June 20, 2010

Testing For Ethical Sensitivity To Responsible Conduct Of Research Among Multi National Stem Researchers

Susie L. Amato-Henderson, *Michigan Technological University*
Tom Drummer, *Michigan Technological University*
Debra Charlesworth, *Michigan Technological University*
Joanna Schreiber, *Michigan Technological University*

Available at: https://works.bepress.com/joanna-schreiber/6/
AC 2010-1787: TESTING FOR ETHICAL SENSITIVITY TO RESPONSIBLE CONDUCT OF RESEARCH AMONG MULTI-NATIONAL STEM RESEARCHERS

Michael Bowler, Michigan Technological University
Susie Amato-Henderson, Michigan Technological University
Tom Drummer, Michigan Technological University
Joseph Holles, Michigan Technological University
Joanna Schreiber, Michigan Technological University
Ted Lockhart, Michigan Technological University
Debra Charlesworth, Michigan Technological University
Jingfang Ren, Michigan Technological University
I. Introduction

The "Four Component Model" of moral behavior developed by Rest, and applied to professional ethics by Bebeau and others argues that ethical behavior is a consequence of four factors: (1) ethical sensitivity, (2) ethical judgment, (3) ethical motivation and commitment, and (4) ethical character and competence. The majority of theoretical and pedagogical research in the field of Ethics Education in Science and Engineering (EESE) has centered on ethical judgment. Significantly, relatively little work has been done on ethical sensitivity in EESE even though ethical sensitivity provides an individual with the range of concrete considerations upon which that person makes ethical judgments. Ethical sensitivity determines an individual's ability to recognize in any given situation what ethical issues are at stake, the possible consequences of actions, the rights and obligations of the individuals involved, and the ethical principles that are pertinent to the situation. In this regard, ethical sensitivity precedes ethical judgment. Moreover, an individual's ethical sensitivity determines to a significant extent the range and boundary of his or her ability to make appropriate ethical judgments.

The Four Component Model has been used extensively in the study of professional ethics and in the design, development and assessment of programs geared toward professional ethics training and education. It is only relatively recently that it has begun to be used in the study of the responsible conduct of research (RCR). In recent years, two instruments for measuring ethical sensitivity in science and engineering have been developed, namely, the Test of Ethical Sensitivity in Science and Engineering (TESSE) developed by Borenstein, et al. and the Test for Ethical Sensitivity in Science (TESS) by Clarkeburn. Although both investigated ethical sensitivity to issues arising from out of science and engineering, neither focused primarily on assessing ethical sensitivity of STEM (Science, Technology, Engineering, and Mathematics) researchers to the responsible conduct of research.

As part of an NSF sponsored project we have designed an instrument for testing the ethical sensitivity of STEM researchers to situations involving RCR. One of the significant hurdles faced by those wishing to develop such an instrument is that STEM researchers, even those residing and working in the U.S., represent a distinctly diverse group. Possible confounding variables include native language, education, or literacy. For example, one important issue in this regard is that STEM researchers’ fluency in English can vary widely. This is crucial in testing for ethical sensitivity given that these tests typically rely on an individual's ability to read, interpret and respond to a vignette. We have designed an ethical sensitivity instrument that attempts to mitigate some of these factors, thus giving us a more accurate assessment of an individual's ethical sensitivity to RCR.

Our key insight in this regard was that measuring ethical sensitivity to RCR by having participants respond in English to vignettes written in English that involve issues related to
research ethics alone was inadequate since one could never tell whether a disparity in score is due to different levels of ethical sensitivity or fluency in English or any number of other extraneous factors. Thus, we designed our test so that an individual’s ethical sensitivity toward RCR is measured by comparing his or her ethical sensitivity to cases involving research ethics with his or her ethical sensitivity toward everyday, non-research oriented ethical issues. Fluency in English and other extraneous factors will equally affect the respondent’s answers to the cases involving RCR and those involving everyday, non-RCR ethical issues. Thus, if one compares the respondent’s ethical sensitivity to RCR cases with his or her ethical sensitivity to the everyday, non-RCR ethical issues, rather than just relying on the first, one achieves a measure of ethical sensitivity that mitigates the extraneous factors. In other words, a given individual is serving as his or her own control with regards to these issues.

This method of mitigating extraneous factors is connected to a more general insight. Namely, that measuring ethical sensitivity to cases involving RCR alone does not paint a complete picture of an individual’s ethical sensitivity to RCR. Ethical sensitivity to cases involving some particular, circumscribed area of ethics (e.g., cases involving issue of professional ethics) indicates an individual’s ability to recognize in those cases ethically relevant aspects such as the responsibility of people involved, the consequences of actions, the rights and obligations of individuals, and principles of action. Certainly this ability is dependent upon a given individual’s general ability to recognize responsibility, consequences, rights and obligations, and principles of action in any given case. Put differently, someone who is generally insensitive to the consequences of actions will be insensitive to the consequences of actions when those actions involve RCR and vice-versa. Therefore, to obtain a complete picture of an individual’s ethical sensitivity to particular areas of ethical concern requires that one compare this, as best as one can, to that person’s general ability to be ethically sensitive. In this manner one can tell whether or not a given individual or group of people is specifically more or less sensitive to research ethics than they are sensitive to everyday cases involving ethical dilemmas. Knowing this would allow one to begin to examine and diagnose whether and why one individual or group of people is more or less sensitive to RCR over and above the fact that they may simply be more or less sensitive to ethical issues per se.

In this paper, we describe the design of our instrument for measuring ethical sensitivity to RCR and examine the results of our pilot study of the instrument among a group of multi-national STEM graduate students. We also investigate possible applications of our instrument to the study of research integrity and for the assessment of pedagogical approaches to responsible conduct of research.

II. Methods

Vignettes

Guided by the principles of responsible conduct of research (see Table 1) our research team created several vignettes with varying occurrences of violating ethical principles, some of which involved RCR issues. For the purposes of this paper, two such vignettes have been tested and scored, thus we will describe only these two below.
Table 1. Principles of Responsible Conduct of Research

<table>
<thead>
<tr>
<th>Principle</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intellectual honesty in proposing, performing, and reporting research</td>
</tr>
<tr>
<td>Accuracy in representing contributions to research proposals and reports</td>
</tr>
<tr>
<td>Fairness in peer review</td>
</tr>
<tr>
<td>Collegiality in scientific interactions, including communications and sharing of resources</td>
</tr>
<tr>
<td>Transparency in conflicts of interest or potential conflicts of interest</td>
</tr>
<tr>
<td>Protection of human subjects in the conduct of research</td>
</tr>
<tr>
<td>Humane care of animals in the conduct of research</td>
</tr>
<tr>
<td>Adherence to the mutual responsibilities between investigators and their research teams</td>
</tr>
</tbody>
</table>

In the first vignette, a fictitious faculty member (Dr. Xiao) who had been ridiculed by a colleague for his accent and how difficult it was to understand him (non-RCR issue) was later in a position of reviewing a NSF grant application submitted by his fictitious ridiculer (Dr. Willard). Without revealing his past associations with the applicant to his review panel (RCR issue; conflict of interest), Dr. Xiao suggests that the review panel reject Dr. Willard's proposal even though Dr. Xiao recognizes that it represented important research (RCR issue; Fairness in peer review). This vignette is referred to as the Xiao scenario for the remainder of this paper.

In the second vignette, a fictitious researcher (Dr. Arnaut) developed a drug that he hopes to patent after conducting a study of its effectiveness. Dr. Arnaut hires Dr. Watson to help with the study and write the final report. Due to difficulty recruiting participants, Dr. Arnaut decides not to tell participants about possible side effects of his drug (RCR issue; protection of human subjects) because he considers them to be minor. The results of the study do not support the effectiveness of the new drug. Therefore Dr. Arnaut fires Dr. Watson (non-RCR issue) despite paying him his full salary. This vignette is referred to as the Arnaut scenario for the remainder of this paper.

Pilot Study

We pilot tested our vignettes using an online survey with students who had applied to and been accepted into graduate programs at Michigan Technological University. A solicitation email was sent to personal and/or school-related email accounts provided by all accepted students (n = 740). The vignettes were contained within a larger study which examined a STEM researcher’s role orientation along several dimensions. Approximately 90 graduate-level students provided at least a partial response to the surveys, which equated to a response rate of about 25% after accounting for “bounced” email solicitations. Despite our attempt to minimize the length of the surveys by creating several versions, a large rate of partial responses resulted nonetheless. We believe the large rate of partial responses was due to the length of the survey. As a pilot study, one purpose was to narrow the potential pool of vignettes to include only those that are most effective. In addition, this survey contained two tools, thus future iterations will be done using single tool surveys that are much shorter in length.
Scoring of Responses

Five raters independently scored participants’ responses to the ethical issues contained within each ethical sensitivity vignette. As described above, each vignette contained from two to four ethically sensitive situations, some of which involved RCR issues. The Xiao scenario contained two RCR issues and one non-RCR issue. The Arnaut scenario contained one RCR and one non-RCR issue. Scores were assigned based upon recognition of the action, consequences, accountability of perpetrator, rights of the victim, and duties of those involved as provided in the written responses by participants (see Table 2 for an overview of the scoring criteria and method). Each issue could thus receive a score ranging from 0 to 6. Scores for each issue were summed for each rater resulting in what we termed an ethical sensitivity score (ES) for each issue. For example, the Xiao scenario resulted in 3 separate ethical sensitivity scores per participant.

Table 2. Ethical Sensitivity Vignette Scoring

<table>
<thead>
<tr>
<th>Criteria</th>
<th>Method of Scoring</th>
</tr>
</thead>
</table>
| Recognition of the action.                               | 0 = no explicit recognition in response  
1 = no explicit recognition in response, but implicit recognition based upon existence of other scored criteria for this issue  
2 = explicit recognition of the issue                      |
| Recognition of possible consequences resulting from the action. | 0 = No recognition in response  
1 = Recognition of issue in response                        |
| Recognition of the perpetrator’s responsibility or accountability for the action. | 0 = No recognition in response  
1 = Recognition of issue in response                        |
| Recognition of the rights of the victim(s) of the action. | 0 = No recognition in response  
1 = Recognition of issue in response                        |
| Recognition of the duties and/or obligations of the individuals involved in the action. | 0 = No recognition in response  
1 = Recognition of issue in response                        |
| **Total Score Range per ethical issue**                  | 0 to 6                                                                                                  |

Inter-Rater Reliability and Agreement Analysis

Scores were totaled for each ethical issue within the vignette resulting in one ES score per issue. We used the ES scores and the first criteria scores (recognition of the action, see Table 2) in the reliability analysis. In addition, ES scores were summed across all issues within each vignette, resulting in a total ethical sensitivity score (TES). For analyses comparing sensitivity of RCR issues to non-RCR issues, we computed an ES difference score, subtracting a participant’s ES score to the non RCR issues from their score on the RCR issues. The resulting ES difference score is a relative measure of ethical sensitivity to RCR vs. non-RCR issues.
We measured agreement between raters in ES scores with Pearson correlations. We measured reliability in the ES score with the intra-class correlation coefficient (ICC) based on a two-way mixed effects ANOVA model. Because the ICC is dependent on variability between subjects and can be misleading, we also tested for differences between raters using the ANOVA model. When we found overall differences between raters (P < 0.05) we compared least square mean scores for each pair of raters. Least square means were used rather than simple means because of the imbalance in the data. For pairs of raters found to be significantly different we report the mean and standard deviation of the difference in ES scores.

To determine sources of disagreement within each vignette, we estimated the ICC for the ES score for each ethical issue within the vignette as well as the individual criterion score regarding the recognition of each ethical issue (score on first criteria, recognition of issue). Failure to recognize an ethical issue would lead to the greatest discrepancy between rater scores.

In one vignette there was some missing data as not all raters scored all responses. We used only those responses scored by at least four raters. In the ICC analysis we used parameter estimation methods that allowed us to incorporate unbalanced data.

### III. Results

#### Inter-Rater Reliability and Agreement Analysis

**Xiao Scenario**

This vignette contained three ethically sensitive situations, two of which were RCR related. Of the 39 participant responses there were 34 complete cases in which all raters scored the response. The highest possible score for this vignette was 18.

Correlations between raters of ES scores ranged from 0.71 to 0.87 and averaged 0.79. The ES score intra-class correlation was acceptable with ICC = 0.80 (see Table 3), but two of the three ethical situations had a lower ICC than desirable. For each ethical situation the ES score ICC was very similar to the ICC for the scores on the recognition of the issue criteria.

<table>
<thead>
<tr>
<th>Item</th>
<th>Result</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total ES ICC</td>
<td>0.80</td>
</tr>
<tr>
<td>Ridicule of Xiao ES score (non-RCR) ICC and recognition of issue criterion score ICC()</td>
<td>0.64 (0.54)</td>
</tr>
<tr>
<td>Xiao nondisclosure ES score (RCR) and recognition of issue criterion score ICC()</td>
<td>0.84 (0.80)</td>
</tr>
<tr>
<td>Unfair rejection ES score (RCR) ICC and recognition of issue criterion score ICC()</td>
<td>0.66 (0.62)</td>
</tr>
</tbody>
</table>
There were significant differences in ES scores between raters ($P < 0.05$). Least square mean scores ranged from 7.46 to 8.99. Two pairs of raters had significantly different mean scores; the highest mean score was significantly different from the two lowest mean ES scores, which were very similar. For these two pairs the mean difference in total ES Scores was 1.53 ($SD = 2.83$) and 1.52 ($SD=2.53$) on the six-point scale.

**Arnaut Scenario**

Three raters each scored $n = 35$ sets of responses. This vignette contained two ethical issues, one of which was related to RCR, with a maximum possible TES score of 12. Correlations between the TES scores of the three raters were $r = 0.87, 0.87$ and 0.76. There was no significant difference between raters ($P = 0.40$) and the Total ES score ICC was 0.83, indicating a high degree of rater consistency. The mean TES scores for the three raters were 5.97, 5.94, and 6.28 (see Figure 1). The RCR ethical issue ES score (warning of potential side-effects) had an ICC of 0.75, with mean rater scores of 3.40, 3.48 and 3.48 (on the 6 point scale) and no significant difference between raters ($P = 0.90$). The differences between raters were larger for the non-RCR issue (means = 2.48, 2.48 and 2.88; $P = 0.05$), although the ICC was acceptable at 0.77. Intra-cluster correlations for recognition of the two issues were 0.79 (RCR issue) and 0.81 (non-RCR issue).

![Figure 1](image.png)

**Figure 1.** Comparison of mean scores for three raters on two vignettes, expressed as a percentage of maximum possible score. Vignette 1 contained two ethical issues and Vignette 2 contained three issues.
Comparison of Ethical Sensitivity to RCR and non-RCR issues

As described in the methods, ES difference scores were calculated to gauge participants’ sensitivity to RCR issues in comparison to their sensitivity to non-RCR issues. First, the mean of the individual raters’ ES scores were calculated for each issue. Next, for each participant the mean ES score on the non-RCR issues was subtracted from the mean ES score on each RCR issue to result in ES difference scores for each RCR issue (see Table 4).

<table>
<thead>
<tr>
<th>Issue</th>
<th>Mean ES score (SD)</th>
<th>Mean ES difference score for RCR issues (SD)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Xiao – ridicule of accent (non-RCR)</td>
<td>3.195 (1.49)</td>
<td></td>
</tr>
<tr>
<td>Xiao – nondisclosure of association (RCR)</td>
<td>1.562 (1.84)</td>
<td>-1.633 (1.90)</td>
</tr>
<tr>
<td>Xiao – unfair peer review (RCR)</td>
<td>2.892 (1.41)</td>
<td>-0.304 (1.14)</td>
</tr>
<tr>
<td>Arnaut – not informing of side effects (RCR)</td>
<td>3.438 (1.62)</td>
<td>0.819 (1.54)</td>
</tr>
<tr>
<td>Arnaut – firing Dr. Watson (non-RCR)</td>
<td>2.619 (1.64)</td>
<td></td>
</tr>
</tbody>
</table>

Using these ES difference scores as a measure of relative sensitivity to RCR issues compared to non-RCR issues, a series of planned independent groups t-tests (alpha = .05) examined ES scores and ES difference scores between native English speakers (n = 28) and non-native English speakers (n = 12). This analysis was conducted to examine our premise that simply measuring ethical sensitivity on RCR issues is not an accurate indication of differences in RCR ethical sensitivity – a measure of relative sensitivity is needed. For example, just examining the measurements of ethical sensitivity to RCR issues alone in the Dr. Xiao vignette (see Figure 1) shows a rather large difference between native and non-native English speakers. This difference could be due to language ability, communicative abilities in general, reading comprehension, or any of a number of other factors including true differences in ethical sensitivity.

In fact, significant differences in ES scores existed between groups on all of the RCR issues (two from Xiao, one from Arnaut), with native English speakers scoring higher (i.e., larger ES scores) than non-native English speakers (See Figure 2). Significant differences were also found on the Xiao non-RCR issues, (t(38) = 2.508, p = .025) with native English speakers scoring higher (m = 3.625, SD = 1.10) than non-native English speakers (m = 2.194, SD = 1.83). On the non RCR issue in the Arnaut scenario, no significant differences existed between groups (t(33) = 1.58, p = .125) with native English speakers (m = 1.04, SD = 1.39) and non native speakers (m = .08, SD = 1.88) both scoring relatively low on the issue.

Examining relative measures of ethical sensitivity using the ES difference score (see Figure 2), the apparent differences between native and non-native English speakers are not so great. Since the influence of language, communicative ability, reading comprehension, etc., have a more or less consistent influence on measures of RCR and non-RCR ethical sensitivity in our instrument, the difference in the relative scores (ES difference scores) should not be due to those factors. One could assume that any differences in relative scores are due to factors impacting one’s ethical sensitivity. Thus, this would provide one with a more accurate indication of differences across individuals in ethical sensitivity to RCR.
In the present case, when the difference scores were examined (i.e., difference ES scores), no significant differences existed on this measure of relative sensitivity between native and non-native English speakers (see Figure 3). This indicates that what appeared to be a large difference in RCR ethical sensitivity between native and non-native speakers when looking at only raw RCR sensitivity scores was likely due to factors other than actual differences in actual sensitivity (i.e., language, reading comprehension, communicative ability, etc.) Thus, the resulting ES difference score, calculated using each individual as their own control by comparing their sensitivity to RCR issues against sensitivity to non-RCR issues, appears to provide a more accurate measurement of differences between individuals in ethical sensitivity to RCR issues. The findings of no differences between groups in the present case should not lead one to assume that there would never be differences. In fact, one would expect, especially between individuals, that there would be differences in this relative measure of ethical sensitivity when applied to difference cases or RCR principles. When such differences are observed, one should assume that they are not due to language, reading comprehension or communicative ability, but instead due to differences in sensitivity.

![Figure 2. Ethical Sensitivity Scores for RCR and non-RCR issues](image-url)

**Figure 2. Ethical Sensitivity Scores for RCR and non-RCR issues**
II. Discussion

Rater consistency was acceptable for both vignettes, although on one vignette we did observe enough variability between raters to warrant further training to decrease rater variability. There was less variability between raters on the vignette with fewer ethical situations which may just be an artifact of higher scores having higher variability. Respondents tended to have slightly higher percentage scores on the two-issue vignette compared to the more complex vignette (Fig. 1).

Perhaps more importantly, our test appears to control for varying language skills, reading comprehension and communicative ability. When scoring just an individual’s sensitivity to RCR issues, differences between native and non-native English speakers existed. When calculating a relative ES difference score, which is a comparison of sensitivity to RCR issues and sensitivity to non-RCR issues within each individual, participants served as their own control for these differences. This method of examining differences in ethical sensitivity should provide a broader picture of ethical sensitivity across various individuals of diverse backgrounds.

Our instrument provides researchers with a valuable tool for assessing differences between individuals’ ethical sensitivity to situations involving RCR versus those that do not involve RCR. Importantly, it is able to give a measure of differences in ethical sensitivity among diverse individuals from various backgrounds. Thus, this tool is applicable to the diverse and multinational group of people who make up the population of STEM graduate students and
researchers. In addition, we believe that our tool can be a model for those designing instruments that measure differences in ethical sensitivity among other groups of people to other situations, especially where there may be a confounding issue involving language, reading comprehension or communicative ability. Also, this tool can be adapted such that it can measure differences in ethical sensitivity while mitigating unforeseen contravening factors. In each of these respects, the instrument that we have designed will help educators assess an individual’s relative ethical sensitivity while controlling for factors that, although unconnected to ethical sensitivity, can significantly affect the measurement provided by traditional ethical sensitivity instruments. This will help educators avoid treating someone as deficient with respect to others regarding ethical sensitivity when in truth it could merely be an issue of language deficiencies, etc. Many of those who have or would have scored lower on traditional ethical sensitivity instruments simply because of language fluency will be afforded a much more accurate and fair assessment of their relative ethical sensitivity.

By measuring ethical sensitivity in relation to a controlling situation, in this case ethical sensitivity to RCR in relation to ethical sensitivity to everyday, non-RCR cases, our tool achieves a more subtle and nuanced determination of an individual’s relative ethical sensitivity to RCR. Rather than simply providing scores for ethical sensitivity to RCR among a group or among individuals, it allows researchers to investigate whether and why certain groups of people or particular individuals are more or less sensitive to situations involving RCR than they are to everyday ethical situations. For example, it tells researchers something important about two individuals who score equally on ethical sensitivity to cases involving RCR when one of them scores equally on ethical sensitivity to everyday ethical situations as they do to situations involving RCR, while the other scores much lower on RCR than they do to everyday situations. In the latter case, the person may exhibit certain “blind spots” to situations involving RCR or may be unable to recognize one or more of the elements that contribute to ethical sensitivity to RCR. For instance, he or she may simply be unable to readily recognize the consequences of certain actions related to decisions involving STEM research or fail to see that a principle of RCR is at stake in a given situation, where in cases involving everyday situations they are able to do these things quite well. The same will apply if our instrument is used as a model for developing ethical sensitivity instruments for other areas or other disciplines.

Of course, this opens up profound possibilities for advancing ethics education regarding ethical sensitivity. It allows educators if they so choose to target in their educational interventions the one or two elements of ethical sensitivity for which an individual or a group shows a deficiency in relation to their ethical sensitivity to everyday ethical situations. For instance, educators could tell that a certain student or group of students simply need to be better educated about the principles of RCR since they consistently exhibit a lower level of recognition that an ethical principle is involved in a particular situation involving RCR than they do in cases involving everyday ethical dilemmas. At the same time, it would also allow educators to determine that no amount of extra training in the principles of RCR is going to help a certain individual or group of people because their ethical sensitivity is not deficient in that regard, but perhaps they need help in seeing the consequences of actions.

Our instrument would also give educators the ability to examine the effects of particular educational interventions. For example, if the result of a student taking a course that specifically
targets RCR were to raise the student’s ethical sensitivity to everyday ethical dilemmas as much
as or nearly as much as it does to cases involving RCR then perhaps a general, non-specific
course in applied ethics would do the same, thus extenuating the need for a whole collection of
ethics courses targeting specific areas. In other words, perhaps the specialized course is in truth
simply raising the student’s general ethical awareness and that is why they are more sensitive to
issues involving RCR. Our tool would give education researchers the ability to detect this and
similar phenomenon. Of course, the instrument can also be used to pinpoint which aspects of
ethical sensitivity an educational intervention is affecting, thus demonstrating some limitations or
advantages of that intervention.

These are just some of the new and important applications that our tool makes possible. And as
a template, it opens up the possibility of applying these methods to other areas and disciplines.

V. Conclusion

In conclusion, we have developed an instrument for measuring differences in ethical sensitivity
to responsible conduct of research for STEM researchers. We have piloted this instrument using
a group of multi-national STEM students accepted to graduate programs at Michigan
Technological University. Rater consistency was acceptable when scoring the individual
vignettes in this instrument. The instrument included both RCR and non-RCR issues in each
vignette. In this manner, an individual respondent's ethical sensitivity toward RCR is measured
compared to his or her ethical sensitivity to non-research oriented ethical issues. Because of its
wide applicability, this instrument may offer a valuable tool for assessing an individual's ethical
sensitivity to situations involving RCR. While our tool does not currently assess every RCR
principle, the basic methodology and scoring provide the foundation for designing a tool that
would measure differences in ethical sensitivity for every RCR principle, or some relevant subset
of principles.

Acknowledgements:
The authors gratefully acknowledge financial support from the Ethics Education in Science and
Engineering program, Division of Social and Economic Sciences, Directorate for Social,
Behavioral & Economic Systems of the National Science Foundation through grant #0832922.

References
   ethics (Hillsdale, NJ: Lawrence Erlbaum Associates).
   389 – 408.