NCS Pearson Performance Scoring Center scores the Washington Assessment of Student Learning (WASL). The pool of employees hired to score received identical training from Pearson, then were assigned to either the WASL or the Ohio Proficiency Test (OPT). Following is an approved synopsis of a report by an actual Pearson scorer, John Koudela III. Mr. Koudela was employed in the Spring of 2003 to score the OPT at the NCS Pearson Performance Scoring Center in Auburn, WA (also often called Pearson Education, Pearson Educational Measurement, and NCS Pearson). The last three pages discuss actual questions and the scoring criteria.

On the Scoring of OPT/WASL Educational Assessments by John Koudela III

Summarized by CURE, Feb 2005

This report contains questions from the Ohio Proficiency Test (OPT) for illustration, criticism, and comment. The OPT 6th Grade Science questions I examined in this report are currently available online at http://www.ode.state.oh.us/proficiency/previous_test/6th_grade/March%2003%206th%20Grade%20Science.PDF
It is the scoring of the assessments that this report is about.

The following August 27, 2000, *Seattle Times* article accurately reflects the process when I scored for NCS Pearson in 2003: "Temps spend just minutes to score state test; A WASL math problem may take 20 seconds; an essay, 2 1/2 minutes." Jolayne Houtz, *Seattle Times* staff reporter.

Qualifications for Scoring: The qualifications for scoring varied depending on the source consulted – the Pearson handbook, newspaper ads seeking to employ scorers, or the web site of The Partnership for Learning (an organization supported by big business which advocates for the WASL). **Most of the individuals hired by Pearson to score did *not* have teaching backgrounds or a degree related to one of the preferred fields.**

My qualifications for the scoring position: My degree is in Recreation and Leisure Studies with a minor in Psychology. I have a 25+ year work history in purchasing and component engineering in the electronics manufacturing industry. However, I do not have a degree in engineering or a comprehensive science and math background. Others with whom I worked had degrees ranging from Physics to Law to Political Science. Most of us were out of work before landing a temporary position with the NCS Pearson Performance Scoring Center.

Hiring Process: I took a test on various subjects and, after an interview, was offered a position to score the 6th grade science OPT. There were so many positions to fill I suspect hardly anyone was turned away. All positions were temporary, including the site manager. I was paid $11.40 an hour. Not all in my group were teachers. Pearson clearly was lenient in their selection to acquire the 350 or so people needed to score the thousands of assessments within their schedules. However, this was not what they told their education customers, teachers, and supporters.

Training at Pearson: Pearson spent two days training us. We worked on the questions we would be scoring after our training by using older answers. We worked through the problems, reviewed scores on the anchor tests (actual questions with real answers). Then we scored practice tests – example questions, not real ones, but with typical answers. The training supervisors then told us what were the likely acceptable answers. Before trainees could score real questions they had to pass three quizzes of about ten questions each and get a score of 80% or better out of those three tries or they would not qualify to score. It was really all about the ability to keep track of what is allowed points and what is not which changed daily.

Working at Pearson: My views about Pearson and these kinds of assessments changed altogether as I learned how unfair test questions were and how the criteria for scoring changed daily. Students could be hung for *how* they were scored. Parents, teachers, and schools were being misled. I knew others needed to learn the truth about what was really going on behind these closed doors at the NCS Pearson Performance Scoring Center in Auburn, WA.

The supervisors were constantly adjusting the scoring guides. Scoring changed daily on most of the questions because not all the variables were known in advance. They were determined as the students' answers were examined against the rubrics. When scorers had questions about how to score a particular answer, the supervisors would relay the inquiry to the "Range Finding Committee." Several of us asked if we would need to re-score answers we did prior to these new decisions. We were told not to worry about that. Unless those answers were re-scored, those students missed out.

"The chair is orange" was the constant daily mantra to all scorers. We were to accept whatever the rubrics and Range Finding Committee decisions were, regardless of our own judgment. We *had* to score according to what the Committee said. As we were repeatedly told, "The chair is orange" if the Committee says it is.

Don't talk to the Press or wear clothing with words on them while working at Pearson. We were told not to speak with the press should they show up in the parking lot. Now that I know what these questions were like and how they were scored – I feel it is absolutely necessary the press and everyone be informed about how they are scored and who scores them. I have recently sent (March 2004) a letter to NCS Pearson stating I am no longer in agreement with anything I signed with them when I worked for them in 2003. I simply will not stay silent about what Pearson is doing to students and to education. The public deserves to know the truth because it will adversely affect their children and their future.

We were also told not to speak about any subject at work that may cause someone else to become emotionally affected – so we had to be careful what to talk about. And we were told to be very careful not to wear any clothing that had words or images on them that could incite adverse emotional responses and reactions to current events in the world.

Reliability of scoring: For reliability, scores were checked by another scorer. As long as two scorers agreed on what the score should be for a question it was considered good. The problem with that kind of reliability is that both scorers could be wrong.

The other scoring reliability check examined whether or not incorrect scores were at least adjacent to the correct scores given on the same answer. An adjacent score was considered only partially correct. For example, answer A could be scored with as many as three different scores: 2, 1, and 0. If the answer's correct score was determined to be 2 points. Then 1 is an adjacent score and thus is a reliability level better than a score of 0. A scorer's reliability was then measured by how many wrong scores, how many adjacent scores, and how many correct scores were given. *Because of the changing variables determining correct and wrong answers, the reliability percentages varied daily!*

We all wondered how Pearson could have any reliability from one day to the next. *One day's answers for a given question were wrong and another day the same answers given by other students for the same question were correct.* I don't believe most of the students who took this assessment were scored reliably on any basis of measurement. Changing variables each day, changing Committee decisions each day, the push for more reads from each scorer – all played a role in the scores each student received and on the overall reliability of scores.

If scorers or supervisors had even the slightest question how to score an answer, they had to check their notes, rubrics (scoring keys) in their binders, and pages of variables daily drawn up and hung up on wall partitions. Supervisors also had to check periodically with the Range Finding Committee for guidance on how many points to give for an answer. *One question on the OPT eventually created over 27 pages of hand written possible answers and point scales,* and nine specific variables in addition to fatal errors caused by invalid parts in answers. (See "Mining in the Desert" question, below.)

What does this mean for student scores? Reliability and scoring processes were questionable. There were very few scorers who had both a high number of reads and high reliability percentages. This resulted in lower scores for many students. Whenever more variables were added to the equation for scoring, we asked if answers we had already scored would be rescored with the new information. We were always told not to worry about it, which left us to wonder if Pearson had any intention at all of giving past answers new scores with the new information. I suspect the same will occur on Washington State's WASL.

Other Observations and Comments:
I never saw who was on the Range Finding Committee or where they were located when supervisors contacted them. At the end of the project – that is the OPT scoring for Science – we had to destroy everything in our binders. We were to throw everything out, including all the anchor tests we took, the actual questions and scoring information, and all our notes. I kept one of the 4 point questions and made notes on other questions. All test answers were essays scanned into a computer for scorers to read on their individual computers and score.

Questions on the OPT I Scored at NCS Pearson:

Mining in the Desert: This is the complete section on this question. Items in bold come from their paperwork.

The Problem:

36. An underground stream provides water for many plants on a certain desert mountainside. Some of these plants are cactuses that are pollinated only by one species of bat. The bats live in a nearby cave and depend on the cactuses for food.

Other plants grow only where the stream reaches the earth's surface. Insects and small mammals find food and shelter among these plants. Because the plants bloom when mountain rains swell the stream, their flowers warn the townspeople below of possible flooding.

Some people want to mine the mountain for metals. This would involve setting off explosives and causing shock waves in the area.

In your **Answer Booklet**, describe four possible impacts these mining activities could have on the plants, animals, and people in the area.

The Rubric : We were given a scoring guide for this question, called a "rubric". Students gave answers in essay form. (The above problem could get up to 4 points. Bold type, brackets, and underlines were on the rubric)

Scorepoint 4:
A 4-point response describes four possible impacts the mining activities could have on the **plants, animals,** and **people** in the area. Reasonable answers include but **are not limited to the following** (students should be given credit for reasonable, deducible answers not included here.)

- Shock waves from the mining could disturb the flow or even the presence of the underground stream, which in turn could remove the water source from the plants, the animals and insects that take shelter in them and get food from them (in particular, the bats – and hence the loss of both cactus and bat species, since they depend on each other); it could change the structure/shape of the land and mountain so that water, rain, or river flows differently or not at all; it could cause erosion of the soil; remove the townspeople's source of "forewarning" for possible flooding (blooming flowers); and cause dust clouds or airborne materials that could negatively affect human and/or animal health.
- Looked at from another angle, the mining could destroy or disturb the caves/habitats that the bats live in, killing them or causing them to leave the area and thus again wiping out the cacti that depend on the bats.

[Major concepts that students should demonstrate understanding of (include any four of the following): relationship between mining activities/shock waves and physical land structure, particularly underground river and/or caves; dependence/interdependence of plants and animals on water sources, or microhabitat; dependence of plant on an animal pollinator in order for plant species' reproduction/survival; dependence on animal(s) on plant as food sources; relationship between certain plant presence and their ability to hold soil or lessen erosion; relationship between certain plant flowering activity and indication of flooding; etc.]

Scorepoint 3: A 3-point response identifies three of the above impacts/demonstrates understanding of three of the above concepts.
Scorepoint 2: A 2-point response identifies two of the above impacts/demonstrates understanding of two of the above concepts.
Scorepoint 1: A 1-point response identifies one of the above impacts/demonstrates understanding of one of the above concepts.
Scorepoint 0: A 0-point response indicates **no** understanding of the concepts/results the mining actions could have on the land, plants, and animals, and people.
Scorepoint A: BLANK – meaning the student did not write any response to the test question.

Problems and New Variables in Addition to the Rubric that Came Up After Scoring Started
This question caused a multiple number of questions from test scorers, leading supervisors to generate over 27 hand-written pages of criteria. OPT test scorers, NCS Pearson, Ohio teachers, and the Range Finding Committee do not know what are acceptable answers for a test question until the tests are taken by the students. This is likely to happen with the WASL. In addition to the 27 hand-written pages there were nine items in the problem that mining activities could affect and these along with the 27+ pages and info about what would invalidate a sentence or be considered fatal errors were all drawn up put up on presentation boards and hung on partitions for us to refer to as we scored the tests. For instance, the nine items that could be affected and for which a student could get points were:

1. Underground Streams
2. Plants
3. Animals
4. Bats
5. Cacti
6. Land/Erosion
7. Forewarning
8. Airborne Materials /Air Pollution
9. Water pollution

Other notes I made follow. They came from the 27+ pages of notes, and from the fatal errors and invalidation paperwork the supervisors wrote. (Other scorers had many more notes. Criteria were added almost daily, as scorers brought various problems to the attention of the supervisors.)

- If the clause is contaminated by the use of the word "people" or shows that "people" are affected then the answer is invalid and the student does not score any points.
- If the student says that mining kills cacti then the student scores 1 point.
- 2nd level consumers (animals) – ok to get 1 point.
- Explosions can't kill bats even if they are in the caves. (We were constantly reminded that bats and people don't die and that if any student said so we were to give NO points for those answers. The Range Finding Committee would not accept such answers. We were to abide by Committee decisions when it came to giving out points.)
- Students would get points if they said that the bats would move. They would not get points if the students wrote 'bats would leave'.
- Chemicals can infect animals, plants and people – this gets 3 points.
- Bats die IF cactus dead 'cause bats have no food – gets 2 points.
- If leaving – gets 0 pts., if moves away – gets pts.
- If plants die and people have no warning – gets pts.
- If flowers died and people have no warning – gets pts.
- If people have no warning – gets 1 pt.
- Animals have no food – ok for pt.
- Kills cacti and bats = 2pts.
- You could starve the animals = 1pt.
- If animals starve, destroy habitat, die of thirst – this affects only one of the nine categories so the response gets 1 pt only.
- Bats may not come back = 1pt
- Bats move or leave area = 1pt (Bats leave area was later changed to not being acceptable.)
- Cacti die, plants die = 2 pts.
- Plants, animals, bats killed by pollution = 4pts.
- If bats can't pollinate then bats are being affected so the response gets 1 pt.
- People might starve – this invalidates the sentence regardless of whatever else is in the response – no pts.
- Mudslides are accepted as land erosion.
- Plants being buried does not mean animals don't get food.
- If the stream gets covered it does not mean animals will die of dehydration.
- No credit if killing animals or humans without an explanation.
- If the response mentioned 'avalanche' then no points. Even a rock avalanche was invalid because Committee said that avalanche can only relate to snow and there is no snow in a desert.
- Explosions NOT explosives cause problems or can kill.
- Any plant mentioned about being a forewarning or warning to the townspeople counts as ok.
- if animals are killed – the response must say what from.
- Plants dying needs no explanation.
- Plants and cacti are interchangeable. (This changed later, depending what the response was about.)
- Plants and animals don't die from the explosions.
- An animal death must be from an environmental reason.
- If the response is that bats will be scared then no pts.
- Habitats destroyed is ok even if animals are not mentioned.
- If people injured then they need to explain why. If people killed though – then no pts.
- Food not lowered – this is too vague – no pts.
- Destroy homes or land – ok. Destroy homes can be taken as meaning animal habitat.

Mario and Lawn mowing

This was about what month Mario, who mows lawns, would make the most money and why. The student was provided with three graphs spanning January to December: Amount of Average Rainfall each month, Amount of Average Daily Sunshine each month, and Amount of Average Temperature each month. There was no graph showing the average amount of pay Mario made each month. A correct answer was June or July or both of these months. Acceptable reasons included: most rain, most hours of sunshine per day, highest temperature causing people to hire Mario instead of mowing their own lawns, or some combination of these answers. Acceptable answers could also include: more rain that caused the grass to grow, more hours of sunshine allowing Mario to mow more lawns per day, higher temperature to dry the rain so Mario could mow the lawn.

The problem included the graphically shown fact that in the month of June, the rain had reached an amount of 10 inches with 15 hours of average sunshine per day and an average temperature of more than 95 degrees. Imagine some of the problems the students encountered with these conditions. Yet we had to mark those answers wrong where students picked months with less rain, lower temperatures, and less hours of sunlight. The Range Finding Committee had decided that the months of June and July were the only correct answers with the highest of rainfall inches, hours of sunlight and highest temperatures. One can only wonder how in the world Mario could mow lawns when it had to have been raining almost every day during the months in June and July, the sun shined for up to 15 hours per day, and temperatures were close to 100 degrees! Scorers had to look for a clue that the student was reading any or all of the graphs so some sort of score point could be given.

Some students claimed Mario made the most money in June because they interpreted the hours of sunshine graph to indicate that Mario made $15 an hour that month. Others, using the same graph, came to the conclusion that Mario made the most in June because he took up to 15 hours to mow a single lawn! Most of the answers the students gave for this question were wrong and it was clear students really struggled with it.

If the Range Finding Committee wanted to see if students were reading the graphs correctly, they should have asked questions directly related to the graphs. Instead, they expected them to assume that to make the most mowing lawns, you have to pick those months with the most hours of sunlight, highest temperatures, AND the most rainfall. I think most adults would agree that these are not the best conditions for mowing lawns. The issue should have been whether or not students could read the graphs. The best answers came from students who knew something about mowing lawns from their own experience. In my view, those who answered May and August, when conditions were moderate and not so extreme, chose the best months. Of course, we had to accept the Committee's decisions ("The chair is orange."), regardless. Conformance was the main issue, not students' ability to logically arrive at their answers or rely on their own experience! This was true of all test scoring of all test answers.