The Alberta Journal of Educational Research *Vol. XLV, No. 4, Winter 1999, 377-391*

Philip Nagy

and

Randall Penfield

The Ontario Institute for Studies in Education of the University of Toronto

A Procedure for Detecting Student Profile Patterns in a Performance Assessment

This study investigates student score profiles of the mathematics component of the 1997 Ontario grade 3 assessment. In addition to an overall score, students are given scores on three knowledge or skill dimensions, and five scores on content strands. The purpose of this investigation was threefold: (a) to assess the extent to which student profiles contain differentially diagnostic information, (b) to examine classroom-level patterns in the student profiles, and (c) to develop alternative methods of analyzing profile data to gain classroom-level diagnostic information. The results show that 70% of the students have the same score on all three knowledge/skill categories (flat profiles) and thus provide no differentially diagnostic information. The profiles for the remaining 30% of the students consisted almost exclusively of contoured profiles in which there was a difference of only one unit between one of the categories and the other two. Using algorithms developed in this article, these profiles were used to assess the relative strengths and weaknesses at the classroom level, as well as examining within-classroom diversity. This investigation found that nearly 60% of classrooms had a relative weakness in comprehension, and 25% of classrooms displayed substantial diversity. Examination of profiles of the five content strands indicated that nearly 69% of the students also had either flat or contoured profiles, whereas the remainder had more complex patterns. Methodologies for interpreting relative strengths and weaknesses on the mathematics strands at the student and the classroom level are discussed.

Cette recherche étudie le rendement d'élèves de troisième année en mathématiques à l'examen provincial en Ontario en 1997. Les élèves y obtiennent une note globale, une note sur trois catégories visant les connaissances ou les habiletés, et cinq notes pour le contenu. La présente étude avait trois buts: (a) évaluer jusqu'à quel point les profils d'élèves contiennent de l'information diagnostique différentielle, (b) discerner les patterns à l'échelle de la salle de classe à partir des profils d'élèves et (c) développer des méthodes alternatives d'analyser les données du rendement pour permettre la cueillette d'information diagnostique à l'échelle de la salle de classe. Les résultats indiquent que 70% des élèves ont eu la même note sur les trois catégories visant les connaissances ou les habiletés (les profils plats); ces résultats ne fournissent donc aucune information diagnostique différentielle. Les profils de l'autre 30% des élèves consistaient presque exclusivement de profils courbés où il y avait une différence d'une unité entre une des catégories et les deux autres. A l'aide d'algorithmes développés pendant l'étude, on s'est servi de ces profils pour évaluer les forces et les faiblesses relatives à l'échelle de la salle de classe et pour étudier la diversité à l'intérieur de la salle de classe. L'étude a révélé que dans presque 60% des salles de classe, il y avait une faiblesse relative au niveau de la compréhension et qu'une diversité considérable existait dans 25% des salles de classe. L'étude des cinq notes pour le contenu a indiqué que presque 69% des élèves avaient un profil qui était soit plat ou courbé, alors que les autres avaient des profils plus complexes. On

---

Philip Nagy is a professor in the Measurement and Evaluation program. His interests are in large-scale assessment and data analysis.

Randall Penfield is a doctoral candidate in the Measurement and Evaluation program. His interests are in psychometrics and educational measurement.

discute de méthodologies servant à interpréter les forces et les faiblesses relatives en mathématiques à l'échelle de l'individu et à l'échelle de la salle de classe.

Introduction

In recent years, large-scale assessment has moved away from multiple-choice tests to more complex performance assessments. One reason for this shift has been criticism of the narrowing curricular impact of multiple-choice testing, much of which can be construed as variations on the theme of inappropriate teaching to the test (Cannell, 1988; Madaus, 1991; Moore, 1994; National Commission on Testing and Public Policy, 1990). Unlike multiple-choice items, performance assessments are seen as *authentic* in their relationship to educational goals.[1] The argument is that if the assessment tasks are truly what we want students to learn, then teaching to the test is turned from a liability to a benefit.

Performance assessments take the form of tasks more in keeping with classroom and real-world demands than simply choosing one of several options to a multiple-choice question. The tasks are usually paper-and-pencil, although laboratory and other manipulative skills are sometimes required. They are often embedded in the curriculum and require more time to administer, but proponents argue that the additional time is instructionally beneficial to students. One important aspect of such assessments is that they require expensive grading by panels of experts, usually teachers. Thus issues of costs versus benefits require close scrutiny (Hardy, 1995).

Elaboration of this issue requires some historical context. Performance (and portfolio) assessments grew out of the movement to understand and improve classroom testing (Wiggins, 1993). They were seen as tools useful in the classroom, in the hands of teachers who already knew the students well, and as "one more piece of information" about the student. In this context they were promoted and continue to be accepted as powerful techniques, but as techniques more of formative assessment and instruction than of summative assessment. As one tool in the ongoing conversation between teacher and student, they were treated, by and large rightly, as not subject to the requirements of basic measurement principles. That is, when coupled with additional information about the student in the teacher's hands, the danger of taking unwise, especially summative, action on inaccurate information was minimized.

Performance assessments were borrowed from this within-classroom context by the large-scale testing movement, and once removed from that context they lost their immunity from examination using the traditional canons of measurement (Bateson, 1993). In the ensuing few years, a variety of technical concerns were raised: problems of low generalizability (Brennan & Johnson, 1995; Gao, Shavelson, & Baxter, 1994); score stability over time (Ruiz-Primo, Baxter, & Shavelson, 1993); and accuracy of individual and even group average scores (Cronbach, Linn, Brennan, & Haertel, 1997).

Beyond technical issues, Black (1994) reported that large-scale performance assessments could falter in the face of teacher resistance, just as multiple-choice large-scale assessments have. Shepard et al. (1996) and Firestone, Mayrowetz, and Fairman (1998) found that the hoped-for beneficial impact on curriculum

was minimal. In summary, the switch from multiple-choice testing has solved some problems, but others have emerged, and the hoped-for benefits appear elusive. Hence cost benefit issues continue to be a concern.

Another context in which multiple-choice testing and performance assessment can be examined are the questions of breadth versus depth and instructional utility. Large-scale multiple-choice testing programs have been suggested as able to yield instructionally diagnostic information. However, if a multiple-choice test covers a substantial amount of curriculum, then the number of items possible on any topic is too small for reliable individual diagnosis of a specific difficulty.[2] For example, a test broad enough to cover the mathematics curriculum for an entire grade could not contain more than a few items on anything as specific as, say, addition of fractions, although collection of items under a broader umbrella, perhaps involving skills rather than content, would ameliorate this problem. Although such a subtest does not usually contain enough items to be highly reliable at the individual level, supporters of such tests argue that at the group (e.g., classroom) level, even a few items can provide useful diagnostic information.

So if results from a three-item subtest on, again, addition of fractions show sufficiently narrow confidence intervals at the classroom level, the argument is that we can draw conclusions about classroom performance on this subtopic. The counterargument is that such conclusions need to be limited specifically to performance on the actual items used and not to a more general category; that is, to addition of these three specific items rather than to addition of fractions in general. This seems to be the crux of the breadth-depth issue with respect to large-scale multiple-choice tests of the type described and of their potential for instructional diagnosis.

The purpose of this study was to examine the potential benefit of performance assessments in providing group-level instructional diagnosis. Scoring systems typically used in performance assessments do not offer the seductive lure of great specificity, for example, addition of fractions. Instead, designers of performance assessments ask scorers to approach student work from a variety of perspectives. For example, the Ontario assessment system under examination produces a score for each of three knowledge/skill categories, *Understanding*, *Applying*, and *Communicating*, on a 4-point scale.

In the rationale for this analysis, we used a limited definition of diagnosis. In keeping with Gipps, Brown, McCallum, and McAlister (1995), we recognize that few teachers, especially good teachers, are particularly surprised by the results of large-scale assessments. Teachers tend to know which students are doing well and which are doing less well. Our definition of diagnosis focuses on those students who show uneven (contoured, in our terminology) rather than even (flat) achievement profiles. Our argument is that large-scale assessment scores are most useful to teachers when they show different levels of skill in different categories of outcomes. In effect our claim is that a flat profile tells a teacher nothing the teacher did not already know, and if large-scale assessment is to contribute to understanding of the student's achievement, it will be more likely to occur when the profile is contoured.

We begin with 3-element vectors representing the scores of each student on these three categories, such as 132 (1 for Understanding, 3 for Applying, and 2

for Communicating). Our argument is that, although flat profiles (111, 222, 333, 444) reveal each student's general level of performance, they do not provide differentially diagnostic information (i.e., a pattern of relative strengths and weaknesses). On the other hand, a pattern such as 112 reveals a relative strength in one area (in this case, Communicating). Our analysis focuses on these contoured profiles, looking for classroom-level patterns in relative strength and weakness. We explore ways to examine trends (i.e., numbers of students in a class showing the same relative strength or weakness) and diversity (i.e., classrooms with a number of students showing strength in an area, and a number of students showing weakness in the same area).

Two different aspects of these classroom vectors can be examined, overall trend and internal diversity. If a teacher has a large number of students with weakness in Applying, and a small number with strength in Applying, then there is a clear trend that the teacher can try to deal with at the classroom level. However, if a teacher has a number of students with weakness in Applying, and a similar number with strength in Applying, then there is some serious diversity that will require considerably more individualization and be much more difficult to deal with. Although all good teachers engage in individualization, our concern is to shed light on the extent to which it is necessary. We sought ways to examine each of these situations separately.

We downplay the fact that a pattern of 111 tells a teacher something different than a pattern of 333 on the grounds that this is something the teacher was more probably already aware of and that such flat patterns do not provide the teacher with diagnostic information on *differences* in relative strengths and weaknesses. The downside of our decision is that patterns showing the same relative strength are treated the same (e.g., 112, 223, 334 all show a relative strength in Communicating), setting aside the information conveyed by the overall levels of performance.

The designers of external tests are always clear that scores on individuals should not be used in the absence of other information about the student, a position that holds as well for performance assessment as for any other assessment (Cronbach et al., 1997; Shavelson, Baxter, & Gao 1993). So we address our questions primarily at the classroom level of data aggregation, and tangentially at the individual level as we aggregate the data.

Although there is a small literature on profile analysis, it does not address performance assessment. The more common definition of a profile arises from a test battery context in which scores from test components are compared (Mehrens & Lehmann, 1991; Sax, 1997). The literature offers little discussion of the use of profiles in interpreting assessment data. Hills (1993) briefly discusses pitfalls often occurring in profile interpretation, and several studies have addressed methods to estimate the reliability of a profile of measures (Rae, 1991; Yarnold, 1984). However, we have been unable to locate any discussion of the extent to which profiles provide differentially diagnostic information at the student, class, or higher level.

Our first set of questions asks the extent to which individual students have flat (e.g., all 3s) or contoured (e.g., a mixture of different values) profiles. Our second set of questions asks whether there are classroom-level patterns in these student profiles. Although there are reasons to be concerned about the

reliability of individual profiles, group results are more reliable, and we take a pattern showing many students in a class with the same strength or deficit as having potential diagnostic value.

Some brief caveats are in order. First, we recognize the benefits of individual student test score information used in conjunction with everything else the teacher knows about the student, even though score reliability might not be all that is hoped for. We also recognize that good teachers always make an attempt to individualize, even though our focus is on group instructional diagnosis. Further, it is not our intention to deal with the substance of remediation of difficulties. Our focus is on the presence and magnitude of differential patterns in the achievement results only. Finally, this analysis is exploratory; one of our goals is to uncover alternative ways of looking at the data.

Data

In the spring of 1997 the Ontario Education Quality and Accountability Office (EQAO) administered to all 130,000 grade 3 students in the province a two-week curriculum-embedded assessment that produced several scores in each of reading, writing, and mathematics. The scores for mathematics are examined in this study. Each student received an overall mathematics score. In addition scores were reported for the following three knowledge/skill categories, Understanding, Applying, and Communicating (U, A, and C) and for the following five content strands: Numeration, Geometry, Measurement, Patterning, and Data Management.

Performance on each category and strand was reported using a four-point anchored scale. For example, the four points for the example knowledge/skill category *Applying procedures* were:

- 1—shows *basic* application of some procedures in *attempting* some simple tasks;
- 2—shows *some* application of procedures to *complete* simple tasks with *accuracy;*
- 3—shows *regular* application of procedures to *complete* tasks with *accuracy;* and
- 4—shows *regular* application of a *wide variety* of *complex* procedures to produce *accurate and complex* responses.

For the descriptions of the four levels of the Geometry and spatial sense strand, the levels were:

- 1—shows *basic geometry concepts and procedures* in some simple tasks;
- 2—shows *some required geometry concepts and procedures* in simple tasks;
- 3—shows *required geometry concepts and procedures* to complete tasks of *some complexity with accuracy;* and
- 4—shows, *beyond what is required, a variety of geometry concepts and procedures* to complete *complex tasks accurately.*

In addition to the four-point scales, students were also categorized as *exempt* (various exceptionalities including ESL) and *no data* (including absenteeism and blank and unscorable responses). These two categories were treated in this analysis as missing data. Our focus is first on the set of three knowledge/skill scores. Then we treat the set of five strand scores in a more preliminary fashion, briefly exploring their possibilities in comparison with the knowledge/skill scores.

Method and Results

The analysis was completed in several stages, with the procedures of each subsequent stage dependent on the results of the previous stage. Thus the results are presented together with the method for each stage of the analysis.

Formation of Individual Score Vectors

The first step in examining the scores was to combine each student's set of three scores for Understanding, Applying, and Communicating in mathematics into a single 3-digit vector described earlier. A missing value on any of the three resulted in a missing value for the vector.

Data from 120,315 grade 3 students were then summarized, revealing that 69.3% of the students had *flat* profiles (approximately 13%, 32%, 21%, and 3% at each of levels 1 through 4); another 30.5% had *contoured* profiles consisting only of adjacent values (e.g., 434, 112). Only 0.22% (265 students) had *unbalanced* profiles consisting of either nonadjacent numbers (e.g., 313) or three different numbers (e.g., 423). In all, 47 of the 64 (4^3) possible combinations appeared in the data. No students had both 1s and 4s.

In order to move from the individual to the classroom level, we needed some decision rules. Any classroom-level differential patterning that might be gleaned from the individual vectors lay in the 30% of cases that had contoured profiles, rather than the large number with flat profiles or the trivially small number with unbalanced profiles. All flat profiles were coded as 0, and the few unbalanced profiles were treated as missing data. The remaining contoured profiles were categorized as showing either a relative strength (one category up, two down) or relative weakness (two categories up, one down) in a category. Using this system, 99.8% of the students could be classified into one of seven independent categories, flat or 0 (which was not used in the analysis), plus these six:

- +U, strength in understanding (211, 322, 433);
- –U, weakness in understanding (122, 233, 344);
- +A, strength in applications (121, 232, 343);
- –A, weakness in applications (212, 323, 434);
- +C, strength in communicating (112 223, 334); and
- –C, weakness in communicating (221, 332, 443).

Note again that the coding for each student is independent of the student's overall performance in that, for example, 443 and 221 are treated the same, relatively weak in Communicating.[3] We then formed a classroom profile consisting of a vector of seven numbers, each representing the number of students in each classroom in each of the categories 0, +U, –U, +A, –A, +C, and –C.

Construction of Classroom Vectors

In order to search for patterns and frequencies at the classroom level, we had to deal with variations in classroom size. We could have changed the 7-digit vector from counts to proportions, but that would have caused us to lose class size information, weight the data from small classrooms too heavily, and possibly introduce distortions. Another possibility, also rejected, was to eliminate classes below some given size, such as 10 students. The problem with this strategy is that just over half the grade 3 students in Ontario in 1996-1997 were in split classes, divided about equally between 2-3 and 3-4 splits. A

decision to eliminate smaller classes would thus eliminate a disproportionate share of the students in split classes. There seems good reason to suspect that patterns of achievement of students in straight grade 3 classes, 2-3 splits, and 3-4 splits might be systematically different, so we decided to keep all the classes. However, the procedures described below are much more likely to reveal patterns in larger classes than in smaller.

To get a general understanding of our data distribution, we summed the number of contoured profiles so that we had for each class a number of flat profiles (0) and a number of contoured profiles (categories +U, –U, +A, –A, +C, and –C added together). Table 1 shows a cross-tabulation of these numbers. For example, in the top left entry, 89 classes had 0-2 students with flat profiles and 0-2 students with contoured profiles. Again, almost in the center of the table, 205 classes have 15-16 students with flat profiles and 7-8 students with contoured profiles. Within the limits of our category widths, every class size[4] in the province can be read from Table 1. The body of Table 1 has been broken into nine regions for ease of discussion.

Consider first the column on the left. From the column totals, 44.8% of the classrooms have four or fewer contoured profiles on which to base a search for differentially diagnostic information. The classrooms in the first row are smaller, varying in size from one to 14. Further, although they contain a substantial *proportion* of students with contoured profiles, there is little likelihood of finding diagnostic information in data from these classes. There simply are not enough contoured profiles in the classes in the first row to show a pattern. Those classes in the second and third rows (left hand column) are larger. Unlike the first row, the proportion of contoured profiles is small. Consequently, like the classes in the first row, there is little chance of uncovering diagnostic information.

The right column of Table 1 contains the classes with the largest numbers of contoured profiles. As shown by the sum of the subcolumn totals, 5.9% of the classrooms have at least 11 students with a contoured profile. The first and third rows are essentially empty. In contrast, the number of classrooms in the second row is large, and the chances of finding potentially diagnostic information in these classes are excellent. There are enough contoured profiles that, when we break them down by the three categories of Understanding, Applying, and Communicating (our next step), useful patterns are likely.

The middle column of Table 1 includes 49.1% of the classes. Those in the top row, some 800 or 12% of classes, offer the best chance of finding diagnostic information, but the middle group, with about one third of students showing a contoured profile, also show promise. The bottom group, with some 300 classes, has a smaller proportion of contoured profiles, and thus less chance of finding diagnostic information.

The analysis in Table 1 shows that despite the fact that only 30.5% of students have contoured profiles, it is possible to identify significant numbers of classes with more than their share of this 30.5%. That is, students with contoured profiles are distributed unevenly across classrooms. If a diagonal line is drawn from top left to bottom right of Table 1, those classes above the line are the more likely to contain sufficient numbers of students with contoured profiles to show classroom-level patterns.

Table 1
Numbers of Classes at Combinations of Flat and Contoured Profiles

Contoured / Flat	0-2	3-4	5-6	7-8	9-10	11-12	13-14	15-16	Total	Total %
0-2	89	49	10	3	0	0	0	0	236	3.3
3-4	258	173	49	13	1	1	0	0	495	7.1
5-6	312	251	109	35	7	1	1	0	716	10.2
7-8	271	262	180	56	18	8	1	0	796	11.4
9-10	169	245	175	75	88	12	7	0	732	10.5
11-12	109	173	181	117	121	54	12	5	730	10.5
13-14	75	134	158	172	48	66	20	6	751	10.8
15-16	38	97	165	205	166	70	19	5	765	10.9
17-18	21	92	206	235	134	64	8	1	761	10.9
19-20	11	92	161	170	93	32	6	0	565	8.1
21-22	8	54	85	79	38	7	2	0	273	3.9
23-24	6	39	44	25	6	2	0	0	122	1.7
25-26	3	8	8	3	1	0	0	0	23	.4
27-28	1	9	2	1	0	0	0	0	13	.2
Total	1,454	1,678	1,534	1,189	711	317	76	17	6,978	100
Total %	20.8	24.0	21.9	17.1	10.1	4.6	1.1	.2	100	

Classroom Indices of Strength and Weakness

The next step in the analysis moved beyond aggregates of contoured profiles to the specifics of relative strengths and weaknesses. Although the analysis coming out of Table 1 might suggest that further work focus on only a subset of classrooms, because this is an exploratory examination we decided to proceed with all classes.

Table 2 shows percentages of classes as a function of numbers of class members with each of the six possible contoured profiles. For example, the +U column gives information for profiles strongest in Understanding: 211, 322, or 433. Some 30% of classes have no such students, whereas about 16% have three or more such students. Table 2 reflects the fact that overall achievement is highest in Understanding and lowest in Communicating. More than half the classes have no students at all with a profile showing relative strength in Communicating (112, 223, or 334), whereas one quarter of classes have three or more students with a relative weakness in Communicating (221, 332, 443). These differences are quite large when viewed in the context that only 30.5% of the students have contoured profiles.

Seeking Trends

To look for trends, the number of students in the category –U was subtracted from the number in +U, yielding a factor, ΔU, that is positive if a classroom has a larger number of students relatively strong in Understanding compared with those relatively weak. The same was done for the two categories of Applying (+ and –) and the two of Communicating (+ and –). In this way, each classroom

Table 2
Percentages of Classes Containing Numbers of Students with Each Indicator Profile

Percent of Classes	*+U Profiles*	*–U Profiles*	*+A Profiles*	*–A Profiles*	*+C Profiles*	*–C Profiles*
0 Students	30.6%	72.3%	52.9%	48.8%	56.5%	27.1%
1 Student	31.7%	22.8%	31.1%	32.4%	29.9%	31.2%
2 Students	20.9%	4.4%	11.5%	13.0%	9.7%	20.7%
3 Students	10.3%	.4%	3.4%	4.3%	3.2%	11.9%
4 Students	4.2%	.1%	.8%	1.2%	.6%	5.6%
5 Students	1.5%	0.0%	.1%	.2%	.1%	2.3%
6 Students	.6%	—	.1%	.1%	—	.8%
>6 Students	.1%	—	0.0%	—	0.0%	.4%

was assigned a 3-element vector (ΔU, ΔA, ΔC) based on the contoured profiles only.

For example, consider a classroom of 22 students, containing 12 students with flat profiles and 10 with contoured profiles, giving a total class distribution of (0, +U, –U, +A, –A, +C, –C) = (12, 4, 2, 1, 0, 1, 2). On conversion to a 3-digit profile, this becomes (2, 1, –1), meaning that in this class there are slight trends to relatively high scores on Understanding and Application and relatively low scores on Communication. That is, students with relative strength in U and A outnumber those with relative weakness, while those with relative weakness in C outnumber students with relative strength.

Preliminary investigation of patterns among these 3-digit classroom vectors revealed that positive ΔUs tended to go with negative ΔCs and vice versa. That is, classes that had relative strength in Communicating had relative weakness in Understanding. The ΔA values tended to be less variable; about one third of classes had ΔA = 0 (equal numbers of +A and –A students) and only 16% had ΔA > 2 or < –2.[5] There is little variability in the relative strength or weakness in Application at the class average level. We therefore decided to focus our attention on the relationship between Understanding and Communicating.

Table 3 shows the cross-tabulation of ΔU versus ΔC. Negative ΔC values were observed for 55.9% of classrooms, indicating that there was at least one more student with a weakness in Communicating (221, 332, 443) than with a strength. Of these classrooms, 35.2% possessed a relative strength in Understanding. On the right hand side, 15% of classrooms have a relative strength in Communicating, and one eighth of these, almost 2% of the entire set, also have a relative weakness in Understanding. There appears to be the potential for drawing useful classroom-level instructional implications from these data.[6]

Seeking Diversity

To look for diversity in classrooms another view was required. The procedure described above cannot distinguish between a class with no students at either +A or –A and one with three students at +A and three at –A, as both yield a trend score of 0. To deal with diversity, each pair of values for each classroom was examined: +U and –U, +A and –A, +C and –C. The lower of each pair was

Table 3
Cross-tabulation, Percentages of Classrooms by ΔU and ΔC

	$\Delta C < 0$	$\Delta C = 0$	$\Delta C > 0$	*Total*
	4.7%	2.5%	1.8%	8.9%
$\Delta U < 0$				
$\Delta U = 0$	16.1%	10.8%	4.8%	31.6%
$\Delta U > 0$	35.2%	15.6%	8.6%	59.4%
Total	55.9%	28.9%	15.1%	100.0%

taken and summed across all three scores. For example, with 22 students in a classroom, the values (12, 4, 2, 1, 0, 1, 2) yield 2 + 0 + 1 = 3; that is, min (4, 2) + min (1, 0) + min (1, 2). This value provides a measure of diversity in the classroom.

The principle behind this measure is simple. Consider pairs of elements for +U and –U from two different classrooms (4, 2) versus (2, 0). Both yield a ΔU score of +2, indicating two more students with relatively high scores on Understanding. However, in the first case the teacher must contend with four students relatively strong in Understanding and two relatively weak, whereas in the second case the teacher has no students relatively weak in Understanding. The two diversity scores of +2 and 0 reflect this difference. One of these two teachers is faced with considerably more individualization.

The diversity scores for all classes are reported in Table 4.[7] Although we cannot determine whether this analysis has uncovered classes with highly divergent student populations, about 25% of classes have a diversity measure of 2 or larger. These classes are more likely to require greater individualization on the part of the teacher.

A skeptical view of the above is that we have ignored reliability issues and, even worse, relied heavily on differences between unreliable numbers. We have made classroom-level recommendations based on data from an unknown subset of the classroom members who did not have flat profiles. This may be true, and it can be examined, at least on a case basis. We isolated individual profiles from the members of three classes, chosen nonrandomly, in an attempt to exemplify the approach.

The first class comes from the subsample of classes with positive ΔU and negative ΔC values (third row, first column, Table 3). In this class 28 students had 15 flat profiles, 11 had contoured profiles, and two students did not respond to a sufficient portion of the assessment to be given scores on which to base our indices. The breakdown is as follows:

- the flat profiles are four at level 2, 10 at level 3, and one at level 4;
- there are no cases of +A, –A, +C or –U;
- there are three cases of +U, two at 433 and one at 322;
- there are eight cases of –C, five at 332 and three at 221.

These results, taken together, show a high-achieving class. Roughly a third of the class have some relative problems with Communicating, and a few have relative strength with Understanding. There are no cases of students showing

Table 4
Frequencies and Percentages of Classes at each Diversity Level

Diversity Measure	*Number of Classes*	*Percent of Classes*
0	3,003	43.0
1	2,216	31.8
2	1,152	16.5
3	426	6.1
4	148	2.1
5	28	.4
6	5	.1

a relative strength in Communicating. Group-instructional implications, although admittedly general, are relatively clear.

The second example classroom comes from the small number of classrooms with negative ΔU and positive ΔC (first row, third column, Table 3). There are 15 flat profiles and eight contoured in the class. The breakdown is as follows:

- the flat profiles are three at level 2, 10 at level 3, and two at level 4;
- there are no cases of +A, –A or –C;
- there is one case of +U, 433, and four cases of –U, two at 233 and two at 344;
- there are three cases of +C, one at 223 and two at 334.

Again, this appears to be a high-achieving class. In classrooms like this, group instructional implications are not so clear-cut. The number of students identified as having a relative weakness in understanding is smaller and balanced by one student with a relative strength.

The third example classroom comes from the small group of classrooms with high diversity scores, as shown in the sixth row of Table 4. In this last class there are 30 students; 11 have flat profiles, 18 have contoured profiles, and one student's score is missing. The breakdown is as follows:

- the flat profiles are five at each of levels 2 and 3, and one at level 4;
- all six possibilities of contoured profiles exist;
- there are four cases of –U (233), balanced with one case of +U, 322;
- there five cases of –A, one at 212, three at 323, and one at 434, balanced by one case of +A at 232;
- there are four cases of –C, three at 332 and one at 443, balanced by three cases of +C two at 223 and one at 334.

Not only is this a highly diverse class, but it is also on average much lower achieving than the previous two examples. Although there are no simple diagnostic prescriptions for a class like the example class, the data give a clear signal that considerable individualization would be required.

Strands

One could make a case that the content strand profiles hold more potential for differential instructional implications than do the knowledge/skill profiles. We conclude this report with some preliminary work on the content profiles. For the content strands, scores for each student on the five content strands of *Numeration, Geometry, Measurement, Patterning,* and *Data Management* were

similarly converted into a 5-digit vector so that, for example, a vector of 32334 meant that a student had 2 on Geometry, 4 on Data Management, and 3s on each of Numeration, Measurement, and Patterning. In this case 498 of the 1,024 (4^5) possible combinations appeared in the data.

Again, a missing value on any strand resulted in a missing value on the overall vector. With five rather than three scores required for the content strand, there were more cases of missing data for the content strands, 9.1% versus 7.1%.

In total, 117,706 students yielded complete data. Of these, 31.6% had flat profiles (e.g., 22222), and another 61.9% had profiles consisting of adjacent values only, such as 22333 or 43444. The remaining 6.5% of students had unbalanced profiles. With five pieces of information to work from rather than three, our decision rules had to be different and more complex.

Coding the Strands

An important difference between the 3-digit knowledge/skill vectors and the 5-digit content vectors is that in the first case, each student could be classified into one of only seven independent categories. Following the same procedure with the 5-digit strands was not possible, as we needed to allow for the possibility of strength or weakness in more than one strand. We decided to categorize each student as relatively strong, average, or weak (+1, 0, or –1) on each of the five strands and then to examine each strand in turn rather than all of them together as in the approach for the skill categories.

Our procedure is illustrated using the second position in the 5-element vector, but our argument is symmetrical for all five strands. Conceptually, our procedure compares each strand score to the mean of the other four, although we actually worked by comparing four times each score with the total of the other four. Consider the two cases 34333 and 44333. We want each of these to result in a +1 for the second content strand Geometry. That is, if a strand has the highest score in the vector, even if tied with one other strand, our methods should give a positive value. Four times the Geometry score, 16, is always greater by at least 3 than the sum of the remaining four strands, which will have a maximum of 13. The sum of the remaining four is 13 for 44333, 12 for 34333, and less if any of the 3s are changed to a lower score. The argument is parallel for 33222 and 22111. If we had 44334, then this case would be scored –1 for the third and fourth strands, and 0 for the other three. The rule for a score of +1, then, is that four times the strand score must exceed the sum of the other four strands by 3 or more.

Consider the two cases 21222 and 11222. We want each of these to result in a score of –1 for Geometry. That is, if a strand has the lowest score, even if tied with one other, our methods should give a negative value. Four times the Geometry score gives 4, which will always be less than the sum of the remaining four by at least three. The sum of the remaining strands is 8 for 21222 and 7 for 11222; the difference is greater than 3 if any of the 2s become larger. The same argument holds for 22333 and 33444. So the algorithm is that if four times a score is less than the sum of the other four by at least 3, the result is –1. If it is greater than the sum of the other four by at least 3, the result is +1. If it is within two of the sum of the other four, the result is 0.[8]

Table 5
Frequencies of Classes by Numbers of Geometry Scores

+1 / −1	0	1-2	3-4	5-6	7-8	9-10	Total
0	595	760	220	51	13	1	1,642
1-2	969	1,545	593	142	74	4	3,276
3-4	412	750	291	60	3	1	1,517
5-6	118	214	62	9	1		404
7-8	38	55	8	1			102
9-10	17	11					29
11-12	1	1	1				3
Total	2,152	3,336	1,176	263	40	6	6,973

At the individual level, 9% of students have a relative weakness in Geometry (–1 as described above) and 11% of cases show a relative strength (+1) . The remaining 80% of students receive a value of 0 according to the algorithm and thus are deemed to have an achievement level in Geometry close to their overall achievement.

To investigate the distribution of these cases across classrooms, we counted the number of cases of +1 and –1 scores in Geometry in each class and tabulated these numbers in Table 5. With only 20% of the students having a nonzero score in Geometry, about 10% of classes have no students at all with scores of either +1 or –1 in Geometry. Apart from this group, the top left hand corner of Table 5, about 1,000 classes (the top row) have clear strengths in geometry, and another 1,500 classes (left hand column) have clear weaknesses.

In general, classes on the right hand side of the table have strengths in Geometry, and those toward the bottom of the table have weaknesses. As we move into the center of the table, and especially toward the bottom right (where, happily, there are few classes), we find the classes with high diversity. Clearly the content strand data reveals potential for yielding diagnostic information.

Summary

This is a preliminary attempt to develop methodology to identify classroom-level patterns in student profiles. The analysis was based on considering each student's score in relation to his or her other scores in the set. In the set of three knowledge/skill scores, some systematic patterns could be found at the classroom level. In the set of five content strand scores, classroom-level differences in relative strength were found in one example strand.

Further work on the content strands is planned. Patterns of strength and weakness in combinations of content areas need to be examined, and levels of error need to be discussed.

Notes

1. This point can be debated.
2. Multiple matrix sampling, in which not all students respond to all items, can solve the breadth-depth issue, but at the price of no longer yielding individual level information.

3. Other decisions could have been made. For example, data from lower-achieving students could be given more importance, or relative weaknesses could be treated as more important than relative strengths. One could also choose to take into account the level of the flat profiles.
4. For split classes we have included only the number of grade 3 students.
5. The question of whether some of these ΔAs of 0 consist of a +A of, say, 6 and a –A of 6 is addressed below when we discuss diversity. Briefly, the answer is no; almost 98% of classes have both +A and –A Δ 3.
6. Table 2 reports data for schools at all levels of ΔA. To examine the possibility of different patterns at different levels of ΔA, we repeated the analysis for schools at each of the five levels of ΔA separately. With minor exceptions results were parallel to Table 2.
7. The diversity index reported is in the spirit of exploratory data analysis. However, we also calculated traditional within-class variances on the student-level raw data. The two approaches correlate 0.93.
8. This algorithm is not perfect. It gives counterintuitive results in extreme cases such as 41111. Refinements of the algorithm are being considered for further work.

Acknowledgment

We wish to thank the Education Quality and Accountability Office for access to the data for this report.

References

Bateson, D. (1993). Psychometric and philosophic problems in "authentic" assessment: Performance tasks and portfolios. *Alberta Journal of Educational Research, 40*, 233-245.

Black, P.J. (1994). Performance assessment and accountability: The experience in England and Wales. *Educational Evaluation and Policy Analysis, 16*, 191-203.

Brennan, R.L., & Johnson, E.G. (1995). Generalizability of performance assessments. *Educational Measurement: Issues and Practice, 14*(4), 9-12.

Cannell, J.J. (1988). Nationally normed elementary achievement testing in America's public schools: How all 50 states are above the national average. *Educational Measurement Issues and Practice, 7*(2), 5-9.

Cronbach, L.J., Linn, R.L., Brennan, R.L., & Haertel, E.H. (1997). Generalizability analysis for performance assessments of student achievement or school effectiveness. *Educational and Psychological Measurement, 57*, 373-399.

Firestone, W.A., Mayrowetz, D., & Fairman, J. (1998). Performance-based assessment and instructional change: The effects of testing in Maine and Maryland. *Educational Evaluation and Policy Analysis, 20*, 95-113.

Gao, X., Shavelson, R.J., & Baxter, G.P. (1994). Generalizability of large-scale performance assessments in science: Promises and problems. *Applied Measurement in Education, 7*, 323-342.

Gipps, C., Brown, M., McCallum, B., & McAlister, S. (1995). *Intuition or evidence?* Milton Keynes, UK: Open University Press.

Hardy, R.A. (1995). Examining the costs of performance assessment. *Applied Measurement in Education, 8*(2), 121-134.

Hills, J.R. (1993). Interpreting profiles. *Educational Measurement: Issues and Practice, 12*(4), 26, 32-33.

Madaus, G.F. (1991). The effects of important tests on students. *Phi Delta Kappan, 72*, 226-231.

Mehrens, W.A., & Lehmann, I.J. (1991). *Measurement and evaluation in education and psychology* (4th ed.). Toronto, ON: Holt, Rinehart & Winston.

Moore, W.P. (1994). The devaluation of standardized testing: One district's response to a mandated test. *Applied Measurement in Education, 7*, 343-367.

National Commission on Testing and Public Policy. (1990). *From gatekeeper to gateway: Transforming testing in America.* Boston, MA: National Commission on Testing and Public Policy, Boston College.

Rae, G. (1991). Another look at the reliability of a profile. *Educational and Psychological Measurement, 51*, 89-93.

Ruiz-Primo, M.A., Baxter, G.P., & Shavelson, R.J. (1993). On the stability of performance assessments. *Journal of Educational Measurement, 30*, 41-53.

Sax, G. (1997). *Principles of educational and psychological measurement and evaluation* (4th ed.). Toronto, ON: Wadsworth.

Shavelson, R.J., Baxter, G.P., & Gao, X. (1993). Sampling variability in performance assessment. *Journal of Educational Measurement, 30*, 215-232.

Shepard, L.A., Flexer, R.J., Hiebert, E.H., Marion, S.F., Mayfield, V., & Weston, T.J. (1996). Effects of introducing classroom performance assessments on student learning. *Educational Measurement: Issues and Practice, 15*(3), 7-18.

Wiggins, G. (1993). *Assessing student performance: Exploring the purpose and limits of testing*. San Francisco, CA: Jossey-Bass.

Yarnold, P.R. (1984). The reliability of a profile. *Educational and Psychological Measurement, 44*, 49-59.