Algorithm Fairness and Respondus Monitor Proctoring: A study using *Casual Conversations*

Scott Klum, David Smetters

Update 2025 - The study that follows was conducted in 2022 to analyze the fairness of computer vision models for Respondus Monitor, an automated proctoring system. An updated analysis in 2025 seemed worthwhile given the enhancements to its proprietary models over the intervening years.

In 2022, we analyzed the 92 false positive errors (0.2%) that occurred during the processing of approximately 45,000 videos. By 2025, Respondus Monitor's face detection models produced only 18 false positive events (0.04%) for the same videos – and only 12 when controlling for poor lighting. While this improvement is significant and welcome, there are now too few false positive errors across age, gender, and skin tone to draw meaningful conclusions.

The 2022 study is still useful in that it shows the rigorous process needed to evaluate fairness, and it offers a publicly available baseline for other proctoring systems. The study also highlights key challenges for face detection systems like poor lighting. By 2025, however, the high level of precision of Respondus Monitor's face detection models has turned the question of algorithm fairness into a footnote rather than a headline.

Respondus' commitment to ensuring fairness doesn't end with this work. Beyond the internal analysis and validation for each model released, Respondus plans to incorporate data from Meta's second release of Casual Conversations which includes variables on disability and physical attributes. The analysis will continue to be publicly available so educators can have confidence in the systems they deploy.

Introduction

Respondus Monitor® is a fully-automated proctoring system used at approximately 1,500 colleges and universities to protect the integrity of online exams. Amongst its features, Respondus Monitor uses face detection technology to determine if an examinee's face is present in the video frame. If an examinee's face cannot be detected for a certain amount of time, the proctoring results show a flag for that segment of the video. These results appear on a video timeline, and the video can be reviewed by the instructor to determine if an exam violation has occurred.

Respondus is committed to using responsible artificial intelligence (AI) and to following industry best practices in computer vision and machine learning. A core aspect of this commitment is ensuring *fairness*. Fairness is a measure of equity across different groups of users (e.g. users of different genders, skin tones, and ages). This paper examines methods used by Respondus to ensure algorithm fairness in its use of face detection in Respondus Monitor.

A primary best practice in the analysis and evaluation of machine learning algorithms is to examine raw data whenever possible[1]. This is made possible with Respondus Monitor when examinees provide Respondus permission to use their proctoring videos for research purposes[2]. However, privacy considerations prevent this data from being shared with other researchers or used for illustrative purposes.

As a solution to this issue, we present an analysis of Respondus Monitor using *Casual Conversations*[3], a publicly available dataset created by Meta AI (formerly Facebook AI). While Casual Conversations was originally designed for research on deepfakes, it provides a close proxy to proctoring data of examinees. Importantly, because it is publicly available, interested parties can examine videos alongside face detection results to gain a better insight into the performance of the system, and to understand the challenges inherent to proctoring data.

What is algorithm fairness?

Algorithm fairness in AI, computer vision, and machine learning is a complex topic. Generally speaking, for an algorithm to be considered fair, it should maximize the equality of some outcome across groups. Algorithm unfairness (also called "bias" or "discrimination") is an oft-discussed topic in the media, particularly with respect to face recognition, gender detection, and ethnicity detection. The stories describe such systems going awry, often because of poor implementation or misuse. In the worst cases, mistakes can result in people being implicated for a crime they did not commit.

Fairness is therefore a critical measure to consider when designing and implementing algorithms. It is not enough to simply consider fairness in the abstract, it needs to be measured and interpreted such that the algorithms in question can be iterated on and improved. Ideally, a diverse set of stakeholders are involved in the process, with the results being made available to users of the system and other researchers. This paper represents one example of how this occurs at Respondus, and to offer a model which other online proctoring companies can follow and build upon.

Face detection vs. face recognition

Face detection and *face recognition* are distinct technologies, yet the concepts are frequently (and incorrectly) used interchangeably. The confusion is undoubtedly due to the similarity of the names themselves as well as the fact that these technologies are often used together. Indeed, one cannot perform face recognition until face

1. Readers are encouraged to examine this and other best practices from Google, Microsoft, and Amazon.

2. Some Respondus Monitor users grant Respondus the right to use their anonymized data for research purposes. No data from the European Union, California, and certain other regions are used for research or product improvement purposes.

3. https://ai.facebook.com/datasets/casual-conversations-dataset/

detection has first occurred. But to confuse face detection with face recognition is like confusing baking powder and baking soda. The white powdery ingredients look alike, sound alike, and are even used together in baking, but their chemistries are entirely different, as are their purposes.

Face *detection* algorithms are used to determine if one or more human faces are present in an image and, if so, where the faces are located within the image. Many algorithms have been designed for face detection over the years. Older algorithms are often based on lighting contrast between different regions of the face. Newer algorithms involve complex neural nets that build a statistical model to identify facial features – features that are often uninterpretable to humans. When face detection algorithms are chosen and implemented properly, they can achieve exceptionally high performance, nearly matching human performance. There are, however, situations where even the best algorithms fail.

Face *recognition*, by contrast, uses the locations provided by face detection algorithms to create a biometric template of the face. These templates can then be compared to determine if two faces belong to the same identity. When an iPhone is unlocked with FaceID, for example, a face recognition algorithm has determined that the person unlocking the phone and the owner of the phone are the same person. Similarly, gender, emotion, and ethnicity detection algorithms take the location of a face provided by a face detector and perform additional analysis.

Most fairness issues raised in media stories relate to algorithms that *analyze* a face, not simply detect the presence of one. Nearly all digital cameras use face detection to locate human faces in the photo frame to determine an optimal focal point, often placing a green square around each face detected. There is little controversy with this technology amongst regulators and the public at large.

Face detection algorithms, not face recognition algorithms, are used within Respondus Monitor to determine if an examinee is present or absent from a proctoring video.

Face detection in Respondus Monitor

If a face cannot be detected for a certain amount of time during a proctored video, a flag for that segment will appear in Respondus Monitor's proctoring results. The instructor can then review the video that corresponds with the flag to determine whether an exam violation has occurred. It's important to note that when Respondus Monitor uses face detection to flag a video segment, it does not mean an exam violation has occurred. It simply means the examinee's face could not be detected in the video during that portion of the exam.

Proctoring videos are different from those often associated with the use of face detection, like surveillance videos. For example, proctoring videos are generally recorded with an inexpensive webcam, indoors, using artificial light, and with the subject's face at about 0.5 to 1 meter from the camera. The unique characteristics of proctoring videos enable Respondus to select algorithms well-suited for such environments and underscores the importance of using real proctoring data to fine-tune such algorithms.

Despite these specialized algorithms, questions arise as to why Respondus Monitor is sometimes unable to detect the presence of an examinee when the instructor can see the person in the video frame. These questions lead to a discussion of algorithm errors.

Understanding errors: false positives and false negatives

Face detection algorithms, such as those used within Respondus Monitor, are evaluated according to their errors. Two types of errors can occur: a face is incorrectly detected in a location where no face exists (a *false positive*), or no face is detected in a location where a face exists (a *false negative*).

Because Respondus Monitor uses face detection to determine if an examinee is *missing* from the video, these definitions are reversed. That is, a false positive in Respondus Monitor occurs when it incorrectly flags an examinee as missing when the person is actually present. A false negative occurs when it incorrectly indicates an examinee is present when the person is actually missing.

Determining whether an error has occurred with a proctoring system's algorithms is more nuanced than it may appear. A computer vision researcher might be pleased that their face detection system can detect a person's face in an extremely dark room, whereas an instructor might wonder why the video wasn't flagged since the face is hardly visible to the human eye. Conversely, if an examinee's head is tilted downward and hair is fully covering the person's face, a researcher wouldn't consider it an error if the face was not detected, whereas an instructor might. Herein lies the art of developing face detection technology for use with a proctoring system.

It's also important to consider both types of errors in concert. For instance, an instructor may be alarmed if the proctoring system fails to detect when an examinee has moved outside the video frame. But if the face detection system triggers many missing flags when the examinee is still within the video frame, that too results in an unsatisfactory experience. In general, Respondus' research team prioritizes the minimization of false positives over the maximization of true positives, primarily because the latter is relatively straightforward to achieve at satisfactory rates. This study focuses on ensuring that false positives, when they occur, are distributed fairly across groups.

Algorithm fairness in Respondus Monitor

With an understanding of errors and associated challenges in mind, we can now discuss the fairness of Respondus Monitor's use of face detection. As alluded to above, we want to minimize the difference in error rates (in particular, false positives) across ages, genders, skin tones, and other criteria (e.g. the presence or absence of eyeglasses).

With proctoring videos, false positive errors largely result from three conditions (see Fig. 1):

1. Face occlusions – Anything that covers a significant portion of the face can cause an error. Examples include hair, hands on the face, clothing (hoodies, baseball caps, burqas), masks, etc.

2. Face cropping – If a webcam is tilted too high, the lower section of the face may be cropped (e.g. the nose, mouth and chin may not be visible) which can cause errors; cropping also occurs when the examinee leans outside of the left or the right of the video frame (common with open book/notes exams).

3. Poor capture conditions – Extreme backlighting can cause an examinee's face to be very dark while extreme forelighting can cause an examinee's face to be very bright. Both of these present challenges for face detection systems. Low lighting

introduces a similar problem, where there is minimal contrast on the face. Poor quality webcams and slow internet connections can also introduce artifacts – these are less impactful than lighting issues, but can still present problems for face detection algorithms because less data is available for analysis.

When these conditions are controlled for, the computer vision team at Respondus has consistently found no significant differences in error rates across age, gender, and skin tone groups when using production data (where approximate labels are added by researchers). However, consistent with existing research[4], there are slightly higher error rates when videos have poor capture conditions (particularly, poor lighting) and examinees have very light or very dark skin tones. This is, in part, due to the lack of contrast between the foreground (i.e. the face) and the background of the video frame. Indeed, in these cases it's often a challenge for a human to make out details of a face. We only know a face is present based on contextual clues.

Respondus Monitor mitigates the impact of the above-mentioned conditions in several ways. One approach is to integrate secondary algorithms that do not involve detection of a face. Similarly, the duration of an event, motion, and other factors can be used to smooth the analysis and reduce false positive results.

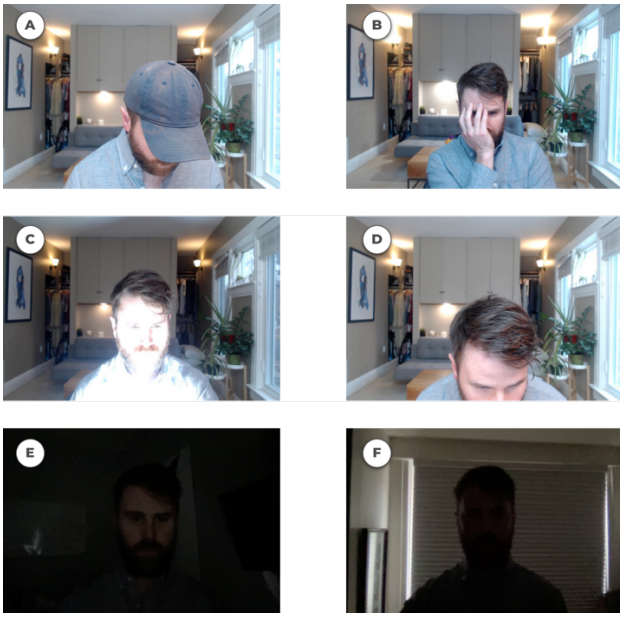


Figure 1: Example conditions that can cause false positive errors in proctoring videos: (a) occlusion from headwear, (b) occlusion from examinee's hands, (c) extreme forelighting, (d) face cropping from examinee leaning forward, (e) dark lighting, and (f) dark lighting with backlighting causing a silhouette.

Meta's Casual Conversations

Respondus has used production data since 2015 to measure algorithm fairness and to reduce false positive event rates arising from poor face detection. However, production data does not allow for analysis by those outside of Respondus. Using Casual Conversations, a dataset provided by Meta AI (formerly Facebook AI), it is possible to analyze the performance of the Respondus Monitor's algorithms in an open and verifiable way. The Casual Conversations dataset is comprised of 45,186 videos from 3,011 paid participants who responded to "random questions from a pre-approved list to provide their 'unscripted' answer". All participants self-reported their age and gender, while trained annotators labeled participants' apparent skin tone using the six-level Fitzpatrick scale[5]. Videos are also annotated as having normal or dark lighting.

Casual Conversations is not a proctoring dataset. It does, however, provide similar enough data to enable analysis and illustration of the conditions conducive to the types of errors described above. Further, it can be downloaded and used by other researchers or interested parties to explore the data or evaluate their own systems. As a result of this study, the Casual Conversations videos will supplement existing regression tests at Respondus to ensure new algorithms used in Respondus Monitor support the goal of maximizing fairness.

Respondus Monitor on Casual Conversations

All Casual Conversations videos were processed using the Respondus Monitor proctoring system. A total of 126 events (i.e. a face not being detected for a significant duration) occurred across the 45,186 videos[6]. Each event was then analyzed by staff on the Respondus computer vision team to determine if it was a false positive. Any event in which the participant was, in fact, missing from the video (i.e. a true positive) was excluded from the error rate calculation. Any event that occurred because the video was excessively blurred was also excluded. This resulted in a false positive rate (FPR) of approximately 0.2% (92/45186)[7].

Lighting

Fig. 2 shows the FPR across three groupings: all videos, videos that have normal lighting, and videos with dark lighting. The statistically significant difference in FPR between videos with normal lighting and dark lighting is intuitive: it's harder to see faces in the dark. The subsequent sections describe FPRs across three demographic variates (gender, age, and skin tone) while highlighting the impact of lighting on each.

---

[4] See section 1.2 "The role of image quality" in https://nvlpubs.nist.gov/nistpubs/ir/2019/NIST.IR.8280.pdf

[5] The Fitzpatrick scale is a six-point numerical schema for classifying human skin tones, with 1 being the lightest skin type and 6 the darkest.

[6] Frames and labels from all events can be found at https://static-storage-cloud.respondus2.com/events/index.html

[7] It was outside the scope of this study to determine if all instances of a person missing from a Casual Conversation video (known as a true positive) were properly detected by the Respondus Monitor system. Respondus uses a set of extensively annotated proctoring videos to ensure a satisfactory true positive and false negative rate. This study focuses instead on the false positive rate of Casual Conversations using the same parameters used for production proctoring data.

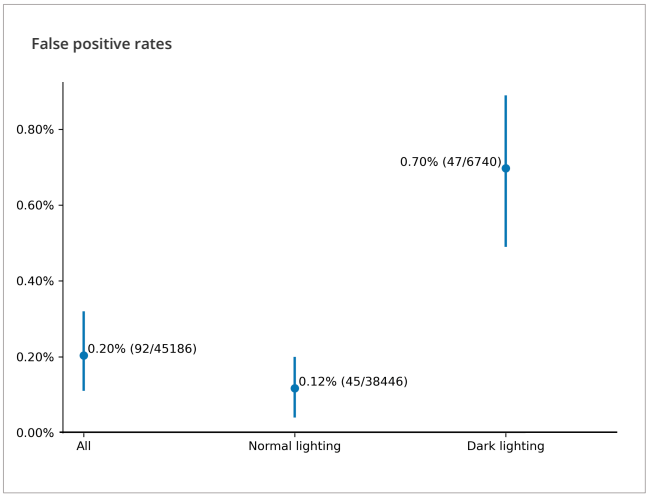
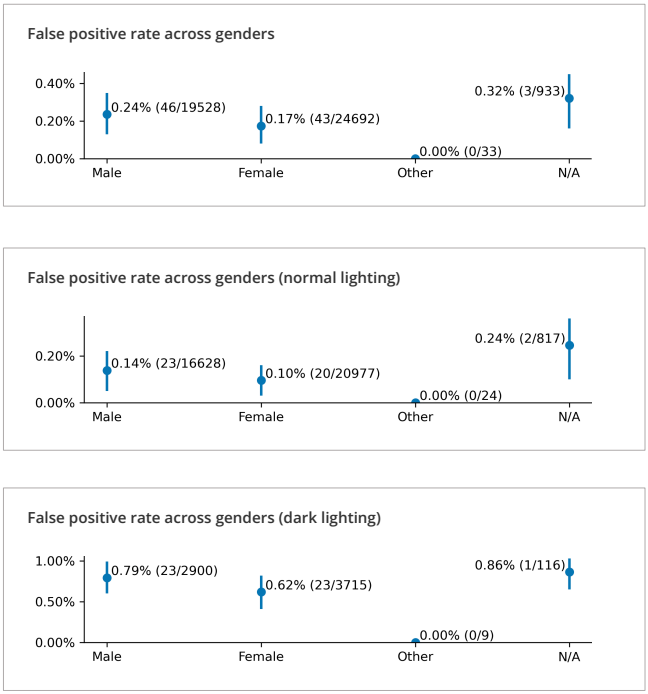


Figure 2: False positive rates on Casual Conversations for all videos, videos with normal lighting, and videos with dark lighting.

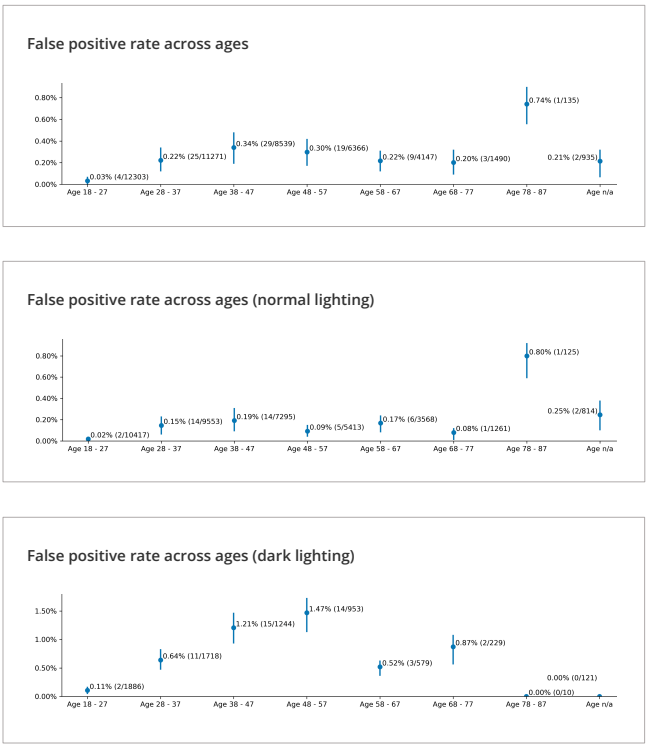
Gender

No statistically significant differences in FPRs were found between videos with “Male” or “Female” participants, nor with participants who did not provide their gender (“N/A”, see Fig. 3). While participants of the “Other” gender had no errors, the total number of videos for this group (33) is insufficient to make any strong conclusions. Notably, dark lighting increases FPR for the “Male”, “Female” and “N/A” genders, consistent with the pattern observed across all videos.



(full size image) Figure 3: False positive rates on Casual Conversations across genders (including subsets of normal and dark lighting). No statistically significant differences are observed between genders but dark lighting increases errors for “Male”, “Female”, and unreported (“N/A”) genders.

Age

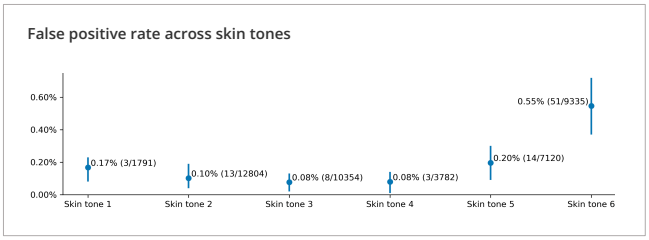


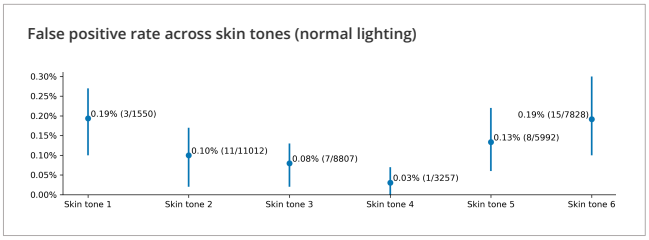
(full size image) Figure 4: False positive rates on Casual Conversations across age ranges (including subsets of normal and dark lighting).

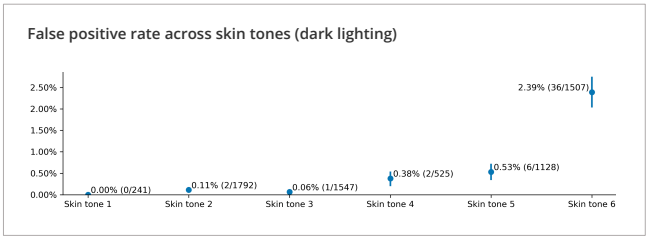
Participants between the ages of 78 and 87 were flagged at a higher rate than all other age ranges (Fig. 4), but this result, while statistically significant, is based on a single error out of 135 videos. It is unlikely this pattern would hold if there were more videos of participants in this age range. As with gender, dark lighting increases the false positive rate of most age ranges. Dark lighting also increases the difference in FPR between age ranges, but no age range has a statistically significant increase in FPR over all others. Further, there isn’t a clear relationship between age and FPR. This indicates that the variability here is more correlative than causal.

Skin tone

Fig. 5 shows FPR across Fitzpatrick scale skin tones. When videos have normal lighting, no skin tone has a higher FPR than all others. However, Skin tone 6 has a statistically significant increase in FPR when there is dark lighting. If dark lighting is not controlled for, this result also occurs with all videos of participants with Skin tone 6. These results are consistent with analysis performed on production data from Respondus Monitor. Example frames from erroneous events can be found in Fig. 6. These examples help illustrate the lack of contrast when a video has dark lighting which can present a challenge for face detection systems.







(full size image) Figure 5: False positive rates on Casual Conversations across Fitzpatrick scale skin tones (including subsets of normal and dark lighting). Dark lighting results in a statistically significant increase in the false positive rate of Skin tone 6.

Figure 6: Examples of false positive errors in videos with dark lighting and participants with Skin tone 6. The low contrast shown in these examples can cause face detection systems to fail.

Discussion and future work

The computer vision team at Respondus does not consider this study exhaustive or fully reflective of real-world proctoring videos. Some variates that exist in traditional proctoring videos, like examinees resting heads on their hands, are not present in Casual Conversations. Annotated proctoring data is therefore essential when maximizing the fairness of systems like Respondus Monitor.

At the time of this writing, Respondus is testing a new generation of algorithms that reduce false positive rates by approximately 90%, with the Casual Conversations data. As researchers, we must balance this substantial reduction in false positive rates with the understanding that Respondus' objective is not to build a low-error face detection system. Rather, the goal is to provide proctoring results that help instructors better identify where an exam violation may have occurred.

Instructors generally want to be alerted to the fact that an examinee's room has extremely low illumination, or that the student's face is mostly outside of the video frame (even when the technology can detect faces in both situations). These proctoring-specific challenges necessitate adjustments to the parameters of face detection algorithms and often merit the use of supplemental algorithms (such as those that evaluate the illumination of the exam environment). As new algorithms are added and complexity increases, so too does the importance of systems that measure these algorithms for fairness.

Conclusions

Respondus Monitor is an automated proctoring solution that uses face detection algorithms to help instructors determine if an exam violation has occurred. It is critical to ensure that the algorithms used by proctoring systems are fair for all users, regardless of age, gender, or skin tone. Using the publicly available Casual Conversations dataset from Meta AI, this work analyzed the fairness of Respondus Monitor's face detection algorithms by comparing false positive rates (FPRs) between variates.

Across all of Casual Conversations, Respondus Monitor had a low FPR with an error occurring in about one in every 500 videos. Regardless of lighting conditions, no gender or age range had a meaningful increase in false positive rate over all others. However, FPRs generally increased in videos with dark lighting. Further, there was a statistically significant increase in FPR in videos of participants with the darkest skin tone compared to other skin tone groups, although this effect disappears when lighting is controlled for. The effect of dark lighting is likely due to the lack of contrast, illustrated by the example frames in Fig. 6. Errors in such environments are often understandable to instructors. Indeed, some instructors expect to be alerted of such events and therefore would not view these as errors.

Respondus is continually developing new algorithms to reduce all types of errors. Regardless, this study supports Respondus' recommendation that examinees take exams in well-lit rooms. Not only does this help instructors see the student and examination environment more clearly, but can mitigate a large percentage of false positive errors.

Acknowledgments

We want to thank Meta (formerly Facebook) for funding the development and maintenance of the Casual Conversations dataset. We also want to convey thanks to the Meta AI team for their quick responses to our questions along the way.

About the Authors

Scott Klum *is Chief Scientist at Respondus. He received his M.S. in Computer Science from Michigan State University and has numerous peer-reviewed publications on the topic of biometrics and computer vision.*

David Smetters *is Founder and CEO of Respondus. Prior to Respondus, he was on the executive team of MicroCase, a developer of statistical analysis software and research methods applications.*