Large-scale Assessments in Education
a SpringerOpen Journal

RESEARCH Open Access

On the use of rotated context questionnaires in conjunction with multilevel item response models

Raymond J Adams[1*], Petra Lietz[2] and Alla Berezner[2]

* Correspondence: r.adams@unimelb.edu.au
[1]University of Melbourne and Australian Council for Educational Research, Melbourne, Australia
Full list of author information is available at the end of the article

Abstract

Background: While rotated test booklets have been employed in large-scale assessments to increase the content coverage of the assessments, rotation has not yet been applied to the context questionnaires administered to respondents.

Methods: This paper describes the development of a methodology that uses rotated context questionnaires in conjunction with multilevel item response models and plausible values. In order to examine the impact of this methodology on the continuity of the results, PISA 2006 data for nine heterogeneous countries were rescaled after having been restructured to simulate the outcomes of the use of different rotated context questionnaire designs.

Results: Results revealed negligible differences when means, standard deviations, percentiles, and correlations were estimated using plausible values drawn with multilevel item response models that adopted different approaches to questionnaire rotation.

Conclusions: The results of the analyses support the use of rotated contextual questionnaires for respondents in order to extend the methodology currently used in large-scale sample surveys.

Keywords: Multilevel item response models; Questionnaire design

Background

A common goal of sample surveys is to measure a latent variable proficiency, an aptitude, an attitude, or the like and then relate that latent variable to other characteristics of the respondents. For example, in an educational context, the relationship that is examined might be the correlation between the latent variable and another characteristic, such as years of schooling, or it might be between-group differences in mean scores for a latent variable. The ultimate aim of the survey is to examine the distribution of the latent variable in the target population and to make inferences concerning the relationships between latent variables and other variables in the target population. In psychometrics, the science of constructing measures of latent variables, it is generally accepted that measures of latent variables are fallible and include random error components that must be taken into consideration when such inferences are being made. Cochran (1968), for example, argues that when measurement error in latent variables is ignored, most statistical tests are vitiated.

Springer

© 2013 Adams et al.; licensee Springer. This is an Open Access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/2.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

The study of statistical models with error invariables is a well-developed area of statistical and psychometric inquiry. Its extensive body of literature dates back to at least Adcock (1878, cited in Gleser, 1981), and from there to Gleser (1981), Anderson (1984), Mislevy (1985), Fuller (1987), and Adams, Wilson, and Wu (1997). Econometricians were the first to extensively study models with errors in the variables, and their use in econometrics became widespread Anderson, (1984). In psychological and educational research, the presence of substantial measurement error resulted in the development of linear structural relation (or LISREL) models (see, for example, Jöreskog & Sörbom 1984; Muthén 2002) and latent regression (also known as multilevel item response theory) models (Adams, Wu, & Carstensen, 2007; Fox & Glas 2002).

In the context of large-scale sample survey studies,[a] multilevel item response theory models have been the method of choice for investigators undertaking appropriate data analysis in the presence of measurement error. There appear to be three primary reasons for this choice. First, the models are scalable; that is, they are methods that have been demonstrated to work well in contexts with many thousands of sampled respondents, many latent variables, and hundreds of manifest variables. Second, they can be integrated with other key components of sample survey methodology, in particular the weighting and sampling variance estimation that is required in structured multistage samples. And, third, they can be broken into discrete steps so that the study developers can construct a database and secondary analysts can then use *standard* and readily accessible analytic tools to analyze the data in ways that properly deal with the impact of the presence of measurement error (Adams, 2002; Adams, Wu, & Macaskill 2007; Gonzalez, Galia, & Li 2004; Mislevy 1990).

Researchers exploring PISA, NAEP, and TIMSS data have used the multilevel item response theory approach to examine the relationships between a small number of latent proficiency variables, for example, three to seven such variables in the case of PISA, and quite a large number of other variables collected via respondent contextual questionnaires. To ensure adequate content coverage of the latent proficiency variables, PISA, NAEP, and TIMSS all use multiple linked test booklets, which means that although each respondent responds to just 60 (NAEP) to 120 (PISA) minutes of assessment material, the total sum of assessment material used far exceeds this amount.

As noted, each of these studies routinely uses linked (or rotated) assessment booklets, a process often referred to as a multiple-matrix sampling design (Shoemaker 1973). However, in order to broaden the assessment while limiting individual response burden, the studies rely on a single set of contextual variables being administered to all respondents. No attempt, as far as we are aware, has been made thus far to apply such a rotated design to the context questionnaires, and thereby extend the number of contextual variables beyond that which can be obtained from a single common questionnaire administered to all respondents. Gonzalez and Eltinge (2007a, 2007b), however, have discussed the possibility of using rotated questionnaires in the US Consumer Expenditure Quarterly Interview.

In this paper, we explore the possibility of administering rotated context questionnaires to respondents in order to expand the coverage of contextual variables in sample surveys that employ multilevel item response theory scaling models. In addition, we examine how a changed methodology might affect the continuity of results with respect not only to the latent proficiency variables themselves but also to their correlations with the context constructs. The specific context for our work is the PISA survey.

Although the idea of having rotated forms of the respondent context questionnaires in order to extend their content coverage is appealing, the situation for these questionnaires differs slightly from that for the test booklets. To illustrate this difference, we provide an overview of the PISA analysis approach and then follow it with an explanation of the difference between using data from the test booklets and using data from the respondent context questionnaire in the multilevel scaling model used in PISA.

The pisa analysis approach

PISA is a cyclical cross-sectional study, with data collections occurring every three years. Four PISA assessments have now been completed (OECD 2001, 2004, 2007a, 2010) and a fifth is being implemented. Here we discuss the third cycle of PISA, the data collection that occurred in 2006 (referred to as PISA 2006). Our focus on the third cycle reflects our decision to use data from PISA 2006 to explore the potential use of rotated questionnaires.

PISA 2006 tested three subject domains, with science as the major domain and reading and mathematics as the minor domains. PISA allocates more assessment time to a major domain than it does to the minor domains, and typically reports subscales for major domains but not for minor ones. During PISA 2006, 108 test items, representing approximately 210 minutes of testing time, were used to assess student achievement in science. The reading assessment consisted of 28 items, and the mathematics assessment consisted of 48 items, representing approximately 60 minutes of testing time for reading and 120 minutes for mathematics.

The 184 main survey items were allocated to 13 half-hour (30-minute) mutually exclusive item clusters that included seven science clusters, four mathematics clusters, and two reading clusters. Thirteen test booklets were produced, each composed of four clusters according to a rotated design. This approach resulted in 120-minute test booklets consisting of two 60-minute parts, each made up of two of the 30-minute clusters and with students allowed a short break 60 minutes after the start of the test.

The booklet design was such that each cluster appeared in each of the four possible positions within a booklet exactly once, and each cluster occurred once in conjunction with each of the other clusters. Each test item, therefore, appeared in four of the test booklets. This linked design made it possible, when estimating item difficulties and student proficiencies, to apply standard measurement techniques to the resulting student response data (OECD, 2008). Student performance results were reported in terms of one overall scale in science, five science subscales, one overall scale for mathematics, and one overall scale for reading.

Fitting a multilevel item response model

The PISA research team used the mixed coefficients multinomial logit (MCML) model, as described by Adams, Wilson and Wang (1997a) and Adams and Wu (1997), to scale the 2006 data, and they used the ConQuest software (Wu et al. 1997) to carry out the process. Details of the scaling can be found in Adams (2002). We provide a limited sketch of the process here so as to contextualize the extension to the methodology that we explore in this paper.

The multilevel scaling model used consists of two components a conditional item response model, $f_x(\mathbf{x};\boldsymbol{\xi}|\boldsymbol{\theta})$, and a population model, $f_\theta(\boldsymbol{\theta};\boldsymbol{\gamma},\boldsymbol{\Sigma},\mathbf{W})$. The conditional item response model describes the relationship between the observed item response vector $\mathbf{x}$ and the latent variables, $\boldsymbol{\theta}$. The $\boldsymbol{\xi}$ parameters characterize the items. The population model, which describes the distribution of the latent variables and the relationship between the contextual variables and the latent variables, is a multivariate multiple regression model, where $\boldsymbol{\gamma}$ are the regression coefficients that are estimated, $\boldsymbol{\Sigma}$ is the conditional covariance matrix, and $\mathbf{W}$ are the contextual variables.

The conditional item response model and the population model are combined to obtain the unconditional, or marginal, item response model:

$$f_x(\mathbf{x};\ \boldsymbol{\gamma},\boldsymbol{\Sigma},\ \mathbf{W}) \ = \ \int_\theta f_x(\mathbf{x};\ \boldsymbol{\xi}|\boldsymbol{\theta})\ f_\theta\ (\boldsymbol{\theta};\boldsymbol{\gamma},\boldsymbol{\Sigma},\mathbf{W})\ d\boldsymbol{\theta} \tag{1}$$

It is important to recognize that, under this model, the locations of respondents on the latent variables are not estimated. The parameters of the model are $\boldsymbol{\gamma}$, $\boldsymbol{\Sigma}$, and $\boldsymbol{\xi}$, where $\boldsymbol{\gamma}$ and $\boldsymbol{\Sigma}$ are the population parameters and $\boldsymbol{\xi}$ are the item parameters.

Directly estimating $\boldsymbol{\gamma}$ and $\boldsymbol{\Sigma}$ from the item response vectors bypasses the problem of having fallible estimates of latent proficiencies, that is, the issue of problems caused by measurement error, as discussed in the introduction. This approach also leads to unbiased estimates of population characteristics being obtained, assuming, of course, that the data satisfy the assumptions of the scaling and regression models.

The item response model used in (1) does not require the same complete list of item responses for all respondents. So, provided that the item response data are missing at random (Rubin 1976), which is the case with the rotated test booklet designs used in PISA, this model is well suited to incomplete designs. However, this is not the case for the population model, which in its PISA implementation requires complete data.

Plausible values

Currently, only a limited range of researchers are able to implement methodologies that permit the estimation of $\boldsymbol{\gamma}$ and $\boldsymbol{\Sigma}$ for the set of contextual variables that are of interest to them.[b] Therefore, to support further analysis, PISA uses the imputation methodology usually referred to as plausible values (Mislevy 1991) during construction of its public access databases.

Plausible values are intermediate values that are used in the algorithm that is implemented in ConQuest to estimate the parameters of (1) (Volodin & Adams 1997). PISA plausible values are sets of imputed proficiencies that are provided, per respondent, for all latent variables included in the scaling. They are thus random draws from the estimated posterior proficiency distribution for each student. Adams (2002) details how the random draws are made. The theory supporting the use of the plausible value approach can be found in Rubin (1987) and Mislevy (1991); Beaton and Gonzalez (1995) provide an overview of how plausible values should be used.

A key feature of plausible values is that they allow the results obtained from fitting (1), in particular the regression coefficients $\boldsymbol{\gamma}$ and $\boldsymbol{\Sigma}$, to be recovered without the need to access the specialist software required to fit model (1). They can also be used to estimate the parameters of any submodel of the regression model used in (1).

In PISA, the regression parameters in (1) are estimated on a country by country basis. Similarly, plausible values are drawn on a country by country basis. But before this

model can be estimated, it is also necessary to select the contextual variables, **W**, that will be used in each country. In PISA and NAEP, these variables are referred to as conditioning variables. The steps used to prepare the conditioning variables in PISA are based upon those used in NAEP (Beaton 1987) as well as in TIMSS (Macaskill, Adams, & Wu, 1998), and are given below:

- *Step 1:* Three variables (booklet ID, school ID, and gender) are prepared so they can be directly used as conditioning variables. Variables for booklet ID are represented by deviation contrast codes (Pedhazur 1997). Each booklet other than a reference booklet is represented by one variable. Variables for school ID are coded using simple contrast codes, with the largest school as the reference school.
- *Step 2:* Each categorical variable in the student questionnaire is dummy coded. Details of this dummy coding can be found in the PISA 2006 technical report (OECD 2008). For variables treated as continuous (including questionnaire indices constructed using item response theory), missing values are replaced with the country mean, and a dummy variable indicating a missing response is created.
- *Step 3:* For each country, a principal components analysis of the dummy-coded categorical and continuous variables is performed, and component scores are produced for each student (The number of components retained must be sufficient to account for 95% of the variance in the original variables).
- *Step 4:* The item-response model is fitted to each national dataset. The national population parameters are estimated using item parameters anchored at their international location, and conditioning variables are derived from the national principal components analysis and from Step 1.
- *Step 5:* Five vectors of plausible values are drawn using the method described above. The vectors provide a plausible value for each of the PISA 2006 reporting scales.

The pool of candidate variables for **W** consists of the variables in the student contextual questionnaire. Until now, these have been limited to the number of variables that can be obtained through the administration of a single 30-minute contextual questionnaire to all respondents. This situation raises the question of whether the multilevel item response theory methodology, which easily handles rotated assessment booklets, can be implemented with rotated contextual questionnaires. The question is asked because, in the case of rotated contextual variables, the set of candidate variables for **W** will differ for different respondents.

The motivation for using rotated questionnaires in order to extend the coverage of the student questionnaire stems from several somewhat related reasons. First, and just as is the case with the multiple matrix sampling for the test, it is desirable to limit the amount of time students are required to concentrate on completing a questionnaire while simultaneously providing opportunity to increase the content coverage of the questionnaires. Second, once various cognitive domains have been assessed, the number of variables or constructs that are thought to be related to performance in the different domains ends up being larger than if only one domain had been assessed. Third, given that in most countries the variance in performance between students exceeds the variance in performance between schools, it is necessary to seek out in the student

questionnaire a greater number of variables for inclusion that can be used subsequently to describe performance differences between students.

However, before going down the path of rotating context questionnaires in PISA, we need to address three questions:

1. Is it possible to develop a methodology that uses rotated contextual questionnaires?
2. Will a change in the methodology to rotated questionnaires have an impact on the continuity of PISA results?
3. Will such a change affect the estimated relationships between the context variables and performance?

In the following section, we explore five alternative approaches to allocating contextual variables to rotated booklets. We also examine the effects of these approaches on the estimated distributions of the latent proficiency. More specifically, we direct our analyses toward an examination of how the means and distributions of the latent proficiency variables are affected when two forms of the PISA student context questionnaire (StQ) are used. PISA rotates the two forms in a way that leads to all students being asked to respond to questions in a common part of the questionnaire and then leads to half of the students being asked to respond to questions in one of the two rotated parts of the questionnaire and the other half to the other rotated part of the questionnaire. Even more specifically, our analyses also address whether results differ depending on how student context constructs are assigned to the rotated forms.

We acknowledge the possibility of using other rotated questionnaire designs. One such design, for example, could involve three rotated forms, with each construct included in two of the three forms. Such an overlap would enable a wider range of subsequent analyses because it would allow calculation of the correlations between the constructs in the different forms. This approach, however, would reduce the additional space gained as a consequence of rotation. Furthermore, because the main aim of this paper is to examine the possible implications of a rotated questionnaire design for the proficiency estimates rather than search for the optimal questionnaire rotation design, we elected to use the two-form design.

Methods

To explore alternative approaches to using rotated context questionnaires, we rescaled the PISA 2006 data (OECD, 2007b) for nine countries after restructuring this information to simulate the outcomes of the use of rotated context questionnaires. During our research, we considered two rotation designs, each of which consisted of two questionnaire forms. In both designs, the two questionnaire forms shared a common set of variables. Each form also contained a variable set unique to it. To achieve this design, we divided the available pool of questions into three mutually exclusive subsets of variables. We assigned the first subset, the common set, to both questionnaire forms, and then assigned the second subset to the first rotated questionnaire form and the third subset to the second rotated questionnaire form.

The difference between the two rotation designs lies in the approach that we used to generate the variable sets in the rotated parts. While the common set was fixed to be the same in both designs, the method of constructing the rotated part differed. In Design 1, we constructed the variable sets so that each had a similar correlation with

science performance (Variable Set 1.1 and Variable Set 1.2). In Design 2, we constructed the variable sets so that one had a lower correlation with science performance (Variable Set 2.1) and the other had a higher correlation with performance (Variable Set 2.2). This second design enabled us to ascertain if the correlations between the questions and performance would be likely if questions were assigned, in actuality, to the rotated parts of a questionnaire. In other words, the two designs allowed us to explore what impact the constructs in the rotated parts of the questionnaire had on various aspects of the proficiency estimates (i.e., means, standard deviations, percentiles, correlations), with that impact dependent on whether the constructs had a similar or a different relationship with performance.

Table 1 provides a summary of the two designs, and therefore illustrates how we allocated the variable sets to the questionnaire forms. Note, in particular, that the variable sets in the rotated parts of the questionnaire included constructs formed from individual variables and that the items forming a construct were not split across the questionnaire forms.

Table 2 sets out the variables contained in the common part of the questionnaire. These variables consisted of the major reporting variables—age, gender, grade, parental occupation and education, immigration status, age at which the student (if an immigrant) arrived in the country, and language spoken at home. The variable called *effort* in the table relates to a question that asked students to indicate the level of effort they put into the PISA achievement test compared with other tests they had taken. The three remaining constructs in the table are based on responses to questions regarding cultural and other possessions as well as educational resources available at home.

The allocation of constructs to the sets in the rotated parts of the questionnaire involved the following steps. We began by calculating, at the student level for each country, correlations between each construct and each of the proficiencies in the content domains (i.e., mathematics, reading, and science). We then used these results to compute the average country-level correlations between each variable set and the performance for all countries and for OECD countries only. Finally, we allocated the constructs to the two sets using the results of the second step so that the average correlations of the two sets with achievement at the student level were similar for rotated Forms 1.1 and 1.2 and differed for Forms 2.1 and 2.2.

Table 3 details which variables were allocated to the variable sets in the rotated parts of the questionnaire, and Table 4 provides the outcomes of that allocation in terms of correlations with science (as the major domain) proficiency in PISA 2006. We allocated the different constructs to the two forms of the questionnaire in such a way that responses from only half the students to each of the four sets of variables were retained,

Table 1 Rotation designs

	Design 1		Design 2	
	Form 1.1	**Form 1.2**	**Form 2.1**	**Form 2.2**
Common part	✓	✓	✓	✓
Variable Set 1.1: Average correlation with performance	✓			
Variable Set 1.2: Average correlation with performance		✓		
Variable Set 2.1: Low correlation with performance			✓	
Variable Set 2.2: High correlation with performance				✓

Table 2 Items and constructs in the common part

PISA 2006 variable name	Variable/construct label
PROGN	Country study program
GRADE	Grade
AGE	Age of the student
SEX	Gender
BMMJ1	Occupation of mother
BFMJ2	Occupation of father
BSMJ5	Occupation of self at 30
MISCED	Educational level of mother
FISCED	Educational level of father
IMMIG	Immigration status
AGECNT	Age arrived in country
LANG	Language at home
EFFORT	Effort thermometer question
CULTPOSS	Classic literature, books of poetry, works of art
HEDRES	Study desk, quiet place to study, computer for school work, educational software, own calculator, books to help with school work, dictionary
WEALTH	Own room, internet link, dishwasher, DVD/VCR, three country-specific wealth items + number of cellphones, TVs, computers, cars

resulting in missing information on this set of variables from the other half of the student sample (see Table 1 above).

When scaling the data from each design, we implemented three approaches:

- *Common part conditioning:* With this reduced conditional model, the information from only the variables in the common part of the questionnaire was used for comparative purposes.
- *Joint conditioning:* Here, the two questionnaire forms were used jointly for each design. In practice, this meant having one conditioning model for each country and then setting the data for one set of variables to missing for one half of the students and setting the data for the other set of variables to missing for the other half of the students. Information from the common part for all students was included in the model.
- *Separate conditioning:* This approach involved using the two questionnaire forms separately for each design, and that, in turn, meant running separate conditioning models, with one using only data from the first rotated form and the other using only data from the second rotated form. The information from the common part of the questionnaire for all students was also included in the model.

When taking the joint conditioning approach, we replaced missing information with the mean for that construct and also included a dummy variable indicating missing. We then replaced the missing information for the categorical variables with the mode for that variable and again included a dummy variable indicating missing. Thus, our analyses involved inclusion of two variables for each background item, one indicating the actual response of a student or the mean/mode if the response was missing, and the other variable indicating whether the response was not missing (=0) or missing (=1).

Table 3 Variable/construct allocations to each set

Variable set 1.1		Variable set 1.2		Variable set 2.1		Variable set 2.2	
Database variable name	*Variable/construct description*	*Database variable name*	*Variable/construct description*	*Database variable name*	*Variable/construct description*	*Database variable name*	*Variable/construct description*
CARINFO	Student information on science-related careers	ENVOPT	Environmental optimism	CARINFO	Student information on science-related careers	GENSCIE	General value of science
CARPREP	School preparation for science-related careers	ENVPERC	Perception of environmental issues	CARPREP	School preparation for science-related careers	INTSCIE	General interest in learning science
ENVAWARE	Awareness of environmental issues	GENSCIE	General value of science	INSTSCIE	Instrumental motivation in science	JOYSCIE	Enjoyment of science
HIGHCONF	Self-confidence in ICT high- level tasks	INTCONF	Self-confidence in ICT internet tasks	ENVOPT	Environmental optimism	PERSCIE	Personal value of science
INSTSCIE	Instrumental motivation in science	INTSCIE	General interest in learning science	ENVPERC	Perception of environmental issues	SCIEACT	Science activities
JOYSCIE	Enjoyment of science	INTUSE	ICT internet/ entertainment use	SCHANDS	Science teaching: hands-on activities	SCIEEFF	Science self-efficacy
PRGUSE	ICT program/ software use	PERSCIE	Personal value of science	SCINTACT	Science teaching: interaction	SCIEFUT	Future-oriented science motivation
SCIEFUT	Future-oriented science motivation	RESPDEV	Responsibility for sustainable development	SCINVEST	Science teaching: student investigations	SCSCIE	Science self-concept
SCINTACT	Science teaching: interaction	SCAPPLY	Science teaching: focus on applications or models	SCAPPLY	Science teaching: focus on applications or models	ENVAWARE	Awareness of environmental issues
SCINVEST	Science teaching: student investigations	SCHANDS	Science teaching: hands-on activities	HIGHCONF	Self-confidence in ICT high-level tasks	RESPDEV	Responsibility for sustainable development
SCSCIE	Science self-concept	SCIEACT	Science activities	INTUSE	ICT internet/entertainment use	INTCONF	Self-confidence in ICT internet tasks
		SCIEEFF	Science self-efficacy	PRGUSE	ICT program/software use		

Table 4 Correlations between variable sets for the rotated parts of the questionnaire and science performance

	Student level
Variable Set 1.1: Average correlation with performance	0.08
Variable Set 1.2: Average correlation with performance	0.11
Variable Set 2.1: Low correlation with performance	−0.02
Variable Set 2.2: High correlation with performance	0.22

Note: The reported correlation is the simple mean across countries of the simple mean across constructs of the student-level correlations within each country.

We acknowledge that in contexts where there is an interest in the estimates of the regression coefficients, this approach can produce biased results (Jones 1996; Rutkowski 2011). However, this line of argument may be partially irrelevant, or less important, in the current context, where concern lies with the outcomes of analysis based upon alternatively derived plausible values, rather than upon the direct estimates of the regression coefficients. In essence, the focus here is on the generation of the plausible values themselves, and not on the estimated regression coefficients obtained from an analysis of the plausible values. So while it may indeed be unwise to use dummy coding to deal with missing data when analyzing the final dataset, it does not follow that dummy coding should not be used when generating the plausible values. Any impact of this way of treating the structurally missing data caused by the rotation will be evident in the obtained results.

An alternative treatment for missing data, which we could have implemented as a fourth scaling approach, would have been to use imputations as a means of replacing missing information with "pseudo-information." However, imputations for missing data are model-dependent draws from the posterior distribution of random variables, conditional on the observed values of other available variables, and requiring use of estimated relationships between the variable that is missing and the remainder of the variables. In order to account for the uncertainty associated with these imputations, we would need to have multiple sets of data, a requirement that would increase the operational burden by a multiplier equal to the number of imputations (often 5). We therefore considered this approach to be a nonviable one.

Each of the above listed approaches to scaling followed the procedures that were implemented in the official OECD analyses of PISA 2006 data (OECD, 2007a). Descriptions of these approaches can be found in the PISA 2006 technical report (OECD, 2008). Our application of the three scaling approaches we used in combination with the two rotation designs (see Table 1) led to five sets of results for each of the three cognitive domains, namely mathematics, reading, and science.

Data

We purposely selected the countries that we included in our analyses because we wanted them to be fairly heterogeneous in terms of level of science performance, culture, language of instruction, and the mix of OECD and non-OECD countries. We considered that this approach would make exploration of the implications of the rotated questionnaire design in very different contexts easier and more valid. The nine countries that we eventually selected are listed in Table 5.

Table 5 Countries in the analyses

Country	Region	OECD member	Science performance		Effective sample	Sample size
			Mean	*SE*	*Size**	
Colombia	South America	No	388	3.4	632	4,478
France	Western Europe	Yes	495	3.4	694	4,716
Germany	Western Europe	Yes	516	3.8	914	4,891
Hong Kong SAR	Asia	No	542	2.5	1,374	4,645
Jordan	Middle-East	No	422	2.8	1,003	6,509
Norway	Scandinavia	Yes	487	3.1	954	4,692
Poland	Central Europe	Yes	498	2.3	1,472	5,547
Russian Federation	Eastern Europe/Asia	No	479	3.7	597	5,799
United States	North America	Yes	489	4.2	630	5,611

Note: *Effective sample size in each country, taking into account the inflation of the total variance due to measurement error and complex sampling design (Design Effect 5). Values for science are taken from Table 11.15 on page 202 of the PISA 2006 technical report (OECD, 2008).

Results

The combination of the two rotation designs and the three scaling approaches led to the following five sets of results:

- *Set 1:* This set of results, pertaining to the common part conditioning, is labelled "common" in the results tables in this section of the paper.
- *Sets 2 and 3:* These two sets of results, for joint conditioning, are labelled "samecorrjoint" and "hilocorrjoint" in the results tables. The former denotes Design 1, in which the variable sets had similar correlations with performance, and the latter denotes Design 2, in which one variable set had high and one variable set had low correlations with performance.
- *Sets 3 and 4:* These two sets of results relate to the separate conditioning. They are respectively labelled "samecorrsep" for Design 1, in which the variable sets had similar correlations with performance, and "hilocorrsep" for Design 2, in which one variable set had high and one variable set had low correlations with performance.

The results that we obtained from fitting the original PISA 2006 multilevel item response model that used all variables in the student questionnaire are labeled "original" in the results tables.

The comparisons that we report below between the results produced from the original PISA 2006 analyses and those obtained from the five rotation models are first those for the proficiency means and standard deviations, second those for the percentiles of the proficiency distributions, and third those for the correlations between proficiency and the context constructs. We considered the differences to be substantive if they exceeded the standard error of the corresponding estimate.

Means and standard deviations

The comparison of means and standard deviations between the plausible values generated from the five rotation models and the original plausible values revealed

Table 6 Differences in estimated means from the original scaling for *mathematics* for each of the alternative conditioning approaches

	Colombia	Germany	France	Hong Kong SAR	Jordan	Norway	Poland	Russian federation	United States
Common	0.0	1.0	−0.8	−0.1	0.0	−0.4	−0.9	−0.8	0.0
Samecorrjoint	0.2	0.5	−0.4	−0.2	0.3	−0.4	−1.3	−1.0	−0.3
Samecorrsep	0.3	0.4	−0.8	−0.3	−0.4	−0.6	−0.9	−0.9	−0.3
Hilocorrjoint	0.1	0.4	−0.1	0.8	0.2	−0.7	−1.0	−0.7	−0.2
Hilocorrsep	0.3	0.9	−0.6	0.6	0.4	−0.9	−1.7	−1.0	0.0

Note: Standard errors of the originally estimated means in mathematics range from 3 to 4 PISA points.

no differences of substantive importance with respect to performance in mathematics, reading, or science. The differences between the estimated means using each of the alternative rotation designs and those originally obtained are shown in Table 6 for mathematics, Table 7 for reading, and Table 8 for science. The differences for standard deviations are shown in Table 9 for mathematics, Table 10 for reading, and Table 11 for science.

We can see from Table 6 that the original PISA means for mathematics performance in the selected countries varied from 370 for Colombia to 547 for Hong Kong SAR, and the standard errors for the means were about 3.0 to 4.0 PISA points. As such, and within this context, we can consider the values reported in Table 6 to be very close to zero and therefore of no substantive importance.

In reading (Table 7), the original PISA means for the selected countries varied from 385 for Colombia to 536 for Hong Kong SAR, and the standard errors for the means in reading ranged from 2.4 (Hong Kong SAR) to 5.1 (Colombia) PISA points. Therefore, as was the case for mathematics, the differences in estimated means between the original results and the results of the alternative conditioning approaches reported in Table 7 can be considered trivial.

In science (Table 8), the original PISA means for the selected countries varied from 388 for Colombia to 542 for Hong Kong SAR, and the standard errors for the means in science ranged from 2.3 PISA points in Poland to 4.2 PISA points in the United States. Because none of the values reported in Table 8 came even close to the lower limit of the standard error of the original mean estimate, we can again consider the differences to be negligible.

In summary, the size of the reported differences between the means generated from the five rotation models and the original means indicates that essentially the same results emerged for each of the three domains.

Table 7 Differences in estimated means from the original scaling for *reading* for each of the alternative conditioning approaches

	Colombia	Germany	France	Hong Kong SAR	Jordan	Norway	Poland	Russian federation
common	0.1	0.4	0.2	1.0	−0.3	0.2	1.1	−0.3
Samecorrjoint	1.0	−2.2	−0.2	1.1	−0.6	0.6	0.6	−0.6
Samecorrsep	−2.6	−2.2	−1.6	0.6	−0.4	−0.1	−0.7	0.3
Hilocorrjoint	0.5	−2.0	−1.5	1.2	−0.8	−0.5	0.6	−0.2
Hilocorrsep	−1.8	−2.3	−0.4	1.0	−0.4	−0.5	−0.5	−0.3

Notes: Standard errors of the originally estimated means in reading range from 2 to 5 PISA points.
Due to an error in the printing of booklets, no reading estimates were available for the United States in PISA 2006.

Table 8 Differences in estimated means from the original scaling for *science* for each of the alternative conditioning approaches

	Colombia	Germany	France	Hong Kong SAR	Jordan	Norway	Poland	Russian federation	United States
Common	−1.4	−0.2	0.6	0.2	0.2	0.0	0.2	0.1	0.2
Samecorrjoint	−0.4	0.1	0.2	−0.2	0.4	0.5	0.1	0.3	0.2
Samecorrsep	−0.9	−0.3	0.5	−0.4	0.0	0.7	0.7	−0.1	0.1
Hilocorrjoint	−0.6	0.5	0.6	−1.0	0.1	0.4	0.5	0.2	0.2
Hilocorrsep	−0.6	−0.2	0.4	−0.1	−0.1	0.7	0.2	0.2	−0.3

Note: Standard errors of the originally estimated means in science range from 2 to 4 PISA points.

Table 9 shows the standard deviations for the differences in mathematics performance between each of the alternative rotation designs and those originally obtained in the PISA database. Here we can see that the original PISA standard deviations in mathematics for the selected countries varied from 84 in Jordan to 99 for Germany, and the standard errors for the standard deviations ranged from 1.2 PISA points in Poland to 2.6 PISA points in Germany. Two of the values reported for mathematics in Table 9 exceeded the upper limit of this range. Both pertained to Colombia and both related to Rotation Design 2, where the correlation between the constructs and performance was higher in one of the rotated forms than in the other form.

Table 10 shows the differences between the estimated standard deviations in reading that resulted from each of the alternative rotation designs and those originally obtained. The original PISA standard deviations in reading for the selected countries varied from 82 in Hong Kong SAR to 112 in Germany, and the standard errors for the standard deviations ranged from 1.5 PISA points in Poland to 2.8 PISA points in France. Nineteen of the values reported for reading in Table 10 exceeded the upper limit of this range.

It is noteworthy that not one of the differences in Table 10 is associated with the common part conditioning (="common") approach, which used only the variables in the common part of the questionnaire. In contrast, all differences exceeded the upper limit for the scaling approach in which the two questionnaire forms were used separately and where one variable set had high and one variable set had low correlations with performance (=hilocorrsep).

In Table 11 (science), the original PISA standard deviations for the selected countries vary from 85 in Colombia to 107 in the United States, and the standard errors for the standard deviations range from 1.1 PISA points in Poland to 2.1 PISA points in France. Seven of the values reported for science in Table 11 exceeded the upper limit of this range. All of these differences were associated with Rotation Design 2, which means

Table 9 Differences in estimated standard deviations from the original scaling for *mathematics* for each the alternative conditioning approaches

	Colombia	Germany	France	Hong Kong SAR	Jordan	Norway	Poland	Russian federation	United States
Common	0.3	−1.5	−0.5	0.3	0.4	−0.7	1.0	0.4	0.0
Samecorrjoint	0.3	−0.9	0.0	1.2	0.8	−1.1	0.7	0.1	−0.6
Samecorrsep	−0.3	−1.3	0.3	1.7	2.6	0.5	1.3	0.4	−0.3
Hilocorrjoint	3.2	−0.7	1.2	0.6	1.4	−1.0	1.2	2.2	0.3
Hilocorrsep	2.8	−1.0	1.0	0.8	1.9	0.8	1.9	2.3	0.5

Note: Standard errors of the originally estimated standard deviations in mathematics range from 1 to 3 PISA points.

Table 10 Differences in estimated standard deviations from the original scaling for *reading* for each the alternative conditioning approaches

	Colombia	Germany	France	Hong Kong SAR	Jordan	Norway	Poland	Russian federation
Common	−0.8	−2.4	−0.1	−1.6	0.7	−1.3	0.8	−1.7
Samecorrjoint	0.6	6.4	1.8	−0.8	1.2	0.5	2.0	−1.6
Samecorrsep	12.5	6.9	5.2	1.0	2.5	4.3	4.5	3.5
Hilocorrjoint	6.3	7.0	3.2	−0.6	2.8	1.7	3.5	1.2
Hilocorrsep	9.8	8.2	6.2	3.5	4.9	5.2	6.0	7.4

Notes: Standard errors of the originally estimated standard deviations in reading range from 2 to 3 PISA points.
Due to an error in the printing of booklets, no reading estimates were available for the United States in PISA 2006.

that the correlations between the variable set and performance in one of the rotated forms were consistently higher than the correlations in the other rotated form. In contrast, with respect to Rotation Design 1, where the constructs in each form had similar correlations with performance, no difference exceeded the upper limit of the standard error associated with the original estimate of the standard deviation.

Percentiles

Although we compared for all countries of interest the percentiles of the distributions of the plausible values based on the five rotation models and the original plausible values, we decided, for the sake of brevity, to report only the results for Colombia and Poland in this paper. Our reason for this choice is that the sets of results for these two countries showed the most variance. Table 12 presents the findings for Colombia, and Table 13 the findings for Poland.

Scrutiny of these tables shows that, in general, the differences between the plausible values drawn using each of the five rotation models and the original plausible values are larger in the tails of the distributions (namely the 5th and 10th percentiles at the bottom end and the 90th and 95th percentiles at the top end) than they are in the middle of the distributions. The largest absolute difference is recorded for the estimates of the 5th percentile in reading for Colombia. However, due to the relatively large standard error associated with these estimates (i.e., 5 to 11 PISA points in reading for Colombia), we can consider the differences between this and the original estimate to be immaterial.

While none of the differences between the original estimates and the estimates based on the rotated questionnaire models is of substantive importance, it is still interesting

Table 11 Differences in estimated standard deviations from the original scaling for *science* for each the alternative conditioning approaches

	Colombia	Germany	France	Hong Kong SAR	Jordan	Norway	Poland	Russian federation	United States
Common	−0.3	1.6	0.6	1.8	−0.2	−0.1	−0.2	0.1	0.0
Samecorrjoint	−0.3	0.2	−0.2	1.5	−0.5	0.3	−0.2	−0.1	−0.4
Samecorrsep	−0.7	1.7	−0.1	0.9	−0.2	−0.3	−0.1	0.3	−0.2
Hilocorrjoint	4.6	0.7	2.3	2.5	1.0	1.1	1.2	3.0	0.6
Hilocorrsep	4.4	1.4	3.1	1.6	1.7	1.7	1.5	2.6	0.6

Note: Standard errors of the originally estimated standard deviations in science range from 1 to 2 PISA points.

Table 12 Differences in estimated percentiles from the original scaling for *Colombia* for each of mathematics, reading, and science, using the alternative conditioning approaches

	5th percentile			10th percentile			25th percentile			75th percentile			90th percentile			95th percentile		
	Math.	*Read.*	*Sci.*	*Math.*	*Read.*	*Sci.*	*Math.*	*Read.*	*Sci.*	*Math.*	*Read.*	*Sci.*	*Math.*	*Read.*	*Sci.*	*Math.*	*Read.*	*Sci.*
Common	−1	−1	−2	−1	−1	−2	1	−1	−1	1	−3	−1	−1	2	−1	0	2	−3
Samecorrjoint	1	−1	0	1	−3	0	1	−1	0	0	1	−1	1	3	0	2	2	−1
Samecorrsep	1	−10	−1	2	−4	−1	2	−2	0	0	1	−2	−1	3	0	1	6	−2
Hilocorrjoint	−5	−12	−11	−3	−9	−8	−2	−4	−4	2	5	2	4	6	6	5	8	4
Hilocorrsep	−5	−17	−10	−2	−15	−7	−1	−6	−3	3	5	2	5	7	6	5	10	4

Note: Standard errors of the originally estimated percentiles for Colombia ranged from 4 to 9 PISA points for mathematics, from 5 to 11 PISA points for reading, and from 4 to 6 PISA points for science.

Table 13 Differences in estimated percentiles from the original scaling for *Poland* for each of mathematics, reading, and science, using the alternative conditioning approaches

	5th percentile			10th percentile			25th percentile			75th percentile			90th percentile			95th percentile		
	Math.	*Read.*	*Sci.*	*Math.*	*Read.*	*Sci.*	*Math.*	*Read.*	*Sci.*	*Math.*	*Read.*	*Sci.*	*Math.*	*Read.*	*Sci.*	*Math.*	*Read.*	*Sci.*
Common	0	−1	−1	−2	1	1	−2	0	2	−1	1	0	0	3	1	2	4	0
Samecorrjoint	−3	−3	−1	−3	−3	0	−2	0	0	−2	2	0	−1	3	0	1	4	−1
Samecorrsep	−2	−7	−2	−4	−6	1	−1	−1	1	−1	2	0	−1	4	0	1	8	−1
Hilocorrjoint	−2	−6	−3	−2	−5	0	−1	−1	0	−1	3	1	0	5	2	2	9	3
Hilocorrsep	−3	−11	−4	−5	−8	−1	−2	−2	−1	−2	2	1	0	7	2	2	9	2

Note: Standard errors of the originally estimated percentiles for Poland ranged from 3 to 4 PISA points in mathematics, from 3 to 6 PISA points in reading, and from 3 to 4 PISA points in science.

to examine the largest difference for each domain. In mathematics, the largest difference of five PISA points is recorded several times in Tables 12 and 13. This five-point difference can be noted in Poland for the 10th percentile between the separate conditioning using Rotation Design 2 (hilocorrsep) and the original estimate. In Colombia, the difference is apparent in both the 5th and 95th percentile estimates for both the joint conditioning and the separate conditioning using Rotation Design 2 and in the 90th percentile estimate for the separate conditioning using Rotation Design 2 (hilocorrsep).

In reading, the largest difference, 17 PISA points, emerged for the 5th percentile estimate in Colombia. This difference was the one between separate conditioning with Rotation Design 2 (hilocorrsep) and the original estimate. In science, the largest difference of 11 PISA points was again found for the 5th percentile in Colombia, but this time the difference was between joint conditioning using Rotation Design 2 (hilocorrjoint) and the original estimate.

Thus, despite none of the differences in percentiles being of substantive importance, we can detect a pattern whereby differences were somewhat larger when Rotation Design 2 was involved. As a reminder, this design involved allocating constructs that were more highly correlated with performance to one of the rotated forms of the questionnaire and assigning the constructs with lower correlations with performance to the other rotated form.

Correlations with context constructs

We calculated, for six of the nine countries under review, correlations between all 28 context constructs and the plausible values drawn using each of the five rotations. These new correlations were thus based on the reduced sets of data, that is, the variables in each of the two forms using responses from only half of the students, with the other half of the student sample set to missing. We then compared these 2,520 new correlations (6 countries × 28 constructs × 5 rotation designs × 3 domains) with the correlations with the original plausible values.

Data on only 24 of these constructs were available for France, Hong Kong SAR, and the United States. In addition, due to an error in the printing of the reading booklets, no reading proficiency estimates were available for the United States. This meant that, for these three countries, 960 correlation coefficients were calculated (2 countries × 24 constructs × 5 rotation designs × 3 domains + 1 country × 24 constructs × 5 rotation designs*2 domains) and compared to the correlation coefficient between a certain context construct and the original plausible values, resulting in a grand total of 3,480 comparisons.

Our summarizing of the resulting information involved two steps. Our intention with the first step was to find out if we could observe a general trend in terms of changes in the sizes of the coefficients between the five rotation estimates and the original estimates that used complete data on all variables. Our aim during the second step was to conduct a review at the construct level in order to identify possible patterns indicating where changes might have occurred.

During the first step, we calculated the mean of the correlations across the constructs for each domain and each country. Next, we computed the ratio of the mean correlations for the five rotation designs to the mean correlation without rotation. We then

Table 14 Correlations of context constructs with proficiencies for original estimates and estimates from five rotation designs

Country		Mathematics						Reading						Science					
		Original	*Common*	*Same-corrjoint*	*Same-corrsep*	*Hilo-corrjoint*	*Hilo-corrsep*	*Original*	*Common*	*Same-corrjoint*	*Same-corrsep*	*Hilo-corrjoint*	*Hilo-corrsep*	*Orignal*	*Common*	*Same-corrjoint*	*Same-corrsep*	*Hilo-corrjoint*	*Hilo-corrsep*
COL	Mean *r*	0.06	0.06	0.06	0.06	0.06	0.06	0.05	0.05	0.06	0.04	0.06	0.05	0.07	0.07	0.07	0.07	0.07	0.07
	Ratio to original	1.00	0.99	0.92	0.91	1.06	1.04	1.00	1.06	1.22	0.93	1.27	1.11	1.00	1.05	1.01	1.03	0.96	1.01
DEU	Mean *r*	0.16	0.16	0.15	0.15	0.16	0.15	0.12	0.13	0.12	0.11	0.13	0.11	0.16	0.16	0.16	0.16	0.16	0.16
	Ratio to original	1.00	1.00	0.97	0.97	1.01	0.99	1.00	1.08	0.99	0.97	1.07	0.96	1.00	0.99	0.98	0.99	1.00	1.00
FRA	Mean *r*	0.21	0.21	0.21	0.20	0.22	0.21	0.18	0.18	0.18	0.17	0.17	0.17	0.22	0.22	0.22	0.22	0.22	0.22
	Ratio to original	1.00	0.99	0.98	0.96	1.02	0.99	1.00	1.02	0.98	0.97	0.96	0.94	1.00	1.00	1.01	0.99	1.00	1.00
HKG	Mean *r*	0.15	0.15	0.15	0.15	0.15	0.15	0.11	0.12	0.11	0.11	0.12	0.12	0.17	0.17	0.16	0.17	0.17	0.17
	Ratio to original	1.00	1.03	1.00	1.03	1.00	1.00	1.00	1.04	1.02	1.02	1.05	1.06	1.00	0.99	0.99	1.01	0.99	0.99
JOR	Mean *r*	0.12	0.12	0.11	0.11	0.11	0.12	0.12	0.11	0.12	0.12	0.11	0.11	0.12	0.13	0.12	0.12	0.13	0.12
	Ratio to original	1.00	0.99	0.97	0.92	0.97	0.99	1.00	0.94	1.00	0.97	0.91	0.92	1.00	1.04	0.98	1.01	1.03	1.00
NOR	Mean *r*	0.13	0.14	0.13	0.13	0.14	0.13	0.11	0.11	0.10	0.11	0.11	0.11	0.15	0.15	0.15	0.15	0.14	0.15
	Ratio to original	1.00	1.08	1.04	1.02	1.05	1.03	1.00	1.06	0.98	0.98	1.03	1.07	1.00	1.02	0.99	1.00	0.98	1.03
POL	Mean *r*	0.11	0.10	0.10	0.10	0.10	0.11	0.09	0.09	0.09	0.09	0.08	0.09	0.12	0.12	0.11	0.12	0.11	0.12
	Ratio to original	1.00	0.92	0.93	0.97	0.93	1.00	1.00	1.01	0.97	1.01	0.93	1.02	1.00	0.99	0.96	0.98	0.95	1.00

Table 14 Correlations of context constructs with proficiencies for original estimates and estimates from five rotation designs *(Continued)*

RUS	Mean *r*	0.07	0.07	0.07	0.07	0.07	0.07	0.06	0.06	0.07	0.06	0.06	0.07	0.07	0.07	0.07	0.07	0.07
	Ratio to original	1.00	1.01	1.00	0.97	1.01	0.98	1.00	1.09	1.13	1.04	1.07	1.10	1.00	1.06	1.07	1.06	1.07	1.04
USA	Mean *r*	0.15	0.15	0.15	0.15	0.15	0.15	*	*	*	*	*	*	0.17	0.17	0.17	0.17	0.17	0.17
	Ratio to original	1.00	0.97	0.99	0.98	0.98	1.01	*	*	*	*	*	*	1.00	1.00	1.02	1.02	1.00	1.03
Grand mean ratio			**1.00**	**0.98**	**0.97**	**1.00**	**1.00**	**1.04**	**1.00**	**1.04**	**0.99**	**1.04**	**1.02**	**1.00**	**1.00**	**1.02**	**1.01**	**1.00**	**1.01**

Note: Due to an error in the printing of booklets, no reading estimates were available for the United States in PISA 2006.

averaged the ratios over countries to obtain a grand mean ratio. Table 14 presents the results.

The table has three main sections mathematics, reading, and science. The first column of each section gives the original correlation across all context constructs with mathematics proficiency, and the next five columns give the correlations across all context constructs for each of the five rotation designs:

- The common part (common);
- The *joint* scaling of the variable sets that had the same correlation with achievement (samecorrjoint);
- The *separate* scaling of the variable sets that had the same correlation with achievement (samecorrsep);
- The *joint* scaling of the variable sets in which one had a high and the other a low correlation with performance (hilojoint); and
- The *separate* scaling of the variable sets in which one had a high and the other a low correlation with performance (hilosep).

If, for example, we look at Colombia in Table 14, it is apparent that the mean correlation across all of the context constructs with mathematics is the same for the original as well as for all five rotation designs, namely 0.06. The same applies in science, where the mean correlation across all context constructs and performance is 0.07, regardless of the rotation design. Slight differences only are apparent in reading, where the original average correlation between all background constructs and performance is 0.05, but is lower (0.04) for the design with the separate conditioning of variable sets with the same correlation (samecorrsep) and higher (0.06) for both of the designs that involved the use of joint conditioning. However, the size of these differences is not substantive.

The second row for each country in Table 14 sets out the differences in the form of ratios between the original correlation and the correlation estimate based on the five rotated designs. In many instances, the ratios are 0.99, 1.00, or 1.01, indicating very little differences between the estimates.

In combination, these results reveal no pattern of upward or downward change between the correlation estimates from the rotation designs when compared with the original correlation estimates across the very different countries in the analyses. Thus, for example, the differences were no more pronounced in a country with relatively higher correlations between the context constructs and performance, such as France, than they were in a country with lower correlations, such as Colombia.

During the second results-summarization step (taken with the aim of reviewing the results at the construct level), we recorded only the 82 instances of the 3,480 correlations where the absolute differences between the correlation coefficients exceeded 0.03. This decision was based on the fact that such a difference would exceed the standard errors of the corresponding estimates, which, in PISA, are usually less than 0.02. Details concerning the differences that emerged appear in Table 15 for mathematics, Table 16 for reading, and Table 17 for science.

As can be seen, the number of correlation coefficients with context constructs exhibiting differences between the rotation results and the original results varies across

Table 15 Summary of differences in correlations with context constructs: mathematics

Country	Construct	Rotation design	Corr. with rotation result	SEr	Corr. with original result	SEo
Colombia	ENVPERC	Samecorrjoint	0.13	0.02	0.16	0.03
	GENSCIE	Hilocorrsep	0.07	0.02	0.11	0.02
	INTSCIE	Hilocorrjoint	−0.08	0.02	−0.12	0.02
		Common	−0.08	0.02	−0.12	0.02
	SCIEEFF	Hilocorrjoint	0.17	0.02	0.21	0.02
France	ENVOPT	Hilocorrsep	−0.27	0.02	−0.30	0.02
	INTSCIE	Hilocorrjoint	0.30	0.01	0.33	0.01
	SCINTACT	Hilocorrjoint	−0.15	0.02	−0.19	0.02
		Hilocorrsep	−0.16	0.02	−0.19	0.02
Germany	None					
Hong Kong SAR	GENSCIE	Samecorrsep	**0.20**	0.02	0.17	0.02
Jordan	JOYSCIE	Samecorrsep	0.18	0.02	0.21	0.02
		Hilocorrjoint	0.18	0.02	0.21	0.02
	SCIEEFF	Hilocorrjoint	0.16	0.02	0.19	0.02
Norway	SCINVEST	Hilocorrsep	−0.26	0.02	−0.29	0.01
Poland	JOYSCIE	Samecorrjoint	0.12	0.01	0.15	0.02
	SCIEACT	Common	0.00	0.01	0.03	0.01
	SCSCIE	Samecorrjoint	0.18	0.01	0.22	0.02
Russian	ENVAWARE	Hilocorrsep	0.29	0.02	0.32	0.02
Federation	SCSCIE	Hilocorrjoint	0.09	0.02	0.12	0.02
United States	None					

Notes:
Only differences between correlation coefficients exceeding 0.03 are reported.
SEr: Standard error of the correlation between the construct and the rotation result.
SEo: Standard error of the correlation between the construct and the original result.
Bolded cell: Correlation coefficient between construct and rotation result is higher than the correlation between construct and original result.

the three subject domains. The smallest number of differences are recorded for mathematics—a minor domain in PISA 2006. Here, only 19 differences are larger than 0.03. For science, 28 differences exceed that size. The domain recording the most differences is reading. The only sizeable difference for a country is that for Hong Kong SAR. Across the countries, differences are apparent for between 4 (Russian Federation) and 13 constructs (Colombia), with a total of 63 differences exceeding 0.03 in reading.

In order to investigate whether any of the constructs or rotation results was more prone than others to being involved in the differences, we summed their occurrences across countries and domains. Awareness of environmental issues (ENVAWARE) was the construct for which most of the differences (15) were recorded. Results for the rotation plausible values based on the common part conditioning showed the smallest number of differences (i.e., five) compared with the original plausible values. In contrast, the largest number of differences involved the rotation results for the 41 occurrences recorded for separate conditioning using Rotation Design 2 and the 35 occurrences recorded for joint conditioning using Rotation Design 2. Hence, the rotation results based on the design in which one form contained constructs that were more highly correlated with performance and in which the other form contained

Table 16 Summary of differences in correlations with context constructs: reading

Country	Construct	Rotation result	Corr. with rotation result	SEr	Corr. with original result	SEo
Colombia	CARINFO	Samecorrjoint	−0.05	0.02	−0.08	0.02
	CARPREP	Samecorrsep	−0.12	0.02	−0.15	0.02
	ENVOPT	Samecorrjoint	−0.29	0.02	−0.32	0.02
		Samecorrsep	−0.27	0.03	−0.32	0.02
	ENVPERC	Hilocorrjoint	0.12	0.03	0.15	0.03
	HEDRES	Samecorrsep	0.30	0.04	0.33	0.02
		Hilocorrsep	0.30	0.02	0.33	0.02
	HOMEPOS	Samecorrsep	0.30	0.04	0.34	0.02
		Hilocorrsep	0.31	0.02	0.34	0.02
	INTCONF	Samecorrsep	0.27	0.04	0.30	0.02
		Hilocorrjoint	0.27	0.02	0.30	0.02
	INTUSE	Samecorrsep	0.04	0.04	0.07	0.02
	PRGUSE	Samecorrjoint	−0.03	0.02	−0.06	0.02
		Common	−0.03	0.02	−0.06	0.02
	SCIEEFF	Hilocorrjoint	0.14	0.02	0.18	0.02
	SCIEFUT	Samecorrjoint	−0.14	0.02	−0.17	0.02
		Samecorrsep	−0.14	0.03	−0.17	0.02
		Hilocorrsep	−0.14	0.02	−0.17	0.02
	SCINTACT	Common	0.02	0.03	0.05	0.03
	WEALTH	Samecorrsep	0.27	0.03	0.30	0.02
France	ENVAWARE	Samecorrsep	0.39	0.02	0.42	0.02
		Hilocorrsep	0.38	0.02	0.42	0.02
	INTSCIE	Samecorrsep	0.26	0.02	0.30	0.01
		Hilocorrsep	0.27	0.02	0.30	0.01
	SCIEACT	Samecorrjoint	0.18	0.02	0.21	0.02
	SCINTACT	Hilocorrjoint	−0.15	0.02	−0.19	0.02
	SCINVEST	Samecorrsep	−0.22	0.02	−0.25	0.02
		Hilocorrjoint	−0.22	0.02	−0.25	0.02
		Hilocorrsep	−0.22	0.02	−0.25	0.02

constructs that correlated less with achievement seemed more likely than the other three rotation results to differ from the original results.

With three exceptions, the correlation coefficients were higher between the constructs and the original results than between the constructs and the rotation results. The exceptions included the construct measuring the general value of science (GENSCIE) and its correlation with mathematics performance in Hong Kong SAR and for the correlations between that same construct and the two different reading proficiency estimates in Germany.

A final finding was that the estimated correlations between the context constructs and performance tended to be smaller for the plausible values that were generated from the five rotation models compared with those generated for the original plausible values.

Table 17 Summary of differences in correlations with context constructs: science

Country	Construct	Rotation result	Corr. with rotation result	SEr	Corr. with original result	SEo
Colombia	CARPREP	Hilocorrjoint	−0.10	0.02	−0.13	0.02
	ENVAWARE	Hilocorrjoint	0.30	0.03	0.34	0.03
		Hilocorrsep	0.30	0.02	0.34	0.03
	ENVOPT	Samecorrsep	−0.35	0.02	−0.38	0.02
		Hilocorrjoint	−0.33	0.02	−0.38	0.02
		Hilocorrsep	−0.34	0.02	−0.38	0.02
	ENVPERC	Hilocorrjoint	0.17	0.02	0.20	0.02
		Hilocorrsep	0.17	0.02	0.20	0.02
	HEDRES	Hilocorrjoint	0.30	0.03	0.34	0.03
		Hilocorrsep	0.29	0.02	0.34	0.03
	HOMEPOS	Hilocorrjoint	0.31	0.03	0.34	0.03
		Hilocorrsep	0.31	0.03	0.34	0.03
	SCINVEST	Hilocorrjoint	−0.15	0.02	−0.18	0.02
		Hilocorrsep	−0.15	0.02	−0.18	0.02
France	ENVAWARE	Hilocorrjoint	0.46	0.01	0.49	0.01
		Hilocorrsep	0.46	0.01	0.49	0.01
Germany	None					
Hong Kong SAR	None					
Jordan	ENVAWARE	Hilocorrjoint	0.40	0.02	0.43	0.01
		Hilocorrsep	0.40	0.01	0.43	0.01
Norway	ENVAWARE	Hilocorrjoint	0.38	0.02	0.41	0.02
		Hilocorrsep	0.38	0.02	0.41	0.02
	SCIEEFF	Hilocorrjoint	0.35	0.01	0.38	0.01
	SCSCIE	Hilocorrjoint	0.34	0.01	0.37	0.01
		Hilocorrsep	0.34	0.01	0.37	0.01
Poland	GENSCIE	Hilocorrjoint	0.26	0.01	0.29	0.01
	SCIEEFF	Hilocorrjoint	0.42	0.01	0.45	0.01
		Hilocorrsep	0.42	0.01	0.45	0.01
Russian	CARINFO	Hilocorrjoint	−0.01	0.01	−0.04	0.01
Federation	ENVAWARE	Hilocorrsep	0.37	0.02	0.40	0.02
United States	None					

Notes: Only differences between correlation coefficients exceeding 0.03 are reported.
SEr: Standard error of the correlation between the construct and the rotation result.
SEo: Standard error of the correlation between the construct and the original result.

Discussion and conclusions

As Rutkowski (2011) notes, despite the fact that latent regression is well established both theoretically and practically as an analytic approach in sample surveys, there is a dearth of literature concerning various implications of and threats to the application of this methodology. This paper has added to that literature by exploring some of the implications of using rotated contextual questionnaires for respondents so as to expand the coverage of contextual variables while still placing a reasonable limitation on respondent time.

Our modeling, using PISA 2006 data, of the potential impact of the use of rotated questionnaires revealed very similar results regardless of whether we scaled the data

using rotated context questionnaires or nonrotated questionnaires. Indeed, differences in terms of estimated means, standard deviations, and percentiles tended to be slight and were therefore nearly all of no substantive importance. Likewise, our comparison of mean correlations across all context constructs with performance between rotation and original results and the corresponding ratios of differences revealed no general upward or downward trend in estimates.

Our analyses furthermore revealed very few substantive differences between the plausible values generated from the rotation models and the original plausible values. This outcome leads to the following conclusions with respect to (a) the possibility of developing a methodology using rotated context questionnaires, (b) the possible impact on the continuity of results, and (c) correlations between context variables and performance.

First, the research shows that it is possible to develop a methodology that uses rotated contextual questionnaires in conjunction with multilevel item response models. The three approaches to scaling that we explored in this paper involved common part conditioning, joint conditioning, and separate conditioning. The common part conditioning used information from only the variables in the common part of the questionnaire, the joint conditioning employed information from the rotated parts jointly, and the separate conditioning used information from the rotated parts separately. We paired these latter two approaches with the two questionnaire rotation designs. In Design 1, the constructs in each rotated form showed similar correlations with performance. In Design 2, the constructs were assigned to forms in such a way that the constructs in one form had relatively higher correlations with performance whereas the constructs in the other form had relatively lower correlations with performance.

Second, the comparison of the results from these five rotation models and the original results showed little, if any, impact on the continuity of PISA results in terms of means, standard deviations, and percentiles. This meant that we found no substantive differences between estimates of the mean based on the original plausible values and estimates of the mean based on the plausible values obtained from the five models using a rotated student-context-questionnaire design in mathematics, reading, or science.

We found a number of relatively robust differences between the standard deviations based on original plausible values and those based on plausible values generated from the questionnaire rotation models. The large majority of these instances emerged in the minor domain of reading and with respect to Rotation Design 2, in which we assigned constructs to forms in such a way that the constructs in one form had relatively higher correlations with performance than the constructs in the other form. Our comparison of the percentiles of the distributions of the plausible values based on the five rotation models and the original plausible values revealed no substantive differences in any of the domains.

Third, our comparison of the estimated correlation coefficients between the context variables and the five rotation plausible values on the one hand and the original plausible values on the other hand revealed some nontrivial differences, ranging from 19 differences in science to 63 differences in reading. Most of these differences were associated with the separate conditioning approach used in conjunction with Design 2 (hilocorrsep). This evidence, combined with the results for the standard deviations, suggests a preference for the Design 1 approach, where constructs are assigned to rotated forms in such a way that their correlations with performance are similar across forms.

There might have been an expectation that excluding a variable from the conditioning model would bias the subsequent estimates of the correlation between that variable and outcomes toward zero, given that the bias is a function of the marginal explanatory power of that variable (see, for example, Mislevy 1991). However, we did not exclude variables from the conditioning during our current analyses and we did not, of course, neglect to carry out conditioning. Indeed, all of the rotation designs that we examined involved some form of conditioning.

Of note with regard to the four designs in which we used responses from only half of the respondents is the fact that we conducted the conditioning using the data from this half and then set the other half to missing, thereby ignoring this information during the analyses. In other words, these four designs excluded from the analyses students who had not responded to the questions, which meant that information relating to them was absent from the conditioning. In this sense, the rotation designs paralleled the original analyses, which included all students who had responded to these questions and included all of this information in the conditioning. In this respect, the only difference between these analyses and ours is that our estimates were based on a smaller number of cases, namely half, which meant that any reduction in the size of estimates would only be a consequence of the smaller number of cases available in the analyses.

Another conclusion that can be drawn from our results is that additional information obtained from the context questionnaire adds very little to the estimation of the latent variable, and consequently has a negligible influence on the plausible values. This, of course, is not surprising because nearly all of the information concerning the latent variables for each respondent came from the two hours of cognitive testing in the content domains during which the PISA latent scales were measured with high reliability.

Indeed, the robustness of the results from our scaling approach that used the common part of the context questionnaires indicates that—for the purpose of obtaining plausible values—the questions in those questionnaires could be reduced to a core set, such as gender, parental education and occupation, migration, home language, and home possessions. However, in PISA, the contextual variables are not included in the scaling as a means of improving the reliability of the plausible values. Rather, they are included so that the contextual factors possibly associated with performance can be subsequently analyzed. Importantly, what we are showing here is the potential that a rotated design has to broaden the range of contextual variables included, and therefore increase the relevance of the assessment for policymakers, educators, and researchers, while simultaneously allowing the respondent time to be kept to approximately 30 minutes a length consistent with most current large-scale assessments.

In terms of which particular rotation design might be preferable, our findings indicate that the population parameter estimates that are based on the rotated forms (i.e., one form containing constructs that are more highly correlated with achievement and the other form containing constructs that correlate less with achievement) are more prone to differ from the population parameter estimates based on the original plausible values than on the population parameter estimates based on plausible values generated using the other rotation models considered in this paper. Thus, it would seem desirable to assign constructs to forms in a way that means the constructs in each rotated form

correlate similarly with performance. This approach could be achieved by basing the assignment of constructs to forms on the results obtained from field trials.

Finally, while further work using other datasets and other types of analyses seem desirable to provide further evidence, the outcomes of our research support using rotated contextual questionnaires for respondents in order to extend the methodology currently used in large-scale sample surveys. We consider such an extension presents good news for researchers and respondents alike, because it would permit a broadening of coverage, a reduction in response time, or both.

Endnotes

[a]Examples include the Organisation for Economic Co-operation and Development's (OECD) Programme for International Student Assessment (PISA), the International Association for the Evaluation of Educational Achievement's (IEA) Trends in International Mathematics and Science Study (TIMSS), and the US National Assessment of Educational Progress (NAEP).

[b]Such tools are, however, available in the public domain and fully described in the literature (Fox & Glas 2002; Sinharay & von Davier 2005; Volodin & Adams 1997; Wu, Adams, & Wilson 1997).

Competing interests
The authors declare that they have no competing interests.

Authors' contributions
RJA, AB and PL designed the analyses, carried them out and prepared the manuscript. All authors read and approved the final manuscript.

Author details
[1]University of Melbourne and Australian Council for Educational Research, Melbourne, Australia. [2]Australian Council for Educational Research, Melbourne, Australia.

Received: 28 August 2013 Accepted: 3 September 2013
Published: 16 September 2013

References

Adams, RJ. (2002). Scaling PISA cognitive data. In RJ Adams & ML Wu (Eds.), *PISA 2000 technical report*. Paris, France: OECD Publications.

Adams, RJ, Wilson, MR, & Wang, WC. (1997a). The multidimensional random coefficients multinomial logit model. *Appl Psychol Meas, 21*, 1–23.

Adams, RJ, Wilson, MR, & Wu, ML. (1997b). Multilevel item response modelling: An approach to errors in variables regression. *J Educ Behav Stat, 22*, 47–76.

Adams, RJ, & Wu, ML. (2007). The mixed-coefficient multinomial logit model: A generalized form of the Rasch model. In M von Davier & CH Carstensen (Eds.), *Multivariate and mixture distribution Rasch models: Extensions and applications* (pp. 57–76). New York, NY: Springer.

Adams, RJ, Wu, ML, & Carstensen, CH. (2007). Application of multivariate Rasch models in international large scale educational assessment. In M von Davier & CH Carstensen (Eds.), *Multivariate and mixture distribution Rasch models: Extensions and applications* (pp. 271–280). New York, NY: Springer.

Adams, RJ, Wu, ML, & Macaskill, G. (1997c). Scaling methodology and procedures for the mathematics and science scales. In MO Martin & DL Kelly (Eds.), *TIMSS technical report* (Implementation and analysis, Vol. II, pp. 111–145). Chestnut Hill, MA: Boston College.

Anderson, TW. (1984). *An introduction to Multivariate statistical analysis*. New York, NY: John Wiley & Sons.

Beaton, AE. (1987). *Implementing the new design: The NAEP 1983–84 technical report (Report No. 15-TR-20)*. Princeton, NJ: Educational Testing Service.

Beaton, AE, & Gonzalez, EJ. (1995). *The NAEP primer. Center for the Study of Testing, Evaluation, and Educational Policy*. Chestnut Hill, MA: Boston College.

Cochran, WG. (1968). Errors of measurement in statistics. *Technometrics, 10*, 637–666.

Fox, J-P, & Glas, CAW. (2002). Modeling measurement error in a structural multilevel model. In GA Marcoulides & I Moustaki (Eds.), *Latent variable and latent structure models* (pp. 245–269). London, UK: Lawrence Erlbaum Associates.

Fuller, WA. (1987). *Measurement error models*. New York, NY: John Wiley & Sons.

Gleser, LJ. (1981). Estimation in a multivariate errors-in-variables regression model: Large sample results. *Ann Stat, 9*, 24–44.

Gonzalez, JM, & Eltinge, JL. (2007a). Mulitple matrix sampling: A review. In *Proceedings of the Section on Survey Research Methods, American Statistical Association* (pp. 3069–3075). Alexandria, VA: American Statistical Association.

Gonzalez, JM, & Eltinge, JL. (2007b). *Properties of alternative sample design and estimation methods for the consumer expenditure surveys*. Arlington, VA: Paper presented at the 2007 Research Conference of the Federal Committee on Statistical Methodology.

Gonzalez, EJ, Galia, J, & Li, I. (2004). Scaling methods and procedures for the TIMSS 2003 mathematics and science scales. In MO Martin, IVS Mullis, & SJ Chrostowski (Eds.), *TIMSS 2003 technical report*. Chestnut Hill, MA: Boston College.

Jones, M. (1996). Indicator and stratification methods for missing explanatory variables in multiple linear regression. *J Am Stat Assoc, 91*(433), 222–230.

Jöreskog, KG, & Sörbom, D. (1984). *LISREL VI: Analysis of linear structural relationshipos by maximum likelihood, instrumental variables and least squares methods*. Mooresville, IN: Scientific Software.

Macaskill, G, Adams, RJ, & Wu, ML. (1998). Scaling methodology and procedures for the mathematics and science literacy, advanced mathematics and physics scales. In M Martin & DL Kelly (Eds.), *Third International Mathematics and Science Study, technical report: Vol. 3. Implementation and analysis*. Chestnut Hill, MA: Boston College.

Mislevy, RJ. (1985). Estimation of latent group effects. *J Am Stat Assoc, 80*, 993–997.

Mislevy, RJ. (1990). Scaling procedures. In EG Johnson & R Zwick (Eds.), *Focusing the new design: The NAEP 1988 technical report (No. 19-TR-20, pp. 229–250)*. Princeton, NJ: Educational Testing Service.

Mislevy, RJ. (1991). Randomization-based inference about latent variables from complex samples. *Psychometrika, 56*(2), 177–196.

Muthén, BO. (2002). Beyond SEM: General latent variable modelling. *Behaviormetrika, 29*(1), 81–117.

Organisation for Economic Co-Operation and Development (OECD). (2001). *Knowledge and skills for life: First results from PISA 2000*. Paris, France: OECD Publications.

Organisation for Economic Co-Operation and Development (OECD). (2004). *Learning for tomorrow's world: First results from PISA 2003*. Paris, France: OECD Publications.

Organisation for Economic Co-Operation and Development (OECD). (2007a). *PISA 2006: Science competencies for tomorrow's world*. Paris, France: OECD Publications.

Organisation for Economic Co-Operation and Development (OECD). (2007b). *Database PISA 2006*. Available online at http://pisa2006.acer.edu.au/downloads.php.

Organisation for Economic Co-Operation and Development (OECD). (2008). *PISA 2006: Technical report*. Paris, France: OECD Publications.

Organisation for Economic Co-Operation and Development (OECD). (2010). *PISA 2009 results: What students know and can do* (Vol. 1). Paris, France: OECD Publications.

Pedhazur, EJ. (1997). *Multiple regression in behavioral research* (3rd ed.). Orlando, FL: Harcourt Brace.

Rubin, DB. (1976). Inference and missing data. *Biometrika, 63*(3), 581–592.

Rubin, DB. (1987). *Multiple imputation for nonresponse in surveys*. New York, NY: John Wiley & Sons.

Rutkowski, L. (2011). The impact of missing background data on sub-population estimation. *J Educ Meas, 48*(3), 293–312.

Shoemaker, DM. (1973). *Principles and procedures of multiple matrix sampling*. Cambridge, MA: Ballinger Publishing Company.

Sinharay, S, & von Davier, M. (2005). *Extension of the NAEP BGROUP program to higher dimensions (RR-05-27)*. Princeton, NJ: Educational Testing Service.

Volodin, N, & Adams, RJ. (1997). *The estimation of polytomous item response models with many dimensions. Paper presented at the Annual Meeting of the Psychometric Society*. TN: Gatlinburg.

Wu, ML, Adams, RJ, & Wilson, MR. (1997). *ConQuest: Multi-aspect test software [Computer program]*. Camberwell, VIC, Australia: Australian Council for Educational Research.

doi:10.1186/2196-0739-1-5
Cite this article as: Adams *et al.*: On the use of rotated context questionnaires in conjunction with multilevel item response models. *Large-scale Assessments in Education* 2013 **1**:5.

Submit your manuscript to a SpringerOpen journal and benefit from:

- ► Convenient online submission
- ► Rigorous peer review
- ► Immediate publication on acceptance
- ► Open access: articles freely available online
- ► High visibility within the field
- ► Retaining the copyright to your article

Submit your next manuscript at ► springeropen.com