Questions about forensic science: Response

Jonathan Koehler
Northwestern University School of Law, jay.koehler@northwestern.edu
Reactions to the Hwang Scandal

IT CAME AS QUITE A SHOCK TO KOREAN ACADEMICS TO LEARN THAT Woo Suk Hwang’s papers on patient-specific stem cells were fabricated. Members of the Korean Society of Molecular and Cellular Biology, the largest life science academic society in Korea, seriously regret that such a fraud could occur. Since the ethical debate over human ovum supply and somatic cell cloning began, our society members have felt very uneasy and frustrated.

Indeed, we decided to establish a charter for scientific conduct with a strong emphasis on the ethical implications of biological research. The life science researcher’s charter has been unanimously acknowledged by our members and was declared officially in October 2005 at the annual congress.

The main points of the charter are as follows. First, we have to consider the impact that research may have upon humans, society, and the ecosystem before initiating that research. Second, we have to ensure and respect the dignity of life within the research objectives, from cells to living organisms. Third, we should not fabricate any experimental results and should be righteous in the distribution of materials and results. Finally, we should be fair in acknowledging authorship and intellectual property of research outcomes.

As the president of the Korean Society of Molecular and Cellular Biology, I sincerely regret that such a fraud occurred. A strong policy to prevent any further similar disgraceful incidents will be established. I believe in the ethical sincerity and academic integrity of our scientists, as suggested in the Charter of Ethics for Life Science Researchers, and that we will continue on in our efforts toward bettering society and human life.

SANG CHUL PARK
President of Korean Society of Molecular and Cellular Biology, Department of Biochemistry and Molecular Biology, Seoul National University Medical School, 28 Yon Gon Dong, Chong No Ku, Seoul 110-799, South Korea.

RECENT REVELATIONS REGARDING THE RESEARCH by Woo Suk Hwang and his colleagues on patient-specific embryonic stem cells created by somatic cell nuclear transfer (“Editorial retraction,” D. Kennedy, 20 Jan., p. 335) in South Korea undermine the credibility of the nascent, and fragile, stem cell field. These unfortunate circumstances may embolden opponents of embryonic stem cell research who have argued against such research based on moral objections and on mistrust of scientists to monitor their own activities and ambition.

Excesses in high-profile biomedical research are regrettably not new. The history of the gene therapy field provides one perspective. Soon after cloning of mammalian genes first became possible, expectations were raised that gene therapy might be used to treat serious genetic disorders, such as hemoglobin diseases, cystic fibrosis, and cancer among others. After a flurry of initial clinical experiments in gene therapy that led to unsubstantiated claims or lack of objective findings, a panel was convened by the NIH Director Harold Varmus in 1995 to assess the state of the field (1). This group described a field in which research findings were oversold, expectations were raised beyond what was reasonable at the time, and scientific rigor was relaxed in the enthusiasm to rush ahead.

If gene therapy and stem cell fields have elements in common, what does recent history suggest for the future? Since 1995, progress in gene therapy has been episodic, yet clearly on a positive trajectory. In an elegant study reported in 2000, Fischer and his colleagues provided evidence for successful gene therapy of X-linked combined immunodeficiency (2). Reconstitution of the immune system was sustained. However, a significant setback was encountered by 2003. Several patients developed leukemia due to insertion of the gene therapy vector in an oncogenic locus, a complication that was anticipated as a rare “side effect” but may be addressable with improved vectors. Fortunately, chemotherapy induced remission in these patients. So, while there are potential serious adverse events associated with gene therapy, they need to be weighed against the lethality of the original condition and the capacity to manage the side effects of therapy. Although progress in the clinical arena hasn’t matched what was hoped for in the early 1990s, conclusive evidence of efficacy and success has emerged 10 years later.

Except for the use of bone marrow transplantation for the treatment of primary hematological conditions, the stem cell field (as related to treatment of human disease) is in its infancy, perhaps similar to the status of gene therapy nearly 20 years ago. Although we may despair of the recent events unfolding in South Korea, we should take solace from the confidence that strict adherence to scientific rigor and reason will ultimately prevail and permit realization of the potential of stem cells to ameliorate the suffering of patients with life-threatening diseases.

STUART H. ORKIN
Howard Hughes Medical Institute, Dana-Farber Cancer Institute, Children’s Hospital Boston, Harvard Medical School, 44 Binney Street, Boston, MA 02115, USA.
IT IS APPROPRIATE THAT SCIENCE SHOULD LEAD the way in reconquering exposure of the fraudulent claims of W. S. Hwang et al. that they developed 11 patient-specific cell lines by somatic cell nuclear transfer (SCNT) (D. Kennedy, “Editorial retraction,” Letters, 20 Jan., p. 335).

The profoundly negative effect of this episode is all the greater because of the way in which the matter was handled from the outset. When the 2005 paper was received in the Science editorial office, it was regarded as a showstopper, something that would make big headlines, with important implications for the treatment of a number of diseases. That much was noted in the News of the Week article “… And how the problems eluded peer reviewers and editors” (J. Couzin, 6 Jan., p. 23), e.g., “immediately, the journal’s editors recognized a submission of potentially explosive importance.” The paper was published in due course and hailed in several quarters as important science. But was its science in any way special?

Even if Hwang et al. had achieved what they described, all they had done was to repeat what others had already accomplished. Cooperation, the impact of this sorry affair might have been much less.

T. JOHN MARTIN
St Vincent’s Institute of Medical Research, 9 Princes Street, Fitzroy, Victoria 3065, Australia.

THE ROLE OF YOUNG KOREAN RESEARCHERS in the Hwang controversy (“How young Korean researchers helped unearth a scandal,” S. Chong and D. Normile, News of the Week, 6 Jan., p. 22) raises important aspects of research misconduct that are long overdue for international action.

It took the actions of an anonymous whistleblower to unmask the deception and dishonesty of Woo Suk Hwang. It is noteworthy that the whistleblower chose to make his allegations anonymously—even though he was no longer working in the laboratory—and to a TV program and not to the university involved or to regulatory authorities.

The central role of whistleblowers in the Hwang scandal affirms the urgent need for (i) whistleblowing of fraudulent activity to be accepted and encouraged as a legitimate duty that is integral to the responsible conduct of research; (ii) institutional policies that protect the rights of all parties, especially junior researchers, to due process and protection from retribution, intimidation, and harassment; and (iii) an international standard of responsible research and definition of research misconduct.

L. STEPHEN KWOK
14 Bedford Street, Willoughby North, NSW 2068, Australia.

Questions About Forensic Science

IN THEIR REVIEW “THE COMING PARADIGM SHIFT in forensic identification science” (5 Aug. 2005, p. 892), M. J. Saks and J. J. Koehler confuse the roles of adversaries in the criminal justice system with those of objective scientists. The “assumption of discernible uniqueness” may seem to be a tenet of forensic science; however, it is not found anywhere in the literature. They claim that “Traditional forensic scientists seek to link crime scene evidence to a single person or object ‘to the exclusion of all others in the world.’” Some analyses can never obtain such resolution, and the practitioners of those disciplines would not claim to be able to do so. Those disciplines that do seek to individualize evidence do not adhere to their invented proposition “when a pair of markings is not observably different, criminalists conclude that the marks were made by the same person or object.” The references they cite [see their (7, 8)] for this proposition contain no such language. Source attribution rarely, if ever, relies on a single marking.

We take exception with the implication that “all” experts have a propensity to fabricate and lie about evidentiary results. In fact, all comparative forensic science fields have a reasonably high frequency of exclusions. This is in conflict with the notion of data manipulation to achieve unique identification. There is as much incentive in obtaining a true result when it is an exclusion as there is in achieving a match. Fudging a match has dire consequences that the overwhelming majority of forensic scientists well appreciate; the true perpetrator is still free preying on innocent victims and the forensic scientist risks having a contrary (legitimate) scientific opinion presented in court.

Errors do occur in any endeavor involving humans. However, Saks and Koehler do not define the types of error that can occur and describe which ones are of consequence and which are not. Instead, they focus on diminishing the weight of evidence based on a hypothetical error rate that does not apply to the case at hand. Saks and Koehler declare that “the practical value of any particular technology is limited by the extent to which potentially important errors arise” as if this potential necessarily decreases the value of the evidence. A known error rate is not a direct measure of the reliability of the specific result(s) in question. The most direct way to measure the truth of the purported results is to have another expert conduct his/her own review (1), as is advocated by the National Research Council for DNA analyses (2).

Saks and Koehler misstate many of the false-positive error rates. For example, microscopic hair comparison is estimated at 12%.

ROCKNE HARMON1 AND BRUCE BUDOWLE2

1Alameda County (CA) District Attorney’s Office, Oakland, CA 94607, USA. 2FBI Laboratory, Quantico, VA 22135, USA.

References

IN THEIR REVIEW “THE COMING PARADIGM SHIFT in forensic identification science” (5 Aug. 2005, p. 892), M. J. Saks and J. J. Koehler assert that error rates in forensic science can be calculated for comparisons performed by human examiners, and that these error rates can then be used to predict the probability that...
an error (false match) occurred and thus assess the probable value of the identification for the jury. In fact, the National Research Council concluded that using error rates in such a predictive fashion (especially error rates gathered from proficiency testing) is inappropriate (1).

The likelihood of committing an error will be dependent on the complexity of the task, the examiner, and various conditions of the task. In forensic casework, the conditions are varied and we are human and fallible. Proper quality control is imperative to reducing (but not eliminating) the chance of error.

The authors indicated that proficiency test errors of fingerprint experts were “about 4 to 5%” false-positive errors on at least one fingerprint comparison. The manufacturer of these proficiency tests did not report a 4 to 5% “false-positive” error rate (erroneous matches), but rather they reported that 4 to 5% of the answers “differed from the manufacturer’s expected results” (2), a critical distinction. If an examiner reports “inconclusive” (perhaps they lacked the training and experience to make the match) or records an answer incorrectly (clerical error), this will be reported as “differing from the manufacturer’s expected results.” This is not a false match as the authors are reporting.

Their fig. 1, which purports to show a disturbingly high incidence of false testimony and forensic testing errors, has not previously been published in any peer-reviewed scientific journal. There is no discussion of the data sampling techniques, methods, or criteria that support this graph.

I have several questions regarding the source of these data: Were the errors attributed to faulty “forensic testing” from a handful of scientists or many? Were these cases and testimonies reviewed by experts qualified to make scientific determinations, or rather by lay people, law students/professors, and Innocence Project volunteers? Of the “forensic testing errors,” were these true testing errors or do they simply reflect the limitations of the tests and technology of the era?

I would invite the authors to perform their own research experiments, attend the identification conferences, and become involved in the community that is already performing the research for which they are calling. They will find a new generation of scientifically gifted and objective scientists, skilled at what we do, but interested in discovering new ways to improve it.

GLENN LANGENBURG
Minnesota Bureau of Criminal Apprehension, 1430 Maryland Avenue East, St. Paul, MN 55106, USA.

References
2. Recent CTS reports are available online at www.collaborative testing.com/forensics/forensics_prints.html.

I was dismayed to find a variety of errors in the Review by M. J. Saks and J. J. Koehler on forensic identification sciences (“The coming paradigm shift in forensic identification science,” 5 Aug. 2005, p. 892). Of chief concern is a spurious fact offered by the authors regarding a paper I co-authored with Bruce Budowle (1). In that paper, we reviewed 170 cases in which microscopic and mitochondrial DNA examinations were conducted on hair samples in casework. We found that out of 170 cases, 133 were sufficient for analysis; of these, in only 9 cases did the hairs have a similar microscopic appearance but different mtDNA sequences (6.7%). Nowhere in that paper do we state that error rates “for microscopic hair comparisons are about 12%” as Saks and Koehler quoted in their article. Moreover, the results of our study, although illuminating, cannot be used as an error rate for all forensic microscopic hair comparisons (2); the authors state this themselves, citing the National Research Council’s publication (2), but then go on to do just that for many forensic disciplines.

MAX M. HOUCK
Director, Forensic Science Initiative, Research Office, Manager, Forensic Research and Business Development, College of Business and Economics, West Virginia University, 886 Chestnut Ridge Road, Morgantown, WV 26506–6216, USA.

References

In their review “The coming paradigm shift in forensic identification science” (5 Aug. 2005, p. 892), M. J. Saks and J. J. Koehler claim that handwriting error rates on proficiency tests for handwriting experts are between 40% and 100%. What they fail to state is that the tests they are quoting from were given between 1975 and 1985. These initial tests were themselves designed as “tests” to create a fair gauge of proficiency that would also accurately reflect a forensic document examiner’s (FDE’s) casework. Even so, those in the early 1980s did not recognize the range of conclusions issued by FDEs. Qualified conclusions on the correct side of the opinion scale were incorrectly deemed errors, creating what appears to be a higher error rate. Saks has previously written that the Collaborative Testing Services (CTS) advisory committee informed him that proficiency tests were not suitable for use in gathering data on a forensic discipline (1). CTS tests given between 1990 and 2005 reveal that FDEs issued proper conclusions 95% to 100% of the time (error rates between 0 and 5%). The lower error rates are not due to CTS “dumbing down” the tests, but due to tests that more accurately reflect casework and the range of conclusions issued by FDEs. The error rates of the contemporary CTS tests are in agreement with Moshe Kam’s proficiency testing studies (2–5).

JAN SEAMAN KELLY
Las Vegas Metropolitan Police Department Forensic Laboratory, Suite 201 B, 5605 West Badara Avenue, Las Vegas, NV 89118, USA.

References

Response
The essential message of our review was that forensic individualization/identification science is on course for a “paradigm shift” in which its future will be more scientifically grounded than its past.

Harmon and Budowle take issue with the simple point that traditional forensic science assumes that markings produced by different people and objects are observably different. The notion of uniqueness is widespread in forensic science writing, thinking, and practice. We added the qualifier “discernible” to the uniqueness assumption to indicate that criminals do not refer to uniqueness in the abstract or as a metaphysical property. They mean that conclusions about object uniqueness are attainable in practice ([1], p. 45 and p. 123).

Harmon and Budowle suggest that we claimed that source attribution “relies on a single marking.” We said no such thing, as is evident in the sentence they quote. Our point was simply that when criminals cannot distinguish between two markings—such as two fingerprints—they assume the markings were made by a single person or object.

Harmon and Budowle misrepresent our Review when they say we implied that “all” forensic science experts have a propensity to lie. As we clearly indicated, the word between those quotation marks is that of Andre Moenssens, a former forensic scientist and lifelong supporter of the field. What we did say was that the organizational setting and culture in which many forensic scientists work can create pressures of the sort Moenssens describes. Recent reports of widespread data fudging and fabrication in forensic science provide additional reason for concern (e.g., [2, 3]).

Letters to the Editor
Letters (~300 words) discuss material published in Science in the previous 6 months or issues of general interest. They can be submitted through the Web (www.submit2science.org) or by regular mail (1200 New York Ave., NW, Washington, DC 20005, USA). Letters are not acknowledged upon receipt, nor are authors generally consulted before publication. Whether published in full or in part, letters are subject to editing for clarity and space.
Harmon and Budowle, as well as Langenburg, believe that error rates are not relevant for predicting the chance that an error will occur in an individual case. We addressed this belief in our Review (pp. 894–895) and elsewhere (4). It is a fallacy to believe that base rates should be disregarded in individual prediction tasks because they are insufficiently case-specific. From a Bayesian standpoint, the probability of error in a particular case requires an assessment of both the prior probability that the error will occur and the individuating features of the target case. Because the error rate informs the prior probability (and will often be identical to it), it is enormously relevant to an estimate of the chance of error in a particular case. This is one reason why forensic scientists should participate in well-designed proficiency tests on a regular basis. As reliable data from these tests accumulate, it should be possible to take advantage of increasingly refined error rate estimates.

Harmon and Budowle assert that we “did not point to one example of the foundations of the disciplines being baseless [but] merely focused on errors having been committed by scientists.” We did not say that forensic science is “baseless.” Instead, we identified a series of issues that go to the heart of the status of the traditional forensic sciences as mature sciences. For example, we pointed to forensic individualization science’s continued reliance on an unproven and likely untestable 19th-century model of uniqueness. We suggested that the field needs to adopt a more realistic, data-based, and probabilistic approach. We also noted the paucity of basic research on assumptions and lack of applied research on procedures.

Langenburg takes issue with our report that “[a]bout 4 to 5% of examiners committed false-positive errors on at least one latent” in fingerprint proficiency tests conducted during the past decade. He says that 4 to 5% actually represents the rate that answers differed from a manufacturer’s expected results. Langenburg is mistaken. Our 4 to 5% estimate is the proportion of analysts who indicated that a latent print matched a finger that it did not match at least one time on the proficiency test. This estimate does not include inconclusives. The proportion of analysts who gave answers that differed from the manufacturer’s expected results (i.e., the proportion of analysts who did not correctly identify all latent prints in a test) is much larger, about 25%.

Consider the most recent of many latent print proficiency tests that we relied upon in the paper (5). In this test, 259 analysts were provided with 11 latent prints plus known prints from 4 relevant individuals (persons A to D). Seven analysts (3%) committed obvious false-positive errors. Of these, two analysts mistakenly said that a print that belonged to person A belonged to person B; three analysts mistakenly said that prints that should have been marked as unidentified belonged to person C; and two analysts mistakenly matched prints that belonged to persons A and B to people who were not even provided on the test. These are false-positive errors.

One cannot sweep away the mistakes that have been committed by suggesting they are mere clerical errors or the cautious “inconclusives” of novice examiners. Proficiency tests have detected, and continue to detect, significant false-positive errors by latent print examiners. The rate at which these and other errors occur should be tracked, published, and studied to help identify the probative value of reports offered by forensic scientists.

Langenburg expresses concern about the data on DNA exonerations cases that appear in our fig. 1. As indicated in the Review, the underlying data were provided to us by the Innocence Project, and we relied on those data when computing the proportions associated with the factors in the figure. These data represent all of the DNA exonerations cases that have been coded by the Innocence Project (n = 86 cases) to date. Dozens more cases remain uncoded. Research on DNA exonerations is obviously in its infancy, and we support calls for a more complete and scientific review of these cases.

Houck complains that the 12% error rate we provided for microscopic hair comparisons “using results of mitochondrial DNA testing as the criterion” (p. 895) is not expressly stated in the Houck and Budowle article we cited (6). The data in the Houck and Budowle article formed the basis of our computations, just as they did for the new calculation that Houck offers in his Letter. Table 2 in Houck and Budowle compares the results of visual and mtDNA testing for 170 pairs of hairs (known and questioned). Each mode of testing yielded four categories of outcomes: association (the hairs match), exclusion (the hairs don’t match), inconclusive, and no exam (unsuitable sample for testing). Omitting the 37 unsuitable pairs, 133 remained. Houck now reports that “in only 9 cases did the hairs have a similar microscopic appearance but different mtDNA sequences (6.7%)” (sic: 9/133 = 6.8%). Even if Houck has sound reasons for deflating the error rate by including 38 inconclusives in his denominator, why not also mention that different conclusions were reached by the two methods 35% of the time (46/133)?

Where ground truth is unavailable, as in Houck and Budowle’s study, a conventional approach is to select what is believed to be the best measure as the criterion (“gold standard”) against which a measure of interest can be compared. Taking such an approach, how do microscopic hair comparisons stack up against the mtDNA gold standard when conclusions were offered by examiners? One way to report such data is to say that of the 26 cases in which the mtDNA found an exclusion, the examiners using the visual approach called an association 9 times. These data indicate a Type I false-positive error rate of 35% (9/26). Another way to look at the data is to report that 9 times out of the 78 times that visual examiners declared an association (12%), the mtDNA technique showed an exclusion. That is the 12% we reported in our Review.

We did not state that handwriting error rates on proficiency tests are between 40% and 100%, as Kelly claims. We said that the error rate has run as high as 100%, and we should have more clearly indicated that the risk of error on subsequent proficiency tests still ran as high as around 40%.

Kelly correctly notes that, in general, examiners made fewer errors on more recent proficiency tests than they made in the past. What accounts for this performance change? A thoughtful student of this matter has commented: “Have handwriting examiners improved abruptly and markedly? Or did the tests become easier? Most likely, the latter. The test manufacturers describe them as more straightforward, they appear to be simpler, and rather than complaining about test difficulty (as examiners did before the 1990s), examiners now commented about how easy the tests were” (7, p. 69).

The difficulty of the writing task could have an enormous impact on examiners’ performance. For example, in one recent test where the items varied in difficulty, examiners were pro-
vided with several handwritten receipts and two known sources (8). On one of the receipts, the signature was that of one of the known sources, but the text was produced by neither source. Of 131 examiners, 127 (97%) correctly concluded that the signature “was” or “probably was” that of one of the known sources. But only 25 examiners (19%) correctly excluded both known sources as persons who did not or probably did not write the text of the receipt.

These and other test data suggest that examiner performance varies markedly with the features of the writing task. We therefore support systematic research aimed at mapping the relationship between the varying attributes of writing problems and the circumstances of the examination, and how well examiners perform under those varying conditions. The fruits of such research will provide exactly the kind of information the Supreme Court says trial courts need, namely, guidance for assessing expert performance in the expert “task at hand” (9, 10).

MICHAEL J. SAKS1 AND JONATHAN J. KOEHLER2

1College of Law, Arizona State University, Tempe, AZ 85287, USA. 2McCombs School of Business, University of Texas, Austin, TX 78712, USA.

References
5. Collaborative Testing Services, Test No. 04-517 and 04-518: Fingerprints.
8. Collaborative Testing Services, Test No. 02-524: Handwriting.