The Qualitative Report 2013 Volume 18, Article 20, 1-22
http://www.nova.edu/ssss/QR/QR18/cho20.pdf

When Pandora's Box Is Opened: A Qualitative Study of the Intended and Unintended Impacts Of Wyoming's New Standardized Tests on Local Educators' Everyday Practices

Jeasik Cho and Brian Eberhard
University of Wyoming, Laramie Wyoming, USA

In the context of a newly adopted statewide assessment system, PAWS (Proficiency Assessment for Wyoming Students), this paper describes intended instructional changes and unintended outcomes in classrooms and schools as a result of an assessment policy involving an innovative online portion of the test. An elementary school was selected and prolonged qualitative fieldwork with in-depth and focus group interviews were conducted for 1½ years. A constant comparative data analysis and interpretation from grounded theory methodology led to the following themes: adaptive implementation policy, teachers' dilemmas, instructional change, and school culture change. While observing an elusive role for teachers that involved external accountability factors, researchers also found a practical hope for future PAWS tests, foreshadowing the need for promptly delivered test results for realistic instructional improvement. Keywords: Standardized Testing, Implementation, Grounded Theory, Instructional Change

Since the 2005-6 academic year, the state of Wyoming has adopted a new assessment system, called PAWS (Proficiency Assessment for Wyoming Students), for students from grades three to eleven (WSDE, 2006). Assessments in PAWS include reading, writing, and math. PAWS involves several innovative ideas intended to not only make Federal requirements effective and accurate at the state level, but also to support teachers through the provision of test results in a timely manner. Such timely and specific feedback is designed to provide concrete instructional guidance to teachers. The timely reporting of results is made possible by the power of technology in the context of an ambitious assessment proposal called ISA (Instructionally Supportive Assessment), through which teachers are given a few condensed, key standards/benchmarks aligned closely with actual standardized tests (Popham, 2001, 2002).

The purpose of the study described in this article was to use qualitative research methods to identify problems perceived by teachers in the early process of PAWS testing. First, the authors explored ways in which teachers made sense of the PAWS testing policy in general, and teachers' perspectives on a newly adopted online portion of the testing in reading and math in particular. Second, a process of how planned instructional and school change evolved over time, whether intended or unintended, at the classroom level and across grade levels was documented.

The lead author teaches an assessment course and a qualitative research course at the undergraduate and graduate levels, respectively. The second author teaches secondary social studies methods courses and serves as a field supervisor of preservice teachers. The name, PAWS, an acronym for "Proficiency Assessment for Wyoming Students," has become a common part of the lexicon used frequently both by the media and faculty members in the university to which we belong. Educators and other stakeholders were told that the assessment included some innovative ideas.

As an instructor teaching an assessment course, the lead author was interested in figuring out what this new statewide assessment was about, hoping to include content related to PAWS in his assessment class. Early in the process, the authors had very minimal knowledge regarding PAWS testing and were not involved in any official relationship with the state of Wyoming. One day, we received a brief overview of PAWS testing from Dr. Hing, a senior professor colleague in our College of Education. This professor was officially involved in the development and implementation of PAWS testing. From this vantage point, and drawing on our colleague's deep understanding of the assessment, we started to do research on the newly implemented PAWS testing.

We chose one elementary school as a convenient sample (Patton, 2002). This school is located near the university to which we belong and has been one of the "partner schools" where our college sends student teachers during their last semester. This school was generally viewed as a typical elementary school. Although this school has been identified as "Title I" (i.e., a majority of parents are classified as low socio-economic status), we thought this school was typical of most elementary schools in areas of student achievement, leadership, school culture, etc. We knew some teachers in this school and we were confident about gaining entry to conduct research. Once the IRB proposal was approved by our university, we met with the principal of this school to obtain permission to collect data in the hope of understanding specific assessment-related issues under the first two years of implementation of the PAWS.

Literature Review

First, we will review literature that deals with two different implementation perspectives on education in general and testing policy in particular. The PAWS testing program falls under the policy requirements of NCLB (No Child Left Behind, 2002-present) initiated by former U.S. President George W. Bush. NCLB is a reauthorization of the previous ESEA (Elementary and Secondary Educational Act) in 1965 that detailed the federal government’s support and monitoring for public school curriculum and instruction. To receive financial support from the NCLB Act, each state was required to adopt a standardized testing program of its own and submit a progress report of test scores, called AYP (Adequate Yearly Progress), in the areas of reading, writing, and math.

Second, this paper reviews literature that reports the nature of computer-assisted testing programs, particularly in the context of statewide standardized tests. Having an online component in the standardized testing context is seen as the state of the art. Reviewing the history and extensive empirical effects of computer-assisted testing goes beyond the scope of this paper. The particular focus of this review is on how this innovative component of standardized testing works in conjunction with a general purpose of assessing student learning and feedback for teachers.

Fidelity Versus a Mutually Adaptive Perspective on Testing Policy Under NCLB

Back in the 1970s, researchers in educational change used the term “fidelity” to describe how much an innovation becomes actualized as intended. Fidelity is defined as the use of an innovation as originally intended, a perspective on determining “the extent to which actual use of the innovation corresponds to the intended or planned use” (Fullan & Pomfret, 1977, p. 340). Recently, this term has been widely used in school, as federally funded reading programs (e.g., Reading First) require teachers to *strictly* follow the intended use of a program in a specific manner. The purpose of effective implementation of standardized-test-driven accountability is that it be practiced as intended.

Key to this "fidelity" perspective is a question of how to make sure certain groups' interests and needs are more highly prioritized. The teacher, as a user, is expected to play a passive role in obtaining desired outcomes as the teacher respectfully follows intended uses of any specific curriculum and instruction. Strictly adhering to the prescribed guidelines in implementing a package of curriculum and instruction for the user is essential in this implementation perspective (Klein, Zevenbergen, & Brown, 2006; Snyder, Bolin, & Zumwalt, 1992). The AYP (Adequate Yearly Progress) is the most important indicator of fidelity to which local educators are required to pay great attention. For instance, in the state of Illinois, under NCLB accountability pressures, the initial intent of the SAC (Standards-Aligned Classroom) initiative has shifted from a focus on aligning classroom activities towards a more coercive, regulatory process imposed on failing schools (Vogel, Rau, Baker, & Ashby, 2006).

There is a general criticism of NCLB given that the current approach to accountability involves more prescriptive requirements that have a direct impact on the everyday events in schools and classrooms. Rigidly prescribing what teachers do and when minimizes the autonomy of teachers and thus negatively affects teacher professionalism (Berliner, 2009). There are some districts in Illinois and Texas that refused to follow the implementation of NCLB, and, as a result, these districts did not receive federal financial support. Nonetheless, observed compliance of NCLB is common and, in turn, many point to some negative consequences for educators. According to Cawelti (2006), there are at least three major side-effects of this accountability system: a skewed curriculum, discouraged teachers, and the intentional misuse of data. For example, certain academic subjects (those being tested) are unethically prioritized over those content areas not being tested; teachers are unwittingly forced to teach to the test rather than to teach creatively in meeting various needs of individuals; and cutoff scores are arbitrarily negotiated by those not connected with local classrooms. While juxtaposing the rhetoric and reality of the force of NCLB on the ideals and processes of education, Gay (2007) boldly contends that the narrowly defined accountability system will create unsound results in this nation:

> If we continue the dangerous precedents and directions set by NCLB we run the risk of exacerbating already dire conditions in U.S. education. Achievement gaps will continue and even expand; more and more children will be victimized and then punished for being victims… Coercive, subterfuge and 'one size fits all' educational reform strategies simply are not reasonable or viable bases on which to build constructive educational futures for a nation in desperate need of new directions that are genuinely egalitarian across ethnic, racial, social, cultural, linguistic and ability differences. (p. 291)

Most would agree that NCLB has generated more controversy than any previous educational legislation. The findings of this and other studies document that the effects of recent federal educational policy (i.e., NCLB) are too influential by restricting local educators from making contextually based decisions about their schools, students, curricula, assessments and instructional delivery. Many alternative claims against this narrowly-defined notion of accountability have been asserted in recent years (Berliner, 2009; Goodman, 2007).

In contrast to the traditional concept of fidelity implementation explained thus far, there are two alternative concepts related to policy, curriculum, and assessment implementation processes: mutual adaptation and enactment.

First, the mutual adaptation perspective on implementation is defined as "an effective strategy of a project and institutional setting to each other" (McLaughlin, 1990, p. 12). This perspective acknowledges the significance of local, contextual factors that influence the

process of implementation. Guiding questions by users that exemplify a mutual adaptation perspective may be: "Should innovations be 'done right,' according to the spirit … of their developers' intents? Or should they be adapted to fit local realties, to permit an even fuller and better implementation?" (Miles & Huberman, 1984, p. 279). To a large extent, it is likely that adaptations always take place as local educators attempt to engage in the betterment of implementation, both directly and indirectly, in ways that they believe will better meet the needs of their specific students (Hord, 2001).

Insofar as the mutually adaptive response-type still keeps the original idea embedded in the program or policy, an enactment perspective (Snyder, Bolin, & Zumwalt, 1992) is different in that it encourages or legitimizes adaptations as a result of local needs. That is, users of an innovative program or policy are encouraged to modify it to fit the needs of their local context. Teachers' participation at multiple levels is more likely encouraged in an effort to meet the needs of a local context (Cho, 2000; Harrison, 1997). By listening to teachers' voices and life stories, one is likely to learn more about ways in which modifications or adaptations are made for the sake of diverse learners (Calderhead, 1987; Kirk & MacDonald, 2001). From the enactment perspective, learning is context- and person-oriented, whereby individual teachers' knowledge, beliefs, and values are emphasized when assessing and accounting for "the mental processes that underlie behavior" (Clark & Yinger, 1977, p. 280). Without taking teachers' subjective worlds into consideration, it is generally noted that educational change is unlikely to occur the way as it is intended (Hord, 2001).

Taken together, the three responses--fidelity, mutual adaptation, and enactment--can be viewed as a continuum in which local educators make choices. Cobb and Rallis's (2008) work on accountability offers a practical example of how local educators respond to NCLB. Given their two-dimensional grid in which "internal-external" and "lateral-hierarchical" continua are crisscrossed, Cobb and Rallis emphasize that an ideal of accountability in educational enterprise does not necessarily mean a single hierarchical relationship between two parties: developers and users. Instead, they propose a two-way relationship as an alternative approach to accountability in which teachers and policymakers are laterally dedicated to "mutual obligation and responsibility to define and refine standards and measures, and to evaluate" (p. 182). "[T]rue dialogue between parties … [or] democratic deliberation" (p. 183) for project implementation must be organized laterally when pursuing an ideal of social justice. Cobb and Rallis assert:

> Moreover, in a political world, not all parties may agree on what is fair and reasonable. Engaging in an inclusive and democratic dialogue is critical for clarifying, then, where is the justice? The policy discourse on NCLB has not considered the variable inputs that comprise equitable and fair distribution or the diverse outputs that result. (p. 184)

Cobb and Rallis (2008) continue to construct five response-types that school districts can demonstrate with regard to the implementation and demands of NCLB in terms of a two-dimensional framework. The five response-types are the *Elites*, the *Opportunists*, the *BandAids*, the *Militants*, and the *Swamped*. Example responses from each response type are as follows:

- The Elites: "What I don't understand is how the NCLB proposes a valid accountability system? … We make our own judgments about where to focus our improvement efforts….we do have the resources to make improvement."

- The Opportunists: We can use NCLB as leverage to get what we have been striving for…. I don't see anything good coming out of NCLB…. Our state test scores have always been good … but we do have our problems."
- The BandAids: "Look, we're just playing the game here … this law has forced us into this. Let's just give them what they want and ride this one out…. to institute an after-school program to help those kids improve their [testing] skills."
- The Militants: "[NCLB] is the law of this land. We are obliged to follow it."
- The Swamped: "We know our students don't do well, and we don't know where to begin." (pp. 187-198)

To sum, the ways in which local educators respond to NCLB are likely to be varied in terms of their different perceptions of connecting policy to practice. As noted above, the *Militants*' views are convergent with a general assumption of fidelity-based implementation. The responses of the *Swamped* may be observed when a system of education begins to fall apart. The *other three responses* looking for a betterment of the existing practices are, to different degrees, adaptive in nature, and their attempts to incorporate students' needs into the present policy implementation process are active. Teachers' voices and those of certain groups of students on the margin are regarded as important in weighing the nature of mutuality toward balancing between external and internal requirements.

Standardized Testing and the Effect of Computer-Assisted Testing

The PAWS uses an online system approach as the primary delivery method. Utilization of this technology approach provides teachers with preliminary results on testing outcomes as soon as the testing window closes. Over the years, the use of computer-assisted testing at district and state levels has increased dramatically in conjunction with the benefit of formative assessment (e.g., Delaware, Georgia, Kentucky, Maryland, North Carolina, Texas, Virginia, and so on; Lynd, 2000; Olebe, 1999). The word, formative assessment, means that teachers can receive their students' test results before the school-year ends, so that teachers can re-teach those whose scores are below proficient. One of the major criticisms about standardized testing is that teachers receive test results during the summer or in the fall when their students have already moved on to the next grade-level. Online testing can speed up grading and statistical analysis, and results can be returned to local educators in as little as two weeks. This compressed timeline greatly benefits teachers' professional reflections on instruction and student learning (Dekker & Feijs, 2005).

One recurring question that many educators raise in relation to online testing is its comparability with paper-pencil testing. Some feel that students encounter higher test anxiety in the online format than they do in the paper-pencil format because online testing is less typical and possibly unfamiliar to youngsters. Others feel that, due to limited space on the computer screen, online testing designs are only focused on simple items or problem solving questions, and therefore, may be easier than paper-pencil testing. Research shows a promising result on this issue of comparability. For example, in Virginia, ongoing efforts have been made since 2001 to establish comparability between the computer-administrated tests and the paper-and-pencil formats.

Recent research indicates that comparability can be obtained, if valid, supportive evidence is provided (USDE, 2005). Additionally, some researchers have identified ancillary benefits that can accompany moves to online testing formats. Thomas (2003) reports,

> Virginia officials say that the online testing program also will have the following benefits: improved Internet access for teachers, greater ability to

> share instructional resources, opportunity to integrate technology into instruction, and increased communication among colleagues. (p. 5)

Computer-assisted tests are intended to measure individual student growth over time and are able to deliver immediate results (Stokes, 2005; Wilson, 2005). By the same token, it is generally believed that students can improve their achievement if given appropriate feedback (Cassady, Budenz-Anders, Pavlechko, & Mock, 2001; Trentin, 1997). Providing proper feedback for students is key in gauging the effectiveness of computer-assisted assessment. To a larger extent, the computer-assisted formative assessment is better viewed as "a learning tool" (Buchanan, 2000, p. 199) by which students are motivated to learn. Living in an age of information technology, students of today tend to not only learn comfortably from computers on a daily basis, but also engage easily in typical, computer-related assessments. Research shows some issues related to high-stakes online tests; the drawbacks of online high-stake tests such as inadequate numbers of computers, the need for increased network security, and the need for technology staff (Schaffhauser, 2011); students with special needs (McHenry, Griffith, & McHenry, 2004); and behavioral problems experienced by students (Landry, 2006). Yet, little is known about how students are internally motivated and what unintended outcomes are likely when students engage computer based assessments during a large-scale assessment period.

In sum, teachers are greatly assisted by receiving formative data they can use to improve their instruction. Lawson (1999) reported that the use of the assessment data from computer-aided assessments helped teachers to improve their instructional practices because diagnosing academic strengths and weaknesses of students occurs immediately when using computer based assessments (Challis, 2005; Chaney & Gilman, 2005). Therefore, the usefulness of the computer-assisted assessment in the context of the state accountability systems is likely to be determined by the degree to which these assessments truly provide useful formative information to teachers that can inform and impact their planning and instruction.

Methods

A qualitative, grounded theory approach is the best method for uncovering teachers' ongoing perceptions of the early implementation of PAWS tests. Although some forms of qualitative research have some a priori decisions in place, a qualitative researcher employing a grounded approach is primarily interested in exploring actual meaning-making processes of a person or a group of people for an event or phenomenon encountered in an everyday context. As researchers, we are attempting to describe teachers' ongoing perceptions involving their values, beliefs, knowledge, or emotions. Therefore, our methodological process combines the development of grounded theory with a qualitative case study research design. A case is defined as a bounded system that includes typical or unique characteristics (Stake, 2000). Accordingly, this case study is labeled as instrumental, because this elementary school is believed to show typical school and classroom characteristics under the implementation of PAWS. Instrumental studies explore typical cases that can illuminate issues that may be common across other similar cases. An Institutional Review Broad proposal was submitted to the university to which we belong, and we obtained approval before getting into the fieldwork stage.

A School and Participants Selected

The school purposefully selected (Patton, 2002) is an elementary school located in a

city in southern Wyoming. It is a Title I school, and many students in the classrooms we visited came from low-income families or single-parent households. Interestingly, this school holds varying reputations in the community to which it belongs over the last few years. One year it received the highest test scores and the following year the lowest scores. Mrs. Anderson (all names used in this study are pseudonyms) has served as school principal for more than a decade. She feels her main role is to be responsible for creating a safe learning environment in which staff and teachers do the important job of educating students. Mr. Reynolds has been in charge of the in-school computer lab over the last four years. The four teachers' profiles are as follows. All teachers voluntarily participated in this research following the recruitment presentation in a staff development meeting held in February 2006.

	Grade Level	Gender	Race	Teaching Experience	WYCAS**
Susan	3rd/5th*	Female	White	7	Yes
Pat	4th	Female	White	8	No
Brad	5th	Male	White	2.5	No
Troy	6th	Male	White	11	Yes

* Fifth-grade teacher during 2007/8

** Previous statewide test, Wyoming Competence Assessment for Students (1998-2005)

Note: We indicated in a faculty meeting that one participant in each grade was needed. Four teachers, one at each grade level, publicly volunteered to join the research.

Data Collection, Analysis, and Interpretation

Our roles as researchers were observers and interviewers. Our aims were to listen, watch, and learn, as we were new to most aspects of this implementation. We were very active in building a relationship with all educators in this local elementary school (Ceglowski, 2000) by frequently visiting with teachers in the teachers' lounge and computer lab, occasionally participating in faculty meetings, and respectfully observing activities in the classrooms. We felt that we naturally arrived at a point where participants brought us into the loop of their ongoing conversations about the school and their classroom practices. For example, one teacher remarked, "you have been in our school so much, it feels like you should be a member of our faculty."

Following is an outline of our research design for data collection:

Spring 2006	I: Pre-observations - First PAWS Field Test - Operational Tests II: Major informal and focus interviews	February 23-March 1, 2006 March 6-March 31, 2006 April 3-April 21, 2006 April 24-May 31, 2006
Fall 2006	III: Classroom observations and formal interviews	Sep to Dec 2006
Spring 2007	IV. Classroom observations (Winter PAWS tests) V. Formal interviews (Spring PAWS tests) VI. Classroom observations and formal interview	January to May, 2007 April to May, 2007 May 2007

Data collection spanned a year-and-a-half time period. During this time, the researchers collected data in the following ways:

- Attended two staff development meetings;
- Visited the computer room nine times (about 6 total hours);

- Observed classroom teaching 19 times (about 13 total hours);
- Conducted 30 total individual interviews with all six participants. The majority of the interviews were done right after the first operational test in 2006;
- Facilitated a one-hour focus group with all participants (This interview was conducted at the end of 2005/6 to obtain substantive information needed to make sense of the teachers' perceptions and responses to PAWS);
- Conducted seven additional informal interviews with other teachers and staff;
- Administered a simply written open-ended survey for students (grades 3 to 6, N=120) and teachers and staff (N=15) working in the school to identify overall perceptions of PAWS testing;
- Document analyses occurred throughout the research process to collect important resources such as materials on the State Department website and other pertinent documents, e.g., school intervention programs, test schedules and lesson plans, at the school and the classroom levels.

During all visits, observations, and interviews, information was recorded in the field notes. All interviews were video or audio-taped for later transcription.

Data analysis was viewed as ongoing, interactive, systematic, and reflexive in an effort to identify emerging themes across diverse views of six participants regarding the nature of PAWS tests. The researchers used thematic coding to search for and identify themes across all data. The two major data analysis and interpretation methods employed by the researchers are described as follows:

First, the researchers read interview transcripts and field notes from the beginning to the end multiple times during the summer and winter breaks searching for general and specific themes. In doing so, specific codes or representative key words were assigned to words, sentences or pages that included unique events or expressions. These analytic processes consisted of four steps:

1. Segmenting in which the researchers read texts, drawing lines to find units of data;
2. Initial coding in which the researchers named the units identified in previous steps and constructed lists of codes or tentative categories;
3. Axial coding that connected categories to one another; and
4. Eelective coding in which analysis was focused and connected to theory (Strauss & Corbin, 1990).

For example, as we explored the emergent theme *implementation problem/experience* by six participants, we first identified a gap between policy and practice as an important unit of data; we then named this unit, and other similar data, *open communication* among classroom, school, and state levels; we connected this unit to the related category implementation problem/experience; and finally, we developed the emergent, grounded theme related to *an open, flexible, or adaptive nature of implementation policy*.

Second, in relation to the above method following a tradition of grounded theory, we incorporated two other methods into the data analysis and interpretation processes. Data triangulation (Denzin, 1989) led us to cross-check sources. For example, we explored interview transcripts, field notes taken during observations, and responses to our open-ended survey questions to determine if themes present in one source also showed up in the others. Researcher reflexivity pushed us to deeply engage ourselves into re/constructing a thematic framework. "To reflect is to contemplate… [and] reflexivity is an interactive and cyclical

phenomenon" (Rossman & Rallis, 1998, p. 39). We constantly reflected on possible misconceptions or our own implicit and explicit biases, trying to minimize them as best as we could. The data triangulation process was conducted in many ways. Two examples are presented.

First, during data collection, we heard about different ways that Stanford Learning First (SLF) – a computer-assisted learning and testing program adopted a couple of months prior to the operational test – was planned to be implemented. The information technology staff and the principal viewed the SLF as preparation for the operational test and as an instrument designed to provide teachers with useful information on student learning. We were baffled by contrasting information shared by a teacher who viewed the function of SLF as solely supporting the needs of the technology staff and the school administration. She did not view the SLF as a valuable formative tool.

As learners and listeners, we found ourselves in agreement with and understanding both perspectives. After reflecting on the different perspectives from both sides, our interpretation was able to encompass both. Under the unprecedented pressure of this implementation, the IT staff was trying to be proactive, cautious, and helpful. Teachers, on the other hand, were sensitive to the changing school climate and the increasing pressure on the teachers and school, and so were resistant to the SLF, not so much because it couldn't be a useful instrument, but instead because it was yet one more new thing to do in an already rapidly changing school context. Each of these perspectives helped us better understand the other (Weick, 1976).

Another example involves a brief explanation of how we came up with four themes as a result of utilizing the selective coding process and reflexivity. Since we began this research with little information on both PAWS testing and this school, it took time and reflection to identify the key themes underpinning teachers' perceptions of their daily contexts. The four teachers that were a focus of this research all had intensified feelings about the unknown about the forthcoming test policies and procedures during the first year of implementation. This led us to use the following selective codes: "sink or swim," "standardized testing measuring simple skills," or "playing the game with policy out there and people in here."

During the second year of implementation, we observed teachers' confusion regarding two back-to-back tests, called the "Winter Operational Test" in January and the "Spring Operational Test" in April. Selective codes shifted from felt difficulties to realistic challenges, (e.g., "maintaining our reputation from the previous year," "winners vs. losers: who made it or who didn't make the winter operational test?" "more drills this year," "test on the unfinished curricula," and "more unnecessary meetings across grade levels," etc.). Eventually, however, our data collection enabled us to find instances where PAWS testing was being viewed as promising and hopeful, and so we decided to build a broader thematic framework that included: adaptive implementation policy, teachers' dilemmas, instructional change, and changing school culture. In both of the examples above, employing data triangulation helped us to deepen and evolve our research interpretations, and ultimately, our study's findings.

In addition to the triangulation described above to bolster the trustworthiness of the findings of this research, emergent themes were shared with participants. Member checking (Lincoln & Guba, 1985) was conducted to determine a relevant interpretation of what research participants think, act, and believe surrounding the intended and unintended consequences of PAWS tests.

Findings

> "Open, open, and I will heal your wounds! Please let me out!" it pleaded…Thus, according to the ancients, evil entered into the world, bringing

> untold misery; but Hope followed closely in its footsteps, to aid struggling humanity, and point to a happier future… (Guerber, 1907)

PAWS tests brought about a number of challenges to educators in Wyoming and, to a large extent, these ongoing challenges can be viewed as typical to implementation. Nonetheless, there are many distinctive consequences of this implementation that would offer insight into other contexts. To us, these consequences related to the implementation of the PAWS standardized testing program can be summarized in an analogy to Pandora's Box. In Greek mythology, the evils of the world were released when Pandora's Box was opened. The only thing remaining in the box was Hope. In the studied context, when the "Box" was opened (the testing program was introduced), many initial problems, concerns, and tensions surfaced. Like the evils in Pandora's Box, these problems seemed insurmountable; however, the issues presented by the PAWS testing implementation did not produce feelings of despair. Instead, a spirit of cooperation and collaboration between the Wyoming State Department of Education and local schools emerged. The "problems" presented opportunities to discuss issues of concern to all involved. What started out as chaos and fear, ended with a hopeful vision for the future. This is why we adopted the Pandora's Box metaphor. It is an analogy that draws our particular attention to both overwhelming problems in practice and hopeful messages that bode well for the future.

Taken together, the perceptions of the six participants, coupled with all other data sources, were convergent with the four significant themes presented in this section: (a) adaptive implementation policy at state level, (b) teachers' dilemmas: "Too much pressure" or "Kids are kids!", (c) instructional change: a mixed feeling of progress, and (d) changing school culture: pros and cons.

Adaptive Implementation Policy at State Level

All the participants of this research expressed feelings of uncertainty about PAWS procedures from day one. Expressions such as "I am not sure" or "I don't know" have ended up being the most frequent responses in formal and informal conversations. Even though significantly updated information was delivered at staff meetings on a weekly basis, teachers appeared to be confused about the whole nature of PAWS tests.

State Department's fast response.

Open-ended statewide surveys for both teachers/administrators and students were conducted one day after the field tests and the winter operational tests, respectively. Some immediate modifications for the following actual tests were made and publicized across the state immediately. For instance, teachers were extremely concerned about the reading tests. Long passages used in the online portion were reused in the written portion of the tests as well. Teachers felt students should take the written test first. This would assist students in completing the online multiple choice test items more effectively. Teachers and the principal from *this* school requested this change. Amazingly, the state decision on this matter (to accept the teachers' recommendation) was made within 48 hours!

Allowing educators to make important decisions.

The decision made by the state regarding two test windows was educational and democratic. In January, 2007, all the students took PAWS tests. In three weeks, online portions of results came out, and teachers anticipated which students would fall below

proficiency level. Meanwhile, there was no predetermined policy for who should or could retake the second PAWS tests in April. The complete results including open-ended tests were not available until a week before the beginning of the second PAWS tests. Despite the anger and anxiety associated with the late notice, teachers were surprised to hear that *each teacher/local school* could decide who should or could retake the tests. In addition, each teacher could allow certain students to take particular sub-tests in each subject. For example, a student could take a Number Sense sub-test in math if that student's score on the sub-test didn't meet the proficiency level. Indeed, teachers thought that this flexible policy was pretty reasonable, and most of teachers in this school had very few students take the second PAWS tests, either partially or entirely, while the majority of students didn't lose any instructional time during the four week testing period.

Teachers' Dilemmas: "Too much pressure" or "Kids are Kids!"

Looking at standardized testing from students' eyes may be worth investigating on the grounds that they can offer some insight into this testing business. Having students in grade three take a standardized test may be seen as too early given student development. In this section, we'll first describe how students perceive this high-stakes test, and second, illuminate teachers' perceptions of students' perceptions and their behaviors in the face of actual tests.

Perceptions of students on PAWS tests.

At some point in our fieldwork, we wondered about how the students felt about PAWS testing. By this point, we realized the teachers felt overwhelmed with all the incoming information. Three simply worded, open-ended questions for students were developed. When asked, "Why do you think you take the PAWS test?, students in this school offered a variety of intriguing answers. The scope of their answers ranged from "self" to "society," and the sequence of their answers across grade levels was from "the internal necessity" to "the external accountability." Taken together, responses of students in this school were grouped into four different themes: "basic test necessity," "diagnostic purpose," "meaningfulness or usefulness," and "accountability" that included certain progressive patterns of thoughts from younger to older students. The pattern of perceptions of students in this school regarding PAWS tests ranges from "diagnostic purpose" through "meaningfulness or usefulness" to "external accountability." To elaborate, third-graders put the highest priority on usefulness of this test, assigning mid-priority on diagnosis of PAWS. The shift of students' perceptions on this test was evident as fourth-graders saw the meaningfulness or usefulness for the test as one single highest priority, which appears to be a plausible answer for the reason why fourth-graders in this school rated their online-based PAWS test experiences relatively higher than did the other grades in this school. For fifth- and sixth-graders who had experienced the previous standardized test called WyCAS, responses for this question were considerably more external than those in the third- and fourth-grade in terms of accountability from a larger perspective. Their use of vocabulary expressing their perceptions was straightforward.
Following are some responses from fifth graders:

> "It tells the teacher, principal, and superintendent if you learned enough to go to 6, 7, 8 and all that"; "Our school is trying to raise their reputation; "The Wyoming teachers association wants to know how Wyoming schools are doing."

Some sixth-graders' comments were as follows:

> "The school district knows how good the teachers are teaching"; "… to see if the teacher does a good job"; "It shows that teachers are doing their job …"

Students in this elementary school encountered PAWS tests from different perspectives. Probably, some of them were guided by their parents, teachers, peers, or on their own. One thing that should be noted is that students are not passive in making sense of this testing ritual, regardless of whether they felt it is interesting or burdensome. To some extent, the way fifth- and sixth-graders interpreted PAWS tests in terms of an external measurement of student responsibility for their teachers, is both reasonable and tragic. It is reasonable because their opinions may be reflections of their parents and beyond. It is unfortunate students were not told the purpose of the test, that is the PAWS tests were solely concerned with assessing the performance of their teachers or school and not individual students.

Teachers' perceptions of students' perceptions/behaviors in the face of actual tests.

During the 2005/6 PAWS Field tests, the problems facing teachers were countless in conjunction with validity, measuring what the assessment is intended to measure, and reliability, showing a consistency of test scores. Three main areas that influenced these concerns involved time, test psychology of students in a social context, and indirect resistance against consecutive tests.

First, PAWS tests are not timed, so students can take as much time as they need. However, back-to-back test schedules (1½ hours for the online portion of math and reading consisting of about 60 test items) were unavoidable during the Field Tests. In fact, students knew that because PAWS tests were not part of their grades, they didn't have to take them very seriously. This was in part caused by parents' negative comments on the necessity of PAWS tests, consciously or unconsciously (i.e., PAWS tests are solely intended for measuring the effectiveness of teacher or school performance). In short, the administration of PAWS tests seemed typical in that students did their job as planned, neither excited nor depressed, but happy to have the opportunity for more frequent breaks during and after the tests.

Secondly, in regard to the test psychology of students, teachers found a sense of "indirect peer pressure or competition" among students in this social context. Students tended to be rushed to finish the tests as they saw other students whose abilities were similar to and lower than theirs finish the tests ahead of them. In particular, this observation came out of online portions of testing. The fourth-grade teacher expressed her frustrated response:

> I had one kid and she said 'I don't like this test. I am the last one done.' She felt as if that was a negative thing to be the last one to be done. It doesn't matter… as long as … if you do it correctly, that was what mattered. She doesn't necessarily see it that way. She sees that… if you can do it FAST and do it RIGHT, then I guess that would be the best! But just being done QUICKLY is important for a lot of kids!! That is unfortunate. (Pat, transcript, May 9, 2007)

In order to help students focus on their computer screen, the IT staff placed paperboards on each side of the computers, which was recognized as one of the *best* practices in administrating online portions of testing in Wyoming. Nonetheless, the fourth-grade teacher mentioned that most of her students, if not all, finished their online tests in a manner

similar to the domino effect. In a quiet computer room, the noise caused by moving chairs appeared to pressure students both mentally and socially.

Third, related to the second concern above, a few students did not take the tests seriously. They looked over the screen, clicked the answer, and moved on to the next screen. The mindless use of the mouse-clicking behavior of students, associated partially with test anxiety, was viewed as anonymous in a public place like a computer room, and more importantly, could be an indirect resistant response of students who are required to take many standardized tests throughout the year. The sixth-grade teacher expressed a feeling of uneasiness as to the necessity of this PAWS test from a student's perspective:

> They want to know, 'Do I get the grade for this?' The truth is no. You don't get the grade for this. I am a little hesitant to tell them, because I don't want them to … I don't get the grade on it. Who cares! For some of the kids, they want to know what benefits it does have for me. I don't know if the kids see what the benefit is. I know some of teachers don't see the benefit. I think that is hard. It is hard to be motivated, if you don't find some intrinsic benefit. I don't think they find … I don't know … Some kids like it. They like to show off that they know how to do everything but some kids who struggle … (Troy, transcript, May 24, 2006)

It would be a surprising fact to know the extent to which students engage themselves in the test taking process given the number of tests they take each month. During the second year of implementation, the fact that 6 out of 16 fifth-grade students used only a piece of scratch paper to take the online math test indicated that most of students mentally did math without even grabbing pencils at hand. Thus, thoughtless mouse-clicking was possible.

Instructional Change: A Mixed Feeling of Progress

Ambiguous validity: TECH-knowledge vs. tech-KNOWLEDGE.

Hope follows a sense of ambiguity. A question of 'what is intended to be measured in the world?' had hovered over the heads of educators in this school from the beginning of PAWS testing. The question of validity in the context of computer-assisted testing kept surfacing. Mrs. Anderson, principal, asserts, "It would be problematic if students were only tested on the mechanics in which the proficiency of the manipulative skills of computer played a major role in measuring actual knowledge and skills" (Personal communication, Feb 25, 2006). What educators were worried about was an unbalanced image of technology-driven measurement between TECH-knowledge and tech-KNOWLEDGE. The former is overly focused on computer proficiency, while the latter focus is on validity. That is, what was being measured? Computer proficiency or knowledge?

Supposedly, teachers should have known more about the nature of SLF (Stanford Learning First) before students took the actual Field Tests. SLF was merely seen as practice intended to get students ready for PAWS tests, focusing on TECH-knowledge. Nonetheless, Pat, a fourth-grade teacher noted with a half-smile in her face, "I thought it gets them ready for the formality of the tests. I just brought kids to the computer lab and came back to my room and got my math ready" (Personal communication, April 25, 2006).

Based upon the state's initial plan, the SLF should have beyond what the teacher just mentioned above. In reality, rather than following the planned technology support, teachers used a commercial test preparation book of their own, providing students with feedback in the preparation for PAWS tests. They seemed to pay less attention to TECH-knowledge. What

they did hope was that the IT staff would take care of that part. Also, they hoped that their own test preparation strategies concerned with validity would work out. "Using the commercial test preparation books was what I have done before and hopefully it worked out for this new PAWS test at this time," said Susan, a third-grade teacher (Personal communication, Feb 15, 2006). Such a hopeful feeling mixed with a sense of curiosity led teachers to peek at test items in the field tests over the shoulders of their students in the computer room to see what test items looked like. They were concerned with the discrepancy between ideal/technical efficiency and practice/validity. To the extent to which test items were likely to be easy and straightforward on the screen and that students seemed to do what a majority of students were supposed to do, teachers felt more or less comfortable with PAWS tests as time went. Later, when receiving relatively high test results, their overwhelming concern with the initial validity question largely evaporated. The online portion of PAWS tests invoked many suspicions in regard to validity but did not enter the subsequent area of heated concern, because PAWS testing in this school was regarded as successful in accurately measuring basic knowledge and skills in reading and mathematics because the scores were high.

Assessment Descriptions.

The PAWS test is a formative-based assessment that is intended to provide teachers with information about their instruction. Embedded within this description is an expectation for improving instruction. To this end, teachers are given an assessment guideline, called the AD (Assessment Description), that targets the standards and benchmarks to be tested by PAWS. This guideline seemed somewhat unfamiliar to teachers. The question is, "To what extent is the AD innovation known at the classroom level?"

The gap between policy and practice seemed wide. It may be naïve to believe that information available on the state website would automatically seep down to the classroom level. Assessment Descriptions as core or essential portions of standards and benchmarks being currently used in the state of Wyoming is one example. That is, given the conceptual framework of ISA (Instructionally Supportive Assessment) in which standards/benchmarks are aligned with assessment, the Wyoming State Department of Education has developed AD that serves as a bridge between classroom instruction and assessment. Accordingly, teachers are expected to keep AD in their minds while planning and teaching to the standards and benchmarks in everyday context which, in turn, will seamlessly prepare students for the PAWS tests. In this school, however, teachers in this study told an entirely different story in the focus group interview. In the discussion that follows, surprisingly enough, the remaining three teachers came to dimly remember being told to visit the state website and print off the AD materials for use:

> Fourth-grade teacher: Assessment Descriptions didn't help me at all.
> Fifth-grade teacher: I have never heard. . .
> Sixth-grade teacher: I didn't know.
> Fourth-grade teacher: It was like… small parts of standards…. That is it. If I am supposed to teach a standard on my grade, even if it is not on PAWS, I will teach it. They will have it. That is unfortunate ... Did I give them broader experiences for my kids? Should I just teach two standards, because they are supposedly on the test? They [entire standards] can do a lot of things and they may never show up on that test. That one little piece [Assessment Description] is right there and

> we are valuing that, but we are not valuing other things … Condensing existing standards down to, say, a dozen, gives a better idea of what will be tested, but it was perceived of as inappropriate, I think. (Focus Group Transcript, April 28, 2006)

The mixed expressions of teachers in this research indicated that AD had been disseminated and implemented in this school, probably just like any other school in Wyoming. Nonetheless, to the extent to which teachers are externally informed and internally made aware of AD as innovation did not seem to get inside the heads and hearts of teachers. Three other teachers appeared to feel that AD was like a document that came from the administrative office, one that they could take a closer look at 'later, not right now!' While more aware of and better informed about AD during the second year of implementation, teachers felt pressured to incorporate it into their existing curriculum and instructional practices.

A spirit of formative assessment is everywhere but nowhere.

PAWS tests are intentionally designed to give classroom teachers feedback useful for improving instruction. There are two dimensions of the feedback that teachers receive from PAWS tests.

First, the feedback is timely. Feedback in the form of official results of online tests is formative in nature because teachers are supposed to receive a set of data within two to three weeks after testing is administered. Regarding the arrival of immediate test results, the year 2007, when two windows of PAWS testing were administrated, teachers experienced a spirit of formative assessment in this regard. Receiving results this fast certainly made teachers excited not only because it fulfilled their curiosity about their classroom results, but also because they thought they could use the data for improving their teaching practices. Susan, a third-grade teacher expressed her excitement:

> Not only can't I wait to see the test results of my class, but since I will be teaching 6 grade next year, it will be very interesting to see test results of fifth-grade students as well to be able to make some adjustments for particular students at the beginning of school year … (Interview Transcript, May 24, 2006)

Second, as suggested by Susan, the feedback is prescriptive. Since students' performances were disaggregated into five to six domains of standards and benchmarks, and were represented on their report cards, teachers during the next school year could make use of these assorted data for their new classroom teaching. More specifically, teachers were given specific information on student performance in several domains in terms of color-coded, traffic-light signs: proficient or above (blue color), basic (yellow), or below-basic (red).

These coded signs could be used at the classroom level in many different ways. Teachers recognized the possible value for these coded signs for future use with their students. It was obvious that teachers benefited from the PAWS report cards to the extent they re-interpreted them. That is, teachers felt that PAWS report cards were useful, not just because the somewhat simplified blue, yellow, and red signs gave such directive indicators for a group of students in their classrooms, but because they could use them as supplementary data in determining the level of student performance as they manually collected more direct assessment data in their own classrooms.

Nonetheless, the basic fact that PAWS tests were based upon a spirit of formative assessment appeared not to surface as planned. It seemed to draw a mixed set of responses from teachers about the PAWS tests. For example, an overall response of third- through sixth-grade teachers in this school about the degree of alignment between PAWS and what they taught was mixed. Some said, "PAWS testing measured what I have been teaching" and the others said, "Unlikely." On top of this validity issue, they also expressed a mixed feeling about the nature of PAWS testing in terms of whether or not PAWS testing actually measured important knowledge/skills aligned with state standards. All four teachers of this study expressed their satisfaction about the online tests as well as PAWS report cards. Yet, teachers did not seem to actually change their instruction based upon the high online test scores or conveniently designed color-coded signs of student performance as reported. It had to do with an uncomfortable, but progressive, reflective feeling about the discrepancy between what was taught in one's classroom and what was actually tested. Troy, a sixth-grade teacher, identified a gap between his teaching and PAWS testing:

> The question we were debating is "Is that cheating?" Last week, I saw that type of question on the Field Test. That is why I am doing this whole thing on THEME now… As far as I am concerned, the word THEME is not there … Am I teaching the test? Am I somehow cheating … by saying it to my kids, right before the actual Field Test? I don't know. I don't think I would not do it, but I can see someone's argument that I am getting too close to teaching exactly what is on the test … sixth-graders have never talked about theme and there should not be on the test about theme. That is not fair … I don't know, maybe something I am going to change in the future … (Transcript, May 16, 2007)

In short, the concept of "theme" that this teacher learned would be on the upcoming test, which he had not yet taught, seemed to make him anxious and frustrated. This issue of the missing link between standardized testing and classroom teaching can easily ignite a high level of reflection from teachers to the extent that they put it in both short and long-term arrangements of teaching and assessment at the classroom level. The last statement, "I don't know, maybe something I am going to change in the future," implies an image of "outside in" change (external force leading to change), as opposed to be "inside out" change (inner reflection leading to change). That is, this kind of inside-out reflection or feedback should occur in the spirit of formative assessment, when teachers receive safe, open-ended feedback right after the test.

Changing School Culture: Pros and Cons

Most of teachers in this elementary school, exclusive of teachers in first- and second-grades, appeared to feel that the PAWS is largely "Everyone's business," rather than "Someone else's business." A spirit of school change is likely to be embedded in "we"-consciousness. A bond between the principal and teachers was close, and in effect, emerging concerns or issues were quickly circulated and discussed. In a sense, the feeling was that "we are all in the same boat!" In the monthly staff development meetings and during the summer breaks, conversations over PAWS were mostly active and sometimes controversial. More consistent educational dialogues were evident throughout the last three semesters from spring, 2006 to spring, 2007. Teacher dialogue within and across grade levels with regard to how to increase the teaching of certain topics and deal with struggling students were made on a daily basis. A fourth-grade teacher, Pat, said, " … Compared to previous years, probably through

my all years of teaching experiences, I have never got involved in this much conversations with my grade level teachers … We meet and talk almost everyday …"

In a nutshell, PAWS tests showed, with proper measure, an adequate yearly progress in Wyoming. This school was not exceptional. For example, a dramatic increase in students rating proficient at fourth-grade is evident over the last three years; in reading, scores increased from 35% to 84% to 95%; in writing, scores increased from 35% to 75% to 77%; in math, scores increased from 39% to 95% to 95%. In late summer of 2006, during the first year of implementation, this school reported an amazing outcome on the PAWS tests: 95% math proficiency at the fourth-grade level. In fall 2006, after having this very surprising result, ongoing discussions between fourth- and fifth-grade teachers were suddenly erupting. The fifth-grade teacher who now had completely math proficient fourth-graders came to wonder if this outcome would happen again in her grade level, "What if we don't make it? We have got a lot of pressure from the community." On the other hand, the fourth-grade teacher was also faced with the same pressure, because she herself doubted test scores for certain students who didn't reach that proficient level in her classroom assessment.

While a mix of feelings about the present and future was prevalent, the cohesiveness of this local school was revealed in the fact that "the pressure from everywhere" created a spirit of collaboration among teachers and administrators. Nonetheless, one teacher was frustrated with these more frequent and intensive gatherings. This teacher felt the meetings affected class prep time and were unnecessary. For this teacher, a focus on concepts or inquiry-based teaching was preferred over teaching that focused on facts or test preparation-based learning. Considering this worry, expectations, and some individual disagreement, the school and local community were again pleasantly surprised with 95% math proficiency at not only the fifth-grade but also the fourth-grade level. The biggest winners were, of course, the three fourth-grade teachers who achieved these back-to-back surprising outcomes for the fourth-grade for two years in a row.

Even if the test scores were not the only measure by which to gauge the success of school change or improvement, it is likely that the teachers in this school felt having PAWS tests were beneficial, not just because the teachers proved themselves, but because they were naturally encouraged to share what they think and do with their colleagues. However, some teachers felt burdened by the obligations of daily and weekly meetings to discuss test score results. Such over-emphasis on collaboration that some teachers felt seems unavoidable in this high-stakes-testing era. Given that this external requirement heavily influences schools, the pros and cons of changing school culture in this school's context will continue with a mix of collective collaboration within individuals' privacy needs on a daily basis.

Discussion and Conclusion

A description of what teachers think, act, and believe during the implementation process of this state-wide assessment sheds light on identifying specific lessons or insights that not only help improve the course of future implementation of PAWS tests, but also helps us understand the changing nature of educational practice fostered by the ongoing pressure of accountability. Following are three aspects of PAWS tests that need to be discussed in conjunction with theory and practice.

First, Cawelti's (2006) concern with side-effects of NCLB's accountability-driven approach in the educational enterprise was evident in that teachers were particularly discouraged with an alignment or validity problem between PAWS testing and their own curriculum and teaching practices. The cause and degree of such discouragements facing at least three teachers (not much for Brad, a beginning fifth-grade teacher) appeared to be moderately philosophical and fundamental as they struggled to determine whether concepts

being tested were accurately interpreted or whether use of vocabulary words in certain test items were grade appropriate. They were discouraged with test validity but felt that these matters were minor compared to considering how to approach test preparation differently next year.

The notion that they would prepare for PAWS testing differently is largely related to one of Cobb and Rallis's (2008) five response-types to NCLB. That is, educators at this school in the State of Wyoming are likely to be seen as the Opportunists who consider a top-down sense of accountability as a way of improving their ongoing problems. While PAWS testing was overwhelming for the beginning teacher, Brad, those who have expressed success with prior standardized testing experiences considered PAWS testing an opportunity to improve their instructional practices by expanding the scope of vocabulary words or by including certain literary concepts in their future instructional plans.

Hence, the above finding of this study is largely in line with those of implementation research studies using a mutual adaptation perspective (Harrison, 1997). Theory and practice are interactive and further negotiated in an effort to produce better outcomes for those participating at varying levels. Reframing the existing values or norms in schools and classrooms is evident in that teachers, staff, and an administrator constantly changed their structure and organization of curriculum and teaching, the computer room, and staff development prior to the testing period and in its aftermath. At the same time, the Wyoming State Department of Education was willing to hear the voices of educators and dared to change its testing format and logistics in the middle of the implementation process.

Despite the general support for these findings from previous implementation studies, however, something was unique in the last two years of the implementation process. Change is generally defined as a process, not an event. That is, "change is a process through which people and organizations move as they gradually come to understand, and become skilled and competent in the use of a new way" (Hord, 2001, pp. 4-5). The meaning of change within this school is a mix of event and process because such events-based change initiatives served as fuel to ignite the very process of change at large. Occasionally, the change process demands momentary events that foster teachers' critical reflections. Given this distinctive meaning of change, for example, switching from two test windows to one test window is one case. Those teachers who wished to have multiple test measurements throughout the year during the change process were faced with unexpected challenges from the teachers themselves. This finding may require further investigation by which the mutual adaptive theory of implementation needs to be more sophisticated in contemporary school contexts.

Second, the dream of reformers that instructional change would occur as long as teachers are given Assessment Descriptions was not evident in this case in the state of Wyoming (Hord, 2001). The literature on teacher thinking or cognition showed that innovation, if not fitted to teachers' values and beliefs, was perceived as inappropriate by teachers. Teachers were hesitant to legitimize the use of standardized testing for measuring accountability at the expense of their own beliefs regarding quality education (Calderhead, 1987). However, teachers were clearly aware of the fact that PAWS tests involved pragmatic value, at least in the way that a new individualized Per-Report Card would be forthcoming. All participants expressed curiosity about how this report card would contribute to inform their instructional practices next year, as if they were waiting for Hope to escape from Pandora's Box.

Third, the initial idea intended to provide proper, timely feedback with teachers in the name of formative assessment is still unfinished business (Buchanan, 2000; Dekker & Feijs, 2005). Receiving results of the winter's operational test one week prior to the spring test resulted in unhappy reactions on the side of teachers. This issue is a clear example occasionally encountered in the literature on implementation in terms of who controls what.

The fact that the Wyoming State Department of Education has little control over this technical aspect is evident, and all the intended outcomes that resulted from the power of technology turned into unintended ones. In this regard, it is likely that the ambitious idea of PAWS testing, to make a typical assessment a useful and meaningful tool for teachers, evaporated. Teachers' work remains unchanged.

Further research needs to be conducted to gauge how elementary school teachers in Wyoming make sense of their abnormally busy classroom contexts on a daily basis including, but not limited to, the inclusion of many content-specific assessments, mandatory tutoring for students at basic- or below-basic levels, or unnecessary professional meetings related to test programs (Lawson, 1999). In addition, further research is necessary in relation to psychological and social aspects of online testing for young, mid-elementary 3rd and 4th graders. Their voices and social behaviors under the circumstance of high-stakes standardized testing are relatively unheard and unknown. These young students are not passive in making sense of the current test ritual and their viewpoints are in part guided by their parents and peers. Teachers are public servants and gain a lot of attention from their local community. Collaborative, community-based action research is needed to explore experiences of these mid-elementary students as well as various community members' expectations of teachers.

In conclusion, some discrepancies were observed between plan and action of this statewide assessment. What is missing from the ideal vision of this new statewide assessment initiative is a space reserved for teachers to control their own world of curriculum and instruction. Nevertheless, teachers were excited to take the PAWS' test scores into consideration, in meeting external requirements, and in obtaining feedback for academically challenged students. One has to find Hope in Pandora's Box. PAWS tests have the potential to improve the state's accountability system and individual teachers' instructional practices in the classroom. The daily professional teaching practices of these teachers, impacted by the relatively balanced pressure from policy and from their colleagues within school, are likely to be progressively connected to curriculum, instruction, and assessment, as teachers get involved in more dialogues at grade level and across grade level aimed at fulfilling specific learning needs of each individual. It is hoped that this progressive connectedness in professional work will continue to be in harmony with teachers' personal zone of comfort where they feel some control on what and how to manage their busy daily schedules.

True, Pandora's Box has been opened with a big hope, but the very hope one wishes seems to have not come yet. No doubt, it will come out soon, allowing it not only to fix external problems facing teachers, but also provide long-term insight into ways to enhancing internal pedagogical and curriculum dialogue for diverse learners in the name of assessment/ accountability.

There is a limitation for this study. Because the main focus is on describing an implementation process of a new statewide testing program in an elementary school in the state of Wyoming, quantitative studies cannot provide answers to applications beyond an individual case. A generalization of qualitative findings may occur, case by case, depending on their similarity (Lincoln & Guba, 1985). Unlike conventional, quantitative inquiry, qualitative inquiry involves an in-depth investigation of a small number of cases. To this end, a qualitative inquirer has an obligation to produce fresh insights surrounding the case being studied, insights that not only help the readers enrich their existing knowledge, but also expand it to go beyond imagination. It can be regarded as naturalistic in the general construction of knowledge by humans (Stake & Trumbull, 1982) or as cognitive in an individual's knowledge construction (Donmoyer, 1990).

References

Berliner, D. (2009). *NCLB outrages: Why rising test scores may not mean increased learning*. Retrieved from http://susanohanian.org/show_nclb_outrages.php?id=3738

Buchanan, T. (2000). The efficacy of a worldwide web mediated formative assessment. *Journal of Computer Assisted Learning, 16*, 193-200.

Calderhead, J. (Ed.). (1987). *Exploring teachers' thinking*. London, UK: Cassell Educational.

Cassady, J., Budenz-Anders, J., Pavlechko, G., & Mock, W. (2001). *The effects of internet-based formative and summative assessment on test anxiety, perceptions of threat, and achievement*. Seattle, WA: The American Educational Research Association. (ED 453815)

Cawelti, G. (2006). The side effects of NCLB. *Educational Leadership, 64*(3), 64-68.

Ceglowski, D. (2000). Research as relationship. *Qualitative Inquiry, 6*(1), 88-103.

Challis, D. (2005). Committing to quality learning through adaptive online assessment. *Assessment and Evaluation in Higher Education, 30*(5), 519-527.

Chaney, E., & Gilman, D. (2005). Filing in the blanks: Using computers to test and teach. *Computers in the Schools, 22*(1-2), 157-168.

Cho, J. (2000). Curriculum implementation as lived teacher experience: Two cases of teachers. *Unpublished doctoral dissertation*. The Ohio State University, Columbus, OH.

Clark, C., & Yinger, R. (1977). Research on teacher thinking. *Curriculum Inquiry, 7*(4), 279-304.

Cobb, C., & Rallis, S. (2008). District responses to NCLB: Where is the justice? *Leadership & Policy in Schools, 7*(2), 178-201.

Dekker, T., & Feijs, E. (2005). Scaling up strategies for change: Change in formative assessment practices. *Assessment in Education: Principles, Policy, and Practice, 12*(3), 237-254.

Denzin, N. (1989). *Interpretive interactionism*. Newbury Park, CA: Sage.

Donmoyer, R. (1990). Generalization and the single-case study. In E. Eisner & A. Peshkin (Eds.), *Qualitative inquiry in education* (pp. 175-200). New York, NY: Teachers College Press.

Fullan, M., & Pomfret, A. (1977). Research on curriculum and instruction implementation. *Review of Educational Research, 47*(2), 335-397.

Gay, J. (2007). The rhetoric and reality of NCLB. *Race, Ethnicity and Education, 10*(3), 279-293.

Goodman, K. (2007). *Resistance to NCLB.* Retrieved from http://choosingdemocracy.blogspot.com/2007/10/resistance-to-nclb-ken-goodman.html

Guerber, H. (1907). *The myths of Greece and Rome*. Retrieved from http://alum.wpi.edu/~p_miner/evilones.html

Harrison, J. (1997). *Implementing a multicultural experiential sociology curriculum: Mutual adaptation and reframing theories of action: Preview*. Chicago, IL: The American Educational Research Association. (ED 412165)

Hord, S. (2001). *Implementing change: Patterns, principles, and potholes*. Boston, MA: Allyn & Bacon.

Kirk, D., & MacDonald, D. (2001). Teacher voice and ownership of curriculum change. *Journal of Curriculum Studies, 33*(5), 551-567.

Klein, A., Zevenbergen, A., & Brown, N. (2006). Managing standardized testing in today's schools. *Journal of Educational Thought, 40*(2), 145-157.

Landry, D. (2006). Teachers' (K-5) perceptions of student behaviors during standardized testing. In B. Stern (Ed.), *Curriculum and teaching dialogue* (pp. 29-40). Charlotte, NC: Information Age.

Lawson, D. (1999). Formative assessment using computer-aided assessment. *Teaching Mathematics and its Application, 18*(4), 155-158.

Lincoln, Y., & Guba, E. (1985). *Naturalistic inquiry*. Beverly Hills, CA: Sage.

Lynd, C. (2000). *The new generation of standardized testing*. Washington, DC: Center for Education Reform.

McHenry, B., Griffith, L., & McHenry, J. (2004). The potential, pitfall and promise of computerize testing. *The Journal, 31*(9), 28-31.

McLaughlin, M. (1990). The Rand change agent study revisited: Macro perspectives and micro realities. *Educational Researcher, 19*, 11-16.

Miles, M., & Huberman, M. (1984). *Qualitative data analysis: A sourcebook of new methods*. Beverly Hills, CA: Sage.

Olebe, M. (1999). California formative assessment and support system for teachers (CFASST): Investing in teachers' professional development. *Teaching and Change, 6*(3), 258-271.

Patton, M. (2002). *Qualitative research and evaluation methods* (3rd ed.). Thousand Oaks, CA: Sage.

Popham, J. (2001). *The truth about testing: An educator's call to action*. Alexandria, VA: Association for Supervision and Curriculum Development.

Popham, J. (2002). White paper: Implementing ESEA's testing provisions: Guidance from an independent commission's requirements. *The Commission on Instructionally Supportive Assessment.* Retrieved from http://www.ioxassessment.com/download/ImplementingESEAsTesting.pdf

Rossman, G., & Rallis, S. (1998). *Learning in the field*. Thousand Oaks, CA: Sage.

Schaffhauser, D. (2011). High-stakes online testing. *The Journal, 38*(6), 28-39.

Snyder, J., Bolin, F., & Zumwalt, K. (1992). Curriculum implementation. In P. Jackson (Ed.), *Handbook of research on curriculum* (pp. 402-435). New York, NY: Macmillan.

Stake, R. (2000). Case studies. In N. K. Denzin & Y. S. Lincoln (Eds.), *The handbook of qualitative research* (2nd ed., pp. 435-454). Thousand Oaks, CA: Sage.

Stake, R., & Trumbull, D. (1982). Naturalistic generalization. *Review Journal of Philosophy and Social Science, 7*, 1-12.

Stokes, V. (2005). No longer a year behind. *Learning and Leading with Technology, 33*(2), 15-17.

Strauss, A., & Corbin, J. (1990). *Basics of qualitative research: Grounded theory procedures and techniques.* Newbury Park, CA: Sage.

Thomas, W. (2003). Status of online testing in SREB states. Atlanta, GA: Southern Regional Education Board. (ED 477360)

Trentin, G. (1997). Computerized adaptive tests and formative assessment. *Journal of Educational Multimedia and Hypermedia, 6*(2), 201-220.

U.S. Department of Education. (2005). Online assessment in mathematics and writing. *Institute of Education Sciences NCES 2005-457*. Retrieved from http://nces.ed.gov/nationsreportcard/pdf/studies/2005457.pdf

Vogel, L., Rau, W., Baker, P., & Ashby, D. (2006). Bringing assessment literacy to the local school: A decade of reform initiatives in Illinois. *Journal of Education for Students Placed at Risk, 11*(1), 39-55.

Weick, K. (1976). Educational organizations as loosely coupled systems. *Administrative Science Quarterly, 21*, 1-9.

Wilson, R. (2005). Targeted growth for every student: When this district wanted an assessment program with practical applications to teaching and learning, it selected a computerized adaptive test that measures student growth over time. *Leadership, 35*(2), 8-12.

Wyoming State Department of Education. (2006). *Proficiency Assessments for Wyoming Students*. Retrieved from http://www.k12.wy.us/SAA/Paws/index.asp

Author Note

Jeasik Cho is an associate professor in the Department of Educational Studies at the University of Wyoming, USA. He received his doctorate at The Ohio State University, USA. His research interests include qualitative research, curriculum theory, classroom assessment, and multicultural education. His peer-reviewed journal articles appear in Qualitative Inquiry, Journal of Qualitative Research, QSE: International Journal of Qualitative Studies in Education, Teacher Education Quarterly, Multicultural Education Review, TABOO: The Journal of Culture and Education, International Journal of Educational Research and Synergy, and so on. Correspondence to the author: Dept. 3374, 1000 E. University Ave. Laramie, WY 82071. 307) 766-3128. jcho@uwyo.edu

Brian S. Eberhard is a doctoral candidate in the Department of Curriculum and Instruction at the University of Wyoming, USA. His research includes social studies education, web-based social studies instructional models, and civic curriculum development and assessment. His email: eb@uwyo.edu

The authors would like to sincerely thank Dr. Dan Wulff for his critical and constructive review of our initial draft. We are also so appreciative for our wonderful research participants for their never-ending support! This project would not be possible without their dedications and wisdoms. Lastly, Jeasik Cho thanks Drs. Francisco Rios, Allen Trent, Kevin Roxas, and Won-Hee Lee for their constant support and useful comments on later versions of this article.

Article Citation

Cho, J., & Eberhard, B. S. (2013). When Pandora's Box is opened: A qualitative study of the intended and unintended impacts of Wyoming's new standardized tests on local educators' everyday practices. *The Qualitative Report, 18*(Art. 20), 1-22. Retrieved from http://www.nova.edu/ssss/QR/QR18/cho20.pdf