Decision-Support Data Analysis: Examining Consistency Among Teachers in Writing Assessment[*]

Doug Archbald

Abstract

Education leaders and much literature exhort teachers and school leaders to use data more often and more effectively to guide planning and decision making – called, "data driven decision making." This term is ubiquitous in literature and reform discourse, but "on the ground," so to speak, practitioners face significant challenges in analyzing, understanding, and applying data to improve practice. Obstacles faced by practitioners include insufficient expertise, tools, and time; also, organizational cultures in schools generally create few incentives for data analysis and as often as not sustain norms inimical to the collaboration and collective action required for data driven decision making. The case reported here illustrates these challenges through the actions of a principal identifying a problem of organizational culture and instructional practice and leading an initiative to promote collaboration, analysis, and reflection to help improve writing instruction in his school.

---

1 Sumario en espanol

Los líderes de la educación y mucha literatura exhortan que maestros y educan a líderes para utilizar los datos más a menudo y más indicar efectivamente planeando y la toma de decisiones – llamado, "los datos manejaron la toma de decisiones". Este término es ubicuo en el discurso de la literatura y la reforma, pero "en el suelo," tan hablar, los facultativos encaran desafíos significativos a analizar, a la comprensión, y a aplicar los datos para mejorar la práctica. Los obstáculos encarados por facultativos incluyen pericia insuficiente, las herramientas, y el tiempo; también, las culturas organizativas en escuelas crean generalmente pocos estímulos para el análisis de datos y tan a menudo como no sostiene normas hostiles a la colaboración y acción colectiva requirió para datos manejó la toma de decisiones. El caso informado aquí ilustra estos desafíos por las acciones de un director que identifica un problema de la cultura organizativa y la práctica y de dirigir instruccional una iniciativa para promover colaboración, el análisis, y la reflejo para ayudar a mejorar escribiendo instrucción en su escuela.

> NOTE: Esta es una traducción por computadora de la página web original. Se suministra como información general y no debe considerarse completa ni exacta.

2 Introduction

Education leaders and much literature exhort teachers and school leaders to use data more to guide planning and decision making – called, "data driven decision making." A substantial literature has emerged with theoretical models and practical prescriptions. Yet typical practice as shown by empirical studies still falls well short of theory-based conceptions embraced by scholars and reformers. Practitioners still face many challenges in analyzing, understanding, and applying data to improve practice.

This case illustrates an application of data analysis in service of standards-based instruction. A school principal is concerned about variable academic standards in his school, particularly in literacy instruction. Teachers have avoided for the most part collaborative planning and there has been little scrutiny or discussion of practice. Seeking a mechanism to change this culture, he organizes a benchmarking activity to examine and rate student writing in 5th grade – an activity he hopes will stimulate teacher conversations about writing, help create greater consistency among teachers in assessing writing, and show that instruction can and should be subject to systematic empirical inquiry.

This module strengthens knowledge and skills in using and analyzing assessment data and applying data analysis toward the aim of standards-based writing instruction.

3 Notes For Use As An Instructional Module

Section I. presents theory and research on data-based decision making and school leadership; the first part provides background on data-based decision making and the second part discusses challenges of school leadership aimed at supporting data inquiry to improve practice. Section I. can be read and discussed with or without supplementary reading (see reference list). Students should discuss ways to connect data with writing instruction, trying to be specific about how to achieve what they propose (e.g., how would you actually do this in your school?).

The first part of Section II. presents the methods and the case – an account of one school's approach to data analysis through a benchmarking activity in writing. The principal is concerned about a multi-year pattern of mediocre writing assessment results at his school. He wants to stimulate inquiry into practice and motivate change. This case describes a data-based benchmarking process.

This module can be read in its entirety followed by discussion, or the discussion leader can treat each section separately.

The second part of Section II. presents the analyses and results. The discussion leader should review each table and figure thoroughly so students understand each one. Some are just descriptive snapshots requiring little interpretation; others contain more information, require more interpretation, and raise discussion questions about their implications or limitations.

Section III. presents discussion questions and exercises to deepen students' understanding of the data (quantitative literacy) and to reflect on implications for leadership, professional development, and instructional change.

Section IV. provides notes for the instructor on the questions and exercises of Section III. as well as a rubric to help evaluate the module's largest assignment: proposing an analysis to supplement or expand on the analysis depicted in the case.

4 Section I.

5 Background: Theory and Research

From US Secretary of Education Arne Duncan addressing the Fourth Annual IES Research Conference (IES, 2009), "Robust Data Gives Us The Roadmap to Reform:"

5.1

I am a deep believer in the power of data to drive our decisions. Data gives us the roadmap to reform. It tells us where we are, where we need to go, and who is most at risk….We will ask millions of teachers to use student achievement data and annual growth data to drive instruction and evaluation.

Secretary Duncan's hope for data-driven schools is shared by many (Bernhardt, 2004; Boudett, City, & Murnane, 2005; DQC, 2009; Kowalski, Lasley, & Mahoney 2008; Mills, 2007). A large literature has emerged on data-based decision making along with annual conferences (e.g., DQC, 2009; MIS, 2010) and a variety of foundation-sponsored initiatives around the country helping strengthen districts' data systems and personnel training. Media accounts with headlines like, "Data-driven schools see rising scores," (Hechinger, 2009) help fuel high hopes for data driven decision making as do portrayals of model schools or districts (Henke, 2005; Dattnow, Park, & Wohlstetter, 2007; Zavadsky & Dolejs, 2007). There is no doubt that the role of data in teachers' and principals' practice has grown over the last decade along with improvements in data quality and data access technologies.

As Kerr, Marsh, Ikemoto, Darilek, and Barney (2006, p. 498) note, there are high expectations and much potential for data use, and many ways data can be brought to bear on planning and practice in schools.

5.2

Most commonly, data are used for tasks such as setting annual and intermediate goals as part of the school improvement process. Data may also be used to visually depict goals and visions, motivate students and staff, and celebrate achievement and improvement. Schools use data for instructional decisions such as identifying objectives, grouping and individualizing instruction, aligning instruction with standards, refining course offerings, identifying low-performing students, and monitoring student progress. School structure, policy, and resource use may be informed by data. Schools have also used data for decisions related to personnel, such as evaluating team performance and determining and refining topics for professional development (see, e.g., Bernhardt 2003; Choppin 2002; Feldman and Tung 2001; Mason 2002; Supovitz and Klein 2003).

A persisting gap exists between theory and typical practice as shown by empirical studies (Bruner et al., 2005; Coburn & Talbert, 2006; Coburn, Toure, and Yamashita, 2009; Ingram, Louis, & Schroeder, 2004; Means, Padilla, & Gallagher, 2010; Wayman 2005). Most schools still find significant challenges in trying to use data effectively. It is not easy to transform teachers' roles that have never in the history of the profession entailed widespread expectations of proficiency and participation in data analysis to examine practice, evaluate outcomes, and guide decision making and planning – *especially in collaborative groups* as commonly espoused today.

The challenge, for the most part, is not *lack* of data. It is not that districts do not have data. Indeed, modern data collection and information technology have filled districts' databases with test scores, grading records, conduct records, health information, demographic data, student transcripts, personnel records, finance and budgeting data, survey data, parent information, and more. Districts, generally speaking, *have*

data. What districts generally don't have are enough staff in each school with the needed expertise, initiative, and tools to turn data into actionable information.

Part of the challenge is capacity and part of the challenge is school culture. Capacity issues include limited expertise, data access, analytical tools, and available time. Beyond these issues of capacity, are school culture issues: many teachers are ambivalent about examining their practice with objective data. While no single "attitude" characterizes all teachers' stance toward data, many teachers are apprehensive about the prospect of spotlighting classroom results and many do not fully understand uses of data for guiding instructional planning and evaluation. In any given school, if enough staff have skeptical or resistant attitudes, it will be challenging for the school's leaders to build what is widely advocated in the literature: a culture of inquiry, reflection, and collaboration in which data plays a key role in planning and decision making (Zavadsky & Dolejs, 2007; Newman, 2006).

School leadership is the key to building a staff culture willing to examine practice and collaborate for improvement. Leadership, ideally, should come from a team – the principal and teacher leaders – and it should be driven by clear and specific purposes. Research shows that leadership teams united by a common purpose can be powerful agents of school improvement (McLaughlin & Talbert, 2006; Vescio, Rossa, & Adams, 2008) and that among the many improvement-focused purposes such teams can serve, focusing on data is an important priority (Chrispeels, Castillo, & Brown, 2000; Young, 2006). For instance, Young (2006), based on case studies of four schools, found that leadership effectiveness was a major variable in teacher buy-in and participation. Here is a description of a school with effective leadership:

5.3

The Hilltop principal's vision centers on teachers' learning about instruction as revealed in accounts of classroom practices and in classroom artifacts, supported by a community that holds its members accountable for learning. The agendas and structured activities that the principal and her leadership council establish for the second-grade team's collaboration time define the data in this setting. Data for them consist both of what teachers reveal of their classrooms, as in war stories and student work samples, and how they measure progress, as in assessment results. These times also give Hilltop teachers collaborative experiences around data analysis that begin to build the principal's desired norms. For example, over several meetings in which second-grade teachers jointly scored student writing, one reluctant team member moved from withholding student work, to sharing writing she had already scored on her own, to finally accepting joint grade-level decisions on certain samples to calibrate her scores with the team's interpretation of the district writing rubric. The second-grade team is thus deepening their collaboration, their professional trust in sharing student work, lesson plans, and formative assessments results, and their sense of joint enterprise (p. 538).

It is neither simple nor easy for teachers in a school to transition from roles of autonomy to teamwork and to face heightened expectations about using data to guide decisions. New roles require new skills and may bring changed routines. For leaders it requires identifying opportunities that will create staff buy-in, but that also promote staff learning and improved practice (Chen, Heritage, & Lee, 2005; Chrispeels et al., 2000; Copland, 2003). At the same time, an initiative can go badly if leaders assign staff to tasks or roles they perceive as unproductive or that threaten pride or professional efficacy. As described above, there are many ways of bringing data analysis into practice and many challenges the leader must recognize and anticipate.

One strategy with potential is assessment benchmarking. As used here, this refers to teachers reflecting on and calibrating their assessment criteria and standards against a pre-established standard. Benchmarking can be as simple as a group of teachers discussing a particular scoring rubric and relating it to their individual assessment criteria and standards. Benchmarking can also be more elaborate, involving systematic procedures to rate student work, record assessment scores, analyze results, and develop action plans. The case presented here shows a method to examine writing assessment scores to identify variation and consistency among teachers and to guide teachers' discussion of data, writing instruction, and assessment. It is an approach that is feasible with typical school data and tools and that does not depend on analyses beyond what is reasonable to expect from professional educators. Managed well, it is a productive learning experience with the potential to improve practice.

6 Section II. Examining Writing Assessment Standards and Consistency at Gilbert Elementary School

6.1 The Principal's Concern With Excessive Variability in Writing Standards and Instruction

"How was the conference Dean?" asked Mary Smith, a 4th grade teacher at Gilbert elementary school. "Great," replied Principal Dean Jansen, "I attended some interesting sessions... and got some good ideas for strengthening our writing instruction."

Jansen was principal of Gilbert elementary – a grades 3-5 school with 470 students from a cross section of backgrounds. Gilbert school's achievement scores in writing had remained stubbornly flat for a long time – too long in Principal Jansen's view. Almost half of Gilbert's students scored "below standard" on the state writing assessment; he was concerned that teachers were becoming resigned to this level of performance. "We've got to turn this around" was one of his last statements at a faculty meeting before leaving for the conference.

Principal Jansen attended the conference to seek strategies to promote greater instructional and grading consistency among the teachers in his school writing instruction. Over the last two years, based in part on classroom observations and in part on conversations with teachers, Jansen had become concerned about excessive variation among teachers in writing instruction and standards. He didn't have objective evidence of this variation or that it might be something the school needs to address, but he observed considerable variation in writing assignments and in teachers' grading standards.

Jansen's concern about writing instruction grew in part from a conversation with Ms. Smith. What he learned gave him a fuller understanding of the degree of contrasts in instruction among different teachers. Ms. Smith described her collaboration with another teacher (Jane Jones) developing lessons connecting writing, reading, and science. Ms. Smith described how she was teaching water cycles in ecology and persuasive writing in language arts; and how she combined these subjects in a project where students composed editorials to the newspaper about street water runoff hurting local marshes. Students did this in groups – researching their topics and sharing and revising drafts of their editorial. Ms. Smith described how she and Ms. Jones teamed with two other 5th grade teachers, so that 5th grade students would review the editorials of the 4th graders and provide feedback before the 4th graders' editorial were sent to the newspaper. The local newspaper published several of the editorials.

Principal Jansen knew most teachers did not do this. In fact, writing instruction in other classrooms typically lacked such inventiveness and cross-subject connections. In other classrooms, writing assignments focused more on spelling, vocabulary, and grammar worksheets, and less on actual writing. When writing as assigned, it was more likely to be summarizing assigned readings or responding to assigned prompts (e.g., "write a page about what you would do if you could fly.") In some classrooms, not much writing was assigned at all. There was much variation from classroom to classroom and not much collaboration among the teachers – a situation not uncommon in schools (Rowan, Harrison, & Hayes, 2004; Smith, Lee, & Newmann, 2001; Spillane, 2004).

Over the past year, Principal Jansen had been trying to promote more collaborative work among teachers and more discussions about instruction. He saw this is a major priority: strengthening the collaborative culture of the school. At several recent faculty meetings, he drew attention to this, saying "we can be a better school if we work as a team." He wanted to see more joint curriculum planning, sharing of instructional strategies, and uniform academic expectations.

Principal Jansen was aware that some teachers were not entirely comfortable with the prospect of greater collaboration, concerned that it meant greater scrutiny of their teaching or sitting through series of unproductive meetings. Privately, many teachers believed, "what you do in your classroom is your business and what I do in my classroom is my business." Jansen was concerned about this mentality, viewing it as a barrier to improvement (DuFour, 2011). In his view, the curriculum belonged to the school, not to each individual teacher. He wanted to foster among teachers more of a shared commitment to all students and a culture of collaboration. Jansen believed that if the school culture and curriculum were going to move in the direction of common standards and methods of instruction, he would need to do more that periodically advocate and encourage; he would need to focus teachers' attention and discussions on evidence related to

practice.

Writing instruction would be the focus. He hoped to foster more regular discussions of instructional strategies and grading expectations, sharing assignments and assessments, and engaging in periodic benchmarking activities to calibrate their performance expectations for students. Jansen had a number of ideas for productive activities and he knew others would also. He shared with the staff several articles related to writing assessment and instruction (Andrade, Buff, Terry, Erano, & Paolino, 2009; Gere, 2010). But the big activity he was going to focus on was examining and discussing data related to writing assessment and grading standards. So he organized a benchmarking activity to help Gilbert's 5th grade teachers calibrate their writing assessment criteria and standards. He planned to engage other grades later helped by his experience with this first benchmarking activity.

6.2 Methods of the Benchmarking Activity

The following describes methods of the benchmarking process used to help calibrate teachers' grade level expectation for writing. It is a systematic way to allow teachers to compare their criteria and standards for assessing student writing.

Step 1) All the 5th grade teachers used a common writing prompt drawn from the state writing assessment rubric. (The writing prompt was available online.) The teachers in their individual classrooms each gave the writing assignment to their students, allowing about 2 hours with appropriate breaks. Each classroom produced about 24 essays.

Step 2) Using the state writing assessment rubric, each teacher graded his/her students' essays and recorded the scores in an Excel spreadsheet. Appendix A shows the kind of rubric used.

Step 3) Several weeks later, two trained teachers with experience in rubric-based writing assessment independently scored all the papers, without knowing students' names (each paper was given an anonymous ID). These two teachers had attended state hosted workshops on writing instruction and standards-based assessment and participated as raters for the state assessment. They independently scored the papers and then used a systematic process to give each paper a single score (the process is used for score calibration in rubric-based holistic writing assessment). Thus, each paper received a "benchmark score."

Step 4) After this process was complete, each essay paper had two scores (teacher's score and benchmark score). The data set had five columns, teacher ID, student ID, teacher's score, benchmark score, and a standardized test reading score (added to the data set for additional information, but not analyzed as part of this case).

- *Benchmark score:* score from the trained assessors (1 − 5 [high score]; based on the rubric, 3 is considered "at standard" for 5th grade). As an approximate frame of reference, each score point can be roughly viewed as a letter grade (5=A; 1=F). This frame of reference is helpful for giving a context to better interpret the range and distribution of writing scores.
- *Teacher score:* score for a paper from each student's own teacher (1- 5).
- *Reading test score*: NCE score on the 5th grade state test in reading.

Key analyses and questions explored in the benchmarking activity include:

- What is the distribution of student scores? How many students are below, at, and above standard in their writing proficiency based on the state prescribed scoring rubric? This requires a frequency analysis showing the raw counts and percentages of students scoring at each of the performance levels, which also shows how many are at or above standard and how many are not.
- How well do the individual teachers' scores match up with the benchmark scores? Are teachers' ratings on average higher, lower, or about the same as compared with benchmark ratings? This requires computing classroom means of the teachers' ratings and of the benchmark ratings and computing a classroom-level deviation score (numerical gap between teacher's mean and benchmark rater's mean for each classroom).

- How much consistency or variation in standards is there among teachers across classrooms? This requires, in addition to the deviation analysis above, comparing within each classroom the teacher's and the benchmark rater's scores to determine the consistency of each teacher's scoring relative to the benchmark score for each student. This provides evidence of the extent to which each teacher is consistent in his/her application of assessment criteria from student to student.

6.3 Analyses and Results

Table 1 shows the number of students in each classroom.

Table 1
Classroom Enrollment Counts

Room	Count
1	25
2	22
3	26
4	26
5	23
6	22
7	24
Total	168

[2]

Table 2 and Figure 1 show the large majority of students scored a 2 or a 3. The teachers' scores and the scores of the benchmark raters differ at the two ends of the scale. The benchmark teachers' ratings produced fewer 5s and more 1s: the benchmark teachers rated 3 papers a "5," whereas the classroom teachers rated 14 papers a "5." The classroom teachers rated 2 papers a "1," whereas the benchmark teachers gave "1s" to 21 papers.

Table 2

Score Distributions

	Frequencies:	
Essay	Counts & (%)	
Score	Teachers'	Benchmark
	Results	Results
5	14 (8%)	3 (2%)
4	18 (11%)	17 (10%)
3	76 (45%)	70 (42%)
2	58 (35%)	57 (34%)
1	2 (1%)	21 (13%)

[3]

---

[2] http://cnx.org/content/m41218/latest/table1.png/image
[3] http://cnx.org/content/m41218/latest/table2.png/image

Figure 1

Histogram of Score Distributions: Teachers' Rating and Benchmark Ratings
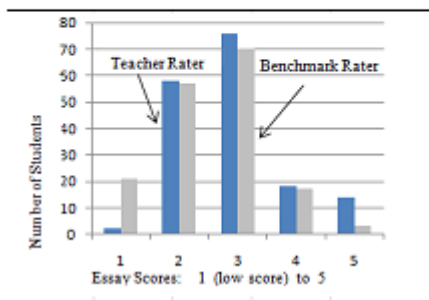

4

Do classroom mean scores on teacher-graded writing correlate with classroom means based on benchmark scores? Table 3 reports two mean scores for each classroom: the classroom teacher's ratings and the benchmark ratings. The scores are sorted from highest classroom mean to lowest based on the teacher-graded writing scores.

Figure 2 shows the classroom means in a scatterplot: the classroom's benchmark score is on the X axis; the teacher score is on the Y axis. Each point is a classroom.

Figure 2

Scatterplot of Classroom Mean Scores: Teachers' v. Benchmark Ratings
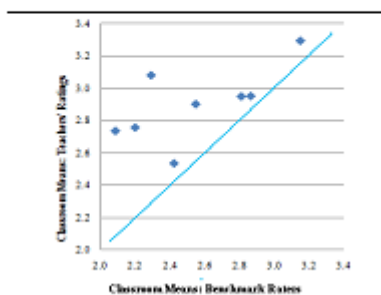
5

Table 3, and Figure 2 (the scatterplot of Table 3's scores) show the teachers' ratings are on average higher than the benchmark ratings. Table 3 shows an overall mean of 2.9 among teachers versus 2.5 for benchmark. The diagonal line on the scatterplot is a reference point; if teachers' scores and the benchmark scores were in perfect agreement, each point would fall on this line. The further away from the line, the greater the deviation of the teacher's score from the corresponding benchmark score for that classroom.

---

[4] http://cnx.org/content/m41218/latest/figure1.png/image
[5] http://cnx.org/content/m41218/latest/figure2.png/image

Table 3
Mean Class Scores: Teachers v. Benchmark

Room	Teacher Score	Benchmark Score
3	3.3	3.1
7	3.1	2.3
6	3.0	2.9
2	3.0	2.8
1	2.8	2.2
5	2.7	2.1
4	2.5	2.4
Overall Mean	2.9	2.5

[6]

While the teachers' scores tend to be higher, they are in fact correlated with the benchmark ratings – a moderate correlation (Pearson "r" correlation = .53). The classrooms with higher teacher ratings tend to be the classrooms with higher benchmark ratings which indicates classroom teachers are more or less consistent in applying the assessment rubric, but, generally err by scoring too high. Classrooms 1, 5, and especially 7 (Table 3) show the biggest departures from the benchmark ratings.

A deviation analysis is another way to summarize how well teacher-graded writing scores for individual students match the benchmark scores. Each student paper has a teacher-rated score and a benchmark score. Thus, for each paper, one can compute a "difference score." This difference score is computed as an absolute value (i.e., the difference score is 1 whether the teacher score is 4 and the benchmark score is 3, or vice versa). Table 4 shows the frequency of occurrence of the difference scores: out of 168 papers, the teacher and benchmark scores matched 86 times (51% of papers); differed by one point 61 times; differed by two points 20 times; and by 3 points once. Thus, 87% of the time, the teacher rater and the benchmark rater were within at least one point of each other in their scoring.

Table 4
Degree of Match in Ratings Between Teacher Scores and Benchmark Scores

Score Dif	Frequency	Percentage
0	86	51%
1	61	36%
2	20	12%
3	1	1%

[7]

Individual, student-level deviation scores can be aggregated to the classroom level and examined for individual classrooms. As described above, each paper has a teacher-rated score and a benchmark score, and so each paper also has a "difference score." Table 5 shows the average, at the classroom level, of the difference scores. A score of zero for a classroom would show that the teacher's scores exactly matched the benchmark score for each paper. (No classroom achieved this.) Classroom #3's scores are very close to the benchmark scores, suggesting this teacher's assessment standards and criteria are highly aligned with those of the benchmark raters. Classrooms #7 and #5 are the furthest off from the benchmarks.

Classroom #4 is an interesting case in that this teacher's mean rating of his/her students' papers is very close to the benchmark raters' mean (Table 3). However, the difference score is relatively large (.73, as shown in Table 5). Thus, even though the means are similar, this teacher's ratings differ often *in both directions* from the benchmark ratings, showing this teacher is not very consistent in applying the rubric. This shows why it is important not just to compare means, but also to compare the difference scores.

---

[6] http://cnx.org/content/m41218/latest/table3.png/image
[7] http://cnx.org/content/m41218/latest/table4.png/image

Table 5
Ave. Difference by
Classroom Between
Teachers' and Benchmark
Scores

Room	Ave.Score Dif
7	1.21
5	1.00
4	0.73
1	0.56
2	0.43
6	0.27
3	0.15
Overall Mean	0.62

[8]

Table 6 shows the scores (teacher-graded and benchmarks) from classroom #7 –the classroom with the largest difference scores. The scores are organized by the size of the gap between the benchmark rater's score for each paper and the teacher's score for each paper. The shading shows visually the extent of benchmark rated v. teacher rated score differences among the 24 classroom papers. The darker the shading, the greater the disparity between the teacher's score and the benchmark score. The teacher in classroom #7 is not consistent in applying the rubric. This suggests s/he does not have a clear understanding of the rubric-based criteria and standards for assessing student papers.

Table 6

Classroom #7's Benchmark Score v. Teacher Score For Each Student

| Benchmark Score | 2 | 3 | 3 | 3 | 2 | 2 | 2 | 1 | 1 | 3 | 3 | 3 | 3 | 3 | 2 | 2 | 2 | 2 | 1 | 3 | 3 | 2 | 2 | 2 |
| Teacher's Score | 5 | 5 | 5 | 5 | 4 | 4 | 4 | 3 | 3 | 2 | 2 | 2 | 2 | 2 | 3 | 3 | 3 | 3 | 2 | 3 | 3 | 2 | 2 | 2 |

[9]

7 Section III. Discussion Questions and Assignments

7.1 Questions and Exercises Related to Section I.

(Exercise #1) There are surveys and rubrics available on the web for assessing school culture and collaborative practices among teachers. Below are web links to a few. Your own school or district may use a survey. Find a survey and select about eight items to capture the main dimensions of a well-functioning collaborative group and the attributes of school culture that support it. Apply those selected items to your own work situation (in a school) or, if you don't work in a school, try and arrange a meeting with a working teacher and review and discuss the selected items (how that teacher views the culture if his/her school with respect to the particular dimensions reflected in the survey items). Record your results and explain whether the ratings you observe are satisfactory or whether practice should attempt to exhibit greater collaboration. Compare your results to those of others doing this same exercise. If there are notable differences between results, discuss what might be the reason for the different results.

- http://schoolreforminitiative.org/protocol/doc/plc_survey.pdf [10]
- http://files.solution-tree.com/pdfs/Reproducibles_BPLC/midyearplcsurvey.pdf [11]

(Exercise #2) Imagine you are the principal of a school with staff and working conditions similar to Gilbert Elementary School – that is, a staff that is quite varied in seniority, working habits, talents, and levels of

---

[8]http://cnx.org/content/m41218/latest/table5.png/image
[9]http://cnx.org/content/m41218/latest/table6.png/image
[10]http://schoolreforminitiative.org/protocol/doc/plc_survey.pdf
[11]http://files.solution-tree.com/pdfs/Reproducibles_BPLC/midyearplcsurvey.pdf

enthusiasm for greater collaboration. You want more collaboration on curriculum planning, instructional strategies, and peer mentoring. Develop either a 400 word memo or a 300 word speech that would be the first communication to the staff about your perception of the need for change. Whatever else your message includes, provide at least three reasons to justify your position. Your message should anticipate responses of skeptical staff members who will wonder – "why should we do this? what's wrong with the way things are now?"

7.2 Questions and Exercises Related to Section II.

(Discussion question #1) The average scores of the teacher-graded papers are .4 higher than the averages of the benchmark papers (Table 4), with much of this gap coming from three classrooms. Do these results indicate standards for assessing and grading writing need to be raised? If so, what is the basis for your conclusion?

(Discussion question #2) Suppose someone claimed that the writing assessment results actually understate the true range of writing proficiency in typical classrooms in the building. The person claims that in reality the true range is even greater than the ratings suggest, and that if there was a different kind of writing task (prompt) and a more elaborate rubric, the results would show a bigger disparity between the high level writers and the low level writers. How would you investigate this possibility? Do you think that distribution of scores in a classroom might be different with a different kind of writing prompt or scoring rubric?

(Discussion question #3) The average rating from teacher Jones of the student papers in his classroom is the same as the average rating of those same papers from the benchmark rater. Does this mean teacher Jones and the benchmark rater are in agreement in their application of the rubric's criteria and scales? Give an example to illustrate your point.

(Exercise #1) Table 6 shows the classroom level deviation scores. Assume you are the principal and you have the data for these tables. Decide how you would communicate this information to the teachers and what you would do on the basis of the information.

(Exercise #2) Develop a specific plan for a 4-hour workshop that would follow after the benchmarking activity described above. This plan should be targeted at an identified grade level. From the literature, identify three or four excellent readings you would assign to workshop participants to prepare.

(Exercise #3) Propose a new and different analysis using additional variables and exploring different questions. The data set used for the benchmarking activity is described in Step 4 of Section II. In addition to the data and variables described in Step 4 above, assume these variables are also part of the data set: marking period grades, gender, race, special education classification, and free-lunch eligibility. Also, you may propose adding additional variables if you have a specific inquiry in mind that could be done with data collection.

The proposal should begin with a "problem statement" – that is, state the issue, need, or concern that motivates doing the analysis and how the information sought can help address the concern. As with the above case's analyses, do not imply that your proposed analyses will provide conclusive evidence; rather, the objective is better information, knowing better key outcomes in the program, and more informed planning and decisions. The proposal should explain the analyses to be conducted, much like the explanations in Section II. above, and it should offer preliminary thoughts on what you would deduce pending different findings. For instance, "A strong correlation between [name of variable] and [name of variable] would be a cause for concern because…." "If a strong correlation is observed, this would invite further inquiry into…."

8 Section IV. Notes for Instructor Related to Discussion Questions and Assignments

(Comments on Exercise #2, Section I.) Theoretical justification for PLCs: (a) strengthen teachers' sense of ownership over curriculum and professional development; (b) improve quality of decision-making by pooling

expertise; (c) raise level of accountability to colleagues. For more information, see DuFour, DuFour, and Eaker (2008), Kilbane (2009), McLaughlin and Talbert (2006), and Mullen & Hutinger (2008).

(*Comments on Discussion question #1, Section II.*) That academic expectations in writing in some classrooms (notably, classrooms 1, 5, and 7) need to be higher is a reasonable conclusion. While the evidence from the benchmarking exercise alone does not constitute incontrovertible proof, the evidence is definitely strong enough to warrant concern about a gap between what the three teachers view as "at standard" writing and the level of proficiency prescribed in state standards as specified in rubrics. It should be emphasized that these data – the evidence generated from an exercise like this – must be interpreted cautiously and discourse among participations should use appropriately qualified language. The evidence should stimulate further discussion about writing instruction, assignments, expectations, and grading.

(*Comments on Discussion question #2, Section II.*) Some experts believe that standardized writing assessments (like the one described here) may have the effect of constraining the range of performance. Imagine, for instance, if the students of an average grade class of 5th were given 24 hours to write an 800 word evidence-based argument on a particular topic (some issue of relevance to 5th graders). The low end papers would not look a lot different from the low end papers in a standardized writing assessment, but the high end papers would be sophisticated essays – fully developed arguments with evidence, definitions, explication of assumptions, examples, and possibly even rebuttals to counter positions. Thus, the range of papers from worst to best would grow because the most able and motivated students would not be constrained by a relatively short time limit and by the five paragraph structure imposed by standardized writing assessment rubrics. Being less constrained, the top students would have more freedom for reading, writing, revising and creating their own argument. So in this sense, relative to the standardized "one period" writing assessment, the observed range in quality of papers submitted would grow. However, what this means is less clear: the 24 hour essay allows more room for other attributes to factor into performance, namely motivation, substantive background knowledge, and information processing skill that arguably fall outside the domain of writing. This raises the question, then, would an assessment based on the 24 hour essay be just a measure of writing, or is it assessing some broader construct. There is no simple answer to this question because it depends on how we think of and define writing. While there is no simple answer, it is instructive to contemplate this question and examine our own conceptions of the construct, "writing."

(*Comments on Discussion question #3, Section II.*) The answer to the question is "No," the two averages being similar does not mean the teacher and the benchmark rater are consistent with each other in how they apply the rubric. It is essential to compare not just the mean scores of one classroom to another, but also to examine and compare deviation scores. For example, the teacher could grade three papers 5, 3, and 1, while the benchmark scores for those same three papers are 1, 3, and 5. Both sets of ratings have an average of 3.0, but, clearly, the teacher is not interpreting the rubric in the same way as the benchmark rater.

(*Comments on Exercise #1, Section II.*) The classroom level (not student level) deviation scores could be shown to teachers either as a group in a meeting or individually in one-on-one conferences; and, if in a group, can be viewed either with teachers identified or not. The leader would want to consider teachers' level of concern about being identified and about whether there would be excessive discomfort in openly discussing individual assessment results. In some schools this would not be a problem, but if a school's leadership and teachers are not practiced with such discussions, it may be better to design the process to avoid inter-teacher comparisons. For instance, code numbers could be substituted for teachers' names in viewing tables showing the full set of classroom results, with later individual conversations to discuss individual results. Ideally, professionals should be comfortable with discussing practice and be accountable to supervisors, colleagues, and clients for performance outcomes; the reality is that culture and practice in many schools does not reflect this ideal.

(*Comments on Exercise #2, Section II.*) Suggested workshop activity: Collaborative grading of papers can be a very productive activity if well planned, organized, and led by an experienced workshop leader with expertise in writing instruction and assessment. The activity involves reading and discussing a range of student papers, discussing the paper's strengths and weaknesses, comparing the attributes of weak papers to better ones, and individual participants explaining to others their grading criteria with examples from papers illustrating these criteria. The workshop can include mini-benchmarking sessions with a small number of

papers (i.e., individual reading and rating of papers followed by comparing and discussing results – discussions should be connected with local or state documents containing approved writing standards). These sessions would be well served by creating tables like Table 7 so participants can examine deviation scores for ratings on individual papers and discuss disparities and outlier scores among the participants. The workshop should culminate with a focus on instructional practice to improve the writing.

(Comments on Exercise #3, Section II.) Many additional analyses are possible. Here are a few examples: (a) Probe more deeply into "error" patterns (deviation scores) in teachers' application of the assessment rubric. (Table 7 shows a deviation analysis for one teacher.) Are teachers more likely to be off randomly or in predictable ways? (b) How big are the differences in performance scores (writing, test scores, grades) among the different demographic groups and do these gaps differ depending upon the measure? (c) Do students with higher standardized test scores also do better on the writing assessment? Get better grades? How strong is the correlation? Do relationships between achievement variables vary by demographic category?

Table 7

Rubric for Exercise #3, Section II: Proposal for Additional Analyses

Dimensions	1 =Unacceptable	2 =Adequate	3 =Target
Problem Statement	Problem identified is not specific; hard to understand; significance not clear.	Problem identified is relatively specific, though may have some unclear aspects; significance is mentioned but not entirely compelling.	Problem identified is specific; significance is compellingly established.
Purpose	The purpose of the analysis is not explained. It is not clear what the purpose is; why the analysis is being done.	There is a purpose statement and it is generally clear though not as specific as it could be. It gives some attention to why the analysis is needed.	The purpose statement is clear and detailed in explaining why the analysis is needed and how the analysis will help planning and decision making.
Data Description	Little or no information provided on data: cases, variables, measures are not clear	Information is provided to explain the variables being examined.	Complete information describes cases, variables, and measures.
Analytical Strategy	No analytical strategy is evident or the strategy presented is inappropriate or difficult to understand. There is little or no clear inferential logic to the design of the analyses. The analysis will not yield useful information.	The analytical strategy is generally appropriate. Elements of the logic may be unclear in places with some additional justification needed, but overall the analyses are justified and will produce useful evidence.	There is a clear, logical, and well-justified strategy. The strategy is appropriate for the questions asked and will provide highly relevant, credible evidence to clarify issues and inform plans and decisions.
Comments on Possible Findings or Decisions	Little or no discussion of possible findings or implications.	Possible findings and implications are considered, though may be somewhat sketchy; few or no qualifications/limitations offered.	Possible findings and implications are considered in some detail and well thought out. Appropriate qualifications and/or limitations are offered.
Overall Coherence and Presentation	The proposal is not organized well. Paragraph coherence is frequently weak; many sentences are unclear; too many sections are difficult to follow.	The proposal is generally well organized and clearly presented, though some sections have weaknesses in clarity, concisions, and/or coherence.	The proposal is consistently well organized and clearly presented. The prose is clear, concise, and well organized.

[12]

9 Reference List

Andrade, H, Buff, C., Terry, J., Erano, M., & Paolino, S. (2009). Assessment-driven improvements in middle school students' writing. *Middle School Journal, 40* (4), March, 4-12

Bernhardt, V. (2003). No schools left behind. *Educational Leadership* 60 (5): 26–30.

Bernhardt, Victoria L. (2004). Data analysis for continuous school improvement. 2nd Edition. Larchmont, NY: Eye on Education.

---

[12]http://cnx.org/content/m41218/latest/table7.png/image

Boudett, K., City, E., & Murnane, R. (2005). *Data wise: A step-by-step guide to using assessment results to improve teaching and learning.* Cambridge, MA: Harvard University Press.

Brunner, C., C. Fasca, J. Heinze, M. Honey, D. Light, E. Mandinach, and D. Wexler. (2005). Linking data and learning: The Grow Network study. *Journal of Education for Students Placed At Risk,* 10, no. 3: 241–267.

Chen, E., Heritage, M., & Lee, J. (2005). Identifying and monitoring students' learning needs with technology. *Journal of Education for Students Placed at Risk, 10* (3), 309–32.

Choppin, Jeffrey. 2002. Data use in practice: examples from the school level. Paper presented at the annual meeting of the American Educational Research Association, New Orleans, April.

Chrispeels, J., Castillo, S. & Brown, J. (2000). School leadership teams: Factors that influence their development and effectiveness. *Advances in Research and Theories of School Management and Educational Policy,* 4:39–73.

Coburn, C. E., J. Toure, & M. Yamashita. (2009). Evidence, interpretation, and persuasion: Instructional decision making at the district central office. *Teachers College Record, 111* (4), April, 1115–1161.

Coburn, C. & Talbert, J. (2006). Conceptions of evidence-based practice in school districts: Mapping the terrain. *American Journal of Education, 112* (4), 469–495.

Copland, M. (2003). Leadership of inquiry: Building and sustaining capacity for school improvement. *Educational Evaluation and Policy Analysis, 25* (4), Winter, 375 − 395.

Data Quality Campaign. (April 2009b). DQC brings together education leaders to urge states to use data systems for continuous improvement. *DQC Newsletter,* 3, Issue 7. http://www.dataqualitycampaign.org/files/DQCNewsletter (accessed May 1, 2009).

Datnow, A., Park, V., & Wohlstetter, P. (2007). Didn't see in the paper... could have missed it? *Achieving with data: How high-performing school systems use data to improve instruction for elementary students.* Los Angeles, CA: Center on Educational Governance Rossier School of Education, University of Southern California.

DuFour, R. (2011). Work together: But only if you want to. *Phi Delta Kappan,* 92 (5), February, 57-61

DuFour, R., DuFour, R. & Eaker, R. (2008). *Revisiting professional learning communities at work: New insights for improving schools.* Bloomington, IN: Solution Tree.

Feldman, J. & Tung, R. (2001). Whole school reform: How schools use the data-based inquiry and decision making process. Paper presented at the annual meeting of the American Educational Research Association, Seattle, April.

Gere, A. (2010). Taking initiative on writing. *Principal Leadership,* November, 37 − 42.

Hechinger, J. (2009, June 12) Data-driven schools see rising scores. *Wall Street Journal.*

Henke, K. (2005). *From vision to action: How school districts use data to improve performance.* Washington DC: Consortium for School Networking (CoSN), 1025 Vermont Avenue NW, Suite 1010.

IES (2009). Robust Data Gives Us The Roadmap to Reform. Address by Secretary Arne Duncan to the Fourth Annual IES Research Conference. Washington, DC: U. S. Department of Education and Institute for Education Sciences. Retrieve from: http://www2.ed.gov/news/speeches/2009/06/06082009.html[13]

Ingram, D., Louis, K. S., & Schroeder, R. (2004). Accountability policies and teacher decision making: Barriers to the use of data to improve practice. *Teachers College Record , 106* (6), June,1258–1287.

Kerr, K., Marsh, J., Ikemoto, G., Darilek, H., & Barney, H. (2006). Strategies to promote data use for instructional improvement: Actions, outcomes, and lessons from three urban districts. *American Journal of Education, 112* (4), 496-520.

Kilbane, J. (2009). Factors in sustaining professional learning community. *NASSP Bulletin, 93* (3), September, 184-205.

Kowalski, T.J., Lasley, T.J. II, & Mahoney, J.W. (2008). *Data-driven decisions and school leadership: Best practices for school improvement.* New York: Pearson.

Mason, S. (2002). *Turning data into knowledge: Lessons from six Milwaukee public schools.* Madison: Wisconsin Center for Education Research.

---

[13] http://www2.ed.gov/news/speeches/2009/06/06082009.html

Means, B., Padilla, C., & Gallagher, L. (2010). *Use of education data at the local level from accountability to instructional improvement.* Washington, D.C.: U.S. Department of Education, Office of Planning, Evaluation, and Policy Development.

McLaughlin, M. & Talbert, J. (2006). *Building school-based teacher learning communities: Professional strategies to improve student achievement.* New York: Teachers College Press.

Mills, G. (2007). *Action research: A guide for the teacher researcher.* Upper Saddle River, NJ: Pearson.

Mullen, C. & Hutinger, J. (2008). The principal's role in fostering collaborative learning communities through faculty study group development. *Theory Into Practice, 47* (4,), 276 — 285.

MIS (2010). Annual Management Information Systems conference, sponsored by National Center for Education Statistics and US Department of Education, Austin, TX.

Newman, J. (2006). Alabama district improves by sharpening data & goals. *Journal of Staff Development,27* (2), Spring, 10-14.

Rowan, B., Harrison, D. & Hayes, H. (2004). Using instructional logs to study mathematics curriculum and teaching in the early grades. *Elementary School Journal, 105* (1), 103-127.

Smith, J., Lee, V., & Newmann, F. (2001). *Instruction and achievement in Chicago elementary schools.* Chicago: IL: Consortium for Chicago School Research

Spillane, J. (2004). *Standards deviation: How schools misunderstand education policy.* Cambridge, MA: Harvard University Press.

Supovitz, J. & Klein, V. (2003). *Mapping a course for improved student learning: How innovative schools systematically use student performance data to guide improvement.* Report by Consortium for Policy Research in Education, University of Pennsylvania, Philadelphia, PA.

Wayman, J. C. (2005). Involving teachers in data-driven decision making: Using computer data systems to support teacher inquiry and reflection. *Journal of Education for Students Placed At Risk*, 10, (3), 295–308.

Vescio, V., Rossa, D. & Adams, A. (2008). A review of research on the impact of professional learning communities on teaching practice and student learning. *Teaching and Teacher Education, 24*, (1), January, 80-91.

Young, V. (2006). Teachers' use of data: Loose coupling, agenda setting, and team norms. *American Journal of Education, 112* (4), 521 − 548

Zavadsky, H. & Dolejs, (2007). Data: Not just another four letter word. *Principal Leadership*, October, 32 − 36.