correct for Interpersonal Incomparability", *Public Opinion Quarterly*, Vol. 74, pp. 201-22.

Kapteyn, A., J. P. Smith and A. Van Soest (2007), Vignettes and Self-Reports of Work Disability in the US and the Netherlands, *American Economic Review*, Vol. 97, pp. 461–73.

Kapteyn, A., J. P. Smith, A. Van Soest and H. Vonkova (2011), *Anchoring Vignettes and Response Consistency*, RAND Working Paper WR-840.

King, G. and J. Wand (2007), "Comparing Incomparable Survey Responses: Evaluating and Selecting Anchoring Vignettes, *Political Analysis*, Vol. 15, No. 1, 46–66.

Kristensen, N. and E. Johansson (2008), New Evidence on Cross-Country Differences in Job Satisfaction Using Anchoring Vignettes, *Labour Economics*, Vol. 15, pp. 96-117.

Kyllonen, P. C. and J. P. Bertling (2013), Innovative Questionnaire Assessment Methods to Increase Cross-Country Comparability, in: L. Rutkowski, M. von Davier and D. Rutkowski (Eds.), *A Handbook of International Large-Scale Assessment Data Analysis.*

OECD (forthcoming), *PISA 2012 Technical Report*, OECD Publishing

Salomon, J. A., A. Tandon, A. and C. J. L. Murray (2004), Comparability of Self-Rated Health: Cross-Sectional Multi-Country Survey Using Anchoring Vignettes, *British Medical Journal*, Vol. 328, pp. 258–61.