Shakespeare by Ear
What Can Intuition Tell Us About What He Wrote?

By Ward Elliott and Robert J. Valenza
April 17, 2008

Abstract: In July 2007 we tested 80 members of the SHAKSPER World Shakespeare Newsgroup for accuracy in distinguishing sonnet-length passages by Shakespeare from passages by others: As individuals, on average, the whole tested SHAKSPER group got almost two out of three unrecognized passages right (63%), and the top 30% got almost three out of four right (74%). After a second round of Golden Ear testing, completed in March 2008, we had a final elite panel of 23, whose members, as individuals, could get 78% of their passages right. As a group, using majority-rule on each question, they got nine out of ten passages right. Computer-aided screening and aggregation raised the accuracy levels from two out of three to nine out of ten – much higher than computers themselves can do for such short passages.

We also asked the Elite Panel to guess whether 20 debated passages were Shakespeare's. Their most interesting answers as a majority-rule group: *A Lover's Complaint*: no. Hand D, *Sir Thomas More*: no. *Arden of Faversham*: no. "As The Dial Hand:" no. "Shakespeare" block of *1H6*: yes. "Nashe" block of *1H6*: yes. 2 "Peele" blocks of *Titus Andronicus*: 1 yes, 1 no. 1 "Shakespeare" block of *Titus Andronicus*: yes. 2 "Shakespeare" blocks of *Edw3*: 1 yes, 1 tossup.

Also of interest were their majority-rule guesses on passages we consider not to be in dispute: *Shall I die*? no. Oxford, Marlowe, Bacon, Chapman, Spenser, Daniel, Drayton, Nashe, Peele, Fletcher, Middleton, Mary Sidney Herbert: none of the above. *Funeral Elegy*: yes.

1. *Computers versus intuition*.

Many Shakespeare buffs are suspicious of stylometric number crunching and would much rather trust their intuitions than someone's counts of authorial quirks. Whenever number crunchers go astray, there is often a chorus of rebukes like those that Ron Rosenbaum delivered to Donald Foster in *The Shakespeare Wars* for "substituting a silicon chip for a tin ear" in misidentifying the *Funeral Elegy by W.S.* as Shakespeare's. "Foster was led astray by the siren song of statistics and the newly fashionable 'science' of 'stylometrics.'"(Rosenbaum, 2006, 173-74).

Foster was our adversary *à l'outrance* in the Shakespeare wars and did not include us or Rosenbaum in his 2002 articles of surrender to Gilles Monsarrat and Sir Brian Vickers (Foster and Abrams, 2002). We don't doubt that Foster deserved his comeuppances from Rosenbaum and Vickers (2002), and we are grateful to both of them for taking the trouble to set him straight. We know of only one person who still believes, as Foster once did, that the Elegy is Shakespeare's, and we are not that person (our 1997, 2001). For everyone else, the *Elegy* war is over.

However, we certainly don't believe that the collapse of Foster's *Elegy* case, which relied in part on bad stylometric evidence, necessarily discredits all other stylometric evidence by association, and we can't recall any assertion of the superiority of intuition, ever, that was actually backed up by any evidence that it worked any better than number crunching. Our own evidence does cast doubt on the notion that it works well on the *Funeral Elegy*. On the other hand, we have long thought that stylometrists might, in fact, have something to learn from intuition, suitably enhanced, and it did not take us long, after validating our computer tests to our satisfaction against passages of known authorship, to try testing the validity of pure intuition as well, both to check against our computers and to fill gaps where computers have fallen short.

2. *Overview of our intuition tests*.

On and off since 1995 we have been experimenting with intuition-testing of various small audiences of Claremont students, using hardcopy surveys. In 2007-08, after many delays, we finally managed to computerize the process and reached a much larger and more focused population of Shakespeare buffs over the internet than we could ever hope to attract in Claremont. In July 2007 we invited members of the SHAKSPER World Shakespeare Newsgroup to take a Round 1 Golden Ear test, which can still be taken on line at http://goldenear.cmc.edu. We ran the test for just ten days and got a gratifying response of 80 respondents. We thanked the respondents and SHAKSPER's moderator, Hardy Cook, and at posted a long report of our analysis at http://www.shaksper.net/archives/2007/0455.html. Readers who would like more detail than this condensed version are welcome to consult this report.

Conclusions: On average, the SHAKSPER respondents, as individuals, could tell Shakespeare from non-Shakespeare two times out of three. The best of them could tell Shakespeare three times out of four. Aggregating the group's answers by majority rule got its accuracy up to three times out of four, and aggregating the best of them brought it to four times out of five. The difference between professionals and amateurs was surprisingly small – 68% right in gross, versus 66%. "Gross" means before subtracting passages they knew, so the actual difference was probably even smaller.

Many initial SHAKSPERian commentators were shocked with its outcomes and spoke of the Golden Ear test as a "humbling" experience. But we were more impressed, perhaps partly because our expectations were lower, but also because we suspected, correctly, that the initial raw accuracy figures could be raised considerably by screening and aggregation. 2/3 accuracy is about what the Claremont students did; it is better than chance, and the final group accuracy, corrected for recognition and enhanced by picking the best respondents and aggregating their answers, was four out of five –not perfect, but high enough to give some long-overdue support to the intuition invokers of the world, and likewise consistent with the Claremont pilot studies. Four out of five is not as good as computers on longer passages (computers can correctly assign 95% or better; see our 2004a, 357 and Section 10 below), but it is much better than computers on the very short, sonnet-length passages we used in the test.

And Round 1 was not the end of it. In March, 2008, using a second online test, we collected and tabulated the responses to Golden Ear Round 2. Almost all Round 2 takers were people who did very well on Round 1, some from the initial 80, some from an unexpected additional 230 who took the online Round 1 test on their own after we stopped counting for record. The additional test enlarged our initial elite panel and permitted us to screen it yet again, cutting it to a final, elite panel of 23 members, all of whom got 70% or better, gross, of all the scorable questions on both tests. On average, this final, double-screened elite, as individuals, could get 77-78% of known passages right, net of passages they recognized. As an aggregated group, they got a remarkable 90% right, enough to provide yet further support for intuition-lovers, and perhaps even enough to justify some rethinking by intuition-skeptics about the usefulness of intuition.

We already knew from Round 1 and from social-science literature (see Surowiecki, 2004) that screening and aggregation can boost accuracy; boosting it to 90% was above expectations, but not shockingly so. Besides testing the elite Round 2 takers on passages of known authorship, we also asked them to tell us whether 20 further passages of disputable authorship sounded to them like Shakespeare. Many of their answers were convergent enough to influence our own thinking on some of the passages. Whether they influence the broader, more conventional world of Shakespeare authorship studies remains to be seen, but, if we had our choice between studying a passage with or without the help of intuition so enhanced, we would not hesitate to consult it.

In the analysis that follows, the three groups most closely examined are our initial Round 1 Panel of 80 ("Round 1 takers), its "rated" initial elite panel of 24 ("Rated Round 1 players") and the double-screened final elite panel of 23, mentioned above, ("Round 2 Elite Panel" or "Final Elite Panel"). Other groups, "Claremont Pilot Studies," and "Round 1 latecomers" will be mentioned separately where relevant.

Table 1 gives an overview of the gross and net accuracy levels found with the three groups with various levels of screening and aggregation. Gross accuracy is accuracy as observed; net accuracy subtracts recognized passages from observed passages. Where available, net accuracy, not gross, is the proper measure of intuitive Shakespeare detection, but it is not always available, and we sometimes have to make do with gross.

Table 1. Average Gross and Net Accuracy Rates, Individual and Group

Round 1

	All, gross N=80	All, net	Rated, Gross N=24	Rated, net
Shakespeare	66%	60%	76%	66%
Non-Shakespeare	66%	66%	81%	80%
All passages	67%	63%	79%	74%

Aggregated (Group)	79%	79%	93%	82%
Double-screened elite, N=19			93%	89%

Round 2 N=23 Double-screened elite only

Shakespeare	76%	71%
Non-Shakespeare	78%	77%
All passages	77%	73%
Aggregated (Group)	100%	100%

Rounds 1 and 2 combined N= R1:19, R2:23 Double-screened elite only

Shakespeare	84%	77%
Non-Shakespeare	80%	78%
All passages	82%	78%
Aggregated (Group)	95%	92%

Table 1. Single-screening, double-screening, and majority-rule aggregation can raise a group's net intuitive Shakespeare accuracy from 63% to 89-93%.

More details below.

3. *Who took the Golden Ear test?*

The first ones to take pilot paper versions of Round 1 were mostly small captive audiences of students from The Claremont Colleges: Valenza's preceptorial class in 1995, a Claremont McKenna College Shakespeare class of 12 in 2002, and the Claremont Rugby Football Club, many of whom volunteered to take the test on a bus tour of New Zealand, also in 2002, and a few other Claremont friends and well-wishers. The takers were mostly young and amateur; the test was fresh; recognition of the passages was almost a non-issue; all of them took the same test at more or less the same time and place under supervision; and their opportunities and inclination to game the test were rock bottom. Paper tests are harder to take, to analyze and to keep track of than electronic ones, but these pilot experiments were enough to show that average individual accuracy was on the order of two in three, that the best of the students could get four out of five, and that screening and aggregation could raise the group's aggregate accuracy to five out of six.

Round 1 involved two waves of takers of our automated Golden Ear Test. The first was 80 SHAKSPERians, members of the SHAKSPER World Newsgroup, who took the test in its first ten days with some mild constraints. We gave them only ten days and asked them not to retake the test without telling us, nor to discuss it in public till the test was over. They mostly complied. The second wave of 230 takers materialized

unexpectedly after we had summarized the results of the first wave on SHAKSPER, and, we suspect, after the test's website was posted on another major Shakespeare website. We paid little attention to it, supposing it to be less pristine and trustworthy than the first wave, and we did not hurry it or ride herd on it to postpone and minimize discussion and retakes, but we did skim it for its best testers and we did give it a cursory look to check its resemblance to the first wave.

The raw outcomes of the two Round 1 waves were almost identical: before screening, aggregation, and discounting for recognized passages the first wave got 18.6 out of 28 (66%) right; the second got 17.6 (63%) right, roughly two out of three for both groups. The average for all 310 takers, early and late, was 17.8 (64%). That the second wave was slightly lower than the first may well tell us that the best and most confident takers (and certainly the higher proportion of Shakespeare professionals, who made up 39% of the first wave, but only 14% of the second) were the least hesitant to take the test. It may also mean, since the first wave marginally outperformed the second, that our concerns about second wave's being less pristine than the first were misplaced, and that whatever upward bias might have sprung from it did not seriously distort its results.

If anyone had tried to game Round 1, we would expect it to be the top-rated players, not the ones with middling scores, but none of the top-rated players could make it to the Final Elite Panel without also doing well on the Round 2 test, where the test was new to all panelists and much harder to game than Round 1. Consistently, the best performers on Round 1 were also the best performers on Round 2, confirming the reliability of their Round 1 scores. The second wave's data, in any case, are available for anyone who wants to check our results for the first wave against those for a larger, but less pristine group.

31 of the 80 first-wave takers considered themselves Shakespeare pros (39%) and 49 (61%) considered themselves amateurs. 26 respondents described themselves as critics, 14 as writers, 33 as artists, including performing artists, and 8 as "other humanities [than literature], or social, or natural sciences." To encourage participation, we did not require test takers to give their names, and most did not. Everyone got their scores and the correct answers before being invited to send us their names. 14 takers (18%) did identify themselves as willing, and, in some cases, eager, to take our Round 2 test. 12 of these were Rated, that is, they scored Bronze or better on the test.

We did not encounter many, if any, A-List Shakespeare celebrities, no Harold Blooms or Stephen Greenblatts among the self-identifiers, nor many, if any, of SHAKSPER's most vocal past advocates of shutting down your computers and listening to your intuition only -- but we can't exclude the possibility that some Shakespeare grandees or intuitionists might have taken the test anonymously. Several A-list people helped us design the test; we did not expect any of these to have taken it. A couple of A-list people, MacDonald Jackson and Richard Proudfoot, did take the online Round 2 at our request. Both of them recognized 90% of the passages on offer, far too many to be comparable to pristine test takers.

Our double-screened final elite panel of 23 was only 40% Shakespeare pros -- lit professors from well-regarded colleges and universities like BYU, Mount Holyoke, Rice, Monash University, Australia, and Claremont (though none were from Harvard or Yale). One was from a Mexican university and not a native English speaker. The remaining pros were stage people -- actors, directors, dialect coaches, or producers. The other 60% could have come from one of those World War II movies where Kowalski, Cohen, Murphy, and Jones go over the top together: we had a Wall Street lawyer; several schoolteachers, a retired school librarian; a mathematician/computer technology worker; a self-educating housewife; two graduate students, one in chemistry; a finance columnist for *Newsday*; a bookstore manager, and a former stringer for the *National Enquirer*. One read Shakespeare to his kids every night for many years. Thirteen were alumni of the Round 1 Rated Group of 24; ten were newcomers, either from the original Claremont panels or from the Round 1 latecomers.

As we have seen, the whole first-wave Round 1 group, on average, got about two out of three identifications right, both in gross figures and in net, since the whole group recognized fewer passages, on average, than the Rated players. Like the Rated players, the whole group did better on non-Shakespeare than on Shakespeare. The average gross score of all 80 takers was 18.6 of 28 (66%); their net score equivalent would be about a point lower, 17.6 (63%).

The 31 pros in the whole group scored between 14 and 25, averaging 18.9 right of 28 questions (68%, gross). The 49 amateurs scored between 14 and 24, averaging 18.4 right (66%, gross). It is not surprising that the pros did better, on average, than the amateurs. It is surprising that the gap was so small, especially considering that these are gross scores uncorrected for passages recognized by the test-taker, which one would expect to be more common among pros than among amateurs. In fact, the average gross accuracy scores of every subgroup – critics, writers, artists, others -- we tested fell into an extremely narrow range, none lower than 18, none as high as 19.[1]

The mean for the Round 1 SHAKSPER 80 was barely a point and a half short of a Bronze. Our preset range boundaries were: Golden Ear, 24-28 out of 28; Silver, 22-23; Bronze, 20-21. None of the original 80 got a tin ear, 12 or less, because chance tends to pull all scores, high and low, toward the mean. It's true that 6 of the 230 Round 1 Latecomers had tin ears, but we would guess that many of these were from partially completed tests. If you don't recognize anything at all and guess at random on every question, you still have 50-50 odds of getting each guess right, and you will much more likely get a 15 or a 16 than a zero. Getting a zero would be a remarkable feat, implying powers of discrimination comparable to what is needed to max the test with a 28. With a 35% average failure rate, about what SHAKSPER's was, you would expect 1.1 Tin Ear (below 12) purely by chance; we got none at all from our first wave. You would also expect 1.1 Golden Ear (24+) purely by chance; we got seven, and would conclude that their success has to be more than pure luck.

4. *Gross and net accuracy enhanced by aggregation and screening.*

See Table 1 and our SHAKSPER report above. Net accuracy, not just gross, is what we were looking for. We approximated net accuracy by subtracting every recognized passage from the test. We tried to avoid familiar passages, identify them, and exclude them where found. "Avoid" means that we tried to pick the least familiar passages, especially Shakespeare passages, we thought we could find, so people would not recognize them and have them excluded. "Identify" means that we asked takers to tell us outright whether they recognized each passage. The responses showed us that, with a group as sophisticated as SHAKSPER, our efforts to avoid familiar Shakespeare passages were not always successful. Our worst choice from this standpoint was a passage from *Twelfth Night*, which was recognized by 45% of all the takers and 75% of the Rated takers. Everyone on the Final Elite Panel recognized it. We should have used something else. Two other play passages and one Shakespeare sonnet got 20-30% recognition from the whole group and 40-50% recognition from the Rated players. The other Shakespeare questions averaged maybe 7% recognition for the whole group, 18% for the ranked group. The overall average recognition rate was three or four times higher for Shakespeare than for non-Shakespeare, and twice as high for rated players as for the group as a whole. That is, on average, 15% of the whole group and 29% of the rated group recognized our Shakespeare passages, and 4% of the whole group, and 8% of the rated group recognized the non-Shakespeare passages. (The Final Elite Group's recognition rates were similar to the Round 1 Rated Group, 31% for Shakespeare, 8% for non-Shakespeare.) By the same token, however, they did *not* recognize 70-85% of our Shakespeare passages, and 92-96% of our non-Shakespeare passages making these fully and properly testable by our methods. See our SHAKSPER posting for further details.

Table 1 is our comprehensive, bottom-line table, which gives the vertical average of individual scores (that is, the sum of all correct answers divided by the sum of all unrecognized takes), above, and the horizontally, majority-rule-for-each question aggregated group score, below, gross figures to the left, net to the right. The key figures are now the net ones. What leaps out from it to our eye is (1) that netting out the recognized answers, unsurprisingly, cuts the Shakespeare accuracy percentages much more than the non-Shakespeare; (2) it narrows the gap between the Rated players' averaged individual scores and those of the whole group slightly, from 12 points to 11 points, thanks mostly to lower net Shakespeare recognition, and (3) surprisingly, it cuts the gap between the two groups' aggregated group scores from 14 percentage points to only three. (On the other hand, the gap between the Final Elite and the whole Round 1 Panel was an impressive 10-13%). Netting for recognition made no difference at all for the whole group's aggregated accuracy score of 79%, but cut the Rated group's aggregated accuracy score from a dizzying 93% to 82% -- only slightly higher than the whole group's despite the rated group's much higher individual accuracy. Double-screening and reaggregation, as we have seen, raised group accuracy to an impressive nine out of ten.

5. *Possible methodological discounts: Honor system, replicability, choice of samples, and sample size*.

There were several important differences between our SHAKSPER panel and our prior student pilot panels. Though most of the students had read or seen several Shakespeare high school favorite plays like *Julius Caesar*, their recognition of our supposedly obscure passages was much, much lower. So were their stakes in the outcome, their eagerness to take the test, their expectations of their own performance, and their overall Shakespeare investments. They had nothing to lose from a low score, no incentive to pump up their scores, and little opportunity to do so either, since they all took the same test on paper at more of less the same time in more or less the same place and didn't get the answers till the tests were all in.

Our SHAKSPER Round 1 group was a different matter. Its members were heavily invested in Shakespeare, often with a conspicuous attachment to one side or the other of a hot debate. More of them had more of a stake in the outcome and more to lose from getting a known low score than any of our students. This means that the incentives not to take the test or, taking it, not to rest content with a low score -- far less to let the results be bruited around -- were much stronger than they were for our students. Online testing, which gave us access to SHAKSPER's coveted, worldwide membership, also rendered the test more subject to gaming than a paper test of the same people in the same room at the same time.

Anyone who offers such a test to an audience like SHAKSPER has to deal with tradeoffs between what you have to do to get people to take the test at all and what you have to do to keep them from giving biased or inflated results. Many social scientists would have wanted us to build in hard controls on bias and inflation: strictly randomize the takers; have a control group; don't tell anyone the answers, make sure they can't easily copy or Google the passages; give everybody a name or a code and put cookies in their computers to make sure they can't take it twice; or, best of all, tell the ones from abroad not to take the test and make the others all come and get tested in the same room at the same time with a timer and a monitor present, just like the College Board, which has excellent reasons to take such precautions for a high-stakes test.

Most of these hard controls seemed to us inappropriate for a group like SHAKSPER, too offputting, too impractical, too pointless, or too easy to get around. We chose soft controls. We tried to make the test as inviting, non-threatening, non-onerous, and rewarding as we could, short, net-based, and with as much anonymity and feedback as anyone could want. We tried to keep the perceived stakes as low as we could. We limited the experiment to ten days. We asked people not to retake the test or discuss the questions online while the test was going on. In short, we relied heavily on the honor system, speed, and soft controls to keep the test one of first impression.

We think we did the right thing, and believe that test abuse was close to zero. We found only two obvious retakes in the Round 1 80 that we analyzed, both innocent, and both later self-identified for us by the takers. The rest look legitimate to us. If there were a few fudged ones, it is extremely unlikely in a test with this many takers that they would change the outcome by more than a percent or so, or, perhaps more important, that the change, if any, would overstate the group's accuracy. None of the comments we got

were concerned in the least with overstating the group's accuracy; everybody was worried about understatement.

The second-wave 230 takers of Round 1 should be more suspect, since they had more time to think about it, eight months, not ten days, were not asked not to discuss the answers or ignore SHAKSPER discussions of the first wave, and we didn't go over their answers looking for retakes. They also could easily have consulted our report on Round 1 posted on SHAKSPER. Yet the results for the second-wave seemed almost identical to those of the first wave -- lower, if anything – and the results of both waves were almost identical to those of the Claremont students, who were in no position to game the test. Moreover, if the test were gamed, the prime suspects would not be the middling or low scorers, but the high scorers, and, as we have seen, none of these could make it to the Final Elite Panel unless they did well on Round 2, which was pristine to all its takers and much harder to game than Round 1. The Rated Takers' accuracy on Round 2 was highly consistent with what they had done on Round 1. The most suspect class of all would be people who scored high in Round 1 but did not identify themselves or volunteer for Round 2. Some of these could have gamed the test, but you would have to wonder why they would do it, if they were anonymous. If they did, they didn't game it enough to raise their group's accuracy any higher than that of other groups whose tests were much less gamable.

In principle, pristineness has to be a major issue with tests like ours. The most reliable takers of the same test have to be the earliest and freshest; later takers are inherently more suspect, especially if their scores are higher; and future replicability of past results is not something that we would promise or expect for any single version of a Golden Ear test. The longer the same tests are out, available, and subject to public discussion, the more familiar and gamable and less reliable you would expect them to be; hence, our decision to look hard at the pristine first wave 80 takers of Round 1, but not so hard at the not-so-pristine second wave, and, wherever possible, to check the results of any given group against some other group that was more pristine or had a less gamable test. In practice so far, apart from the two Shakespeare sachems who recognized almost every passage from memory, we have found no evidence that pristineness or gaming were problems, and results have been similar regardless of how fresh or controlled the test has been.

We did encounter complaints, discussed more fully in our 2007 SHAKSPER report, that the test was too long or the passages too short or insufficiently distinctively Shakespearean . But the same number of longer passages would have made the test unbearably long, and too subject to non-intuitive gaming by stylometric counting; fewer passages would have made it less reliable and more vulnerable to recognition; more distinctive passages, such as "Friends, Romans, countrymen" would be too familiar to pass as tests of intuition, not memory. If we were to do this again, we would still use very short passages, get rid of the one everyone recognized, and try harder to find ones that most people would not recognize – and still realize that some archpros, like MacDonald Jackson and Richard Proudfoot, would recognize even the most obscure ones we could muster and that their intuitions could not be fairly tested with our methods.

7. *Identification Hits.*

As with our student panel, most of our SHAKSPER answers to each question, Shakespeare or non-Shakespeare, right or wrong, showed very high intra-group agreement as to whether or not the passage was by Shakespeare, and also showed high agreement between the whole group and the Rated group. No more than 7% of the aggregated Round 1 answers, and 10% of Round 2 answers, look like tossups. The other 90-96% show majorities of 57% or more. If numbers like these were reported in a national election, everyone would consider it a landslide (Table 2). Elliott, our political scientist, calls this consensus; Valenza, our mathematician, calls it convergence.

Table 2. Group Consensus, Round 1: Very High, but Not Always Correct
All figures net accuracy

	Shakespeare	Non-Shakespeare	All
High consensus, questions answered correctly			
Full panel	9 (68-80% maj)	11 (59-100% maj)	20 (59-100% maj)
Rated only	11 (64-88% maj)	12 (57-100% maj)	23 (57-100% maj)
High consensus, questions answered incorrectly			
Full panel	3 (64-67% maj)	3 (57-69% maj)	6 (57-69% maj)
Rated only	2 (73-78% maj)	2 (57-58% maj)	4 (57-78% maj)
Tossups, all incorrect			
Full panel	2 (51% maj)	0	2 (51%)
Rated only	1 (53% maj	0	1 (53%)

No Round 1 tossups were correct.

This means that both panels had high consensus on 26 or 27 out of the 28 questions and were closely divided on only one or two. Looking at high-consensus answers only, the full panel got 20 of 26 (77%) firmly right, in gross, and the other six firmly wrong. The Rated panel got 23 of 27 firmly right (85%) and the other four firmly wrong. The Final Elite panel got 25 of 28 Round 1 questions firmly right (89%), and three firmly wrong. We'll skip the details of the impressive Shakespeare "firmly rights" and go straight to the equally impressive Non-Shakespeare "firmly rights."

Table 3 shows that none of the three panels analyzed had much trouble with most non-Shakespeare authors represented.

Table 3. Eleven Non-Shakespeare Round 1 Hits with three different panels

Passage	Percentages who thought it non-Shakespeare (full panel/rated only/elite panel only)

Listed in declining order of Rated percentages. All percentages net.

Bacon poems	87/100/100%
Middleton	89/100/79%
Chapman	82/100/94%
Spenser	78/96/89%
Fletcher	67/91/95%
Daniel	75/87/100%
Marlowe II	71/83/81%
Shall I Die?	65/82/70%
Earl of Oxford	59/79/89%
Marlowe I	70/78/84%
Jonson	60/72/72%

Table 3 shows very high convergence among all three panels, large, small, and smallest, in rejecting all these 11 Round 1 passages as Shakespeare's. In general, the more elite the panel, the higher the convergence.

All of these seem like solid hits to us, both according to what we see as the orthodox consensus and according to what our computer evidence has done to confirm it. None of these tested passages seem likely to be Shakespeare's. Not everyone agrees with us or the orthodox consensus on every passage, but the important point here is that remarkably few of our test-takers thought these passages sounded like Shakespeare. We would hardly consider numbers like these a humbling outcome for the group that produced them. *Shall I Die?* was the only one of these widely recognized (by 31%/54% of the two R1 panels, 47% of the final elite panel), but few of those who did not recognize it thought it was Shakespeare's. Would longer passages have greatly enhanced these landslides? We doubt it; they are already so lopsided it's hard to imagine longer passages changing things much, even if they should be easier to identify. Did they signal that the group was in any way befuddled by too-short passages? It doesn't look like it.

8. *Identification Misses*.

However, two convergent outcomes, one Shakespeare, one non-Shakespeare, were clear misses for all three groups. Another four outcomes showed disagreement among the three groups, and one of these – the *Funeral Elegy* – is a clear miss for the Final Elite Panel, which was otherwise the most accurate of the three groups. Table 4, summarizing non-Shakespeare misses, shows a clear miss for one passage, and a division among the three groups for three more passages.

Table 4. Non-Shakespeare misses

Passage	Percentages who thought it non-Shakespeare (Full panel/Rated/Elite Final) All percentages net

Oldcastle	31/42/21%
Drayton, *Idea*	43/43/ but 67%
Funeral Elegy	41/but 57/36% (!)

The *Oldcastle* and Drayton passages, one recalling a beleaguered-stag scene from *As You Like It*, the other a sonnet from Drayton's *Idea*, suggest that even strong majorities of all three groups can sometimes be fooled by well-turned, vivid, image-rich passages by other writers than Shakespeare. The Final Elite Panel could tell that Drayton's passage was not Shakespeare, but they were even more in the dark than either Round 1 group on *Sir John Oldcastle* and the Elegy![2]

What about the *Funeral Elegy*? Donald Foster relied in part on computer tests to prove that the Elegy "couldn't not be Shakespeare," and spoke of intuitive "sniff tests" with a hint of disdain. When Brian Vickers' crushing countercase, *Counterfeiting Shakespeare* (2002) loomed, and Foster abandoned his Shakespeare ascription, the supposedly dull, pious, pedestrian Elegy of the eye instantly became Exhibit A for those who say you should always trust your gut instincts, never anyone's computers.

If that were really so, Foster should have stuck to his guns on authorship and not sniffed at sniff tests. Only 6% of the whole panel recognized our Elegy passage, and 59% of those who didn't thought it was Shakespeare's! So did an astonishingly convergent 74% of our wondrously accurate Final Elite Panel! Only the Round 1 Rated Panel were aggregate doubters. A net 57% of them thought it was *not* Shakespeare.[3] Our computer tests say that Foster did the right thing to concede, and that the Elegy is on a different statistical planet from Shakespeare, though it could easily be by Ford, exactly as almost everyone now thinks (our 2001, Vickers, 2002). Not only were the whole Round 1 group and its elite at odds with each other on this one, but the two elite panels are also at odds with each other. These conflicts tell us that the best of ears can be fooled, and, in particular, that they can be fooled on the Elegy, and that Rosenbaum's harsh condemnation of Foster's tin ear may be overdrawn. On balance, Rosenbaum to the contrary, we can hardly call the Elegy a great success story for detection by gut instinct.

Table 5 shows one clear Shakespeare miss for all three panels, and three divided outcomes where the whole Round 1 group was closely divided or wrong, but the Final Elite Panel got it right with room to spare.

Table 5. One Shakespeare miss and three divided panels

Passage	Percentages who thought it Shakespeare (Full panel/Rated/Final Elite Panel) All percentages net
The Rape of Lucrece	38/22/35%
Pericles Act V	33/47/but 67%

Love's Labor's Lost	49/but 64/but 67%
Venus and Adonis	49/but 59/but 71%

Only one Round 1 taker recognized our passage from *The Rape of Lucrece*, and very few of the others thought it was Shakespeare's. Seven takers recognized our passage from *Pericles*, Act. V, but two-thirds of the whole panel, and 53% of the Rated panel, thought it was not Shakespeare's. *Pericles* is generally considered co-authored by Shakespeare and George Wilkins; scholarly consensus gives Acts 3-5 to Shakespeare, and our tests agree with it. The Final Elite Panel got *Pericles*, *LLL*, and *Venus and Adonis* right, but two-thirds of them missed the passage from the *Rape of Lucrece*.

9. *Three shots in the dark.*

Three passages on the Round 1 test were not scored, since scholarly consensus as to who wrote them is not settled. But we tested them anyway in case we found the group's instincts helpful in determining actual ascriptions. This is wholly uncharted territory, but, if we had a computer test that looked like it might be 80% accurate – let alone 90% -- we might not bet a thousand pounds on it, as we have on some of our computer tests, but we certainly would not want to sheathe it with an undeeded blade. The same may be said for ascription by gut instinct. With tweaking, we know it can reach up to 92% group net accuracy for passages of known authorship, and we can't imagine our readers not being curious as to what it says about passages of unsettled authorship. Table 6 gives the outcomes:

Table 6. SHAKSPER's Group Ascriptions for Three Doubtful Passages

Passage	Percentages who thought it Shakespeare (Full panel/Rated/Final Elite Panel) All percentages net
1H6	87/89/66%
A Lover's Complaint	42/26/12%
Edward III	68/73/47%

14%, 25%, and 36% , respectively, of the three panels recognized another beleaguered stag scene from *1H6*, Talbot before Bordeaux. 87/89/66%% of those who did not recognize it thought it was Shakespeare -- highly consensual, we would think. Gary Taylor (1995), Marina Tarlinskaja, and Sir Brian Vickers (2007a) all assign the scene, 4.02, to Shakespeare; Paul Vincent thinks it is co-authored by Shakespeare and "Author Y." Marcus Dahl could find no hand but Shakespeare's in the whole play. 4.02-04 looks like a Shakespeare could-be by all our regular tests, improbable by one new test, but not improbable enough to make it a couldn't-be for us. We haven't yet tested 4.02-.05 as a block, Vickers' and Tarlinkskaja's preferred new division. The passage itself is much too short for our tests. Every intuitive group's judgment in this case is consistent with all five stylometric views of *1H6*, though one could argue from the *Oldcastle* passage that

the groups' judgment on non-Shakespeare beleaguered-stag passages has some weak spots.

Perhaps surprisingly, only one person recognized the passage from *A Lover's Complaint*. Could it be more discussed these days than read? Of the many who did not recognize it, 58/74/88% thought it was not Shakespeare. MacDonald Jackson, Kenneth Muir, and most scholars of the late twentieth century have assigned *LC* to Shakespeare, reversing an earlier consensus that it was not Shakespeare. Our best guess (our 1997 and 2004), and Brian Vickers' (his 2007), and Marina Tarlinskaja's (2004) is that it is not. All three groups' judgment favors the doubters, but, again, one could argue from their problems with the passage from *The Rape of Lucrece* that their judgments on Shakespeare poems outside the Sonnets might not be watertight.

Five, eight, and five percent of the three panels recognized the Countess Scene passage from *Edward III*, which we consider a recent addition to the consensus Canon. The scene as a whole is also a Shakespeare could-be by our computer tests. Our two Round 1 panels seem to say it could be Shakespeare, with 68/73% of the two Round 1 panels in agreement – but only 47% of the Final Elite Panel – a tossup, in our judgment -- thought it was Shakespeare. On balance, we would think it supports the Shakespeare ascription, but not as strongly as it would have with clearer support from the Final Elite Panel.

10. *How the three intuition-tested groups compare in accuracy with stylometric tests.*

Most of the samples we used in our Golden Ear tests have no more than 150 words, far shorter than any for which we have dared to validate any of our quantitative tests. For comparison, our current estimated composite accuracy rates for longer, single-authored passages look something like Table 8:

Table 8. Accuracy of computer tests and intuition tests on samples of various lengths

Text	Shakespeare	Non-Shakespeare
Whole plays	100%	100%
Poems, 3000 words	100%	100%
Play Verse, 3000 words	95%	100%
Poems, 1500 words	100%	100%
Play Verse, 1500 words	96%	88%
Poems 750 words	93%	71%
Play Verse 750 words	97%	75%
Poems, 470 words	92%	73%
Fin. Elite Panel 150 wds	94-95%	89%

Source: Our 2004a, 257.

Table 8. Intuitive recognition of very short passages, enhanced by double-screening and aggregation, can reach accuracy levels that computers can reach only on much longer passages.

After enhancement, group intuition seems to be almost as accurate on short passages as stylometrics are on long ones – and far better on short passages than any stylometrics we know of. Individual intuition is still error-prone. Only two of our Final Elite Panel individuals got better than 90% right, roughly equaling the group's aggregate score; 6 got better than 85%; the Elite Panel's overall individual average was 77-78%. It is worth noting that our stylometric accuracy levels are no less enhanced than our intuition levels; both are fortified by screening and aggregation. We screened our stylometric tests to find the best individual tests, just as we screened our Golden Ear respondents, and we combined individual test scores to get the most accurate composite scores, more or less as we did with our screened Golden Ear respondents.

11. *Round 2 guesses considered.*

For Round 1, as noted, we included three "shots in the dark" of disputed authorship and found that the Final Elite Group was evenly divided on a "Countess" passage from *Edward III*, strongly agreed that the Bordeaux scene from *1H6* sounded like Shakespeare, and strongly agreed that the passage from *A Lover's Complaint* did not sound like Shakespeare. None of these conflicted with our stylometric evidence.

Only ten questions in Round 2 were of settled authorship and used for scoring. We thought they were, if anything, harder than Round 1, but the Final Elite Panel made short work of them, getting all ten right, in the aggregate, nine of them by lopsided margins. The one non-lopsided ascription was a passage from *King John*, where only 52% thought it sounded like Shakespeare, correct, as it happens, but for us much more of a tossup than a landslide.

The other 18 were all shots in the dark, passages of disputed authorship, too many to consider in detail here, but listed without commentary in Table 11 and available on request to anyone interested in looking at them, with a line or two about how we think each one fits with our other evidence. We are working on a chapter on the Shakespeare Fringes – that is, supposedly co-authored works from the Apocrypha and Dubitanda -- and expect the panel's guesses to make the most sense when weighed against other available evidence in deciding how to assign the passages. But here are a few first-impression highlights.

The Elite Panel's guesses are consistent with our doubts about the Hand D section of *Sir Thomas More* (only 24% who didn't recognize it thought it sounded like Shakespeare) and a "Shakespeare" scene from *Arden of Faversham* (15%). They are also consistent with our doubts about a "Peele" block of *Titus Andronicus* and a "Nashe" block of *Henry VI, Part I*. 57% and 82%, respectively, thought the passage sounded like Shakespeare. Only 20% thought that "To The Queen" sounded like Shakespeare.

Jonathan Bate and Eric Rassmussen think it's Shakespeare. It's too short for us to have an official, stylometry-based position on it.

But our intuitive panel also cast a splash of cold water on several of our own proposed ascriptions. The first was our proposed reassignment of two blocks of *Edward III* (our forthcoming). Only 35% thought our "Shakespeare could-be" passage from *Edw3* sounded like Shakespeare; 57% thought our "not-so-Shakespeare" passage did sound like Shakespeare. The Panel also roundly repudiated our unpublished doubts about a stylometrically discrepant passage from *Henry VI, Part III* (87% thought it was Shakespeare). They thought that two of our "Shakespeare" passages from *Henry VIII* didn't sound like Shakespeare (35 and 38%), and they were evenly divided on a passage from *Two Noble Kinsmen*, which we think tests resoundingly like Shakespeare, and which mainstream conventional scholarship also assigns to Shakespeare.

It's too early to predict whether or how these judgments will affect conventional wisdom, but not too early to suggest that some of them may call for another, closer look at the pertinent passages, both by us, with our new optics, and by mainstream, old-optics scholars, who have had no reason previously to concentrate separately on the passages we deem problematic. Our initial inclination in close cases like these would be to weigh our Elite Panel's guesses at least as heavily as one of our better single tests and to give them some extra weight where it matches mainstream opinion, but not ours, but not so much weight when it conflicts with both our evidence and mainstream consensus. In general, we would suppose that challengers to the mainstream consensus should bear the burden of proof, and that, new-optics or old, good negative evidence normally trumps good positive evidence, in the same way that "couldn't be" normally trumps "could be (Our 2004a, 337-341)." Hence, for the passages mentioned above, we would be even more tentative than we were before about our *Edw3* and *3H6* speculations, but we would not pull in so much sail on the *TNK* and *H8* ascriptions, where the Panel's intuition conflicts both with conventional wisdom and stylometric evidence which seems strong to us.

Table 9 gives a summary of what the Final Elite Panel thought of various passages:

Table 9. Final Elite Panel's Shakespeare Guesses, Disputed Passages

Passage	Auth	Panel	SH%
Henry VI, Part 1, 4.02.43-56	SH?	SH	66%
A Lover's Complaint, 166-175	NS?	NS	12%
Countess Scene Edw3, 2.01.766-783	SH?	?	47%
Henry VI, Part I, 4.06.28-41.	SH?	SH	69%
Henry VI, Part I, 2.03.35-47	?	SH	72%
Two Noble Kinsmen, 5.01.77-98	SH?	?	48%
Henry VI, Part III, 1.02.22-34	NS?	SH	87%
Titus Andronicus 3.01.219-233 (=3.01b)	SH?	SH	76%
Henry VI, Part III, 2.02.144-162	SH?	SH	75%

Edward III, 4.05.1-8, 14-18, 28-38	SH?	NS	35%
Henry VIII 1.01.100-114	SH?	NS	38%
Henry VI, Part I, 1.05.19-32	SH?	SH	82%
Titus Andronicus 1.01.442-455 (1.01b)	?	SH	57%
Henry VI, Part III 3.01.82-93	SH?	SH?	53%
Henry VIII, 3.02.166-179	SH?	NS	35%
Sir Thomas More, Hand D, lines 127-139	?	NS	24%
Titus Andronicus . 2.02.1-10	?	NS	37%
Edward III, 4.04.149-162	?	SH	57%
To The Queen	?	NS	20%
Mary Sidney Herbert, *Clorinda*	NS	NS	13%
Arden of Faversham, scene viii, p. 24 lines 12-30	NS	NS	15%

12. *Conclusions*

In some ways, it is astonishing, given the frequency and fervency of declarations that intuition can outperform stylometry – or that stylometry can outperform "sniff tests" -- that no one we know of has ever tried to see whether, and to what extent the first proposition is actually so. After 20 years of testing stylometry, and 12 years of testing intuition, we can see why. Both call for more patience, persistence, and acquired computer technique than most real Shakespeare fans can muster. Snap judgments do have admirers, such as Gladwell, 2005, and a legitimate place in life where time is short and there is nothing better at hand, or, as often happens, getting the decision made is more important than what is decided. But they don't deserve much deference if they are wrong, the outcome of the decision is important, and there is a better alternative available (Groopman, 2007). Any individual's claim to settle an authorship question conclusively with nothing but a quick, snap-judgment sniff test has to be suspect. Most individual Shakespeare buffs are not much better at telling Shakespeare from non-Shakespeare than they are at other social-science teasers, such as determining whether or not people are gay or are telling the truth. As individuals, they are lucky if they get such questions right two times out of three.

It is now apparent that some people are markedly and consistently better than others, though it is not at all self-evident that the best ears are always connected to the most insistent mouths. To find the best ears, you have to do a lot of cumbrous testing, and, to get the best intuitive accuracy from them, you have to do a lot of cumbrous screening and aggregation, which, for practical purposes, is impossible without computers. Once you do that, you can raise group accuracy to nine out of ten, not high enough, probably, to trump consensus or computers where they are strong and consistent, but high enough to take note of in the many instances in which both consensus and computers are inconclusive. We are pleased to have such a tool in our tool box, but most of its sharpness comes not from the raw snap judgments it starts with, but from the computer enhancements which bring it into clearer focus. We have often used the term

"new optics" to describe our quantitative techniques for identifying authorship; it could as well describe the way we have tried to move intuitive recognition from blurry to sharper.

We are, as always, grateful to generations of Shakespeare Clinic students for their bold initial development of our new computer-aided optics. We are also grateful to students since the Clinic who have tried in various ways to computerize our Golden Ear test so it can be taken by many and its results conveniently analyzed. Many have tried this over the years, but only the last three have succeeded. Ryan Wilson, Claremont McKenna College '07, gets the credit for finally getting Round 1 up and running, and Kevin Williams '09 and Andrew Robb '09 for Round 2. All three were advised in the process by CMC's mild-mannered new computer-science professor, Art Lee. Almost none of what is analyzed and reported here would have happened without the help of these four individuals. We are also grateful to our volunteers, the Claremont students in the early years, and the glorious profusion of SHAKSPER buffs for Round 1 and 2. Almost none of what is analyzed and reported here would have happened without their help, either. We are much in their debt.

Only time will tell what to make of enhanced intuition as a regular tool of authorship identification. We now have a tested panel which we have already consulted in close cases; others are welcome to look at their answers and make their own judgments as to how they should best be weighed. We shall probably continue to consult them in the future, and we can't exclude the possibility that someone will come up with their own new wrinkles on measuring and enhancing which could confirm, deny, or extend what we have found. We're ready to help, and we have a sense that intuition testing is coming into vogue in psychology departments (for example, Myers, 2002). Every adventure is a reconnaissance for the next. We know that Shakespeare buffs are not overfond of computers, but we hope that our latest adventure with computer-enhanced intuition will not be the last of its kind.

References

Elliott, W. E. Y. and R. J. Valenza (1997). "Glass Slippers and Seven-League Boots: C-Prompted Doubts about Ascribing *A Funeral Elegy* and *A Lover's Complaint* to Shakespeare." *Shakespeare Quarterly* **48**: 177-207.

Elliott, W. E. Y. and R. J. Valenza (2001). "Smoking Guns and Silver Bullets: Could John Ford Have Written the *Funeral Elegy*?" *Literary and Linguistic Computing* **16**(No. 3): 205-32.

Elliott, W. E. Y. and R. Valenza (2004). "Did Shakespeare Write *A Lover's Complaint*? The Jackson Ascription Revisited." *Words that Count: Early Modern Authorship; Essays in Honor of MacDonald P. Jackson*. B. Boyd. Newark, NJ, University of Delaware Press: 117-40.

Elliott, W. E. Y. and R. Valenza (2004a). "Oxford by the Numbers: What are the Odds that the Earl of Oxford Could Have Written Shakespeare's Poems and Plays?" *Tennessee Law Review* **72**(1): 323-454.
http://govt.cmc.edu/welliott/UTConference/Oxford_by_Numbers.pdf

Elliott, W. E. Y. and R. Valenza (forthcoming). "Two Tough Nuts to Crack: Did Shakespeare Write the "Shakespeare" Portions of *Sir Thomas More* and *Edward III*?" *Literary and Linguistic Computing*.
http://www.claremontmckenna.edu/facultysites/govt/FacMember/welliott/select.htm

Foster, D. W. and R. Abrams (2002). "Abrams and Foster on 'A Funeral Elegy'" http://www.shaksper.net/archives/2002/1484.html.

Gladwell, M. (2005). *Blink : the power of thinking without thinking*. New York, Little Brown and Co.

Groopman, J. E. (2007). *How doctors think*. Boston, Houghton Mifflin.

Myers, D. G. (2002). *Intuition its powers and perils*. New Haven, Yale University Press.

Rosenbaum, R. (2006). *The Shakespeare Wars: Clashing Scholars, Public Fiascoes, Palace Coups*. New York, Random House.

Simonton, D. K. (1999). *Origins of genius : Darwinian perspectives on creativity*. New York, Oxford University Press.

Surowiecki, J. (2004). *The wisdom of crowds : why the many are smarter than the few and how collective wisdom shapes business, economies, societies, and nations*. New York, Doubleday

Tarlinskaja, M. (2004). "The Verse of *A Lover's Complaint*: Not Shakespeare." *Words That Count: Essays on Early Modern Authorship in Honor of MacDonald P. Jackson*. B. Boyd. Newark, NJ, University of Delaware Press: 141-58.

Vickers, B. (2002). *'Counterfeiting' Shakespeare: Evidence, Authorship, and John Ford's Funerall Elegye*. Cambridge, UK ; New York, Cambridge University Press.

Vickers, B. (2007). *Shakespeare, A lover's complaint, and John Davies of Hereford*. Cambridge ; New York, Cambridge University Press.

Vickers, B. (2007a). "Incomplete Shakespeare: or, denying co-authorship in 1 Henry VI." *The Shakespeare Quarterly* **58**(Fall): 310-52.

Notes

[1] Gross Accuracy Scores of Identified Subgroups

Subgroup	Number	Gross Average Score	Gross Average Accuracy %
Professionals	31	18.9	67.5
Amateurs	49	18.6	66.4
Critics	26	18.7	66.8
Writers	14	18.9	67.5
Artists	33	18.6	66.4
Other	21	18.0	64.3

Some of the categories overlap. "Other" is mostly people who declined to state a category.

Would net accuracy differ greatly from these gross accuracy scores? We don't know, but it seems improbable.

One might imagine that writers and artists would be more intuitive, and critics more analytical (see Simonton, *Origins of Genius,* 1999), but the average accuracy of the three categories looks virtually identical.

[2] It is possible that Drayton, who Henslowe says co-authored the play *Sir John Oldcastle* (1600) with Anthony Munday, Robert Wilson, and Richard Hathway, could have written both of the confounding passages. A second edition of *Oldcastle* ascribed it to Shakespeare, and it was included in the 1664 Folio and Brooke's *Apocrypha,* but we know of no one today who seriously ascribes it to Shakespeare, and our tests say it's very unlikely to be Shakespeare's work (our 2004a, p. 402). No takers, incidentally, recognized the *Oldcastle* passage and only one recognized the one from *Idea*.

3 In principle, we could get the Elite Final Panel out of this jam by one much more radical screening, cutting the panel, say, to its top one or two players. As it happens, number 1, who got 37 out of 38 passages right, didn't think the Elegy passage sounded like Shakespeare, but recognized it and can't be counted. Number two got 36 out of 38 right, didn't think the passage sounded like Shakespeare, didn't recognize it, and in principle could be counted. But counting just one top player would throw out all the benefits of aggregation, which means in principle that a composite guess of 40 people with 55% accuracy should be better than one of 20 people with 60% accuracy, let alone just one with 95% accuracy on just the first 38 questions. We might look at this differently if we gave our top two another 38 questions and they also got 95% of those right, but, from what we know now, we would go with aggregation and would bet with much more confidence on the top 23, aggregated, than on just the top one or two.