Computational typologies of multidimensional end-of-primary-school performance profiles from an educational perspective of large-scale TIMSS and PIRLS surveys

Ali Ünlü[1]
Technical University of Munich (TUM)
Centre for International Student Assessment (ZIB)
TUM School of Education

Michael Schurig
IFS, TU Dortmund University

Recently, performance profiles in reading, mathematics and science were created using the data collectively available in the Trends in International Mathematics and Science Study (TIMSS) and the Progress in International Reading Literacy Study (PIRLS) 2011. In addition, a classification of children to the end of their primary school years was conducted in accordance with these performance types. To create performance profiles and classifications, multidimensional item response theory and latent profile analysis were used. The focus in this study is on the comparison and usability of clustering methods in their application in large-scale assessments. In a first step, the cluster solutions of classic approaches such as k-means, fuzzy c-means and hierarchical procedures are compared to one another and assessed in terms of their proximity to the reference typology of latent profile analysis. In the second step, the results of the model-supported latent profile analysis are compared directly with the findings of the classic model-free cluster analyses by means of appropriate measured values. The result is a high consistency in the classification of invariant ranked profiles. In the last step, the calculated "quantitative" cluster solutions are compared to the "qualitative" typology of the pupils derived from the content-based benchmarking of the competency levels in mathematics using the TIMSS guidelines. It is evident that, as a cluster solution, the benchmarking breakdown of the sample in the five competency levels does not show a high goodness of fit with the available data.

Introduction: TIMSS and PIRLS Studies

Every five years, fourth grade students from around the world are compared on their performance in reading comprehension through the Progress in International Reading Literacy Study (PIRLS; e.g., Bos, Tarelli, Bremerich-Vos, & Schwippert, 2012a). Another international test that evaluates student academic performance is the Trends in International Mathematics and Science Study (TIMSS; e.g., Bos, Wendt, Köller, & Selter, 2012b), which assesses fourth and eighth grade students on their performance in mathematics and science on a four-year cycle. The International Association for the Evaluation of Educational Achievement (IEA) is responsible for conducting both studies (www.iea.nl).

Data collection for both studies coincided for the first time in 2011, making it possible to connect the two studies. Thus, the focused subject areas of reading, mathematics and science could be mutually investigated while considering their interactions. As was the

[1] Corresponding author. Prof. Dr. Ali Ünlü (ali.uenlue@tum.de), Technical University of Munich (TUM), Arcisstr. 21, 80333 Munich, Germany.

 Current Issues in Comparative Education 18(1), 6-25.

case in 37 other countries and regions, a common random sample of pupils for both PIRLS and TIMSS was surveyed in Germany (Martin & Mullis, 2013). This special data set opens up new possibilities for analysis and allows for an overarching view of pupils' performances at the end of the fourth grade in reading, mathematics and science (e.g., Bos et al., 2012c).

From a methodical viewpoint, these data have particular demands in terms of the appropriate handling of the multidimensional structure (Foy & O'Dywer, 2013). Multidimensional procedures have thus been increasingly used in recent years to address the demands of modeling such complex data. They can provide more realistic explanations for observed data and include, in particular, the use of multidimensional *item response theory* (IRT) models, which allows for the determination of measurement error controlled correlations between latent dimensions. For the specifics of multidimensional IRT, see Reckase (2009) or van der Linden and Hambleton (1997). In this study, a general multidimensional IRT model was used to analyze the TIMSS and PIRLS data.

In both TIMSS and PIRLS, performances are depicted along a scale that has a mean value of 500 and a standard deviation of 100. From the international comparative perspective, however, the interpretation of inter-individual differences in the dimensions primarily occurs via the achievement of international benchmarks. In other words, the achievement of qualitatively set criteria is defined separately for each dimension (Martin & Mullis, 2013). In TIMSS and PIRLS, the benchmarks divide the performance scales into five performance ranges with a width of 75 points. In Germany, reporting specifically refers to the ranges between the benchmarks, called competency levels. Test values below 400 points correspond to competency level I; values between 400 and 475 points represent competency level II; the range of values between 476 and 550 points is competency level III; competency level IV covers the range of values between 551 and 625 points; and values above 625 points comprise competency level V (Wendt, Tarelli, Bos, Frey, & Vennemann, 2012, p. 62f.).

In order to allow for a contentual interpretation of the competency levels achieved by pupils, the problems or exercises assigned to an individual level are constructed in such a way that a student possessing that level of competence can typically solve these problems. As an example, the qualitative criteria for the achievement of the third competency level (High International Benchmark) for PIRLS 2011 are listed in Table 1.

In addition to Figure 1, we can report that 1.5% of the pupils achieved competency level V in all three areas and 0.6% of the pupils achieved competency level I in all areas. 10.5% of pupils achieved competency levels IV-V in three domains, excluding the performances with three domains all at either competency level IV or V. 11.4% achieved competency level IV in three domains. 32.5% achieved competency levels III-V in three domains. 13.1% achieved competency level III in three domains. 19.2% performed at competency levels I-III in three domains. The remaining 11.2% are distributed across smaller intermediate stages. 55.9% of the pupils were able to achieve at least competency level III and a higher level in one domain. Only 2.9% showed "savantism" or "learning disabilities" in one domain. These were operationalized by pupils who were indexed at one competency level of I or II in two domains and a competency level of IV or V in one domain, or at one competency level of IV or V in two domains and one competency level of I or II in one domain.

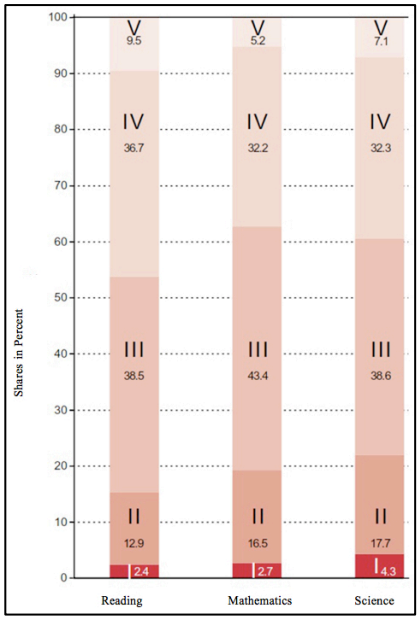
Table 1: PIRLS 2011 High International Benchmark of reading achievement

High International Benchmark	
550 to 476	**When reading Literary Texts, students can:**
	• Locate and distinguish significant actions and details embedded across the text • Make inferences to explain relationships between intentions, actions, events, and feelings, and give text-based support • Interpret and integrate story events and character actions and traits from different parts of the text • Evaluate the significance of events and actions across the entire story • Recognize the use of some language features (e.g., metaphor, tone, imagery)
	When reading Informational Texts, students can:
	• Locate and distinguish relevant information within a dense text or a complex table • Make inferences about logical connections to provide explanations and reasons • Integrate textual and visual information to interpret the relationship between ideas • Evaluate content and textual elements to make a generalization

Source: IEA's Progress in International Reading Literacy Study – PIRLS 2011

In Germany, the achievement of competency levels across the domains was distributed as shown in Figure 1.

Figure 1: Distribution of competency levels achieved by pupils



Source: Bos et al., 2012a, p. 231

As mentioned earlier, since data collection for TIMSS and PIRLS coincided for the first time in 2011, a joint study of all dimensions is possible and allows for the interpretation of interactions. However, a consolidation has to first occur in the scale. It also makes sense to scrutinize the classification that is used in the benchmarks.

Classification Within the Multidimensionality – Motivation

The reliable determination of a valid performance typology is essential. On the one hand, benchmarks and competency levels offer an improvement of the interpretation ability of the IRT scale results; on the other hand, they are the actual parameters for communication purposes, such as for educational policy (e.g., Bos et al., 2009).

In international reports, countries are measured and compared using the respective percent shares of pupils in the individual competency levels. An example of this can be found in the well-known Programme for International Student Assessment (PISA) study conducted by the Organisation for Economic Cooperation and Development (OECD, 2010). The formation of "intrinsically homogeneous" and "amongst themselves heterogeneous" competency groups contributes to the interpretation and significance of IRT estimated numerical values. This is especially true if the categorizing performance types are described in accordance with background characteristics, which are relevant and theoretically important from an educational science perspective. These background characteristics include factors such as cultural and socioeconomic characteristics, migration background and family language, learning disabilities or physical frailties, or subject-related attitudes and self-concepts.

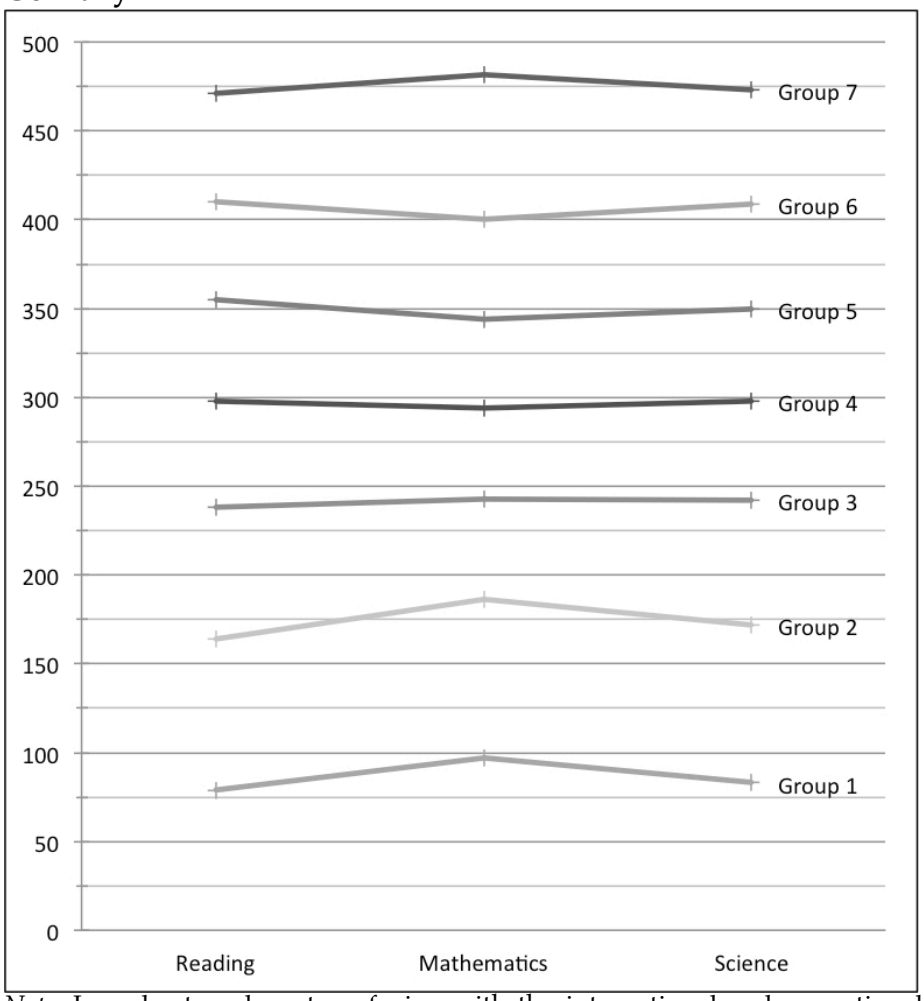
In the context of multidimensionality, however, the approach in terms of benchmarks and competency levels used to date presents a problem. This is because incorporating the complex data structure and taking into consideration the high interdependence of the observed constructs in the classification are possible only in a limited manner. In previous works, the multidimensional IRT scaled performance values are solely included in studies if multidimensional modeling is considered at all. This is to be differentiated from the actual objective of this paper. We aim to develop and examine multivariate data-analytical procedures for the "automated" joint discretization of all competency dimensions while considering their interdependences. To date, the discretization of the dimensions in large-scale assessments is conducted through the qualitative setting of so-called "cut-off points" separately for each individual dimension.

A way in which to bypass this problem is the derivation of latent profiles that are calculated across all the jointly scaled content areas, where the latency rests in the class affiliation and is thus discrete. An example of this can be found in Bos et al. (2012c) who used the TIMSS/PIRLS 2011 data set for Germany containing 3,928 pupils for the consolidation of multidimensionality and classification process. Performance test values and performance profiles across the competency areas were regarded using latent profile analysis.

In line with Bos et al. (2012c), we carried out calculations based on the estimated performance results using the *Latent GOLD* software (Vermunt & Magidson, 2005). Solutions with 1 to 20 profiles were compared using information criteria and log likelihood comparison tests via bootstrapping. A 7-profile solution proved to be the best. There were distinct performance clusters in a strict invariant parallel structure (Figure 2). This structure corresponds to an invariant arrangement or ranking of the performance types across the three dimensions (cf. Croon, 1990). For the present study

the results were reconstructed with congruent outcomes.

Figure 2: Parallel plot of the latent profile analysis of performance values of pupils in Germany



Note: In order to rule out confusion with the international scale, a national scale with a mean value of 300 and a standard deviation of 100 was used for the analysis.

The performance types can be described as shown in Table 2. In addition, the performance profiles can be clearly distinguished by the distribution of background characteristics between the profiles (for specifics, see Bos et al., 2012c).

Table 2: Description of the performance profiles

			Reading				Mathematics				Science			
Group	n	%	M	(SE)	SD	(SE)	M	(SE)	SD	(SE)	M	(SE)	SD	(SE)
7	165	4,3	471	4,4	46	2,3	482	4,8	55	3,1	473	4,7	50	4,5
6	533	13,5	410	2,3	45	1,9	400	3,2	58	2,1	409	2,8	50	2,8
5	900	22,8	355	2,1	43	1,5	344	3,0	59	2,5	350	2,5	51	1,7
4	1003	25,5	298	2,0	44	1,4	294	2,6	60	1,8	298	2,7	52	1,6
3	753	18,9	238	1,7	45	1,4	243	3,0	60	2,0	242	2,2	49	1,8
2	444	11,3	164	3,3	51	2,9	186	3,6	57	3,1	172	3,9	49	2,8
1	130	3,8	79	5,3	60	3,8	97	8,0	60	4,8	83	6,0	55	5,2

Research Questions and the Contents of this Study

The latent profile analysis is based on model assumptions, and the estimation of its parameters under these assumptions is based on maximum likelihood methodology. In this sense, latent profile analysis, as a model-based, inference-theoretical approach (probabilistic formulation), can be compared with the classic cluster analysis as a "model-free", computational approach (descriptive distance calculation).

In order to conduct a computational cluster analysis on a data set, a variety of standards have to be defined (e.g., with regard to the algorithm used and the similarity or dissimilarity metric) that are not possible or necessary in the flexible manner under the maximum likelihood approach. The question arises whether alternative computational clustering methods, without such additional model assumptions and estimation procedures, can be conducted. That is, one may ask if comparable results can be achieved and if similar profiles can be derived with more "simple" methods, when classification occurs in the measurement error controlled, estimated performance values (not raw data). This gives rise to the following research questions that we would like to examine in this study:

> *Q1*. How do empirically derived, multidimensional latent classifications look in comparison to classically derived clusters?
> *Q2*. How do empirically derived classifications across all dimensions, whether latent or computational, look in comparison to the groups obtained from the international benchmarking?

More specifically, in this study we would like to test if, and to what degree, profiles can be derived more simply on the basis of the model-free classic clustering methods (*k*-means, fuzzy *c*-means and hierarchical cluster analysis). For this purpose, a national, multidimensional IRT scale is performed, profiles are constructed using latent profile analysis, and the results of the classic clustering methods are compared to the results of the latent profile analysis based on measures of quality and agreement. In doing so, the cluster number is set to the number of the latent profiles (reference typology) in one case and freely estimated in the other. The exemplary comparison to one individual content area occurs as well to the benchmarking typology in the domain of mathematics. We expect that computational clustering methods, as compared with qualitative classification of pupil performances based on predefined competency levels, will lead to

more homogeneous groups. This paper will conclude with a discussion of future research directions, as well as a recommendation for using computational analyses as classic methodologies to complement the model-based approach.

Clustering Methods

We review some basic concepts of cluster formation, which is the derivation of groups with similar properties (e.g., Ester & Sander, 2000; Everitt, 2011).

For a data set with N objects in C groups, the number of the grouping possibilities, i.e., the number of the C-element partitions of the N-element set is a Stirling number of the second kind:

$$\frac{1}{C!}\sum_{j=0}^{C}(-1)^{C-j}\binom{C}{j}j^N$$

Thus, where $N = 20$ and $C = 3$, there are over 550 million possible combinations. The identification of the contentually significant clusters is important. A large number of heuristic methods have been developed for this purpose. In principle, all cluster analytical methods are algorithms to sort individual empirical observations. The selection of the appropriate cluster solution is conducted by way of intuition and theoretical considerations based on preliminary results, as well as the statistical comparison thereof. All methods attempt to create homogeneous groups that are heterogeneous to one another, whereby the similarity or dissimilarity distance measures used assume great significance. If the distances, or reciprocally the similarities, are combined in pairs, a distance matrix $D = (d_{mn})$ or a similarity matrix $S = (s_{mn})$ is obtained. There are no uniform results for the "correct" selection of the distance measure (Everitt, 2011). When cases are clustered, a distance measure that takes into account the measurement level of the data can be used as a basis, such as the Euclidean distance in the case of metric data.

Classic cluster analytical methods are computational and exploratory. Freedom in the selection of the computational components provides flexibility (comparable to the rotation freedom/problem of factor analysis models). Due to the high degree of interpretive freedom, however, general skepticism toward the uncovered structure is appropriate as well. This should be viewed with caution. The applied methods will be presented briefly in the following sections; specifics can be found in the respective accompanying references. Possible evaluation criteria are explained as they are used in the course of this study.

Latent Profile Analysis

The method that was used to obtain the performance profiles is *latent profile analysis* (LPA), e.g., see Gibson (1966) or Lazarsfeld and Henry (1968). The local stochastic independence is presupposed. It is assumed that the answers of the objects of a *given* class, in this case the pupils in one class or profile, are random. If the latent class was introduced as a control variable, systematic variations between the answers would disappear. Each class continues to be characterized by their specific conditional answer probabilities. The formal structure, discretized here for simplification (sum instead of integral), is shown as follows:

$$P(x_1, \ldots, x_I) = \sum_{c=1}^{C} \pi_c \prod_{i=1}^{I} P(X_i = x_i | c)$$

Whereby π_C describes the relative class size of a class c from 1 to C, and $P(X_i = x_i \mid c)$ stands for the conditional probability of a (continuous) realization $X_i = x_i$ in a variable i from 1 to I in the (discrete) class c. The LPA model parameters are estimated from the data using the maximum likelihood method. The estimation can be carried out via expectation-maximization or EM (e.g., Bacher, Pöge, & Wenzig, 2010, p. 360; see also Dempster, Laird, & Rubin, 1977). The cluster number in LPA is selected using statistics for model quality (e.g., information criteria). Bootstrapping methods are also recommended for significance testing (Bacher et al., 2010, p. 365).

In our application, the IRT scaled performance test values ("plausible values", see below in Section *Multidimensional Scale*) in the domains of reading, mathematics and science correspond to the variables (I = 3), and the classes are represented by the uncovered performance profiles (C = 7). The corresponding solution of the specific LPA calculations can be found in Figure 2 or Table 2.

Hierarchical Cluster Analysis

The family of hierarchical clustering methods (e.g., Everitt, 2011; Kaufman & Rousseeuw, 1990) consists of various agglomerative ("bottom up") or divisive ("top down") algorithms. For exemplification we use a popular agglomerative method (Ward's algorithm; see the following paragraph). Other agglomerative or divisive methods are conceivable if necessary. The clusters are constructed around cluster centers based on the mean values of the cluster objects. In a first step, each object forms its own cluster. In each subsequent step, a cluster pair for which a given distance criterion (see the following paragraph) is minimal is sought and merged. Additionally, the value of the cluster center is recalculated for the newly formed cluster. These steps are repeated until all clusters have merged into one. The cluster number is not specified; but rather the possible solutions, portrayable as the structural order of a dendrogram, are tested against one another. Thus, a deterministic allocation of the objects is given on every possible sectional plane of the dendrogram.

Ward's (1963) algorithm was used in our study. This method consolidates those clusters whose resulting clusters minimize the growth of the sum of the average errors within the clusters. The squared Euclidean distance is the basis for the dissimilarity calculation. In general, the method generates relatively homogeneous clusters, but it is sensitive with regard to outliers (see Everitt, 2011).

k-Means

One of the most common clustering methods is the *k*-means analysis (Forgy, 1965; MacQueen, 1967). It is based on the Euclidean distance. The algorithms of this method allocate objects to the arithmetic means, medians or medoids as the cluster centers (Kaufman & Rousseeuw, 1990). The goal is to subdivide the objects into clusters in such a way that the sum of the squared deviations from the cluster centers is minimal. The sum of squares within the clusters can be interpreted as a type of "error scatter".

The method essentially differentiates four iterative steps: random allocation of the objects into a *predefined* number of clusters; recalculation of the cluster centers; reallocation of the objects via minimal squared Euclidean distance; test to see if the

minimization of the sum of squares of the scatter changes with the allocation. If this is not the case, or a stop criterion is reached, the algorithm terminates itself. The cluster number is determined in corresponding indices based on selected threshold values via comparisons of repeated applications of the method with a variety of numbers of clusters, each with multiple start values.

Fuzzy Methods
The fuzzy cluster analysis has a "blurred" nature. It describes the uncertainty that an object *g* belongs to a class *c*, and thus it distinguishes itself from the previously presented "deterministic" methods (see Bacher et al., 2010; Everitt, 2011). In essence this method can be understood as a generalization of *k*-means.

The *c*-means algorithm for continuous data was used for the fuzzy analysis (Bezdek, 1981, 1983). One criterion based on the weighted sum of the distance squares is used for data vectors x_i of *N* objects in *C* clusters:

$$\sum_{t=1}^{C} \sum_{i=1}^{N} u_{it}^{\nu} \, d^2(x_i, m_t)$$

Whereby m_t is the center of cluster *t*, and the weights (degrees of belonging) u_{it} add up to 1 for all $i = 1, \ldots, N$. In addition, $d(x_i, m_t)$ is the Euclidean distance between data point and cluster center, and v is referred to as the "fuzzification", in other words the "fadedness". The degrees u_{it} of the group membership are unknown and must be determined using the data. The cluster number can be selected via a comparison of the number of iterative steps to convergence (e.g., convergence under minimal iteration).

Methodical Approach
In the first step of the sequential approach, a three-dimensional scale is taken as a basis for the common TIMSS and PIRLS 2011 sample data. Classifications based on the three-dimensional scale for this sample, by means of latent profile analysis and the international benchmarking method, are used as comparison values or reference typologies in the second step. Every other partitioning could equally well be used as a reference typology if it seems useful. Examples are typologies that are derived from substantive theories or those that are based purely on (other) statistical methods. The elaborations presented can then be applied by analogy. In the present study, latent profile analysis and the international benchmarking method are used as our reference typologies.

In the second step in the sequential processing, the population estimators are classified using hierarchical cluster analysis (HCA), *k*-means, and fuzzy methods. First, a comparison is conducted with cluster numbers fixed in accordance with the result of the LPA. Then, a comparison is made with specific optimal cluster numbers. Finally, a comparison is made of the optimal solutions to the performance levels defined in the international benchmarking in the domain mathematics.

Data Set
The entire sample for Germany comprises 4,229 pupils in 197 primary schools. The joint test of PIRLS and TIMSS was administered to the sample for a duration of two consecutive days. In the first day, approximately 50% of the sample took TIMSS while

the other half took PIRLS. The two groups switched tests on the second day (Wendt et al., 2012). The sample size of complete cases includes 3,928 pupils and 197 schools. Each student in the sample was administered tests to measure performance in all three competency areas: reading comprehension (PIRLS), mathematics and science (TIMSS).

Multidimensional Scale

To allow a competency domain accounting for the full range of performance test values in the first step of the sequential approach, the survey data were jointly scaled in a multidimensional IRT model. The mixed-coefficients multinomial logit model of Adams et al. (1997) was used with the *ConQuest* software (Wu et al., 2007). This model is a generalization of the Rasch model (Fischer & Molenaar, 1995; Rasch, 1960). A national background model that includes information from the pupil and parent questionnaires, as well as from the school questionnaires and the parameters to the cognitive abilities was incorporated using latent regressions. Population describing person parameters ("plausible values", PV) were calculated in this manner (see Mislevy, 1991; cf. also von Davier, Gonzalez, & Mislevy, 2009).

A three-dimensional model based on the three primary competency domains was used, even though an eight-dimensional model based on the sub-domains would be conceivable (see Bos et al., 2012c). Both models are preferable to a one-dimensional model. The reading, mathematics, and science competency domains represent the three dimensions of the preferred IRT model. As shown in Table 3, there are strong correlations between the areas.

Table 3: Latent correlations for the three-dimensional domain-related mixed-coefficients multinomial logit model

	Reading	**Mathematics**	**Science**
Reading		.54	.74
Mathematics	.54		.66
Science	.74	.66	

Results

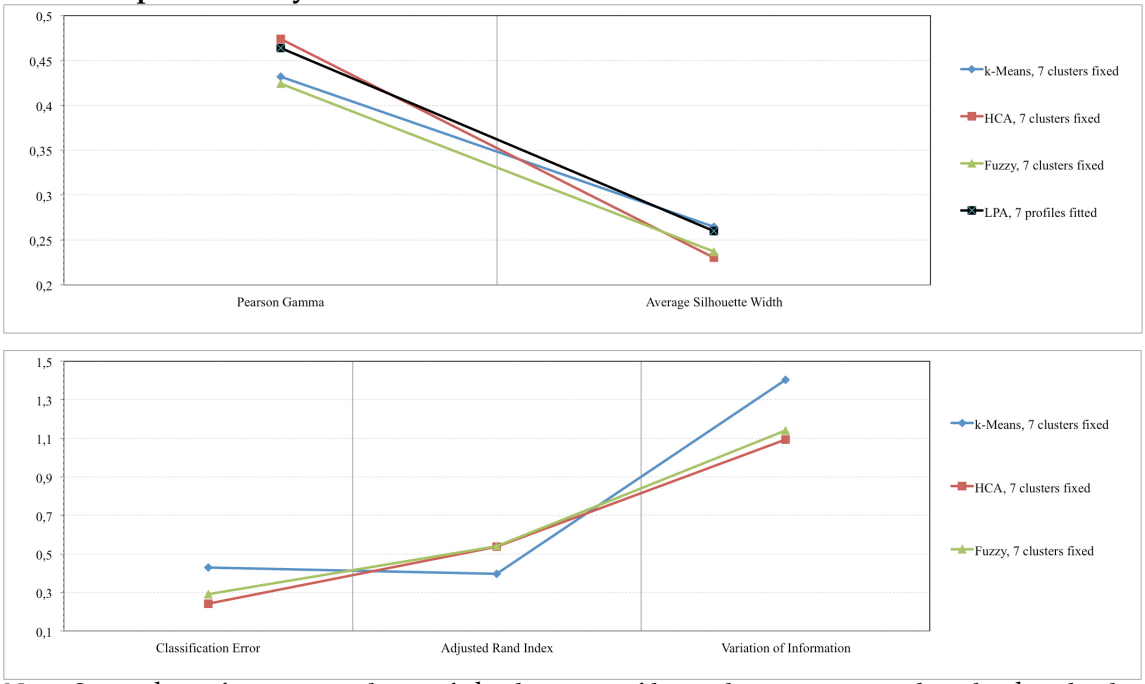
The calculations were primarily conducted in the open-source software *R* (R Core Team, 2014, www.R-project.org). The hierarchical and *k*-means cluster analyses were calculated with the *R* package *cluster* (Maechler et al., 2014), the fuzzy cluster analyses with the *R* package *e1071* (Meyer et al., 2014) and the cluster statistics with the *R* packages *fpc* (Hennig, 2014) and *mclust* (Fraley et al., 2014).

In general, it can be noted that no clusters with sizes $n < 100$ were observed. In every analysis, larger clusters basically form in the center of the performance continuum, with smaller clusters forming on the edges. To allow a comparison, the mode for the fuzzy soft clustering was formed from the degrees of class allocation for hard coding.

LPA Reference Typology and Fixed Cluster Number

In order to conduct a comparison of the cluster solutions based on different evaluation criteria, the results and the specified indices are plotted in Figure 3.

Figure 3: Absolute quality criteria (top panel) and relative measures of agreement (bottom panel) for cluster solutions with a fixed number of clusters $k = 7$ compared to the latent profile analysis as reference



Note: Since the reference typology of the latent profile analysis is assumed, only the absolute quality criteria Pearson Gamma and average silhouette width are useful and plotted here for the latent profile analysis.

The classification error indicates the relative number of misclassifications out of a fourfold table, where the reference typology of the latent profile analysis is assumed. Generally, *smaller values* in this measure are interpreted as a *better match* to the reference solution. Here, the classifications organized according to the latent profile analysis fall between 56.9% (*k*-means method) and 75.8% (hierarchical cluster analysis).

The Rand index (Rand, 1971) is a measure of the similarity of two cluster solutions or partitions (Everitt, 2011; Kaufman & Rousseeuw, 1990). It measures the share of the similar allocations. The adjusted Rand index of Hubert and Arabie (1985) is aligned for accidentally correct allocations under a null hypothesis. This adjusted index is the normalized difference of the Rand index and its expected value under the null hypothesis of a generalized hypergeometrical distribution for the confusion matrix / contingency table, which is defined for two cluster solutions Cl_i and Cl_j as:

$$Rand^*(Cl_i, Cl_j) = \frac{\left[\sum_{i=1}^{I}\sum_{j=1}^{J}\binom{n_{ij}}{2}\right] - \left[\sum_{i=1}^{I}\binom{n_{i.}}{2}\sum_{j=1}^{J}\binom{n_{.j}}{2}/\binom{n}{2}\right]}{\left[\left[\sum_{i=1}^{I}\binom{n_{i.}}{2} + \sum_{j=1}^{J}\binom{n_{.j}}{2}\right]/2\right] - \left[\sum_{i=1}^{I}\binom{n_{i.}}{2}\sum_{j=1}^{J}\binom{n_{.j}}{2}/\binom{n}{2}\right]}$$

The recommended threshold values are 0.22 for a good solution and -0.14 for an accidental correlation. Generally, *higher values* of this index represent *better similarity* of partition pairs. All cluster solutions in our analyses lie clearly above the threshold value,

and therefore the similarities are to be labeled as "good". *Rand** for the *k*-means cluster analysis lies at 0.40, the hierarchical cluster analysis at 0.54 and the fuzzy cluster analysis at 0.54. Though they all can be considered as good solutions conditional to the latent profile analysis, the hierarchical and the fuzzy cluster analysis emerge as particularly good solutions (as was the case previously in the classification error).

As an information theoretic measure of similarity between two cluster solutions or partitions, Meilă's (2007) variation of information criterion measures the quantity of lost and obtained information during the change of a cluster solution Cl_i into a different Cl_j. Formally it is represented as follows:

$$VI(Cl_i, Cl_j) = 2H(Cl_i, Cl_j) - H(Cl_i) - H(Cl_j)$$

Whereby *H(Cl.)* is the respective entropy associated with the clustering and *H(Cl$_i$, Cl$_j$)* is the joined or common entropy of the two cluster solutions. It can be written as the sum of the two conditional entropies *H(Cl$_i$|Cl$_j$)* (information loss about cluster solution *i* during the change) and *H(Cl$_j$|Cl$_i$)* (information gain about cluster solution *j* during the change) and can be interpreted as the information quantity that differentiates the two cluster solutions from one another. At $Cl_i = Cl_j$, and only then, the measure takes the value 0. Also, *VI* is a metric. Thus, *higher values* generally indicate that two cluster solutions *differ more strongly* in their information content. The highest value here is achieved with *k*-means (*VI* = 1.40), while the hierarchical cluster analysis (*VI* = 1.09) and the fuzzy method (*VI* = 1.14) take on lower, that is more advantageous, values. This is in agreement with the findings that we obtained before in the other two criteria as well.

In summary, the following ranking among the three clustering methods, which is consistent across all three criteria, can be recorded. When the cluster numbers are fixed, with LPA as the reference typology, HCA delivers the best match to the LPA solution. HCA is followed closely by the next best match of the fuzzy method, while the *k*-means method shows significantly poorer agreement with LPA.

These parameters are *relative measures of agreement* in the sense that they compare the agreement between two given classifications or partitions of an object set without falling back on the data. In contrast, it is also possible to consider *measures of absolute quality* that quantify the "match" or "quality" in the data for a given classification of an object set. The Pearson Gamma (Halkidi et al., 2001) is one such measure. For a cluster solution, it describes the correlation between the paired dissimilarities or distances and a binary 0/1 vector (0 for the same cluster and 1 for a different cluster) for object pairs, which is interpreted in the same way as the Bravais-Pearson correlation coefficient. This measure emphasizes the approximation of the dissimilarity structure by a clustering in the sense that observations in different clusters have a high correlation with greater dissimilarity or distance values. As a measure of quality, it quantifies the separation of the clusters of a solution and is generally interpreted as *"the greater the value, the better"*. All values take on characteristics between 0.42 and 0.48, whereby *k*-means and fuzzy take on comparable values, as do HCA and LPA.

The average silhouette width is a further measure of quality (Rousseeuw, 1987), and based on dissimilarity measures across all clusters of a solution, it describes on average the tightness of the connection between formed groups. This measure emphasizes the separation of clusters with their neighboring clusters. *Smaller values* of this index are

generally interpreted as *poorer quality*. According to Kaufman and Rousseeuw (1990), a value less than 0.25, as a suggested ad-hoc rule of thumb, is considered to be a "weak split". Under this criterion, the *k*-means solution (0.27) and the solution of the latent profile analysis (0.26) are of better quality, in which all values lie close to one another above the 0.25 heuristic threshold. However, the solutions of the fuzzy method and HCA are considered "weak splits" as they yield similar values below the suggested threshold value.

In summary, it can be noted that only LPA consistently takes on the greater values across the two quality criteria. All three computational methods each fall below the quality values of LPA in at least one of the two criteria.

LPA Reference Typology and Free Cluster Number

In the second step, optimized cluster numbers for the different methods were determined. The comparisons were then repeated. The optimal cluster numbers and the criterion used to select the cluster number are shown in Table 4.

Table 4: Optimal cluster numbers

Method	Criterion	Number of Clusters
k-Means	Variance ratio	7
HCA		6
Fuzzy	Convergence under minimal iteration	8

The variance ratio (VR) criterion, also referred to as the Caliński-Harabasz index (Caliński & Harabasz, 1974), is calculated like the *F* statistic. Formally this is described as follows:

$$VR_{C,N} = \frac{SS_B/(C-1)}{SS_W/(N-C)}$$

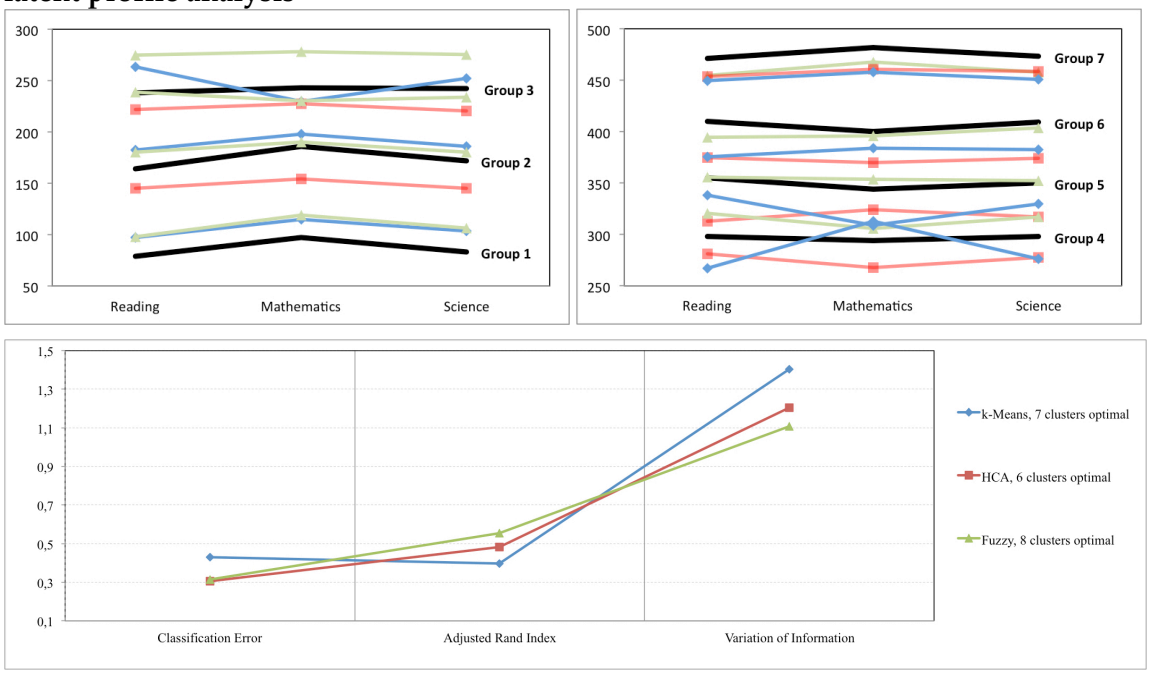
SS_B is the total variation between the C clusters and SS_W is the variation within the clusters, for a sample size of N clustered objects. The comparisons of the free cluster solutions and the classification from the LPA are shown in Figure 4, with the courses of their profile lines in the top panel and the measures of agreement in the bottom panel (the corresponding absolute measures of quality in the data will be summarized later in Figure 5).

The *k*-means method again yields a higher error rate of 43.1%, while the hierarchical and fuzzy cluster solutions are pegged as better solutions with error rates of 30.5% and 31.4%, respectively.

For each solution, an adjusted Rand index that lies significantly above the threshold value for a classification is a good cluster solution in the LPA reference model. The hierarchical cluster solution (0.48) and the fuzzy cluster solution (0.56) have the most favorable values. The variation of information for the *k*-means solution (1.40) is less

favorable than the variation of information for the hierarchical solution (1.21) and the fuzzy solution (1.11).

Figure 4: Parallel plots of the different optimal cluster solutions of performance values of pupils in Germany (top panel) and measures of agreement (bottom panel) of these cluster solutions with freely determined number of clusters compared to the latent profile analysis



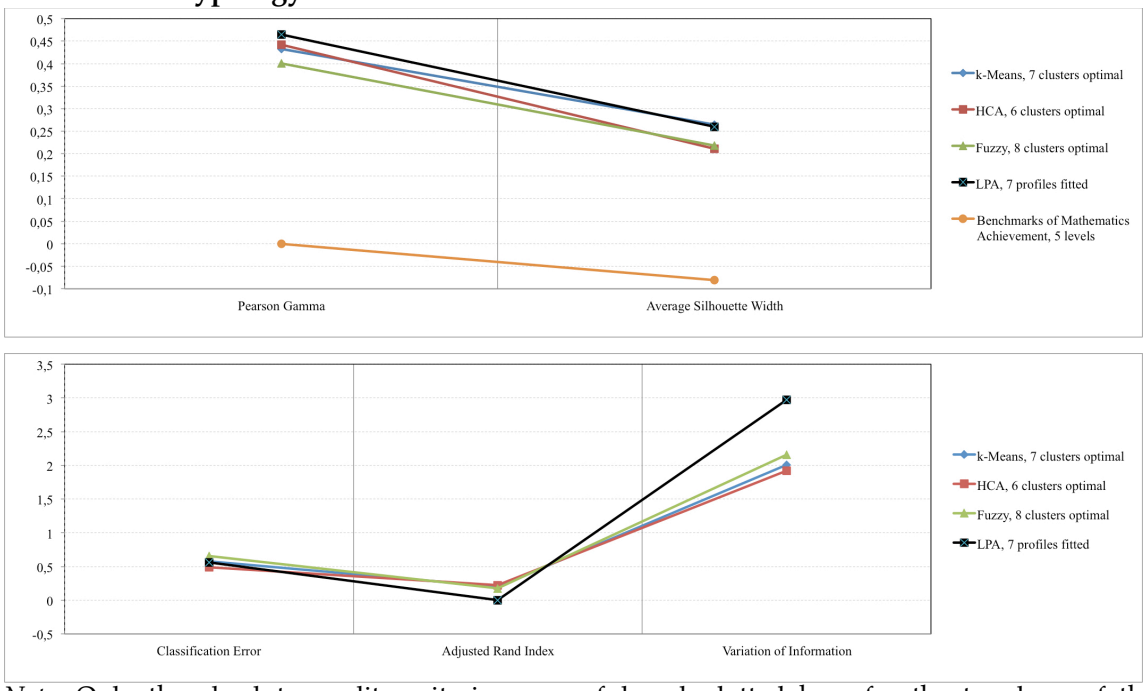
Based on the three measures of agreement, it can be noted that compared with the LPA reference typology, the allocations in the free fuzzy analysis are only minimally changing with just one more cluster. The HCA, in comparison, is getting slightly worse with one less cluster. In a free estimation of the clusters, the *k*-means method yields the previously 7 fixed clusters, which according to the LPA specification is optimal. In comparison with the hierarchical and fuzzy analyses, however, the *k*-means method comes off worse in terms of the measures of agreement. All in all, the optimal cluster numbers of the three model-free methods are around 7. Given these findings, fuzzy emerges as the best solution, followed by HCA and the *k*- means method.

The ranking of these methods is reflected in the courses of the profile lines in the top panel of Figure 4. Evidently, the invariant ranking of the clusters, which yields a parallel structure, is maintained almost completely for each of the methods. As shown, the *k*-means method with the worst match to the reference typology exhibits crisscrossing. It is also worth noting that plotted aggregated mean value profiles provide no differentiated information on the variation within a cluster. Variation bands, for example, could be computed, or the individual cases could be plotted and observed separately by means of interactive graphics. These and other strategies for "diagnostic post-analysis" of competing cluster solutions in the presented comparative education context are to be examined in further research.

Benchmarking Reference Typology

A comparison of the cluster solutions obtained in Section *LPA Reference Typology and Free Cluster Number* with the classification resulting from the benchmarking specification within the domain mathematics is presented here. The competency level allocations for mathematics from the TIMSS study are used to represent the presumed reference typology (see Section **Introduction: TIMSS and PIRLS Studies** and Figure 1). The corresponding results are shown in Figure 5.

Figure 5: Quality criteria (top panel) and measures of agreement (bottom panel) for cluster solutions with freely determined number of clusters (including LPA) compared to the benchmarking performance levels within the domain mathematics as the reference typology



Note: Only the absolute quality criteria are useful and plotted here for the typology of the mathematics performance levels.

The measures were calculated for LPA, HCA, *k*-means and fuzzy based on the specified groupings of the pupils by using all three domains and the corresponding estimated test values. For the reference typology, the computations were based on the qualitative level groups, separately specified in the domain mathematics. In the latter case, the determined mathematics performance value was used as the only characteristic variable. The absolute measures of quality ("standardized" with values in [-1, 1]) were calculated accordingly in all three domains either at the same time for the candidate methods or only in the domain mathematics for the benchmarking reference. The relative measures of agreement compare the respective partitions that are all based on the same object set, i.e., the same sample of pupils.

The results are revealing. The first striking finding is that the benchmarking partitioning of the student sample into the five TIMSS competency levels, according to the reference typology defined above, does not have high quality as a cluster solution in absolute terms (top panel of Figure 5). The measures of quality, Pearson Gamma and average silhouette width take on very small values. From a computational data-analytical

perspective, the clusters specified by the benchmarking within the domain mathematics exhibit a small amount of separation to one another. In terms of the Pearson Gamma, this means either that quite a few pupils with comparable mathematics performances in the test are categorized in different competency levels (first situation) or that pupils with different mathematical abilities are classified into the same level (second situation). As can be seen on the average silhouette width criterion, neighboring competency groups in the first situation are also not being separated sufficiently well. Quite a few cases come to lie in the boundary regions between two levels. With respect to the distribution of the pupils' performance test values, larger gaps between the "congested areas" within individual levels will appear in the second situation. In both cases, from a substantive perspective, further research should examine if it would be possible to realize differentiated partitions or contentual interpretations of the competency levels.

In comparison, the two quality measures take on noticeably high values for all statistical methods. For distance-based methods and LPA, the Pearson Gamma lies between 0.40 and 0.47, with the highest value of 0.46 associated with the LPA solution. Thus, the LPA stands out, even if only relatively (weak split), in the average silhouette width as well. The fact that the TIMSS benchmarking solution cannot be clearly retrieved in the data is obvious from the computed comparison or agreement values, which are consistently worse (bottom panel of Figure 5). The benchmarking partitioning of the pupil sample and the partitions derived from data analysis differ markedly. In comparison to the other methods, LPA stands out again with regard to its distance to the benchmarking specification. The quality of the classification specified by the LPA is the best developed across both quality measures and – presumably for this reason – it deviates the most in its agreement with the benchmark solution with regard to relative comparative criteria.

In light of these findings, it can be noted that the quality of the benchmarking clustering can and should be improved by accompanying statistical considerations, such as the additional optimization of the cluster criteria. Further research is necessary – for example, based on specific restricted latent class models and other large-scale assessment studies and data – to make reliable assertions or create valid benchmarking specifications. This can be important from an educational policy perspective.

Discussion and Summary

The subject of the present study is the testing of different cluster methods for the identification and acquisition of discrete performance profiles in large-scale school performance comparison studies, illustrated here by using the large-scale assessment studies of TIMSS and PIRLS 2011. The probabilistic model-based latent profile analysis, the distance-based computational *k*-means, hierarchical and fuzzy *c*-means cluster analyses were used on performance test values determined through plausible value generation within the framework of a three-dimensional IRT scale. The solutions obtained via the model-free computational methods were compared to the results of the latent profile analysis. In addition, the pupils grouped via the competency levels in mathematics using the official TIMSS benchmarking were compared with the groupings suggested by the cluster methods.

The classic *k*-means, hierarchical and fuzzy cluster analyses, based on the measurement error controlled performance data, yielded results comparable to the model-supported latent profile analysis. With the appropriate framework parameters, the fuzzy cluster and the hierarchical cluster analyses have yielded the best results compared with the

latent profile analysis (reference typology) across the board of a number of comparison indices. Since there are no model assumptions or complex estimation routines, the classic cluster methods appear to be practicable, easy access alternatives for the handling of complex large-scale assessment data. The presented comparison of the cluster methods is not only useful from a methodical perspective; but it is also important from a substantive, practical point of view. The results of the comparison of the quantitative cluster methods to the qualitative benchmarking specification of the large-scale assessment study have shown that the data-analytical methods lead to significantly more homogeneous groupings that exhibit better quality in the data.

Conclusion

We conclude with observations on possible future research directions in this field. Of course, the data analysis demonstrated here can be applied to other studies as well, for example PISA. Different large-scale assessment studies often survey similar competencies and traditionally are compared based on the latent correlations among their continuous performance dimensions. As a supplement to this common approach, the comparison between different large-scale assessment studies using their respective cluster partitions of the pupil sample would be interesting. We suspect this study will make it possible to comparatively investigate different large-scale assessments in terms of their similarities and differences in a more sophisticated, substantially expanded way.

Indeed, more than three dimensions, such as the sub-facets of the respective primary domains, can be clustered as well. We expect that higher dimensional solutions will yield more in-depth insights. This is because the primary dimensions of reading, mathematics and science could be assessed at a more granular level, for example based on the content-related sub-domains of their comprehension processes (PIRLS) and cognitive requirements (TIMSS). As in our study, analytical models that use more than three dimensions can be expanded with continuing or comparative cluster analyses.

As is evident from the average silhouette width, which is low throughout, the specific cluster number is hard to determine. This can be ascribed particularly to the structure of the data. In other words, the highly developed test methods allow a "tight knit" multidimensional normality. The calculated criteria for the cluster numbers often lie in tight proximity. Thus, a low number of clusters appear to be empirically plausible for the TIMSS/PIRLS 2011 data. Depending on the focus of the research, if it is theoretically well-founded, an increase of the cluster number can be appropriate. For this reason, it will be of great interest in the future to examine how much further the granularity of the clusters can be refined to identify small sub-groups or extreme groups such as "savants" or "learning disabled" in a targeted or empirically appropriate manner. The examination of such relatively rare cases could possibly piggy-back on statistical error analyses in model-supported reference typologies, and thus also allow differentiated assertions about misallocations in the partitioning.

In conclusion, we would like to note that these methods as "data analysis" have their weaknesses. Therefore, it is important to regard and understand the different methodical approaches, be they computational or model-based, as perspectives that complement one another. Although the model-based approach is the centerpiece, its utility and importance can easily be expanded with accompanying and supplementary computational analyses, as shown in the current study.

*Corresponding author**: Prof. Dr. Ali Ünlü**, ali.uenlue@tum.de.*

Acknowledgments – Dr. Ünlü wishes to sincerely thank the editor-in-chief Ryan Allen for his work on this paper. Dr. Ünlü is also deeply indebted to Edward Choi and Phoebe Doan for thoroughly reviewing the paper. Their critical and valuable comments, suggestions, and corrections have improved the work greatly.

References

Adams, R. J., Wilson, M. R., &. Wang, W. L. (1997). The multidimensional random coefficients multinominal logit model. *Applied Psychological Measurement, 21*, 1–24.

Bacher, J., Pöge, A., & Wenzig, K. (2010). *Clusteranalyse: Anwendungsorientierte Einführung in Klassifikationsverfahren*. Munich: Oldenbourg.

Bezdek, J. C. (1981). *Pattern recognition with fuzzy objective function algorithms*. Boston, MA: Springer.

Bezdek, J. C. (1983). *Advances in fuzzy information processing*. London: Academic Press.

Bos, W., Tarelli, I., Bremerich-Vos, A., & Schwippert, K. (Eds.) (2012a). *IGLU 2011: Lesekompetenzen von Grundschulkindern in Deutschland im internationalen Vergleich*. Münster: Waxmann.

Bos, W., Voss, A., & Goy, M. (2009). Leistung und Leistungsmessung. In: S. Andresen, R. Casale, T. Gabriel, R. Horlacher, S. Larcher Klee, & J. Oelkers (Eds.), *Handwörterbuch Erziehungswissenschaft* (pp. 563–576). Weinheim/Basel: Beltz.

Bos, W., Wendt, H., Köller, O., & Selter, C. (Eds.) (2012b). *TIMSS 2011: Mathematische und naturwissenschaftliche Kompetenzen von Grundschulkindern in Deutschland im internationalen Vergleich*. Münster: Waxmann.

Bos, W., Wendt, H., Ünlü, A., Valtin, R., Euen, B., Kasper, D., & Tarelli, I. (2012c). Leistungsprofile von Viertklässlerinnen und Viertklässlern in Deutschland. In: W. Bos, H. Wendt, O. Köller, & C. Selter (Eds.), *TIMSS 2011: Mathematische und naturwissenschaftliche Kompetenzen von Grundschulkindern in Deutschland im internationalen Vergleich* (pp. 269–301). Münster: Waxmann.

Caliński, T., & Harabasz, J. (1974). A dendrite method for cluster analysis. *Communications in Statistics, 3*, 1–27.

Croon, M. (1990). Latent class analysis with ordered latent classes. *British Journal of Mathematical and Statistical Psychology, 43*, 171–192.

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society Series B (Methodological), 39*, 1–38.

Ester, M., & Sander, J. (2000). *Knowledge discovery in databases: Techniken und Anwendungen der Wissensextraktion*. Berlin: Springer.

Everitt, B. (2011). *Cluster analysis*. Chichester, West Sussex: Wiley.

Fischer, G. H., & Molenaar, I. W. (Eds.) (1995). *Rasch models: Foundations, recent developments, and applications*. New York: Springer.

Forgy, E. (1965). Cluster analysis of multivariate data: Efficiency versus interpretability of classifications. *Biometrics, 21*, 768–780.

Foy, P., & O'Dywer, L. M. (2013). *TIMSS and PIRLS 2011 relationships report. Technical Appendix B – School effectiveness models and analyses.* URL timssandpirls.bc.edu/timsspirls2011/downloads/TP11_Technical_AppendixB.pdf. Retrieved January 21, 2015.

Fraley, C., Raftery, A. E., & Scrucca, L. (2014). *mclust*: Normal mixture modeling for model-based clustering, classification, and density estimation. R package version 4.4. URL CRAN.R-project.org/package=mclust.

Gibson, W. A. (1966). Latent structure analysis and test theory. In: P. F. Lazarsfeld & N. W. Henry (Eds.), *Readings in Mathematical Social Science* (pp. 78–88). Chicago: Science Research Association.

Halkidi, M., Batistakis, Y., & Vazirgiannis, M. (2001). On clustering validation techniques. *Journal of Intelligent Information Systems, 17*, 107–145.

Hennig, C. (2014). *fpc*: Flexible procedures for clustering. R package version 2.1-9. URL CRAN.R-project.org/package=fpc.

Hubert, L., & Arabie, P. (1985). Comparing partitions. *Journal of Classification, 2*, 193–218.

Kaufman, L., & Rousseeuw, P. J. (1990). *Finding groups in data: An introduction to cluster analysis*. New York: Wiley.

Lazarsfeld, P. F., & Henry, N. W. (1968). *Latent structure analysis*. New York: Houghton Mifflin.

MacQueen, J. B. (1967). Some methods for classification and analysis of multivariate observations. In: L. M. Le Cam & J. Neyman (Eds.), *Proceedings of the 5th Berkeley Symposium on Mathematical Statistics and Probability* (pp. 281–297). Berkeley: University of California Press.

Maechler, M., Rousseeuw, P., Struyf, A., Hubert, M., & Hornik, K. (2014). *cluster*: Cluster analysis basics and extensions. R package version 1.15.3. URL CRAN.R-project.org/package=cluster.

Martin, M. O., & Mullis, I. V. S. (Eds.) (2013). *TIMSS and PIRLS 2011: Relationships among reading, mathematics and science achievement at the fourth grade. Implications for early learning*. Chestnut Hill, MA: TIMSS & PIRLS International Study Center.

Meilă, M. (2007). Comparing clusterings: An information based distance. *Journal of Multivariate Analysis, 98*, 873–895.

Meyer, D., Dimitriadou, E., Hornik, K., Weingessel, A., & Leisch, F. (2014). *e1071*: Misc functions of the Department of Statistics (e1071), TU Wien. R package version 1.6-4. URL CRAN.R-project.org/package=e1071.

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

OECD (2010). *PISA 2009 results: Overcoming social background – Equity learning opportunities and outcomes.* Paris: OECD Publishing.

R Core Team (2014). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing.

Rand, W. M. (1971). Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical Association, 66*, 846–850.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests.* Copenhagen: Nielsen & Lydiche.

Reckase, M. D. (2009). *Multidimensional item response theory.* New York: Springer.

Rost, J. (2004). *Lehrbuch Testtheorie – Testkonstruktion*. Bern: Huber.

Rousseeuw, P. J. (1987). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics, 20*, 53–65.

van der Linden, W. J., & Hambleton, R. K. (Eds.) (1997). *Handbook of modern item response theory.* New York: Springer.

Vermunt, J. M., & Magidson, J. (2005). *Latent GOLD 5.0*. Belmont, MA: Statistical Innovations.

von Davier, M., Gonzalez, E., &. Mislevy, R. (2009). What are plausible values and why are they useful? *IERI Monograph Series: Issues and Methodologies in Large Scale Assessments, 2*, 9–36.

Ward, J. H. (1963). Hierarchical grouping to optimize an objective function. *Journal of the American Statistical Association, 58*, 236–244.

Wendt, H., Tarelli, I., Bos, W., Frey, K., & Vennemann, M. (2012). Ziele, Anlage und Durchführung der Trends in International Mathematics and Science Study (TIMSS 2011). In: W. Bos, H. Wendt, O. Köller, & C. Selter (Eds.), *TIMSS 2011: Mathematische und naturwissenschaftliche Kompetenzen von Grundschulkindern in Deutschland im internationalen Vergleich* (pp. 27–68). Münster: Waxmann.

Wu, M. L., Adams, R. J., Wilson, M. R., & Haldane, S. (2007). *ACER ConQuest 2.0. Generalised item response modelling software.* Camberwell: Australian Council for Educational Research.