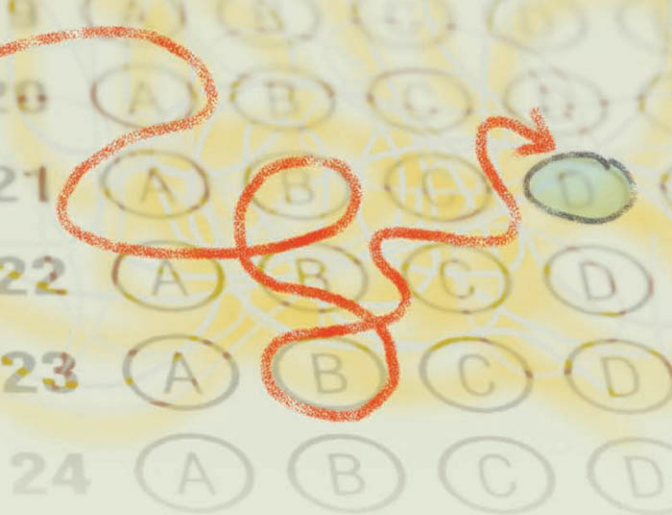

Making the Grades

My Misadventures in the Standardized Testing Industry

Todd Farley

An Excerpt From

Making the Grades:
My Misadventures in the Standardized Testing Industry

by Todd Farley

Published by Berrett-Koehler Publishers

Contents

Preface

WHEN I FINALLY, *finally*, for-real-this-time, I'm-no-longer-kidding, cut all ties to my cushy job in corporate America, it was to the East Village of New York City that I fled. There, in Alphabet City, amid the ghosts of the angry youth who had helped foster the punk rock movement, and surrounded by the memories of the communists, biker gangs, and homeless who used to regularly make their stand against The Man by rumbling with the New York City police in Tompkins Square Park, I hunkered down to forget how I had been earning my money and to figure out my life.

Standardized testing? *Me*? How did it happen?

The standardized testing industry had never particularly interested me, and at no time in my life had I been scared of its consequences or inspired by its possibilities. In grade school, I remember having to take those statewide multiple-choice tests, and I recall thinking only that they had little to do with my life. In high school, on both occasions I took the SATs (the second time at my mother's insistence), I remember most vividly the fact that the test was seriously inconveniencing my weekly game of Saturday-morning Home Run Derby with Shawn and the Druding boys (when Shawn and I showed up at the testing site the first time, we had baseball gloves but no pencils). Years later I ripped open the seal of my

GRE test booklet and thought to myself, "Perhaps I should have studied? Huh . . . ."

Maybe I was being naive, but I always believed that course work, grades, and the professional opinions of the educators who knew me would matter more to colleges or graduate schools than would some random number produced by a mysterious testing company. I couldn't imagine—and I didn't want to be a part of—any institution of higher learning that would ignore my years of classroom work to instead make a decision about me based on a single Saturday's performance. That didn't make any sense to me, so never as a student did I have to fret about standardized testing.

For the last 15 years, during which time nearly every cent I've earned has come from the standardized testing industry, the topic has not interested me any more than it did when I was a student. I've always had other plans for my life—world travel, writing—and testing has been no more than a way to make a living. While I've rather conscientiously attempted to do a decent job in the business, my heart has certainly never been in it: I've just been some guy doing a boring job to pay the bills. Not, perhaps, a philosophy you would hear many teachers espouse, but it was how I got through the days.

The problem is that the testing world is changing. For years, the work I was a part of seemed innocuous. The tests were written, were taken by students, and were then scored by my ilk at a testing company before some state or federal agency used the results to drive curriculum or formulate education policy. It didn't seem to me, however, that I was involved in deciding the future of individual students, teachers, or schools. It just felt like I was involved in some inscrutable statistical dance that didn't specifically mean very much to me or anyone else. While I was vaguely disappointed in myself for not doing what I wanted in life, I certainly never believed by working in the

business that I was doing anything wrong or unethical. It was a job I could make my peace with.

Today, however, my peace is harder to come by. Seemingly every day a different news story shocks me with the increasing importance of standardized testing: lawsuits against the College Board, ETS, and Pearson Educational Measurement over misscored SAT tests that led to students not getting into their preferred colleges; lawsuits from parents in the state of California or the city of New York against the state tests that are keeping their children from being graduated or promoted; lawsuits from the National Education Association (NEA) against the implementation of the No Child Left Behind Act and the massive testing system it advocates.

We are fast approaching a point where the graduation of high school seniors or the promotion of any students will result not only from their classroom work or the opinion of the teachers who spend every day with them but will also hinge on their performance on a single *standardized test* (e.g., the California High School Exit Exam). We are nearing a point where teacher pay and teacher hiring/firing will not be linked to an educator's skill or experience as much as it will be tied to his or her students' *standardized test scores* (as is happening in the Houston and Denver school districts). We are facing a world of education where districts and states are awarded federal funds based not on population or need but instead on regional *standardized test scores* (No Child Left Behind).

Perhaps in theory these are good ideas. Perhaps. However, if you knew what I knew about the industry, you would be aghast at the idea of a standardized test as the deciding factor in the future of even one student, teacher, district, or state. I, personally, am utterly dumbstruck by the possibility. The idea that education policy makers want to ignore the assessments of the classroom teachers who spend every day with this country's

students to instead hear the opinion of some testing company (often "for-profit" enterprises) in a distant state is, in my opinion, asinine. It is ludicrous.

If you knew what I knew, you'd agree: I have seen testing companies regularly forgo accuracy and ethics in the name of expediency and profit; I have seen psychometricians who barely speak the language making final decisions about our students' understanding of English; I have seen hordes and hordes of mostly unemployed people being hired as temporary workers to give the scores that will ultimately decide the futures of our students, teachers, and schools. I have seen it all through more than a decade and a half in the business, and does anyone really want me and my kind—for-profit types working at for-profit companies—making decisions about their kids' futures? Hell, even I don't want that, and I'm pretty good at the job.

As far as I'm concerned, it's one thing to use standardized testing to take an overall snapshot of America's students at various grade levels, but it's something else entirely when you're talking about making decisions about individual students, teachers, and schools based on the work I do. That is something else indeed, and it ain't a pretty something else.

You don't believe me? You don't think that the development and scoring of large-scale standardized tests is nothing but a theater of the absurd?

Then let me tell you a story.

PART

Wage Slave

CHAPTER

1

Scoring Monkey

I BEGAN TO DOUBT the efficacy of standardized testing in 1994, about four hours into my first day scoring student responses to a state test. At the time I was a 27-year-old slacker/part-time grad student at the University of Iowa, and my friend Greg had referred me to NCS (National Computer Systems, a test-scoring company in Iowa City) as a good place to get decent-paying and easy work. Soon thereafter, after a perfunctory group interview that entailed no more than flashing my college diploma at an HR rep and penning a short essay about "teamwork" (an essay I'm pretty sure no one read), I had myself a career in "education."

On my first day, we new employees, as well as dozens of more experienced scorers, met at the company's rented property on the north side of Iowa City, a warren of tiny rooms filled with computers in the dank downstairs of an abandoned shopping mall. Within 10 minutes of sitting down, the gent sitting next to me—named Hank, a floppy leather hat perched on his head, a pair of leather saddlebags slung across his shoulder—

confessed he had worked at NCS for years and regaled me with stories of his life. In no time he told me how he had overcome his nose-picking habit (a dab of Vaseline in the nostrils) and offered to show me the erotic novella he was writing, beginning to pull it from a saddlebag. I politely declined and wondered what I'd gotten myself into.

Other than Hank, around me was a bunch that looked no better. I had dressed how I thought appropriate for the first day of a new job (a pressed pair of khakis, loafers, a buttoned-down blue shirt), but all around my colleagues were slumped like bored college students and mid-1990s slackers in sweat pants and ripped jeans. A whole lot of heads seemed like they had not lately been shampooed; lots of faces looked groggy and uninterested.

The building itself also failed to inspire. We were below-ground, 12 people sitting in our small room around two islands of six computer monitors each, the only windows about eight feet in the air and offering a view of the tires on the cars out back. Occasionally we could see the shoes of people walking by. The room was lowly lighted by phosphorescent bulbs and smelled antiseptic, like cleaning products and the musty industrial rugs that covered the floors. I couldn't imagine I could continue to work there, a man of my grandiose literary ambitions. My only hope was that the job itself would prove interesting.

After perhaps an hour's worth of idling about, waiting for management to seat everyone and file paperwork and start the computers, we began our task: the scoring of student responses to open-ended questions on standardized tests. The six people at my island of computers would score a fourth-grade reading test from a state on the Gulf of Mexico, the tests of those 9- and 10-year-olds from the Deep South being scored by this group of mostly white, midwestern adults. Before we began, however, we were trained on the process by our supervisor/"table leader," Anita.

Anita first showed us the item the students had been given, a task requiring them to read an article about bicycle safety before directing them to make a poster for other students to highlight some of those bike safety rules. Some of us mentioned it seemed like an interesting task, having the students use their creativity to show their understanding of bicycle safety by drawing a poster instead of asking them multiple-choice questions. I nodded to myself, smiling, approving that this first standardized test question I'd seen in years was open to so many possibilities. The question was definitely not rigid or stringent, and it allowed the students to respond in myriad ways.

Next, Anita explained the rubric we would use to score the student work (a rubric, or "scoring guide," is the instructions given to the professional scorers on how they should mete out credit to the student responses). She pointed out how easy the task would be to score, as it was a dichotomous item where students were given either full credit or no credit. If a student's poster showed a good example of a bicycle safety rule (like riding with a helmet or stopping at a stop sign), full credit was earned. If a student's poster showed a poor example of bicycle safety rules (like riding with no hands or riding two abreast in the road), no credit was earned. Finally, Anita showed us training papers, actual student work that had earned either full or no credit. She showed us 20 or 30 "Anchor Papers," examples of posters that had earned the score of 1 and others given the score of 0. Eventually she gave us unscored papers to practice with, reading the responses on our own and individually deciding what score to give. After we discussed the Practice Papers as a group and Anita was convinced we all understood the scoring rules, it was time to begin.

At that point I was operating under the impression the item was relevant and interesting. I also thought the rubric was absolutely clear and would be a breeze to apply. And from my experience scoring the Practice Papers, I expected to have

absolutely no difficulty scoring the actual student responses. At that point, it was all so clean and clear and indisputable I would certainly have been counted among the converts to the idea that standardized testing could be considered "scientifically based research" (to which the No Child Left Behind Act alludes more than 100 times). At that point, I had no doubt I was involved in important work that could produce absolute results.

And then we started to score.

The thing about rubrics, I discovered (and would subsequently continue to discover over the years), is that while they are written by the best intentioned of assessment experts and classroom teachers, they can never—never!—come remotely close to addressing the million different perspectives students bring in addressing a task or the zillion different ways they answer questions. If nothing else, standardized testing has taught me the schoolchildren of America can be one creative bunch.

I bring this up because the very first student response I would ever score in my initial foray into the world of standardized testing was a bicycle safety poster that showed a young cyclist, a helmet tightly attached to his head, flying his bike in a fantastic parabola up and over a canal filled with flaming oil, his two arms waving wildly in the air, a gleeful grin plastered on his mug. A caption below the drawing screamed, "Remember to Wear Your Helmet!"

I stared at my computer screen (the students filled out their tests and those tests were then scanned into NCS's system for distribution to the scorers), looked at my rubric, and thought, "What the #@^&$!?!" In preparing to score the item, we'd all agreed how to apply the rubric and had addressed what seemed like simple issues: credit for good bicycle safety rules, no credit for bad ones. It had seemed so clear.

Looking at my screen, I muttered to myself, held both hands in the air in the universal sign of "Huh?" and flipped

through the Anchor and Practice Papers while awaiting a revelation. Certainly the student *had* shown an understanding of at least one bicycle safety rule (the need for a helmet), which meant I was to give him the score of 1. On the other hand, the student had also indicated such a fundamental misunderstanding of a number of other cycling safety rules—keeping a firm grip on the handlebars, not biking through walls of fire—I couldn't see how I could ever award him full credit. I was actually more worried about the student's well-being than I was concerned with his score.

I held my palms up. I mumbled. I flipped through the training papers. Eventually Anita stood behind me, looking at my screen.

"What are you going to do here, Todd?" she asked.

"Good question," I said.

"Does the student show an understanding of a safety rule?" she asked.

"One safety rule," I said.

"And that means you're going to give it what score?" she asked.

"A 1?" I said, looking over my shoulder at her.

She nodded. "Yup."

"*Really*?" I asked her. "We don't care that as a result of following these 'safety rules' the student is almost certainly going to die?"

She laughed. "I think he was having fun, and he certainly knows how important helmets are."

"Yes, he does," I agreed. "Now let's hope he's wearing a nonflammable one when he crashes his no-hands bike into the burning oil."

She smiled, but less enthusiastically. "We don't make the rules, Todd, we just apply them. The state Department of Education says understanding one safety rule earns the student full credit, so we give them full credit."

I shook my head. "We don't care about the context? We count one good safety rule among three bad ones the same as we do one good rule?"

Anita smiled, perhaps ruefully. "One good safety rule earns full credit," she said. She turned to head back to her own computer, and I watched as she walked away. Hank looked at me, shrugged his shoulders, and smiled. One of the other scorers leaned in toward me and grinned.

"Basically," he said, "we are a bunch of scoring monkeys. No thought required."

"Just click," Hank added, making a motion with his mouse finger. "Just click."

They each nodded to me, shaking their heads slowly up and down, bemused looks on their faces. I realized the two of them had definitely drunk the NCS Kool-Aid.

So as Anita insisted, and for reasons that were clear to me but also hard to believe, I clicked on the 1 button, and the response was scored. In the parlance of NCS, I had officially become a "professional scorer," which seemed a slightly exaggerated title for the work I was doing. The poster of the helmeted daredevil slid off my screen and was replaced by another.

Many of the student responses *were* easy to score. Most students simply showed one safety rule (a biker stopped at a stop sign, another using hand signals to indicate their direction), and I would give those responses full credit. Others ignored safety rules entirely (showing a biker doing a wheelie in the middle of the street, for instance, or drawing *unhelmeted* cyclists jumping over fiery moats), and I gave those responses no credit. Other students earned no points for using the blank poster only as an opportunity to sketch, and there were enough doodles of family pets and "best friends forever" to reconsider the brilliant idea of having fourth graders draw pictures as a part of their tests.

Many of the student responses, however, were befuddling, and we scorers might not know what safety rule was being addressed. Sometimes the handwriting was hard to decipher, and for lengths of time the group would unsuccessfully ponder over a word like *grit* before giving up (later someone would yell "right," their mind having subconsciously solved that puzzle even though a score had long ago been given to the response). Other times the drawings were impossible to interpret, and whether we were looking at a biker or surfer or equestrian was not completely clear. On innumerable occasions the scanning of the tests made it incredibly hard to even see the student responses, leaving us leaning forward to squint at vague and fuzzy lines. Some of the drawings did include a caption to emphasize the safety rule ("Use hand signals!" or "Ride single file!"), but others let the drawings stand alone, leaving us confused. We would usually mull over the response on our screens by ourselves before eventually giving up.

"Is this poster indicating bikers should use hand signals?" someone would ask the group. We would huddle around his or her screen.

"I think so," someone would answer.

"No," I might say, "I think they're waving to a friend."

"No," another scorer would disagree, "I think that biker is giving someone the finger!" And we would laugh, but who really knew what that fourth-grade drawing was getting at?

"Really," the scorer sitting there would say, getting frustrated, "is this acceptable or not?"

The rest of us would begin to disperse.

"I . . . ."

"Well, . . ."

"Good luck with that . . . ."

And we would scatter back to our own desks, back to our own screens of problematic, fourth-grade, bike safety hieroglyphics.

Anita would always try to solve the problem. "Is there a clear bike safety rule?" she would ask. "If there is, credit it. If not, don't."

"What if we're not sure?" someone asked. "This *might* be a good rule."

"A clear bike safety rule gets credit," she said. "If not, it doesn't." Anita was a very efficient woman, very direct, and frankly, I liked her less with each passing minute. She acted like it was all *so* obvious, and meanwhile I was attempting to interpret the Crayola musings of a nine-year-old.

Anita's major contribution to our test scoring was in the form of backreading. As we scored the student responses, she would randomly review on her computer screen a small number of the scores each of the six of us were doling out, checking to see we were applying the rules correctly and in a consistent, standardized form. At times this was helpful, as Anita would call us up to her desk to show us a response we may have misscored.

"Remember, Todd," she might advise, pointing to a student response on her computer screen, "we *are* crediting 'riding in single file' as an acceptable safety rule. You gave this response a 0, but it should be a 1."

"Of course," I'd apologize, "Sorry. I'll credit it next time." Her advice was often helpful in remembering the rules and improving my scoring, so in general I was not averse to heeding her counsel. No one necessarily likes to be told they are wrong, but I understood what Anita was telling me was part of my learning curve at the new job. I stoically soldiered on.

Other times I thought Anita was nuts. Near the end of my first day, she called me up to her desk.

"You gave this the score of 0," she said. "How come?"

"I gave it a 0 because it doesn't show any bicycle safety rules," I said.

"That's not a bike at a stop sign?" she asked.

"No, that's a truck at a stop sign," I told her.

"And what's behind the truck?"

"Well," I said, feeling the blood rushing to my cheeks, "behind the truck is a flat-bed trailer, and securely fastened to that trailer by heavy chains is a bike without a rider." The other scorers began to giggle, laughing at my description and realizing Anita and I were on the verge of a small spat. They began to mill around the screen to look at the disputed student response.

"You cannot be telling me that poster shows any understanding of a bike safety rule."

"Yes, I can," Anita said.

"That might be a car-driving rule," I argued, "but it's not a bike safety rule."

"No way, Anita," someone chimed in, "there's not even a rider on the bike."

"Look," she said, her voice starting to rise, "the rule we've been adhering to is that a bike at a stop sign earns full credit."

"It's not a bike," I said. "It's a truck at a stop sign!"

"There's no one on the bike!" someone mumbled.

"Don't worry about it," Anita said. "Remember, all we can do is apply the scoring rules the state gave us. They said a bike at a stop sign is acceptable, so we credit it."

We headed back to our desks, considerable bitching going on along the way. I shook my head but had to laugh.

"My God," I continued, "we're going to wipe out the entire population of elementary students in that state. They're going to be riding into fires thinking they'll be saved by their helmets, going to think they only have to stop their bikes at stop signs if they're strapped to the back of a truck."

"Enough," Anita said. "It's been a good first day, so let's wrap it up. Just score the response on your screen, and then shut your computer down."

I shook my head, smirking. What did it all mean? Could the 1 or 0 that I was punching into the computer really tell anyone

anything about these students? It all seemed so random. I decided to score that one final response on my big, first day before I could head home to take out my frustrations on the soccer field. One more response, I told myself, just do one more.

I looked at my screen, and I was amazed.

I'll say this: I do love the students of America. They are often a fascinating and unique bunch. Many of the responses we scored, of course, were pedestrian and predictable, but many others were absolutely captivating. The kid who made the safety poster showing the bicycle strapped on the trailer behind the truck? It's 15 years later, and I still don't know what that little Picasso was thinking. Seriously, was his poster *really* an example of bike riders needing to stop at stop signs? Was it the work of a kid who wanted to draw a picture of his father's pickup but who threw in a bike at the last minute only to remain on task? Was his poster a sly commentary on the fallibility of standardized testing? Who knew?

I thought this as I looked at the screen to score the final student response of my first day at NCS. I could do nothing but laugh at what I saw. The poster, without any caption, showed a stop sign next to an abandoned road, and in front of that stop sign lay the crumpled remains of a bicycle, all twisted tires and busted frame. The poster had nothing else, no other words or bikes or vehicles or people.

What did it mean? Was the student a demented genius, a confused reader, maybe just a bad bike rider? Did the poster mean "You better stop at stop signs or you will crash?" or did it mean "Even stop signs can't put an end to accidents?" Who knew?

Unfortunately, what I did know was what Anita would tell me to score it. I was convinced she would tell me to give it a 1, because she seemed to think any bike at any stop sign should always earn full credit.

I shook my head. I just shook my head and shook my head. No friggin' way, I thought. I am not giving this response full

credit for its "understanding" of bicycle safety rules when all it shows is a bike that has been horribly crushed by a horrific accident of some kind.

In front of a stop sign or not.

Uh-uh, I thought. I did the only thing I could do, following the single valuable piece of advice Hank had given me that day: I clicked the X in the top-right corner of the computer screen, shutting it down. That student response would get kicked backed into the system and would, I prayed, get directed to another scorer by the time we began the next day.

I went home, and after playing soccer—after blowing off steam through extended sprints and hearty collisions and many a frustrated howl—I spent the rest of the night railing against NCS and "the system" to a friend of mine, a sweet girl who listened with less interest as the night went on. This girl was a sly one, and eventually she said, "You don't have to work there, you know. You could go back to the university."

I grumbled under my breath, because she knew better. My last job in Iowa City had been at the University of Iowa's business office, where I earned $5.75 an hour to file invoices. While I'd taken secret pleasure in the job because I felt it was reminiscent of Charles Bukowski toiling away at the U.S. Post Office (he and I a couple of great writers struggling against The Man), I quickly left that for NCS because test scoring paid $7.75 an hour. Not only was that two more bucks an hour, it was a nearly 35 percent raise, and there was pretty much no way I was giving that up.

So the next day, for that extra two bucks an hour, I reported back to NCS.

I did it for the money. It would become a familiar refrain.

CHAPTER

2

Numbers

HAVING DISCOVERED my 1994 price—ethics included!—was just $7.75 an hour, I reported back for my second day at NCS and began to see what the gig was really all about. After listening to a cursory, 10-minute review of the bicycle safety item and scoring rules, we began to read and score student responses again. I never again saw or heard of that problematic poster with the busted-up bike "stopped" at the stop sign that had so annoyed me the day before. It had been directed to one of my fellow scorers, and one of them—apparently without much thought, because no one ever mentioned it—scored it, either giving it full credit for so *obviously* understanding the value of stop signs, or maybe giving it no credit because he or she believed (as I did) the poster provided absolutely no clear understanding of anything. While I was curious what score my unknown coworker might have given the response, I was even more intrigued to think what some test developer/education expert imagined that score meant. To me the poster was absolutely unfathomable, but someone out there believed it was an indicator of student learning.

Huh, I thought? Well, I guess they know what they're doing.

At that point I may have wondered about the scoring of test items, but never did I waver from the idea there existed virtual legions of education experts—surely in white lab coats, wearing glasses and holding clipboards, probably at some bastion of Ivy League learning—that could make perfect sense of it all. I didn't expect to understand their genius, but I still had faith in its existence. I decided not to worry and continued to score responses.

The more responses we scored, the more comfortable we became. The rules became embedded in our minds, and no longer did we need to review the rubric. Many of the responses remained unsolved mysteries ("What does that word say?" "What is this biker doing?") that we tried and failed to figure out as a group, but plenty of responses were clear. Good bicycle safety rules got credit, and bad ones did not. "Click, click, click" went our computer mice, 1, 0, 1, 0, 1, 1, 0, 0, 0, 1, 1, ad infinitum. Most of the posters were on our screens for no more than seconds. The ones that were confusing we may have looked at longer, but eventually we just clicked a score button, any score button, to make them go away. At that point none of us believed we were making singularly important decisions about students' lives, so whether we gave some confusing bicycle safety poster a 0 or 1 was not the sort of dilemma that slowed us down.

"Click . . . click . . . click" went the computer mice.

Occasionally we would hear advice from Anita ("Remember, a bike on the left side of the sign is credited for stopping; a bike on the right side is not because it has gone past!"), but we heard less and less from her as the days passed. Perhaps this was because we were becoming more capable and confident "professional scorers," or perhaps it was because we had all become so entrenched in our various positions.

Anita knew, for instance, she could give me advice when I'd made a clear oversight with my scoring ("Don't forget to credit two hands on the wheel, Todd!"), but she also knew I was less liberal than she was when interpreting the posters. She knew I probably wasn't going to credit everything she would, so she let those more aberrant responses pass without mention.

If, for instance, I saw a poster with a biker *near* a stop sign ("Full credit!" I could hear Anita screaming) but *with the student's feet still on the pedals*, there was no way I credited that as "stopping." I would score that a 0 without a second thought—regardless of Anita's assertion that the state committee wanted me to give it credit—because I did not believe it showed any student understanding. I'd have no qualms about giving such a poster a 0, because I really believed mine was the truest interpretation of the rubric. Anita may have disputed that, but I never gave her the chance, clicking away without her advice or consent. I was paid to make scoring decisions, so I made them.

Chances were incredibly good in those cases I would never see or hear about the response again. I would click the score I wanted, and unless Anita happened to see it in her random backreading review, the response was gone forever. The score would be recorded.

The only other way such a response *might* be seen again is if it was fed to another scorer to be "second-scored." Second-scoring is the process whereby a certain number of student responses (sometimes 10 percent for an item, sometimes 25 percent, at times even 100 percent of the student samples) are scored by two different readers, so a statistical comparison of "reliability" can be made. Reliability numbers simply show the agreement between readers, which can be shown to the customer/state as proof of the standardization of the scoring process. Without acceptable reliability numbers, there is no evidence the scores are being meted out in any consistent, standardized way. But, if we scorers had a group reliability of 80 percent on some item

(meaning we agreed with each other on 8 of 10 samples that had been second-scored), it was proof positive we were doling out the points systematically.

In truth, the reliability numbers could have been the real reason we began to hear less from Anita. Because our group reliability on the bicycle safety item was above the required 80 percent (not that difficult since there were only two available scores to give), Anita had no worries. As long as the reliability remained above the required threshold, all of us scorers and all of our scores were considered good, and all we had to do was slog our way through the remaining 60,000 student responses.

In theory, the system was working wonderfully, although I quickly figured out if 20 percent of those posters were second-scored (which they were for that item), that meant 80 percent of them were not. I realized 8 out of every 10 posters I scored would be seen by me and me alone. There was a good chance, as busy as Anita was with six scorers and all her administrative work, I would be the only person who ever saw most of the student responses appearing on my screen. I recognized at that point I could pretty much score the responses any way I wanted, Anita be damned. Using the training I had been given, as well as my own interpretation of what I thought the rubric really meant, I scored and scored. Then I scored some more, click, click, click.

The mood around my island of computers was resigned, sometimes dark.

"If my friends could see me now," someone would say. "I've finally made it!" We would laugh at the ridiculousness of it, a bunch of college graduates—many with advanced degrees—trying to figure out fourth-grade cartoons.

I discovered my colleagues were an interesting bunch. Hank, it turned out, was a graduate of the Iowa Writer's Workshop, an impressive fact diminished only by his position in life at age 50, a temporary employee earning hourly wages at a test-

scoring factory, a guy without a car who seemed to live at the mercy of his hectoring wife and juvenile delinquent daughter. Hank worked at NCS for the simple reason it was the most money he could hope to make in Iowa City: reformed nose pickers who peddled porn and ambled about with leather saddlebags slung over their shoulders weren't worth a lot on the local job market.

Around our scoring island was also a newly minted pharmacist (she spent her break time scanning the newspaper ads for what she called a "real job"), a graduate student doing postdoctoral work in biomedical engineering (he planned on spending only four weeks at NCS, long enough to earn the money he needed until his fellowship kicked in), and a guy studying to take the bar exam. In fact, not only was Vincent studying to take the bar exam, but he was studying during our work day: he kept a law text open on his lap all day, every day, and as many times as Anita asked him to shut it to concentrate on his scoring work, Vincent said, "I'm scoring, I'm scoring" (he did continue to click the scoring button day after day, even as he flipped pages in his texts). He possessed a quite imperious countenance, Vincent did, and ultimately his textbook stayed put. After he later passed the bar exam, never again did we see ol' Vinny.

Other than me, the last fellow in our scoring sextet was Terry, who frankly didn't seem quite right. Terry wasn't dumb, as he spent all his breaks reading one massive tome or another, but he still seemed a bit off. His shirts were normally misbuttoned, and he cinched his pants high above his waist. Although in his 20s, his mismatched socks poked out of orthopedic sneakers, and while he didn't wear a pocket protector, it wouldn't have been surprising if he added one to his ensemble. Terry had just started working at NCS, but he was already beginning to doubt the choice. He wondered aloud, to anyone who would listen, if maybe he shouldn't have taken the job at the

cereal factory in Cedar Rapids, which paid less but was so much closer to the home he shared with his mother.

To confirm my diagnosis Terry was a little off, he believed test scoring at NCS could become a career, when in truth it was nothing more than an employment stopgap for three or four months of the year. That was common knowledge because many of the scorers had been working at NCS every fall and spring for a number of years, but not one of them ever ended up with a full-time job. The reality of the work was that most everyone was there only as a way to bide their time until a real career came along, whether that meant at the pharmacy, in court, or (I hoped) in the glossy pages of a national magazine.

Even if Terry did possess a four-year college degree (the NCS HR department was conscientious about this one fact, as if a four-year degree were indicative of having survived some terribly taxing intellectual gauntlet), he seemed too fragile for the job. He was certainly an industrious worker, and he probably cared more about the scoring than anyone else. Unfortunately, Terry also believed he was deciding some poor student's fate with each click of his mouse, which made him more than a little gun-shy. Often, in fact, Terry was so worried about screwing up he was unable to click the scoring button all by himself. Again and again Terry would ask Anita for help ("Is this a hand signal?" or "Do we credit 'S-T-O-O-P' as 'Stop'?"), or he would ask someone sitting beside him what they thought: "Would you give this credit, Vincent?"

Without looking up from his computer screen and/or law text, the attorney-to-be would usually mutter, "Your problem, Ter." Then Vincent *would* raise his eyes toward Terry, as he clearly enjoyed the look of terror that would inevitably sweep across his neighbor's face.

Terry would be at his desk each morning long before our 8 A.M. start time and would remain there until precisely 4:30 P.M., getting up only for his allotted 15-minute breaks twice a day and

his 30-minute lunch. He took each announcement from Anita as if it were the word of God, often writing notes to himself about her imparted wisdom (scribbling on a notepad, "What did you say, Anita, about filling out our time sheet today?"). When I say about Terry that something wasn't quite right, I mean he was humorless and worried too much, about each tiny scoring comment Anita made but also about all of NCS's Byzantine rules.

The rest of the scorers in the building, however, were basically going mad with monotony, insane with frustration over our overly regimented lives. We would trudge into work as close to 8 A.M. as possible and would start wrapping up our day as soon as we could. Was 4:20 P.M. too soon to shut down our computers, or should we wait until 4:25? Working all the way to 4:30 was inconceivable. We needed to be out the door at *exactly* 4:30 P.M.

On our two 15-minute breaks and during our 30-minute lunch, most of the roughly 100 scorers would rush upstairs and outside to try to suck in some fresh air and catch a glimpse of the sky, perhaps to share a cigarette or talk to a fellow scorer. For a moment, we could discuss things we cared about, whether music, politics, books, or sports. Still, we couldn't leave the work completely behind, and we laughed bitterly about the seeming randomness of the scoring rules, mocked the students ("No, Suzy, we cannot credit 'Don't chew gum' as a bicycle safety rule"), and complained about the table leaders who kept such a close eye on us, watching to the minute the time we took to go to the bathroom or use the phone. Back inside, as the day progressed, the basement would grow warm and stagnant, a lack of air flow and the combined heat of maybe 100 bodies and 100 computers filling the air, a stench coming from the tiny bathrooms in the hallway that were shared by so many.

I don't mean to whine, as I realize the conditions I'm recounting don't exactly evoke images of the Industrial Revolution.

Sitting on my butt all day and barely having to lift one finger isn't the worst way to make a living. But at the same time, the conditions weren't exactly as exalted as NCS's literature had seemed to describe. In interviewing for the job, I'd seen a brochure that NCS produced—used as a tool both to hire scorers and to show customers the sanctity of the work—showing a huge, open, airy room of smiling test scorers, properly diverse (an African American in one row, an Asian in the next, a Hispanic smiling from the back, although that's not exactly the racial breakdown of *Iowa*), each fascinated by his or her job. Those scorers' obvious good cheer and heartfelt participation in the scoring process—nay, their desire to contribute to the American education system and help students!—nearly leapt off the page.

I, meanwhile, remained stuck in a stinky, subterranean cave with a bunch of bitter slackers who were counting the minutes 'til quitting time. Had you suggested we were involved in "scientifically based research," I would have enjoyed a big belly laugh. The only experiment I could imagine was one testing the limits of human dignity: How far could you degrade a college graduate for the princely sum of eight bucks an hour? Could you return him to elementary school, allowing him to use the restroom only with permission? Could you get him to sit and work in a hot, swampy environment rife with stench? Could you take from him all conscious thought? All free will? All but any ability except to do exactly as told, moving naught but the mouse finger?

Apparently so.

My pal Greg, who turned me on to the job, became a fount of work wisdom for me. He'd been at NCS for years (off and on), and he currently worked as a table leader. Greg had spent a year after college in New Orleans and another in Seattle, and he lived in Holland and Belgium before ultimately settling into Iowa City. There he earned his money from NCS and/or unemploy-

ment checks while concentrating most of his efforts on the short films and large canvases he was producing in the basement of his girlfriend's home. Like me, Greg was imagining some sort of glorious and lucrative future in the arts, and neither of us was particularly gung-ho about landing any serious, full-time work. Instead, we killed time at NCS, struggling through the long days, spending our work breaks talking about things we actually cared about, like the championship chances of world soccer powers Ajax of Holland and Manchester United.

When I searched Greg out a week into that first scoring job, I found him supervising a group of six scorers who were arguing about what they were looking at on a computer screen.

"It's a dog," one guy said.

"That can't be a dog. That's a guy."

"But it has fur and a collar!" the first replied.

"Some guys are furry," a young woman said.

"That's not a collar. I think it's a hat."

"What, a bowler?"

"Could be a bowler. Maybe a homburg."

"But it's not a dog because it's standing on two legs."

"Plus," a girl replied, "it's shaking a guy's hands."

"Exactly," the first guy replied jubilantly. "Dogs shake! My dog Pepper—"

"Hold on," Greg interrupted. "Take it easy, people. Let's get an objective opinion here." He pointed to me.

I leaned forward and stared at the screen. I wasn't exactly sure, but I thought I figured it out. "I'd say what you've got there is a bear in a fez introducing himself at a convention."

"It's not a bear!" the first guy howled. "It's a dog!"

"That's not a dog, you mo—"

"Time for break, people," Greg yelled. "Back in 15 minutes!"

Greg's scorers dispersed, continuing their debate along the way, as he and I sat down by his computer. He asked me how things were going.

"Well, at least I get to look at Anita all day."

"The pretty blonde girl?"

"Yup. What's her deal?"

"No idea," Greg said. "She just started."

"What?" I had assumed Anita was some testing expert, but Greg told me it was her first project.

"She's a temp?" I asked.

"Of course," he said. "We all are."

"Everyone?"

"The scorers are temporary," he said; "we table leaders are temps; the two women I answer to are temps."

"There are no full-time employees here?" I asked.

"I think the computer guy's full-time," he said.

I pondered that for a moment. "I can't believe anything gets done without real employees around. I can't believe we temps are going to successfully complete this whole project."

"Depends on your definition of 'successfully.'"

"Yeah," I shook my head. "I can't believe we get acceptable reliabilities, either. Half the responses are bizarre, and some of the scorers seem like idiots."

"Reliability," Greg scoffed. "Don't worry about the numbers. I can make statistics dance."

I looked at him with wide eyes. My break was over and I had to return to my computer, but that was something to consider on the way back to my desk.

With each passing day, what I perceived as the variables of test scoring became more pronounced. Even non-dolts like me made some serious errors. Anita would call me up and point to a bike safety poster on her screen. "What do we have here, Todd?"

"Looks like a poster of a bicyclist with both hands firmly gripping the handlebars over a caption that says 'Keep Your Hands on Tight,'" I'd answer.

"And that deserves what score?" she'd ask.

"Clearly a 1," I would say.

"So why'd you give it a 0?"

"I didn't. I wouldn't. I'd definitely give that a 1. You're saying I gave that a 0?"

And then Anita would point out on her screen both my scorer ID number and the score I had already given: a 0.

"Really?" I'd say. "I can't believe that."

And I couldn't believe it. I couldn't believe I'd so clearly erred when scoring a student response. I always thought the computer had somehow screwed up (either mine or hers), and of course I never remembered seeing the response anyway. Hell, we were each scoring nearly 200 of the posters an hour, so it's not like I could recall that *one* particularly. Or *any* one particularly, for that matter—they were all just a blur. Even if I knew the scoring rules, it seemed occasionally I would flub one simply because they were coming on screen so fast.

Mind you, it's not unreasonable to score that many responses in an hour. If a poster appears in front of you of a girl and her cat, it takes all of a millisecond to click the 0 button. If a poster appears on screen of a boy on a bike wearing a helmet while stopped at a stop sign, it takes only a millisecond to click the 1. And even the more confusing responses didn't take that long to score, because after a couple of seconds' worth of study you just made them go away: click. The only problem with scoring that many responses was first your eyes began to hurt, then you'd become bored and addled and maybe dizzy, and eventually you'd want to cut out your own heart to make your disappointment in life go away ("*This* is what I do for a living?").

Along with the obvious errors made by scoring so quickly, other problems came to light over the weeks. Hank, for instance. And Terry. Those fellows didn't exactly seem to be grasping all the rules. Most of the scorers did (I had a hard time imagining the dude writing his PhD thesis on some arcane

medical engineering subject wasn't grasping our two-point rubric), but not all of them.

"Hank," Anita would say, "don't forget to credit 'riding in single file' as a bike safety rule."

"Yes," Hank would say. "Of course."

Then he would look at me, lower his voice, and ask, "Since when did we start crediting that?"

"Day 1," I'd tell him.

"Oops," he'd say. Looking first at Anita over his shoulder, he'd turn back to me with his finger over his mouth, "Shhh . . . ." His eyes would dance, and you knew that Hank had once been a fun and funny guy. He just wasn't a very good test scorer, although he'd been at NCS for years and would remain for many more (he stayed until the day he tried to foist his erotic novella on to the wrong woman, a mistake that led to his being dragged away for good by the HR department).

In contrast, Terry knew the rules, but he had so much extraneous information written on his rubric he could never keep them straight. "Terry," Anita would say, "you failed to credit hand signals again on this poster."

"Oh, no," Terry would blurt, "oh, no." He would scurry up to Anita's desk, shuffling papers as he went, trying to find his instructional note about the difficult "hand signals" problem. "I'm sorry, Anita," Terry would say, on his face a look of sheer horror indicating his belief he may have both ruined a child's life *and* cost himself a job. "I know I have it on here somewhere," Terry would shudder, flipping pages left and right.

Vincent enjoyed these moments enough that he would look up from his textbooks. "Did you look on the rubric, Terry?" he might ask. "Hand signals is the first example given! The first!" Vincent would shake his head with mock indignation, and Terry would blanch.

Given the countless examples I'd seen of bad scoring, including my own, I wondered how the reliability of our group

remained above 80 percent. I surmised at some point maybe Anita was fudging the numbers, but that seemed improbable (regardless of what Greg had implied) because the reliability was produced by a computer program, not human calculation.

It should be remembered, however, reliability is no more than a number representing the percentage of *agreement* between scorers. If a group decided to score every student response a 0, they would end up with a reliability agreement of 100 percent. That doesn't mean 100 percent of the papers would be scored correctly, only that they were scored 100 percent consistently.

In the case of our bike safety group, our reliability was above 80 percent even though I knew I was making occasional obvious mistakes (as were Vincent and the PhD-to-be and the pharmacist, which we knew from Anita pointing out the errors), and we also knew what Hank and Terry were doing was pretty much a crap shoot. Sometimes they knew the rules and sometimes they didn't. Still, even if Hank completely muffed the "riding in single file" rule for three straight days, or if Terry was wildly inconsistent with the "hand signals" rule (and maybe the stop sign rule? The single file rule? The two hands on the wheel rule?), only 20 percent of their responses were being second-scored in the reliability pool, so with any luck those particular screwups wouldn't mess with our numbers. I mean, you still had to figure that Hank and Terry scored at least *half* of their posters correctly. That possibility, along with the fact there were only two scores (0, 1) on the bike safety item, made it seem like our reliability percentage would be pretty high by default alone, right?

Wrong.

Greg solved the Mystery of the High Reliability for me near the end of the project. After we finished a postwork game of soccer with the Hawkeye Club, he and I retired to the bar at Joe's Place for a pitcher of Leinenkugel and a chat about our job.

"It's simple," he told me. "If I go into the system and see you gave a 0 to some student response and another scorer gave it a 1, I change one of the scores."

I listened.

"The computer counts it as a disagreement only as long as the scores don't match. Once I change one of the scores, the reliability number goes up."

"The reliability numbers aren't legit?" I asked.

"The reliability numbers are what we make them," he said.

"So," I said, "the only number the customer cares about is a number that can be manipulated by temporary employees?"

He clinked his glass against mine. "It means we'll always have work."

I can't say I was surprised, given Terry and Hank (and me). I sipped my beer. I shook my head. "Man," I said, "I don't mean to sound naive, but I thought we were in the business of *education*." I emphasized that last word as if it were holy.

Greg nodded, thinking for a moment. "Maybe," he said. "But I'd say we are in the *business* of education."

The *business* of education, I thought? Never heard of such a thing.

this material has been excerpted from

Making the Grades:
My Misadventures in the Standardized Testing Industry

By Todd Farley
Published by Berrett-Koehler Publishers

For more information, or to purchase the book,
please visit our website
www.bkconnection.com