European Journal of Science and Mathematics Education
Vol. 7, No. 2, 2019 , 83-92

Profile analyses as feedback by evaluating the balance in exam scores

Monika Vaheoja,[1]* N. D. Verhelst,[2] and T.J.H.M. Eggen[3]
[1]Association of Applied Universities in the Netherlands, The Hague, The Netherlands;
[2]Eurometrics, The Netherlands;
[3]BMS-OMD, University of Twente, Enschede, The Netherlands
For correspondence: vaheoja@10vdl.nl

Abstract
In this article, the authors applied profile analysis to Maths exam data to demonstrate how different exam forms, differing in difficulty and length, can be reported and easily interpreted. The results were presented for different groups of participants and for different institutions in different Maths domains by evaluating the balance. Some significant unbalanced profiles were found based on pre-education and institutions without ranking.

Keywords: profile analysis; test score evaluation; assessment results

Introduction

In computer-administrated tests where item banks provide questions, students with different abilities answer a different set of items, often varying in difficulty, discrimination and number of items. Instructors need to consider all these aspects when interpreting the students' total score. This aspect is even more important in terms of providing feedback on students' strengths and weaknesses. Most importantly, these comparisons should be straightforward. While there are different methods forgiving feedback on test scores in large-scale assessments, these usually involve complicated statistical models, such as multidimensional IRT models, and the results are hardly accessible to a large, non-specialised audience. (Reckase, 2009).As Li and De Luca (2014)have noted, the process of feedback should include cross-disciplinary and cross-section dialogues that provide more holistic pictures of assessment feedback. In other words, the results from any assessment should be easy assessable and functional, for only then can such results be used to encourage frank and judicious dialogue between students and teachers (Bennet, 2011). Thereafter, they mayeven be used as an input to evaluate the need for curriculum adjustments.

Verhelst (2011, 2017) proposed a profile analysis for evaluating scores without ranking the student, wherein results from different test forms can be combined. In profile analysis, the strengths and weaknesses of students on different parts of the exam are analysed by evaluating the balance: are the results for an individual, a small group or large groups, as well as for different meaningful parts of the exams, in balance with each other? *In balance* means that a particular student scores higher on easier parts of the exam and lower on more difficult parts of the exam, as can be expected from the students' total score. As this method is suitable for any number of groups, researchers can combine different grouping variables for even more detailed results (Verhelst, 2017). The innovation of this particular profile analysis is the possibility of aggregating the profiles of all respondents based on background variables, thereby combining these variables and statistically comparing the deviations. Regardless of the complexity of the

Table 1. An Example of the Observed, Expected and Deviation Profile Scores

	Part 1 of the exam	Part 2 of the exam	Total score
Observed profile	4.000	2.000	6
Expected profile	4.406	1.594	6
Deviation profile	-0.406	+0.406	0

statistics in the computation of the scores, we will demonstrate that the results are straightforward.

In this article, the authors show an application wherein groups of students are composed based on three background variables (i.e., number of attempts taking the exam, pre-education and teacher-training institutes) on national exam data for teacher-trainees in the Netherlands coordinated by 10voordeleraar (*www.10voordeleraar.nl*).[1]Here, we are interested in identifying special needs groups to master basic knowledge in Mathematics.

Profile Analysis

Profile analysis is a technique that analyses the partial scores of an individual or a group of examinees on an exam (Verhelst, 2011, 2017). A profile is a set of partial scores of an individual student on meaningful parts of the exam, i.e. partial scores on a few categories of items (see Table 1). A profile can either be observed or expected. The observed profile is the sequence of actual partial scores of the examinee, while the sum of the partial scores is the total score of the examinee. The expected profile is a sequence of partial scores: the average or expectation of what a student will obtain in each category, given his or her total score. This average or expectation is a mathematical function of the parameters of the IRT measurement model that has been used to calibrate the exams. The details of this function will not be discussed; instead, the only important thing to remember is that the sum of the expected scores equals the sum of the observed scores. The deviation profile is the difference between the observed and expected profile. The interpretation of the deviation profile in Table 1 is clear, wherein the respondent performed worse than expected on items in part 1 and better than expected on items in part 2.

The researcher is often interested in explaining the differences in deviation profiles via background variables. Therefore, individual profiles as shown in Table 1 should be aggregated to average profiles based on the grouping variables. To statistically evaluate these deviation profiles, the researcher uses software to compute these aggregated conditional expected scores, thus providing a variance-covariance matrix to compute the standard errors of these aggregated means. This software to run the profile analysis, Profile-G, is available (see Verhelst, 2011).

Aggregating the scores across different exam versions. Computer-based assessments often use different versions of the exams. While it is possible that the number of items varies across versions, this does not complicate the interpretation of the results of profile analysis, for the interpretation of the average profiles across different exams, even with a different set of questions, remains the same. The validity of the interpretation of a profile analysis depends on the justification of the partition of the item sets in the exam. For instance,

Table 2. List of All Profiles Resulting in a Test Score of Six

S1	S2	p	S1 × p	S2 × p	S1 ×S1 × p

[1]*10voordeleraar* is translated as "An A (highest grade) for the teacher"

0	6	<0.0005	0.000	0.000	0.000
1	5	0.002	0.002	0.010	0.002
2	4	0.026	0.052	0.104	0.104
3	3	0.141	0.423	0.423	1.269
4	2	0.348	1.392	0.696	5.568
5	1	0.361	1.805	0.361	9.025
6	0	0.122	0.732	0.000	4.392
	sum	1.000	4.406	1.594	20.360

Suppose that the items of a mathematics exam are either algebra or geometry; then, the meaningfulness of the partition emerges in correctly identifying all items in all exam forms or belonging to the algebra and geometry category.

Statistics of the profile analysis. The profile analysis is carried out in the following way: for a given total test score and a given partition of the items in categories, all possible partial score combinations are generated, and, for each profile, its probability is computed. Table 2 includes a small example of an exam with two item categories and a test score equal to six.

The column in Table 2, labelled as S1 and S2, gives partial scores for the first and second category, respectively. In the column labelled (S1 *x p*), the outcomes computing the expected score for S1are shown. The expected score for S2 follows readily, as the sum of both expected scores must equal the given test score (6 - 4.406= 1.594). In the rightmost column, labelled (*S1 xS1 xp*), the outcomes are shown for computing the variance of partial scores for the first category *S1*(20.360 – (4.406)[2] = 0.947). Expected scores and variances for other test scores can be computed similarly; this also holds for cases where there are more than two categories.[2]

The average deviation for a category *j* for some group *G* of a number of test takers $n_G$is then given by:

$$\bar{d}_{jG} = \frac{1}{n_G}\sum_{v\in G}\left[S_{jv} - E\left(S_{jv} \mid S_v\right)\right] \qquad (1)$$

Where S_v is the total test score of test taker v, and S_{jv} is the partial score on category j of v. For each test taker, we have the variance of the *j*th component of the profile, which is denoted by σ^2_{jv}.

The variance of $\bar{d}_{jG}$ is given by:

$$Var(\bar{d}_{jG}) = \frac{\sum_{v\in G}\sigma^2_{jv}}{m^2_G} \qquad (2)$$

Moreover, the standard error of $\bar{d}_{jG}$ is the square root of the variance.
If the group size is not too small, we can use the central limit theorem and assume that the average deviation is normally distributed. The null hypothesis is that the average deviation is zero: the test takers in group *G* behave on average as predicted by the measurement model with a test statistic:

$$z_{jG} = \frac{\bar{d}_{jG} - 0}{\sqrt{Var(\bar{d}_{jG})}} \qquad (3)$$

in which the denominator is the standard error (SE) of the average deviation$\bar{d}_{jG}$. Rejection of the null hypothesis takes place if the test statistic is $|z_{jG}| > 1.96$ (two-sided, with a significance level of 5%).

[2] One should not be misled by the simplicity of this example. In some cases, the number of possible profiles associated with a given test score can be quite large, often more than 5000.

The Example

We applied the profile analysis four times on the calibrated exam responses based on three background variables. Two of the background variables are measured on the student level and can be used to identify the strengths and weaknesses of a special needs group. The third variable is the institution, demonstrating a way in which the profile analysis can be used to evaluate large-scale assessment results. We also combined two background variables for an in-depth evaluation, analysing the exams with OPLM software to calibrate the exams and to estimate the item difficulty parameters and discrimination indices needed to compute the individual expected scores (Verhelst & Glas, 1995; Verhelst, Glas & Verstralen, 1993).

The context of the data. The data used in this example are from the computer-administrated mandatory exams from the public secondary school Maths teacher-training programs in the Netherlands (www.10voordeleraar.nl). The exams are based on a blueprint with a fixed number of items per subdomain. Table 3 shows the number of items for five content domains: analysis, geometry, algebra, statistics and other. Because of the small number of items in the 'Other' category, only the first four domains are used in the categorisation of the items for the profile analysis. The sequence number corresponds to the calendar time of the exam's administration, exam 1 being the first. The length of the exams was shortened after exam 6.The exams share a common set of items making the number of unique items across 11 exams equal to 490.

Background variables. The researchers used three background variables to group and evaluate the profiles on the content categories: pre-education, retakes and the institution. These background variables emerge from respondents' self-reports and are filled in before the examination begins. The background variable pre-education has six categories, *Havo*: Senior general secondary education; *Mbo*: Senior secondary vocational educating; *Hbo*: University of applied sciences; *Vwo*: University preparatory education; *Wo*: Master's degree and *Other*: such as foreign pre-education or adult education arrangement. Retake has two categories. First attempt*:* taking the exam for the first time and retake: respondent taking the exam for the second or more times.[3] The background variable institution has nine different institutions.

Table 3. The Number of Items per Categorisations Based on Content Domains

	Number of items for each domain					
	Analysis	Geometry	Algebra	Statistics	*Other*	Total
Exam 1	16	13	14	12	*5*	60
Exam 2	17	13	14	12	*4*	60
Exam 3	17	13	14	12	*4*	60
Exam 4	17	13	14	12	*4*	60
Exam 5	17	13	14	12	*4*	60
Exam 6	17	13	14	10	*6*	60
Exam 7	16	10	12	8	*4*	50
Exam 8	17	10	12	9	*2*	50
Exam 9	16	10	12	8	*4*	50
Exam 10	17	10	12	9	*2*	50
Exam 11	17	10	12	9	*2*	50
Unique items in total	149	98	116	92	*35*	490

[3] Note that if a student has had a second attempt, his or her response for the first attempt is also in the data file; in the analyses, such a student is treated as two (or more) different students.

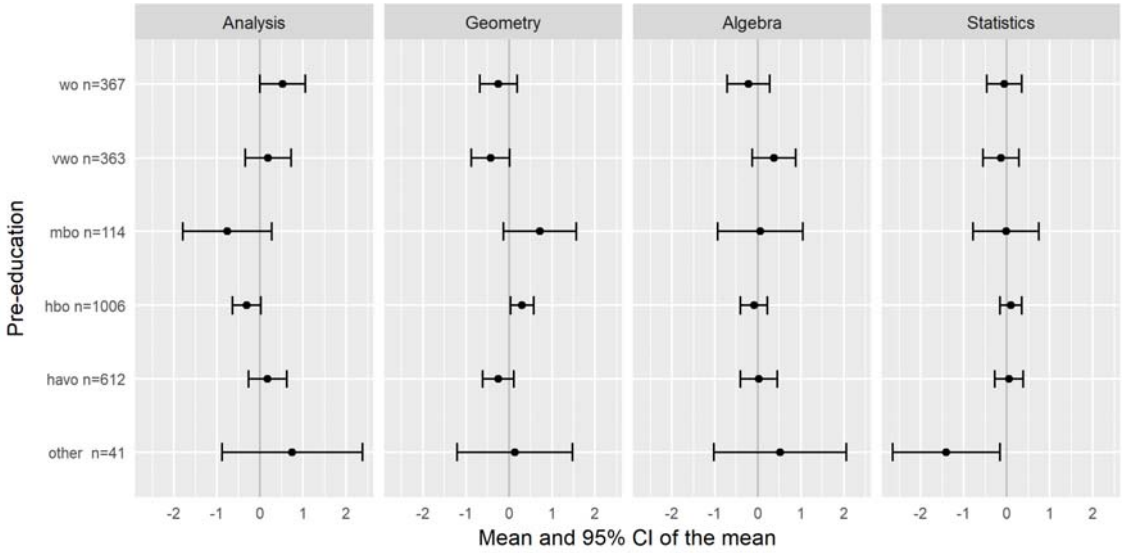

Figure 1. Average deviation profiles on Maths domains based on pre-education. Note: The dots represent the average of a given group, and the error bars are the 95% confidence intervals of the average.

Results

This section presents the results of four profile analysis: pre-education, retake, and different institutions of teacher-traineeships. As last are the results presented in which the variable pre-education and retake are combined. The results are presented for aggregated deviation profiles: the differences between expected and observed profiles.

Average deviation profiles for pre-education. Figure 1 shows the examinee's aggregated deviation profiles on Maths domains based on pre-education. The average deviance profiles are given on the x-axes. To evaluate the statistical significance of the aggregated deviation from zero, only a visual inspection of the figures needed. If the confidence interval does not include zero, the mean difference deviates significantly from it.

Students with *hbo* and other pre-education show a significant imbalance in their profiles. The *hbo* pre-education students score significantly higher ($z = 2.21$; $p < .05$) on geometry ($M = 0.306; se = 0.138$) and tend to score lower on analysis than can be expected based on their total score on the exam. The students with otherpre-education score significantly lower on statistics than expected:

($M = -1.406$; $se = 0.636$; $z = -2.21.$; $p < .05$).

Average deviation profiles for first attempt and retake. Figure 2showsthe examinee's aggregated deviation profiles on Maths exam based on the attempt on the exam. None of the profiles deviated significantly from zero: first-take students tend to score higher on domains on which retake students score lower and vice versa. However, this is a mathematical accessity of the deviance scores, because the sum of the average deviations for each content domain is zero.

Average deviation profiles for different institutions. Table 4 shows the examinee's aggregated deviation profiles on Maths domains for different teacher-training institutes in which six

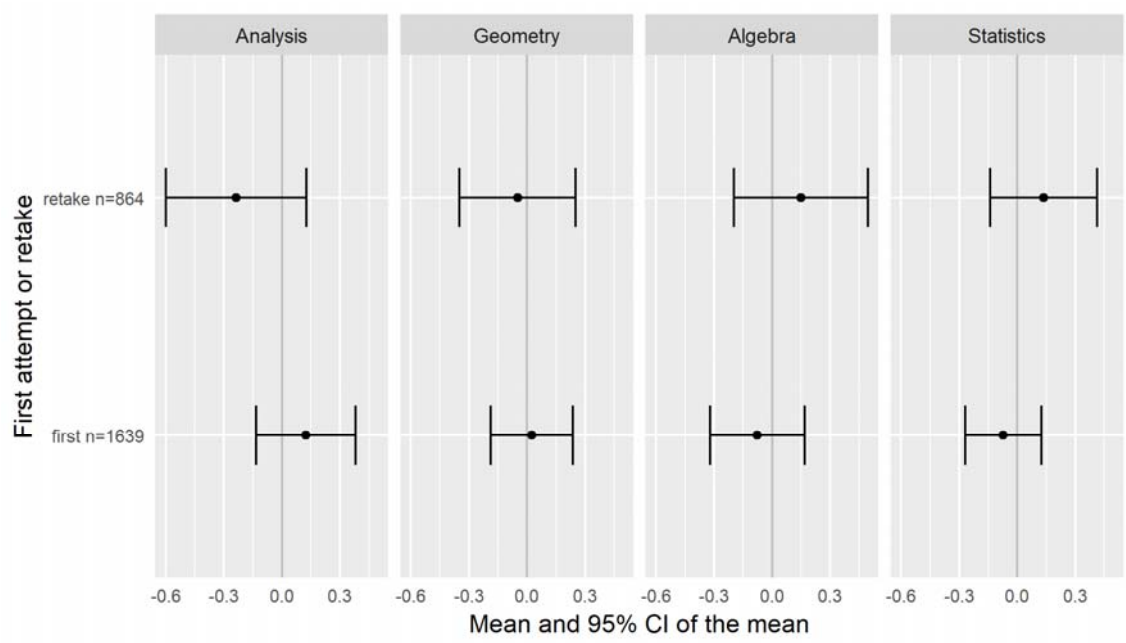

Figure 2. Average deviation profiles on Maths domains based on an attempt on the exam.

Table 4. Mean and SE of Deviation Profiles on Maths Domains for Different Institutions

	Maths domains			
	Analysis	Geometry	Algebra	Statistics
	Mean (SE)	*Mean (SE)*	*Mean (SE)*	*Mean (SE)*
A	0.212 (0.312)	-0.528 (0.257)*	0.723 (0.299)*	-0.406 (0.237)
B	-0.862 (0.210)***	0.373 (0.173)*	0.449 (0.199)*	0.040 (0.163)
C	-0.234 (0.272)	0.427 (0.224)	0.061 (0.257)	-0.254 (0.207)
D	1.087 (0.499)*	-0.851 (0.415)*	-0.872 (0.474)	0.636 (0.379)
E	0.666 (0.301)*	0.120 (0.248)	-0.454 (0.289)	-0.331 (0.226)
F	-0.303 (0.292)	-0.072 (0.241)	-0.232 (0.277)	0.608 (0.226)*
G	1.706 (0.388)***	-0.538 (0.321)	-0.967 (0.370)**	-0.201 (0.295)
H	-0.129 (0.408)	-0.210 (0.339)	-0.002 (0.386)	0.341 (0.312)
I	0.162 (0.822)	-0.089 (0.687)	-0.372 (0.780)	0.299 (0.648)

* $p< 0.05$; ** $p< 0.01$; *** $p< 0.001$

institutes show significant imbalanced profiles. To treat the institutions anonymously, the number of students per institution is not shown. Institution B shows a quite strong imbalance in Figure 3, as three of the four average deviations are significant, wherein the observed score is much lower than expected given their total scores ($M = -0.862; se = 0.210; z = 4.093; p < .001$). The institutions A, D and G show two significant average deviations in their profiles and institutions E and F have only one significant average deviation in their profiles.

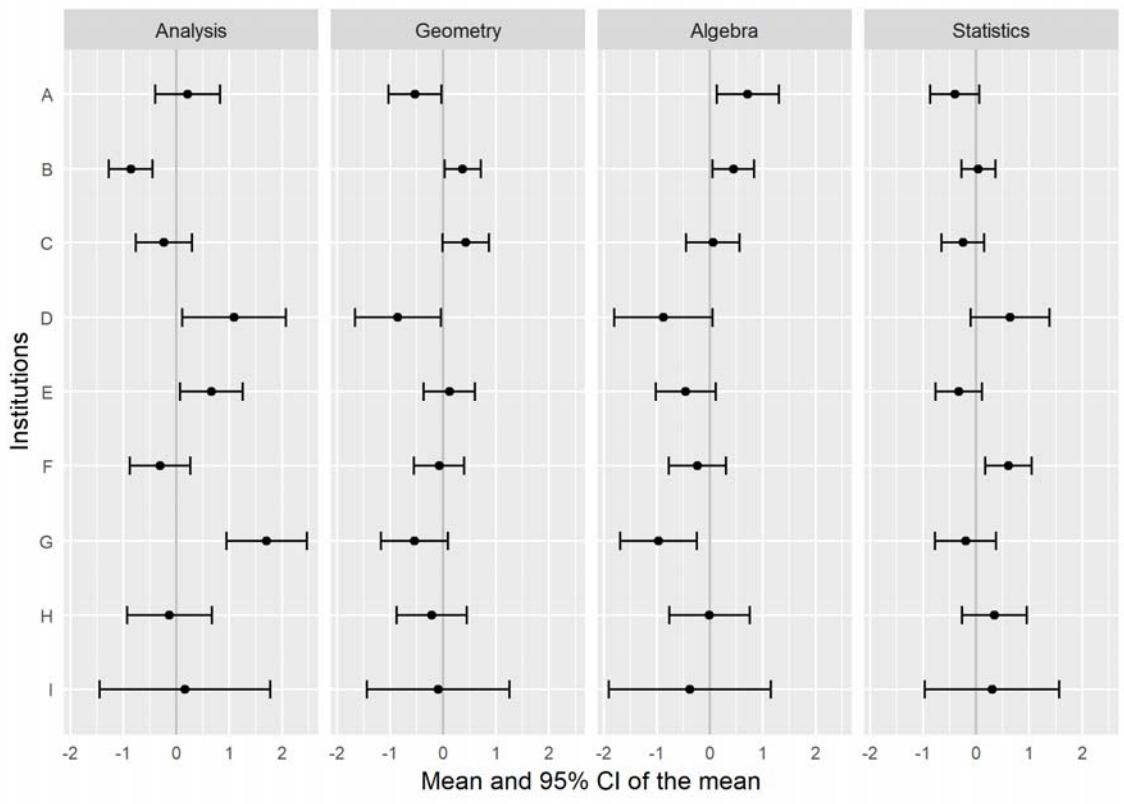

Figure 3. Average deviation profiles on Maths domains for different institutions.

Average deviation profiles for pre-education and attempt. Previously, we showed the main effects of the three grouping variables on four Maths domains. In the following, we show the results of pre-education and attempt as combined grouping variables.

Table 5 shows the average profiles given for first and retake students for each pre-education group. The amount of groups is now 12 (2 x 6), and the groups are much smaller. In the upper part of the table are the profiles for first-take students and, in the lower part, are the profiles for retake students. Even though there was no significant imbalance found for background variable attempt on the exam, this imbalance is identified when the attempt is combined with pre-education. There are three significant unbalanced profiles: *wo*-first take, *hbo*-first take and students with the *mbo*-retake group. Students with *wo* pre-education score significantly higher on analysis, wherein students with *mbo*-retake score significantly lower.

However, if we are interested in the significant difference between the first attempt and retake group for different pre-education groups, we need to compute the difference between the average profiles. For an easy comparison of the results, the differences between profiles for attempt are given in Figure 4. A positive difference indicates a higher deviation score for the first-take group; a negative difference indicates a higher deviation score for the retake group. The confidence interval of the difference between first and retake deviation profiles from zero first-take group; a negative difference indicates a higher deviation score for the retake group. The confidence interval of the difference between first and retake deviation profiles from zeroare given too.[4] There is only one significant difference in deviation profiles, in which students with wo pre-education from the retake group tend to score higher on domain statistics. An interesting aspect is also that students with *mbo* and other

[4] To compute the standard error of the difference, we used the statistical rule that the variance of the difference of two means is just the sum of the two variances.

pre-education from the retake group tend to score lower on domain analysis, which might point out a weakness.

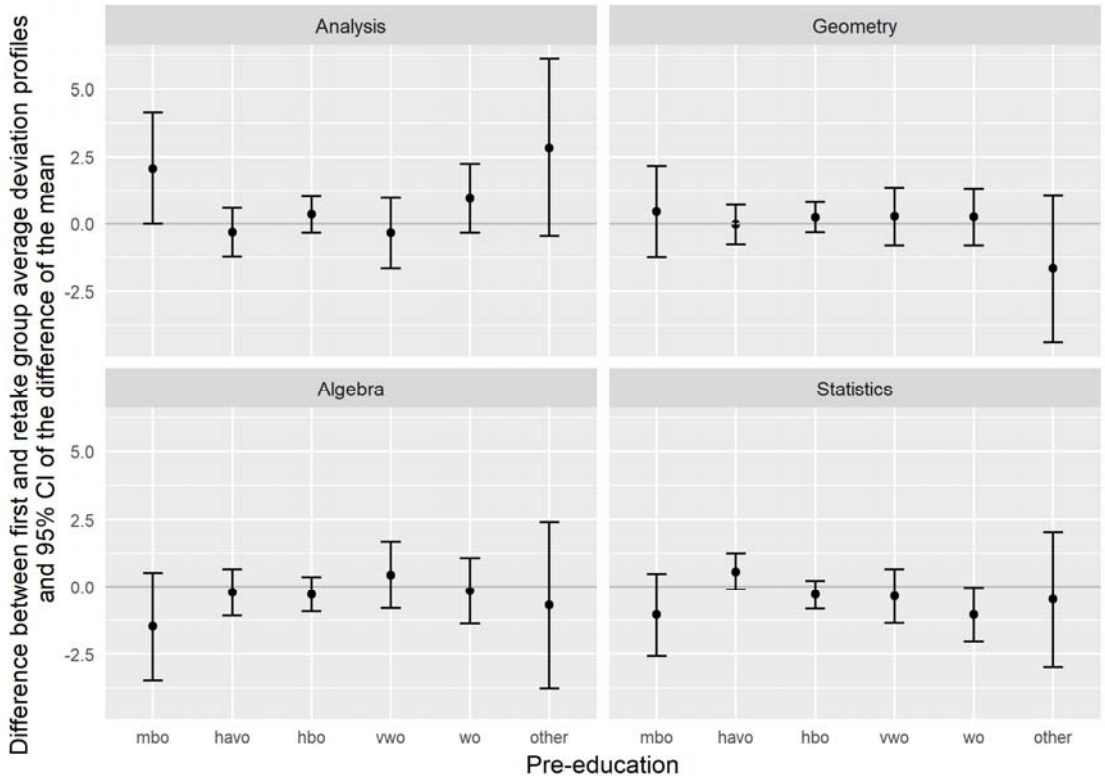

Figure 4. Difference between deviation profiles of first-take and retake students on Maths domains for different pre-education groups.

Table 5.*Mean and SE of Deviation Profiles on Maths Domains for Pre-Education and Attempt*

			Mathsdomains			
Pre-education			Analysis	Geometry	Algebra	Statistics
		n	*Mean (SE)*	*Mean (SE)*	*Mean (SE)*	*Mean (SE)*
First attempt	havo	365	0.052 (0.295)	-0.262 (0.243)	-0.069 (0.284)	0.279 (0.220)
	mbo	61	0.199 (0.720)	0.936 (0.588)	-0.632 (0.685)	-0.503 (0.532)
	hbo	627	-0.180 (0.210)	0.396 (0.173) *	-0.199 (0.199)	-0.017 (0.162)
	vwo	279	0.109 (0.309)	-0.366 (0.255)	0.475 (0.292)	-0.218 (0.240)
	wo	285	0.737 (0.301)*	-0.189 (0.249)	-0.257 (0.284)	-0.292 (0.236)
	else	22	2.059 (1.113)	-0.634 (0.918)	0.202 (1.040)	-1.627 (0.854)
Retake	havo	247	0.363 (0.354)	-0.231 (0.292)	0.150 (0.340)	-0.282 (0.267)
	mbo	53	-1.872 (0.775) *	0.476 (0.637)	0.850 (0.752)	0.546 (0.570)
	hbo	379	-0.529 (0.276)	0.159 (0.229)	0.087 (0.261)	0.284 (0.213)
	vwo	84	0.454 (0.594)	-0.636 (0.491)	0.045 (0.565)	0.138 (0.452)
	wo	82	-0.214 (0.588)	-0.440 (0.486)	-0.093 (0.553)	0.748 (0.452)
	else	19	-0.767 (1.252)	1.028 (1.033)	0.887 (1.188)	-1.149 (0.951)

* $p < 0.05$

Conclusion and Discussion

In this article, the authors applied profile analysis (Verhelst, 2011, 2017) toMaths exam data for teacher-trainees in the Netherlands to demonstrate how different exam formsin large-scale

assessments can be reported and easily interpreted. The results were presented for different groups of participants on different partsof exams by evaluating the balance.

The researchers used three background variables: pre-education, number of exam attempts and different teacher-training institutions. Also, two combined background variables demonstrated an in-depth evaluation. Significant unbalances in profiles were found for variable pre-education and institutions. When the attempt was combined with pre-education, the authors identified oneextra unbalanced profile.

The application of the profile analysis in the current article demonstrates its strength in educational assessment purposes. Despite the complicated computations of the conditional scores and the variance-covariance matrices, interpretation of the results is straightforward. A significant imbalance in the deviance profile suggest a potential need for further investigation of this aspect, which, in turn, might lead to curriculum adjustment. Some domains may have too little time scheduled but actually need more course time for mastery. Moreover, when the profiles from different institutions are to be compared, such data may be used as an input to encourage discussion between different institutions about the design of curricula.

Even though this study presents the current results for different background variables, these results can also be presented on the individual level. Deviation profiles on an individual level can be helpful for students wishing to balance their test scores based on their own strengths and weaknesses. In addition, these findings can be helpful in encouraging dialogue between the student and the teacher. As the results show, students with *Other* and *hbo* as pre-education showed a significant imbalance in their deviation profiles, implying that these students missed some courses. Providing more training for these students in domains where the significant weaknesses are detected might help to balance their profiles.

However, more balance in the average scores does not automatically mean that these students have obtained a minimally competent level of performance. The focus in profile analysis is on the balance of the exams' partial scores, given the total performance on the exam. There is no information given about one performance relative to others; this also means that no ranking between different groups is made. Moreover, exam scores can be evaluated independently of the cut-score set on exam forms. The focus in the profile analysis is not on the percentage of passing performance, but, instead, on the balance. Profile analysis is therefore an excellent method by which to provide feedback to participants, or group of participants, of their whole performance on an exam based on the total score. From a psychometrical perspective, profile analysis is also a generalisation of differential item functioning, wherein the focus is not only on one particular item,but also on a set of items measuring the same construct or items that have the same characteristic (Yildirim, Yildirim & Verhelst, 2014). For example, it might be that students from different backgrounds answer differently on items with numerical answers than multiple-choice items. Such a hypothesis can be answered more generally by grouping items based on these characteristics and by evaluating the deviance profiles of the students, thus providing a lot of information for the test constructors.

For realistic applications, further research is needed to evaluate how the participants receive the results of profile analysis. In particular, researchers should determine if these results are indeed helpful in gaining more insight into exam performance and if the practitioners from different fields find the results straightforward to interpret.Moreover, it might be that the way the results are demonstrated in the current article is the most straightforward way to present the results of different computer-based exam forms whereby results from different exam forms may be evaluated.

References

Bennet, R. E. (2011). Formative assessment: A critical review. *Assessment in Education: Principles, Policy & Practice*, 18 (1), 5 -25.

Li, J.,& De Luca, R. (2014). Review of assessment feedback. *Studies in Higher Education*, 39, 378-393.

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

Verhelst, N. D. (2011). Profile analysis: A closer look at the PISA-2000 reading data. *Scandinavian Journal of Educational Research*, 56, 315- 332.

Verhelst, N. D. (2017). Balance: A neglected aspect of reporting exam results. In M. Rosén et al. (Eds.), *Cognitive abilities and educational outcomes* (pp. 273-293). Springer International Publishing.

Verhelst, N.D., &Glas, C.A.W. (1995). The one parameter logistic model. In G. H. Fischer & I. W. Molenaar (Eds.),*Raschmodels*. New York, NY: Springer.

Verhelst, N. D., Glas, C. A. W., & Verstralen, H. H. F. M. (1993). *OPLM: One parameter logistic model [Computer software manual].* Arnhem.

Yildirim, H.M., Yildirim, S. & Verhelst, N.D. (2014). Profile analysis as a generalized differential item functioning analysis method. *Educational and Science*, *39*, 49-64.