Large-scale Assessments in Education
a SpringerOpen Journal

RESEARCH Open Access

Considerations for correlation analysis using clustered data: working with the teacher education and development study in mathematics (TEDS-M) and other international studies

Sabine Meinck[1*] and Michael C Rodriguez[2]

* Correspondence:
Sabine.Meinck@iea-dpc.de
[1]IEA Data Processing and Research Center, Hamburg, Germany
Full list of author information is available at the end of the article

Abstract

The Teacher Education and Development Study in Mathematics (TEDS-M) of 2008 focused on how teachers are prepared to teach mathematics in primary and lower-secondary schools in 17 countries. The main results were published in 2012, and the associated public-use database provides a valuable source for secondary analysis of the collected data. The data originate from complex samples and present a hierarchical structure. With future teachers embedded in programs embedded in institutions, various types of cluster effects can be observed. Complex methods, including the use of sampling weights and replication methods for variance estimation, are therefore required for data analysis. This paper focuses on the aspects that need to be considered during any exploration of relationships between variables. Correlation analysis may produce misleading results if attention is not paid to the structure under which the data were collected. We illustrate our points with exemplary analysis of TEDS-M data and propose some guidelines to address the issue.

Keywords: Large-scale assessments; International comparative study; Higher education; Teacher education; Methodology; Correlation analysis; Weights; Variance estimation

Introduction

The Teacher Education and Development Study in Mathematics (TEDS-M) 2008 was the first study of post-secondary education conducted by the International Association for the Evaluation of Educational Achievement (IEA). Seventeen countries, reflecting a variety of teacher education systems around the globe, participated in the study, which focused on how teachers are prepared to teach mathematics in primary and lower-secondary schools. Major study results were published by Tatto et al. (2012). The public-use database, accessible free of charge from <http://rms.iea-dpc.org/>, provides a valuable basis for secondary analysis of the collected data.

TEDS-M relied on nationally representative samples. Because it targeted different populations, the TEDS-M research team developed a multipurpose international

sampling plan and adapted it to the specific circumstances of the participating countries. One important feature of the sampling plan was that either complex cluster samples or censuses of individuals, still belonging to clusters, were surveyed. But what were these clusters? For the purposes of TEDS-M, clusters were defined as "programs." This concept plays a key role in the organization of teacher preparation and was common to all participating countries. A program is a specific pathway of teacher education that exists within an institution, requires students to undertake a set of subjects and experiences, and leads to the award of a common credential or credentials on completion (Tatto et al., 2008). During TEDS-M, teacher preparation institutions, including all programs they were providing at the time, were selected in a first step. In a second step, future teachers within these programs were selected. The structure of the final datasets pertaining to future teachers thus reflected a two-level hierarchy, with future teachers nested within their respective programs[a].

The TEDS-M design enables researchers to investigate important questions about teacher-preparation practices and outcomes from the collected data, but it also poses a challenge to analysts. The use of complex cluster samples and the cluster nature of the data in this study have various implications for subsequent data analyses. Many other large-scale assessments, such as IEA's Trends in International Mathematics and Science Study (TIMSS), IEA's Progress in International Reading Literacy Study (PIRLS), and the OECD's Programme for International Student Assessment (PISA), allow for the use of hierarchical linear modeling (HLM) as a powerful tool for disentangling effects at various levels of hierarchical data. However, this analytic method cannot be recommended for TEDS-M data because the preconditions for HLM in terms of sample and cluster sizes were not met in most (if not all) of the participating countries[b]—a point we will return to in the following sections. Our main aims in this paper are to describe how to adequately address the hierarchical data structure of TEDS-M when analyzing and interpreting results and to highlight the particular issues that need to be considered, given the structure of the TEDS-M data, when performing correlation analysis of this information.

Correlation analyses are appropriate for questions about associations among variables. Many questions of this type can be answered using TEDS-M data, such as those concerning associations between characteristics of teacher-preparation programs and the outcomes of those programs. These questions become particularly interesting when the program outcomes are measured at the student (future teacher) level. For example, what is the association between programs that focus more or less on pedagogical practice and future teacher mathematics pedagogical knowledge? We pose similar questions below to illustrate the ideas regarding appropriate structuring of the data for correlational analyses with hierarchical data.

In addition, we recognize that most researchers are interested in more complex questions, typically involving regression models that consider multiple explanatory variables simultaneously. The issues regarding clustered data in correlation analyses are presented here because of the simplicity of the correlational models and our ability to illustrate the effects clearly. All of the concerns and issues raised in this paper apply equally to regression-based analyses, which are correlational in themselves.

Correlation analysis and organizational research

The problems that may arise when performing correlation analysis in research pertaining to individuals in groups have been frequently addressed. As early as 1950, Robinson (1950) demonstrated in a classic paper known to most sociologists that the risk of misinterpretation is very large if one infers individual characteristics from statistics based upon aggregated data, an effect known as the "ecological fallacy." Many authors have subsequently contributed to the topic (e.g., Galtung, 1967; Knapp, 1977) by further pointing out that inferences about the nature of group-level relationships cannot be made from individual-level statistical data. Cronbach (1976) argued strongly for carrying out both within-group and between-group analyses rather than paying attention to the total group level, given that correlation coefficients at that level are always confounded by between- and within-group effects.

Later generations of researchers (among them, Klein & Kozlowski, 2000; Mossholder & Bedeian, 1983; Van Mierlo et al. 2009) have explored the conditions under which the composition of group-level constructs from individual-level survey data are possible. As Mossholder and Bedeian (1983, p. 548) so aptly put it, "the use of aggregated measures is neither good nor bad." The two authors also emphasized the importance of a sound rationale for interpreting individual measures as functional surrogates of macro-constructs.

We can illustrate the source of potential misinterpretation of correlation coefficients when analyzing large-scale assessment data by drawing on data collected from the TIMSS 2007 Grade 8 sample of students. To gain an understanding of eighth-graders' views about the utility of mathematics and their enjoyment of it as a school subject, TIMSS created an index of positive attitudes toward mathematics (Mullis et al. 2008). In Figure 1, we display the relationship between this index variable and the average mathematics achievement in all participating TIMSS 2007 countries, using data aggregated at the country level.

A clear positive relationship can be seen in the figure, confirmed by the large[c] positive correlation coefficient of 0.7***[d]. However, on considering the coding of the

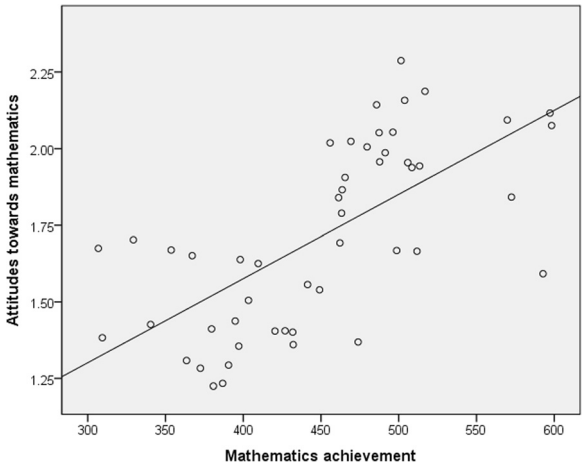

Figure 1 Example TIMSS 2007, Grade 8: Mathematics achievement scores and index on attitudes towards mathematics aggregated at country level.

index (low values represent highly positive attitudes toward mathematics, while high values point to negative attitudes), one could infer that students who like mathematics do not perform well in this subject and vice versa. This inference is actually incorrect as can be seen if we look at the correlation coefficients calculated separately within each country and using individual students' data. In fact, the correlation coefficients are negative in all participating countries, although the coefficients are often small (-0.2 on average over all countries). Consequently, the TIMSS investigators pointed to a positive association between personal attitudes toward mathematics and mathematics achievement (Mullis et al., 2008).

An earlier TIMSS report also refers to the reverse phenomenon (which we can see on looking across the participating countries in Figure 1), wherein a demanding mathematics curriculum may lead to high achievement but little enthusiasm for the subject matter (Mullis et al., 2001). Of course, there are other considerations regarding response styles, particularly those that are culture (country) specific, as pointed out by a reviewer of this paper (see Buckley, 2009, for an example exploration of this issue in the PISA program).

In order to expand our example, let us look more closely at the data for Romania as an exemplary TIMSS 2007 country (i.e., with an average correlation coefficient of -0.2***). If we consider the specific sampling design of TIMSS, we observe that entire classrooms of Grade 8 students rather than individual students were sampled (Joncas, 2008), a practice that we consider can influence the result of our correlation analysis. The reason why we think this is that students who take their mathematics courses together are exposed to the same teachers, teaching methods, and environments. They may therefore share, based on this experience, more similar attitudes on the subject, as well as more similar achievement levels[e]. A correlation analysis conducted with individual-level data is hence actually a mixture of pure individual-level and class-level effects—as far as such effects exist. Therefore, we may gain information when disentangling the two effects.

Investigators interested in the correlation between the two variables at the individual level ("within clusters" effects) while controlling for the group effect can perform group-mean centering. This technique is frequently used in multilevel modeling (see, for example, Kreft & De Leeuw, 1998; Raudenbush & Bryk, 2002). It involves subtracting the respective group means from each individual value for the variables of interest, which results in the group effect (classroom mean) being removed and the individuals set onto the same scale, as shown in Figure 2 for the Romanian data. Calculating the correlation coefficient from the group-mean-centered data gives us the relationship between the two variables at the individual level with the group effect removed, which is -0.3***.

For the analysis of the group-level effect (the "between-clusters" effect), we can use data aggregated at the group level (classrooms), but this time utilizing the average scores of students for attitudes toward mathematics and mathematics achievement as group-level estimates. As is evident in Figure 3, no relationship between the two variables can be observed at group level for Romania (correlation coefficient = 0.0).

We do not want to elaborate in detail why there is no such relationship at the group level even though a relationship exists within groups: a variety of reasons could be hypothesized. Rather, the example should illustrate our point that the effect size and

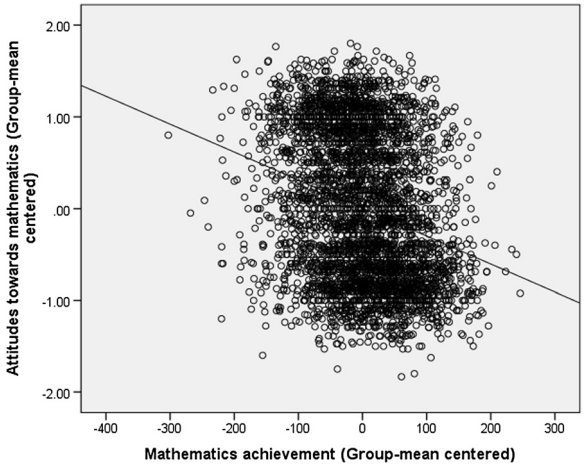

Figure 2 Example TIMSS 2007, Grade 8, Romania: Mathematics achievement scores and index on attitudes towards mathematics, both variables group-mean centered.

even the direction of relationships can differ depending on whether we look at specific group levels (classes, countries) or at the individual level, thereby keeping or removing the group-level effect. Our point holds true for any explored relationship.

We acknowledge, though, that it is not always necessary to disentangle the within- and between-group-level effects. If, for example, the group-level effect is negligible, simple individual-level analysis that ignores the group effect may suffice. Also, we might very well be interested in the composite of both effects, in which case a separation is, of course, meaningless.

Relevance for correlation analysis with TEDS-M data

According to Kreft and De Leeuw (1998), Raudenbush and Bryk (2002) and Snijders and Bosker (1999), multilevel modeling is the method of choice when dealing with cross-level effects. The application of multilevel modeling, however, requires the data

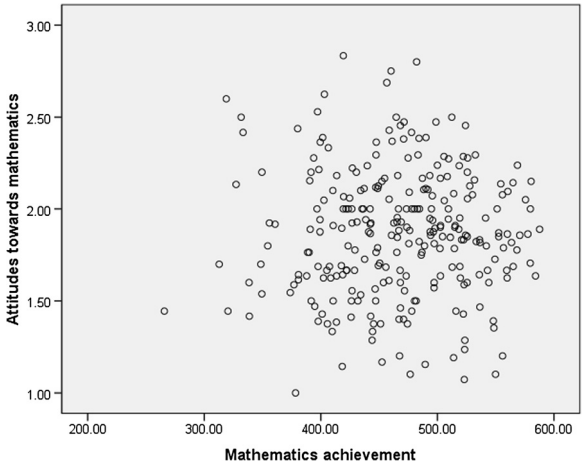

Figure 3 Example TIMSS 2007, Grade 8, Romania: Mathematics achievement scores and index on attitudes towards mathematics aggregated at class level.

to fulfill certain preconditions. To name one, the numbers of units at the different hierarchical levels must be large enough to achieve parameter estimates with acceptable precision and accuracy. As a rule of thumb, numbers of 30 clusters and 20 individuals per cluster are frequently contemplated as minimum sample sizes for this kind of analysis[f]. These preconditions were not met in many of the countries that participated in TEDS-M: only seven countries surveyed more than 30 programs for future primary teachers and only six countries surveyed more than 30 programs for future lower-secondary teachers. In addition, within these countries, many of the surveyed programs contained fewer than 10 future teachers (Dumais & Meinck, in press, a). As a further caveat, the cluster sizes (number of future teachers surveyed per program) varied greatly. For example, the variation of future primary teachers tested per program in the Russian Federation ranged from between 7 and 89 individuals in 49 surveyed programs.

In order to answer research questions that deal with correlations of TEDS-M variables and that acknowledge the clustered data structure, we suggest a four-step procedure:

1. Formulate a research question that addresses the multilevel data structure;
2. Check preconditions, familiarize yourself with the variables of interest;
3. Generate graphs and perform correlation analysis;
4. Interpret results.

In the following sections, we perform the proposed steps with two examples, both of which involve use of TEDS-M data.

Future teachers participating in TEDS-M had to complete a questionnaire that focused on (1) general background information, (2) opportunities to learn, and (3) attitudes and beliefs about the teaching profession. They also had to complete a knowledge test from which two main outcome scores were derived: a mathematical content knowledge score (MCK) and a mathematics pedagogy content knowledge score (MPCK).

There is good reason to assume with regard to TEDS-M that an aggregation of individual future teacher responses on certain variables will give a satisfactory approximation of a feature of their program. We can make this assumption because the future teachers' responses are representative of the responses from all future teachers in the program.

This is certainly the case for variables that provide information about the opportunities to learn, because such opportunities are very often fixed for all future teachers within a program. However, this assumption may not hold in countries such as Chinese Taipei where programs consist of large proportions of elective courses. Also, the average MCK or MPCK score may be seen as a good approximation for the average outcome of a program. It is pertinent to note here the commonly held view that cognitive scores vary predominantly at the individual level. However, they can also be affected by clustering. For example, programs with outstanding mathematics educators may produce future teachers who, on average, have higher MCK scores than their peers in other programs. Consider, also, the likelihood of institutes with a strong reputation for quality education and facilities attracting high proportions of particularly good students. In the absence of experimental design or longitudinal data, the causal

components cannot be evaluated. The point that we are stressing here is that clustering plays a role in understanding individual and program or institutional variation.

If we want to reveal whether there is a correlation between, for example, different opportunities to learn and outcomes of teacher education using TEDS-M data, we need to take all these considerations into account. We can conclude from the example in the preceding section that a simple correlation analysis with data from individuals belonging to groups always depicts a mixture of individual and group effects (as far as there are any such effects). However, as discussed above, the questions may be more complex, in terms of being beyond simple bivariate correlations and involving regression analyses with multiple predictors. Our intention in presenting the correlational example that follows is to clearly illustrate the relevant issues for analysis, which apply equally to regression analysis, which is a correlational model.

Analysis Example 1: Relationship between opportunities to learn and content knowledge outcome—between- and within-program effects

For the purposes of our example, assume that we are interested in determining whether there is a relationship between the opportunities given to future primary teachers to connect classroom learning to practice and their mathematics pedagogy content knowledge. The TEDS-M future teacher questionnaire asked for such opportunities to learn (OTL) through eight different items that contributed to a Rasch score included in the final dataset (Tatto et al., 2012). This Rasch score is called the OTL index "School Experiences" in the TEDS-M dataset. As mentioned before, TEDS-M represented mathematics pedagogy content knowledge by a score (MPCK) derived from a comprehensive achievement test. When conducting our analysis for this example, we used future primary teacher data from the Russian Federation.

Step 1: Formulate a research question that addresses the multilevel data structure

The research question can be phrased as follows: *"Is the opportunity to connect classroom learning to practice related to the mathematics pedagogy content knowledge of future primary teachers? If there is a relationship, is it driven by individual or group level effects, or both?"* In the case of the Russian Federation, we consider there is no reason that would prevent the use of individual data aggregates as approximations for program-level constructs.

Step 2: Check preconditions and familiarize yourself with the variables of interest

The OTL index School Experiences is a composite scale score, derived from the item pool shown in Figure 4. The OTL index was measured at the level of future teachers, indicating the possibility that School Experiences are not experienced similarly across students within the same program. OTL might be conceived of as a program-level variable, but as we found in TEDS-M, there is variability in the measurement of OTL at the future-teacher level. Nevertheless, there is also a great deal of dependency or consistency in future-teacher reports of OTL within programs. Ignoring this fact leads to the issues under examination here.

The variable MPCK is a scale score with an international mean of 500 and a standard deviation (SD) of 100. Both variables deviate somewhat from a normal distribution. As a check regarding the sensitivity of correlations to nonnormality, Spearman's rho was also calculated; however, these coefficients were very close to Pearson correlations (within 0.05), indicating the departure from normality was not distorting bivariate

During the school experience part of your program, how often were you required to do each of the following?

Check one box in each row.

		Never	Rarely	Occasionally	Often
A.	Observe models of the teaching strategies you were learning in your <courses>	☐1	☐2	☐3	☐4
B.	Practice theories for teaching mathematics that you were learning in your <courses>	☐1	☐2	☐3	☐4
C.	Complete assessment tasks that asked you to show how you were applying ideas you were learning in your <courses>	☐1	☐2	☐3	☐4
D.	Receive feedback about how well you had implemented teaching strategies you were learning in your <courses>	☐1	☐2	☐3	☐4
E.	Collect and analyze evidence about pupil learning as a result of your teaching methods	☐1	☐2	☐3	☐4
F.	Test out findings from educational research about difficulties pupils have in learning in your <courses>	☐1	☐2	☐3	☐4
G.	Develop strategies to reflect upon your professional knowledge	☐1	☐2	☐3	☐4
H.	Demonstrate that you could apply the teaching methods you were learning in your <courses>	☐1	☐2	☐3	☐4

Figure 4 Item pool used to derive opportunities to learn index "School Experiences".

associations. For most correlational models, including the general linear model, the resulting statistics are fairly robust to violations of normality. More guidance on determining risky deviations from nonnormality can be found in Howell (2011).

Step 3: Generate graphs and perform correlation analysis

Figure 5 plots future teacher-level data for the two variables of interest against each other. The plot does not distinguish between effects at the individual and at the program level. The corresponding correlation analysis gives a coefficient of 0.1*, pointing to a small positive correlation.

In a next step, we aggregated the future teacher data at the program level, using the means as a program-level feature. The same relationship, now at program level, is displayed in Figure 6. As we can see, a clearer positive relationship seems to exist. The

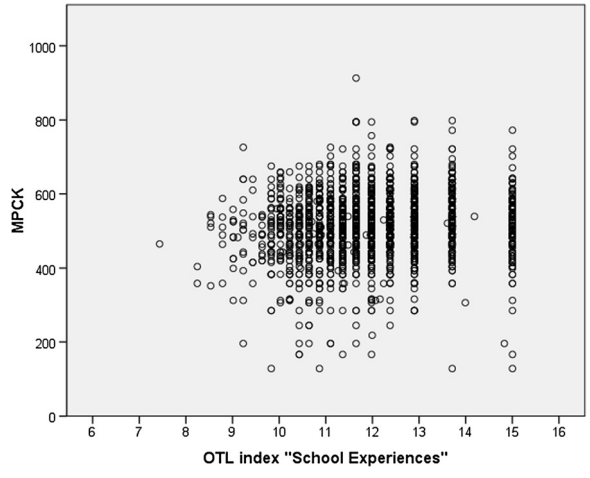

Figure 5 Future teacher-level data, scatterplot of variables MPCK and opportunities to learn index "School Experiences".

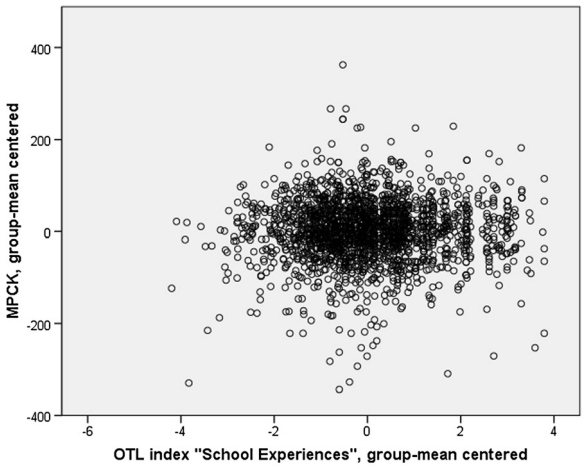

Figure 6 Program-level data (based on future teacher aggregates), scatterplot of variables MPCK and opportunities to learn index "School Experiences".

correlation coefficient based upon the aggregated data is 0.3***, suggesting a moderate positive relationship[g].

In a last step, we examined whether a relationship would also exist at the individual level if we controlled for the program-level effect. In order to do that, we performed group-mean centering, which resulted in the plot in Figure 7. The correlation coefficient of 0.0 (insignificant) for the group-mean-centered data is in agreement with the plot; no connection between the two variables can be observed once the group-level effect is removed.

Step 4: Interpret results

In the Russian Federation, the average MPCK of future primary teachers was higher in programs that provided more opportunities to connect classroom learning to practice. The effect was visible only at the program level; no relationship could be observed at the individual level once the program-level effect had been removed. Although the data

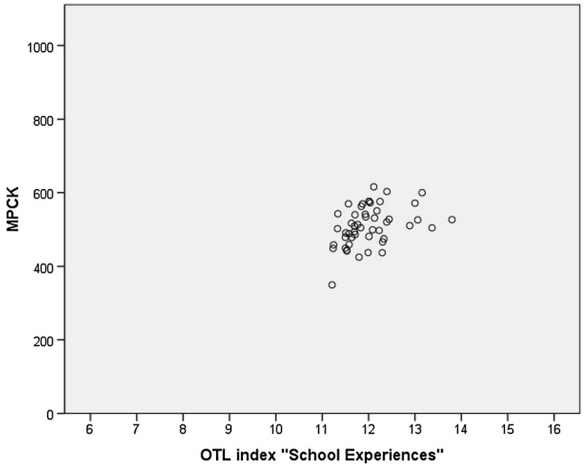

Figure 7 Future teacher-level data, scatterplot of variables MPCK and opportunities to learn index "School Experiences," both variables group-mean centered.

are cross-sectional and no causal conclusions can be drawn, the result suggests that, in the Russian Federation, providing more opportunities to connect classroom learning to practice will have a positive association on future primary teachers' pedagogical content knowledge and therefore presents a matter potentially warranting further consideration when developing program curricula.

Analysis Example 2: Relationship between MCK and MPCK—between- and within- program effects

We could argue that a future teacher with a good understanding of mathematics concepts does not necessarily possess the pedagogical skills to communicate these concepts to students. Therefore, a matter of interest is whether there is a connection between mathematics and MPCK and, if so, how strong that connection is and whether the program affiliation could influence this relationship. For the analysis relating to this example, we again used future primary teacher data from the Russian Federation.

Step 1: Formulate a research question that addresses the multilevel data structure

The research question can be phrased as follows: *"Is mathematics content knowledge associated with mathematics pedagogy content knowledge? If there is a relationship, is it driven by individual or group level effects, or both?"* Again, we used aggregated individual data as approximations for average program-level outcomes concerning content knowledge scores.

Step 2: Check preconditions and familiarize yourself with the variables of interest

Similarly to the variable MPCK, the variable MCK is a scale score with an international mean of 500 and a SD of 100. Both scales depart somewhat from normal distributions, a situation that signals the need to proceed with caution and evaluate the results accordingly.

Step 3: Generate graphs and perform correlation analysis

The graph in Figure 8 suggests a clear positive relationship between the two content knowledge domains, confirmed by the correlation analysis of the individual-level data

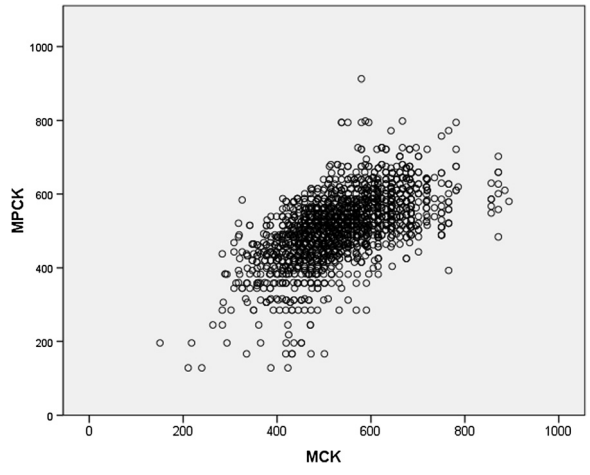

Figure 8 Future teacher-level data, scatterplot of variables MCK and MPCK.

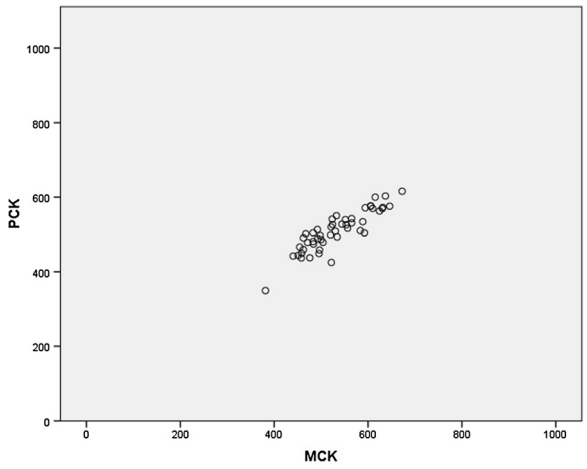

Figure 9 Program-level data (based on future teacher aggregates), scatterplot of variables MCK and MPCK.

(r = 0.6***). In order to answer the group-level-related aspect of the research question, we needed to separate program- and individual-level effects. Hence, we used aggregated future teacher scores to approximate average program-level outcomes in terms of MCK and MPCK. Figure 9 shows the relationship between the two variables when only the program-level effect is considered. Computing the correlation coefficient for the aggregated data confirmed the close relationship: r = 0.7***.

In a last step, we performed group-mean centering for the two variables in order to focus on the effect at the individual level while controlling for the effect of the program. As can be seen in Figure 10, a relationship still exists despite removal of the program-level effect. The connection appears to be weaker, however (r = 0.3***).

Step 4: Interpret results

A large positive correlation between the content knowledge domains of mathematics and mathematics pedagogy could be observed for the TEDS-M future primary teachers

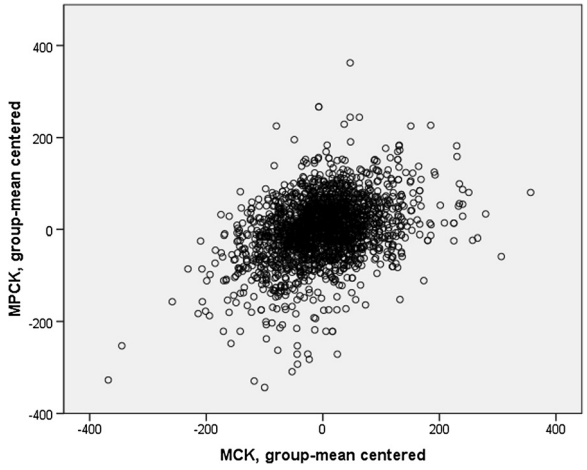

Figure 10 Future teacher-level data, scatterplot of variables MCK and MPCK, both variables group-mean centered.

in the Russian Federation. This effect became considerably smaller at the individual level, however, once we controlled for the program effect. Hence, program quality influences both knowledge domains in similar ways. Programs that produce future teachers with high average MCK also tend to produce future teachers with higher levels of MPCK. At the same time, and irrespective of the program to which belonged, the future teachers tended to combine higher (or lower) levels of both content knowledge domains.

Again, we cannot make causal inferences from the data. Further research may reveal the mechanisms that stand behind these interesting findings. Our results could mean that teachers with an indepth understanding of mathematics are also superior in transferring mathematics knowledge to (primary) students. TEDS-M has shown, however, that in many countries the curricula of future primary mathematics teachers do not focus on mathematics content much beyond that included in the school curriculum of primary school students (Tatto et al., 2012). Countries might be willing to reconsider this approach if a causal relation could be demonstrated.

Distortion of correlation coefficients due to structural features of the countries' systems of Teacher Education

Given the large variety of teacher education systems surveyed in TEDS-M, the main investigators of the TEDS-M study deemed simple country-by-country comparisons to be insufficiently meaningful. The TEDS-M design therefore also included designated program-groups that described the level of education to which a given program intended to prepare teachers, including four program-groups at the primary level (Lower Primary Grade 4 Maximum, Primary Grade 6 Maximum, Primary/Secondary Grade 10 Maximum, Primary Mathematics Specialist) and two program-groups at the secondary level (Lower Secondary, Lower/Upper Secondary)[h]. For example, Poland provided generalist and specialist programs, thereby enabling graduates from both groups to teach students regarded as "primary students"[i]. The next example shows how this structural feature of Poland's teacher education system might influence the results of correlation analysis.

Analysis Example (3): Relationship between opportunities to learn and content knowledge outcome—effect of program-groups

The question that informed this analysis was the extent to which the specific opportunities to learn (OTL) provided by programs in Poland related to the mathematics content knowledge (MCK) of the country's future primary teachers. We created data aggregated at the program level, using future primary teachers' mean MCK scores as an approximation of the program outcome, and using future primary teachers' means of the OTL indexes that were related to MCK[j] as an approximation of the program characteristic in terms of OTLs provided.

When computing the correlation coefficients among these variables at the program level for Poland, we found strikingly high positive correlations between the OTL indexes and the MCK scores (i.e., coefficients of between 0.5*** and 0.8***). However, when we looked at the scatterplots, we identified clustering patterns. Figure 11 illustrates this example for the relation between the OTL indices "Tertiary Level Math—

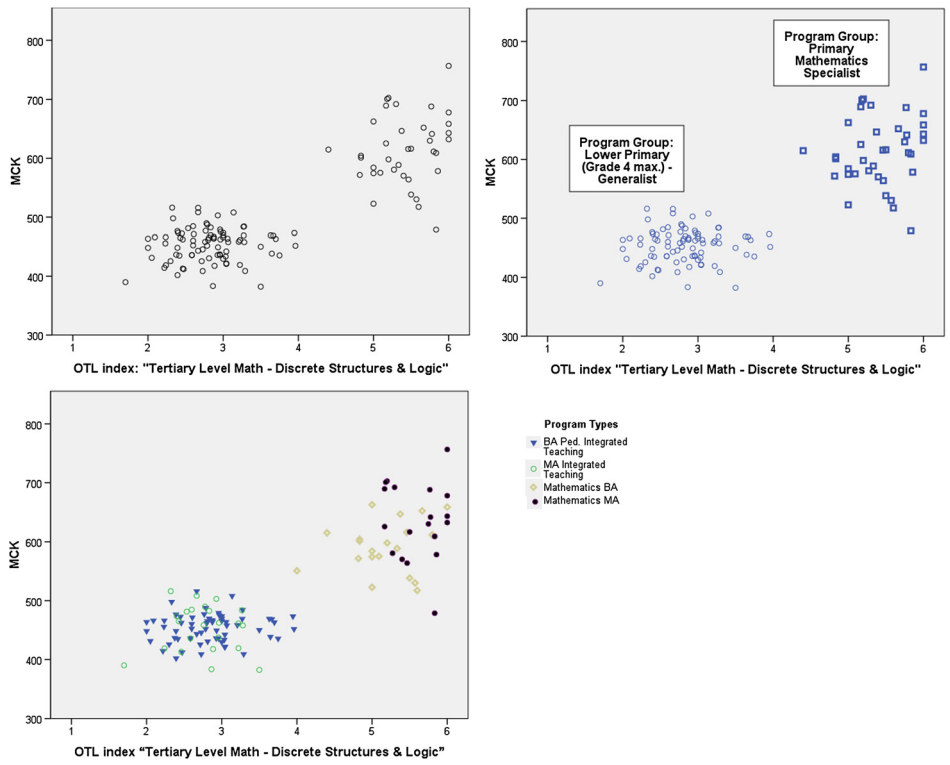

Figure 11 Program-level data (based on future teacher aggregates), scatterplot of variables MCK and opportunities to learn index "Tertiary Level Math—Discrete Structures & Logic".

Discrete Structures & Logic" and the MCK score. Indeed, when highlighting and analyzing the data separately by program group (separating generalist and specialist programs), we found the correlation coefficient dropped below 0.2 (insignificant) within both groups (see the right-hand graph in Figure 11).

We can thus see that the specialist programs were providing more opportunities to learn tertiary-level mathematics and were also producing graduates with higher MCK scores. However, there seemed to be no or only a very weak relationship between the considered OTL index and the knowledge domain: our exploration of this relation for programs within the same program-group produced a correlation coefficient that was insignificant and close to zero, although there was still considerable variation in OTLs and knowledge outcomes between programs of the same group. We found similar patterns when doing the same analysis for other OTL indexes. Therefore, a key question for further analysis of Poland's data might be: What other factors, then, explain variation in MCK outcomes among programs of the same group?

Note that some of the TEDS-M countries provided, even within the designated program-groups, different program-types. In such cases, and even if we were to analyze data separately by program-group, the cluster effects of program-types might distort the analysis results. For example, in Poland, we found bachelor's and master's programs within the same program-group as shown in Table 1. Interestingly, a further separation by program-types did not change the results, as can be seen in the lower left-hand graph of Figure 11. There therefore seems to be only minimal differences between the considered OTL index and the MCK scores comparing bachelor's and master's

Table 1 Program-groups and program-types offering education to future primary teachers in Poland

Program-type	Duration (years)	Grade span	Specialization	Program-group
• Bachelor of Pedagogy Integrated Teaching	3	1–3	Generalist	Lower primary (Grade 4 max.)
• Master of Arts Integrated Teaching	5	1–3	Generalist	
• Bachelor of Arts in Mathematics	3	4–9	Specialist	Primary mathematics specialist
• Master of Arts in Mathematics	5	4–12	Specialist	

Source: Exhibit 2.1 in Tatto et al. (2012).

programs within the same program-group. This matter might be a topic worthy of further research, but it is beyond the scope of this paper.

When countries are compared in terms of these associations, program-type must be taken into account. Countries should not be compared directly: the comparison should be only through similar program-types[k]. Not all countries have all program-types in their teacher preparation institutions. TEDS-M is an important comparative study, but the complexity, both in terms of the data structures we have described here and the program structures introduced in this third example, requires paying additional care and attention throughout analysis and interpretation. In that spirit, we offer additional cautions that must be heeded in any correlational comparative analyses. These are briefly described next.

Cautions when undertaking correlational comparative analyses

Among other objectives, an international study usually aims for comparisons across participating countries. The designated levels and groups of TEDS-M participants (countries, program-groups, primary versus lower-secondary future teachers, educators, etc.) invite comparative analyses focused on the key question of whether the relations among key variables vary across specific groups of participants.

A common comparative approach is to estimate correlations for each subgroup of interest, for example, for program-groups within a country (across institutions) or institutions within a country, or program-groups across selected countries (and many other possible arrangements). Inferences can then be made regarding differences in correlations. Significance tests also exist concerning the difference between two correlations (Howell, 2011).

When contemplating the comparison of correlations for any purpose, there is a need to offer additional cautions and take other considerations into account. At least three functional characteristics of correlations must be addressed in order to support comparative inferences for correlations. These are score distribution shape, degree of linearity, and range variation (Howell, 2011).

When testing the significance of a correlation, we need to ensure that the scores are relatively normally distributed. Nonnormality does not affect the estimation of the magnitude of the correlation. However, a more important requirement is that scores be linearly related. Recall that the Pearson correlation estimates the magnitude of the linear component of a bivariate association. If the variables are not linearly related, the Pearson correlation will underestimate the strength of the association, assuming such an association exists. Other correlations (e.g., Spearman) may better represent the magnitude of a nonnormal association, as they are a function of rank order and not dependent on

Table 2 Descriptive statistics for lower-primary generalists in a sample country

Sample	Measure	Minimum	Maximum	Mean	*SD*
Full	MCK	208	750	457	66
	MPCK	98	751	453	89
Restricted	MCK	238	579	455	55
	MPCK	98	751	451	89

linearity. A quick method to examine the degree to which scores are linearly related includes examination of a scatterplot, much like those presented earlier in this paper.

Finally, if there is range variability among subgroups and if product-moment correlations are estimated for each subgroup, correlations may differ because of the change in variability across the subgroups. Correlations are based on the covariance between two variables, which is a function of the variances of the two variables. When there is less variance, and all else being equal, the covariance, and thus the correlation, is attenuated or reduced. Another similar factor is the presence of measurement error, which similarly attenuates correlations. Corrections exist for adjusting correlations due to range restriction and measurement error (see, for example, Hunter & Schmidt, 2004; Sackett & Yang, 2000). Such corrections help to justify comparative inferences.

As an illustration, consider the correlation between MCK and MPCK. Assume that for one country within the Lower Primary Generalist programs, we find a correlation of 0.54 (Table 2 shows the means and standard deviations). We find, when using data from all the participating countries, that the MCK SD = 66, whereas when we consider only the sample country, the MCK SD = 55 (a 30% reduction in variance). This reduction results in a correlation between MCK and MPCK of 0.48 for the restricted sample. Note that we constructed this example from the TEDS-M database by restricting the range on a single country, so reducing the range of scores on MCK, which in original form ranged from 208 to 750, but in the restricted range included only the scores from 238 to 579.

Further methodological background: sampling weights and variance estimation

In this section, we look at one important aspect tied to the complex sampling design applied in the TEDS-M study and not previously highlighted in this paper. The TEDS-M sampling design actually has two particular implications for all types of data analysis, that is, also for correlation analysis.

Firstly, varying selection probabilities of programs and individuals make the use of sampling weights absolutely critical in order to achieve unbiased parameter estimates (e.g., correlation coefficients)[1]. But the graphs in all preceding figures in this paper display only the relationships observed in the sample. When making inferences about the population, we have to use sampling weights. As a matter of fact, the data points in the graphs may have different weights, or, in other words, they may contribute to the analysis with different magnitudes. This is true for both individual and aggregated data. Note, though, that we took the sampling weights into account for the computation of the correlation coefficients presented throughout the whole paper.

The second aspect requiring careful consideration is the estimation of sampling variance. Formulas for computing sampling errors applied in many standard statistical software packages such as the base module of SPSS assume simple random

sampling. Applying these formulas to TEDS-M data, which originate from cluster samples, can lead, in many instances, to underestimation of the sampling error (and consequently underestimated *p*-values). For the correct estimation of sampling errors, TEDS-M employed Fay's variant of balanced repeated replication (BRR) (Fay, 1989; Judkins, 1990; McCarthy, 1966). Statistical software packages that feature BRR (e.g., the IDB Analyzer[m] or WesVar[n]) have to be used to obtain correct estimates for sampling variances. We applied BRR in order to obtain the sampling errors and the presented significance levels in the TEDS-M examples above.

The TEDS-M public-use database provides all files along with the correct estimation weights for individual and aggregated data. Further, all files carry the necessary variables for variance estimation, using balanced repeated replication. Brese and Tatto (2012) explain in detail the correct use of the weight variables and the steps needed to correctly estimate sampling variances when analyzing TEDS-M data.

Conclusions and final remarks

The examples given in this paper show that correlation coefficients calculated at individual level (future teachers) differ in amount and meaning from those calculated at program level or computed separately for different program-groups. Due to the possible differences between individual and group effects in magnitudes or even direction, we need to take the clustered data structure into account when phrasing research questions, analyzing the data, and interpreting the results. Sound hypotheses must drive the rationale for using individual data or aggregated future teacher data as surrogates for program-level constructs in order to derive correlation coefficients. If all of these aspects are suitably considered, the hierarchical structure of the data can even improve the possibilities of extracting valuable information from the data.

It is also important to consider the complex cluster sample design of TEDS-M when carrying out statistical analyses of TEDS-M data. Sampling weights have to be used to calculate population-based parameter estimates, and balanced repeated replication has to be applied to generate population-based variance estimates for any estimated parameter. These requirements hold true whether the model is a bivariate correlation, a regression model, or a general linear model.

With regard to the sample analyses given in this paper, many further questions arise that might also be answered with TEDS-M data. For instance, would the findings hold for the future secondary teacher population in the Russian Federation, or in other participating countries? And what are the reasons for the program-level effects? Are they perhaps associated with the intake of the program? Or are they connected to certain selection criteria, or opportunities to learn given in the program? And so on. Although we examined only the case of correlation analysis in this paper, subsequent analysis may call for different methods—regression analysis, for example. Also, be aware that even though this matter is not addressed in this paper, other types of analysis may be influenced by the clustered data structure.

As is the case with all data analytic strategies, we need to remember to follow a systematic approach to analysis. The distributional properties of the variables should be understood and the assumptions of the intended statistics assessed. For correlations,

the normality of the distribution of each variable as well as the extent to which the variables are linearly related need to be evaluated.

Finally, when considering the use of correlations for comparative purposes, we need to evaluate the summary statistics for each measure in conjunction with making comparative statements about the magnitudes of correlations across multiple groups. All else being equal, variation in subgroups will directly affect the magnitude of correlations. It should be recalled that, in some cases, samples differ from group to group, a situation which also influences correlations. Overall, we consider that the TEDS-M international database provides a very valuable source for new findings in the field of mathematics teacher education, offering scope for valid results, but only if the attributes of the datasets are properly considered during analyses.

Endnotes

[a]Teacher preparation institutions and mathematics educators comprised further target populations of TEDS-M but were not within the scope of this paper (see Brese & Tatto, 2012).

[b]For detailed descriptions of the participating countries' sample designs, see Meinck and Dumais (in press).

[c]Using Cohen's (1988) scale of magnitudes of correlations.

[d]In the following, significance levels are indicated as given here: * $p = 0.05$, ** $p = 0.01$, *** $p = 0.001$.

[e]This general particularity of cluster samples can be measured by the intraclass correlation coefficient and calls for specific variance estimation methods. For more information, see Joncas (2008).

[f]For further reading, refer to Afshartous (1995), Bell et al. (2010); Maas and Hox (2005), Meinck and Vandenplas (2012), Mok (1995), and Snijders (2005).

[g]Note that the correlation coefficients of raw and aggregated data cannot be directly compared (see section "Cautions in Correlational Comparative Analyses").

[h]For more details, please refer to Chapter 2.2 of Tatto et al. (2012).

[i]TEDS-M considered students in ISCED Level 1 to be primary school students. ISCED stands for International Standard Classification of Education. See for example, >http://www.uis.unesco.org/Education/ISCEDMappings/Pages/default.aspx< for country mappings to ISCED levels.

[j]TEDS-M reported on six OTL indexes pertaining to school-level or tertiary-level mathematics.

[k]The program-types of all participating countries are described in Chapter 3.3 in Tatto et al. (2012), and a matching of program-types to program-groups is given in Exhibit 2.1 of the same source.

[l]Lohr (1999) and Dumais and Meinck (in press, b) are recommended for further reading on the topic.

[m]International Association for the Evaluation of Educational Achievement (IEA) (2012).

[n]Westat Inc (2008).

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

Authors' contributions
SM carried out the example analyses and drafted major parts of the manuscript. MCR drafted smaller parts of the manuscript. All authors read and approved the final manuscript.

Author details
[1]IEA Data Processing and Research Center, Hamburg, Germany. [2]Department of Educational Psychology, University of Minnesota, Minneapolis, USA.

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

References

Afshartous D (1995) Determination of sample size for multilevel model design. In: Williams VS, Jones LV, Olkin I (eds) Perspectives on statistics for educational research: Proceedings of a workshop (pp. 2–21) (National Institute for Statistical Sciences Technical Report Number 35). National Institute for Statistical Sciences, Triangle Park, NC, Retrieved from http://www.niss.org/sites/default/files/pdfs/technicalreports/tr35.pdf

Bell BA, Morgan GB, Schoeneberger JA, Loudermilk BL, Kromrey JD, Ferron JM (2010) Dancing the sample size limbo with mixed models: How low can you go? (SAS Global Forum 2010 Posters Paper 197-2010)., Retrieved from http://support.sas.com/resources/papers/proceedings10/197-2010.pdf

Brese F, Tatto MT (2012) TEDS-M 2008 user guide for the international database. Amsterdam, the Netherlands: International Association for the Evaluation of Educational Achievement (IEA)., Retrieved from http://www.iea.nl/teds-m.html

Buckley J (2009) Cross-national response styles in international educational assessments: Evidence from PISA 2006., Unpublished manuscript retrieved from https://edsurveys.rti.org/PISA/documents/Buckley_PISAresponsestyle.pdf

Cohen J (1988) Statistical power analysis for the behavioral sciences, 2nd edn. Lawrence Erlbaum, Hillsdale, NJ

Cronbach LJ (1976) Research on classrooms and schools: Formulation of questions, design and analysis. Stanford Evaluation Consortium, School of Education, Stanford University, Stanford, CA, Retrieved from http://eric.ed.gov/?id=ED135801

Dumais J, Meinck S Sampling design. In: Tatto MT (ed) Teacher Education and Development Study in Mathematics (TEDS-M): Technical report. International Association for Educational Achievement (IEA), Amsterdam, the Netherlands (in press, a)

Dumais J, Meinck S Weighting and variance estimation. In: Tatto MT (ed) Teacher Education and Development Study in Mathematics (TEDS-M): Technical report. International Association for Educational Achievement (IEA), Amsterdam, the Netherlands (in press, b)

Fay RE (1989) Theoretical application of weighting for variance calculation. In: Proceedings of the Survey Research Methods Section of the American Statistical Association. American Statistical Association, Alexandria, VA, pp 212–217, Retrieved from http://www.amstat.org/sections/srms/proceedings/papers/1989_ 033.pdf

Galtung J (1967) Theory and methods of social research. Columbia University Press, New York, NY

Howell DC (2011) Fundamental statistics for the behavioral sciences, 7th edn. Cengage Learning, Belmont, CA

Hunter JE, Schmidt FL (2004) Methods of meta-analysis: Correcting error and bias in research findings, 2nd edn. Sage, Thousand Oaks, CA

International Association for the Evaluation of Educational Achievement (IEA) (2012) International Database Analyzer (Version 3.0) [Computer software]. IEA Data Processing and Research Center, Hamburg, Germany

Joncas M (2008) TIMSS 2007 sample design. In: Olson JF, Martin MO, Mullis IV (eds) TIMSS 2007 technical report. Boston College, Chestnut Hill, MA, pp 77–92

Judkins DR (1990) Fay's method for variance estimation. J Off Stat 6(3):223–239

Klein KJ, Kozlowski SWJ (2000) From micro to meso: Critical steps in conceptualizing and conducting multilevel research. Organ Res Meth 3:211–236

Knapp TR (1977) The unit-of-analysis problem in applications of simple correlation analysis to educational research. J Educ Stat 2(3):171–186

Kreft I, De Leeuw J (1998) Introducing multilevel modeling. Sage, London, UK

Lohr S (1999) Sampling: Design and analysis. Duxbury Press, New York, NY

Maas CJM, Hox JJ (2005) Sufficient sample sizes for multilevel modeling. Methodology 2005 1(3):86–92

McCarthy PJ (1966) Replication: An approach to the analysis of data from complex surveys. Vital and Health Statistics (Series 2, No. 14). National Center for Health Statistics, Hyattsville, MD

Meinck S, Dumais J (2999) Characteristics of national samples: Implementations of the international sampling design in participating countries. In: Tatto MT (ed) Teacher Education and Development Study in Mathematics (TEDS-M): Technical report. International Association for Educational Achievement, Amsterdam, the Netherlands, in press

Meinck S, Vandenplas C (2012) Evaluation of a prerequisite of hierarchical linear modeling (HLM) in educational research: The relationship between the sample sizes at each level of a hierarchical model and the precision of the outcome model, IERI Monograph Series Issues and Methodologies in Large-Scale Assessments, Special Issue 1. IEA Data Processing and Research Center, Hamburg, Germany

Mok M (1995) Sample size requirements for 2-level designs in educational research. Multilevel Model Newsletter 7(2):11–15

Mossholder KW, Bedeian AG (1983) Cross-level inference and organizational research: Perspectives on interpretation and application. Acad Manage Rev 8:547–558

Mullis IVS, Martin MO, Gonzalez E, O'Connor KM, Chrostowski SJ, Gregory KD, Garden RA, Smith TA (2001) Mathematics benchmarking report TIMSS 1999, eighth grade: Achievement for U.S. states and districts in an international context. Boston College, Chestnut Hill, MA

Mullis IVS, Martin MO, Foy P (2008) TIMSS 2007 international mathematics report: Findings from IEA's Trends in International Mathematics and Science Study at the fourth and eighth grades. Boston College, Chestnut Hill, MA

Raudenbush SW, Bryk AS (2002) Hierarchical linear models: Applications and data analysis methods, 2nd edn. Sage, Newbury Park, CA
Robinson WS (1950) Ecological correlations and the behavior of individuals. Am Sociol Rev 15(3):351–357
Sackett PR, Yang H (2000) Correction for range restriction: An expanded typology. J Appl Psychol 85(1):112–118
Snijders T (2005) Power and sample size in multilevel linear models. In: Everitt BS, Howell DC (eds) Encyclopedia of statistics in behavioral science, vol 3. Wiley, Chichester, UK, pp 1570–1573
Snijders T, Bosker R (1999) An introduction to basic and advanced multilevel modeling. Sage, Thousand Oaks, CA
Tatto MT, Schwille J, Senk S, Ingvarson L, Peck R, Rowley G (2008) Teacher Education and Development Study in Mathematics (TEDS-M): Policy, practice, and readiness to teach primary and secondary mathematics, Conceptual framework. Teacher Education and Development International Study Center, College of Education, Michigan State University, East Lansing, MI, Available online at http://www.iea.nl/teds-m.html
Tatto MT, Schwille J, Senk SL, Ingvarson L, Rowley G, Peck R, Bankov K, Rodriguez M, Reckase M (2012) Policy, practice, and readiness to teach primary and secondary mathematics in 17 countries: Findings from the IEA Teacher Education and Development Study in Mathematics (TEDS-M). International Association for the Evaluation of Educational Achievement (IEA), Amsterdam, Available online at http://www.iea.nl/teds-m.html
Van Mierlo H, Vermunt JK, Rutte CG (2009) Composing group-level constructs from individual-level survey data. Organ Res Meth 12(2):368–392
Westat Inc (2008) WesVar, Version 5.1: Replication-based variance estimation for analysis of complex survey data [Computer software]. Westat Inc, Rockville, MD

doi:10.1186/2196-0739-1-7
Cite this article as: Meinck and Rodriguez: **Considerations for correlation analysis using clustered data: working with the teacher education and development study in mathematics (TEDS-M) and other international studies.** *Large-scale Assessments in Education* 2013 **1**:7.