Leveraging the Crowd to Evaluate Empathy with Virtual Humans

Andrew Cordar
Computer and Information Science and Engineering
University of Florida
Gainesville, FL 32611 USA
acordar@cise.ufl.edu

Benjamin Lok
Computer and Information Science and Engineering
University of Florida
Gainesville, FL 32611 USA
lok@cise.ufl.edu

CHI'13, April 27 – May 2, 2013, Paris, France.

Abstract
Interpersonal skills training is a key concept in medical education. Unfortunately, components like empathy are difficult to measure. In many cases, experts analyze medical interviews for empathetic opportunities and measure the responses. Empathetic opportunities can also be used with conversational virtual humans to elicit empathetic responses. Measuring empathy during these interactions still requires experts.

Leveraging crowdsourcing, we were able to obtain empathy ratings from two groups: "health professionals" and "laypeople." We found the crowd agrees on high/low empathetic responses; however, frequently, empathy was perceived differently between groups. We propose analyzing histograms of ratings due to complexity in empathy perception.

Author Keywords
virtual humans, empathy, crowdsourcing

ACM Classification Keywords
H.5.3 [Group and Organization Interfaces]: Web-based interaction.

Introduction

Conversational virtual people have been shown to be useful in soft skills training [3]. Soft skills provide the ability for people to appropriately interact with others. Soft skills can be trained in a variety of ways; however, conversational virtual humans provide the advantage of consistent interactions in a safe environment. While virtual humans can provide a safe environment for soft skills training, there is still difficulty in how to evaluate some components of interpersonal communication. One component, empathy, is a complex concept. Empathy has many scales, each measuring empathy differently. These scales can be self-reported surveys [5] or require experts to analyze vides or transcripts [2].

There are many definitions of empathy; however, empathy is usually described as the ability to interpret and understand the emotions of another person.

Ideally, empathy would be rated algorithmically; however, to our knowledge, no algorithm exists. Empathy evaluation algorithms development is further complicated by the fact that empathy can be interpreted differently by different people. With so many interpretations, a large crowd of people interpreting should lead to a consensus. Using crowdsourcing, we conducted a study to obtain empathy ratings.

In this study, transcripts were used from health profession students interacting with virtual patients presenting with dysphagia (difficulty swallowing). Each interaction between a student and virtual patient contained one or more "empathetic opportunities" in which the virtual patent communicates to the student a fear or concern the patient would like addressed (Ex. Figure 1). Empathetic opportunities have been used in prior virtual human research, and were shown to elicit empathetic respnses from the participants [3]. Using empathetic opportunities, we obtained ratings on how empathetic the response was to the virtual patient. The empathy ratings came from two groups: the "layperson" group, which was composed of Amazon.com's Mechanical Turk works, and the "health professional" group, which was composed of doctors and nurses acquired through the Survey Monkey Target Audience service.

Student: When did the problem start?
Jackie: My problem started after my supraglottic laryngectomy.
Student: What medications are you on?
Jackie: Lozol, Crestor, CPAP, Nexium, and Allegra
Student: Does smoking make you feel worse?
Jackie: I feel like the smoking relaxes me.
Jackie: I am really worried about the swallowing problems being a symptom of my cancer coming back. I've heart that if you get cancer a second time it's almost always fatal. Do you think this is cancer?
Student: Is there any pain?

Figure 1: Example of Transcript with Empathetic Opportunity and Response

With this approach, we found that:

- There was a consensus in the layperson group for a majority of the ratings (29/36). These ratings included consensus for both high and low ratings. A third category of ratings (7/36) shows that the crowd cannot come to a consensus on every response; however, we attribute this to each individual's interpretation as to what is empathetic. There was consensus in the health professional group for 15/36 of the ratings.
- There is a significant difference in what defines "empathy" between both groups. While health professionals and laypeople agree on some high and low empathy responses, a third category of responses is viewed differently by each group. The health professional group perceived medically relevant responses as much more empathetic than the layperson group perceived the same responses. These difference reflect empirical evidence showing that doctors and patients have a significantly different perception of the doctors' communication skills [6]. With the assumption that the layperson group more likely represents the patient population, our results correlate with their findings.
- Based on the previous two points, mean analysis for empathy ratings is not enough in determining an appropriate measure. By analzying the ratings with

a histogram, a picture of empathy perception is formed. This approach is effective because the data indicates that empathy is in the "eye of the beholder."

Motivating Application

An important concept that must be taught during medical education is empathy. Being empathetic is listed by the Association of American Medical Colleges as a key learning objective [1]. Empathy can establish trust with a patient and improve patient satisfaction. Empathetic doctors have been shown to be associated with increased patient satisfaction and compliance [7]. Being more empathetic, according to Kim et al., is possibly one of the best ways to improve satisfaction and reduce the number of medical malpractice lawsuits [7].

Related Work

To rate empathy with crowdsourcing, we leveraged prior research on human computation, empathy rating research, and algorithmic approaches to empathy.

Human Computation

While computers are useful for performing complex tasks, there are some tasks in which computers do not perform well. These tasks can be solved with human computation, a relatively new area in which humans act as computers to solve problems. An example problem is the ESP Game [10], which generated accurate labels for images using a game with a purpose.

Since 2005, Mechanical Turk is a service offered by Amazon which allows "Requesters" to offer tasks for money in which "Workers" can sign up and complete these tasks. Tasks can include identifying the main object of an image, transcribing audio, or categorizing images. Mechanical Turk has provided an easy way to crowdsource tasks and has been used frequently in academic research.

Morris and Picard looked at using Mechanical Turk to generate cognitive reappraisals (interpreting an event with a different perspective) [9]. Part of this process included asking Mechanical Turk workers to rate how empathetic people were being with their reappraisals. These ratings were only acquired for determining the effects on a cognitive reappraisal when giving two groups two sets of instructions.

Expert-Rated Empathy

Since empathy is an important skill for medical students, it needs to be evaluated. Empathy rating has been done by experts who, in many cases, manually code videos/transcripts of medical interviews.

Bylund and Makoul created the Empathic Communication Coding System (ECCS) which, in addition to categorizing an empathetic response, also identifies patient created empathetic moments [2]. Empathetic moments are manually detected moments in which there is a clear statement of emotion, progress or challenge from the patient. From these moments, the ECCS is designed to help code the responses to the empathetic opportunities. ECCS requires significant time and discussion in analyzing videos for empathetic opportunities as well as coding the responses to those opportunities.

Background

Virtual Humans

Virtual humans are used in a wide variety of situations such as training rapport building or bedside manner (both of which require empathy). Virtual humans have frequently been used in the medical field.

Virtual humans can provide training for medical students' interactions with patients. While medical students can interact with standardized patients (trained actors), virtual patients are a supplement which provide some advantages that standardized patients are unable to simulate. For example, cranial nerve palsy is difficult for standardized patients to show; however, virtual humans have been created to simulate various cranial nerve palsies [8]. In addition to simulating a complex medical condition, this system helps medical students practice a differential diagnosis.

Empathy

Empathy has been defined in the context of patient care. Hojat defines empathy as "a predominantly cognitive (rather than emotional) attribute that involves an understanding (rather than feeling) of experience, concerns and perspectives of the patient combined with a capacity to communicate this understanding" [4].

An important note is that there is a distinction between sympathy and empathy. Sympathy is defined as having pity or sorrow for another person. Sympathy does not necessarily require someone to understand another's feelings but to just recognize that those feelings are present. For example, if someone is having a bad day because they failed an exam, a sympathetic response would be "I'm sorry you had a bad day." An empathetic response would be "I understand how tough that is. I remember how bad I felt when I failed a test." The reason this response is empathetic is because 1) the person states they understand it can be difficult to handle a failed test and 2) the person validates their understanding by remembering how bad they felt when they failed a test.

Study Conditions

The study had two conditions. The first condition involved obtaining empathy ratings from the layperson, composed of Mechanical Turk workers. The second condition involved obtaining empathy ratings from health professionals, composed of doctors and nurses recruited through Survey Monkey's Target Audience service.

Layperson Group

Amazon Mechanical Turk's tools were leveraged to create Human Intelligence Tasks (HITs). Each HIT was a subset of the transcript between a student and a virtual patient. Each HIT included instructions on how to correctly rate the empathy of the student's response. The workers were given definitions for both empathy and sympathy to distinguish the two concepts. They were instructed to only consider empathy when rating the responses, and workers were asked to rate their agreement with the statement: "I think this is an empathetic statement" with ratings on a 7-point Likert scale (1 - Strongly Disagree, 7 Strongly Agree). An example HIT can be seen in Figure 2.

A total of 131 HITs were posted on Mechanical Turk at $0.03 a HIT. 5400 ratings were contributed for all 131 HITs. There were between 39-45 ratings per HIT.

A qualification test was created to make sure the worker had a clear understanding of empathy, and understood the task. The test was also designed to help filter out abusers who were not legitimately completing the HITs (either not taking the HIT seriously, or intentionally providing wrong answers). The qualification test consisted of five questions formatted similar to an actual hit.

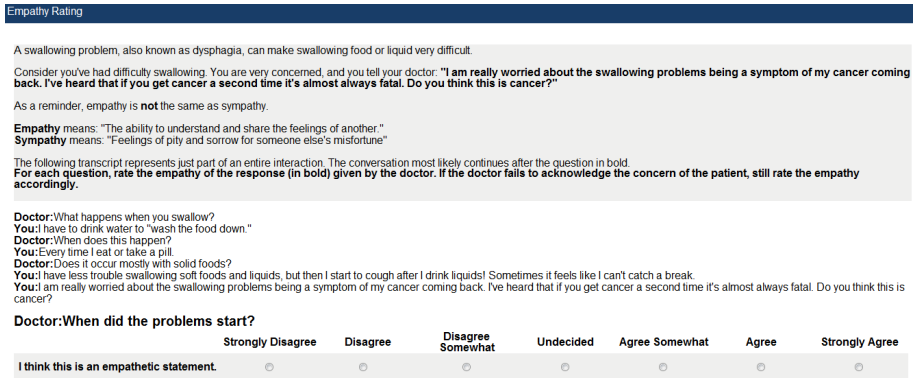

Figure 2: Example HIT

In order to pass the qualification test, the workers must have been able to correctly rate two control questions to Empathetic Opportunity 1 (Table 1). These responses were created on the ECCS scale of a high and low empathy response [2]. The other 3 questions of the qualification test were taken from the original 134 empathetic responses gathered in the transcripts.

Health Professional Group

To obtain a large number of health professional based empathy ratings, we used a service offered by Survey Monkey which provides specific target audiences to complete surveys. The specific target audience criteria chosen for this study were doctor and nurse.

49 surveys were completed. The survey consisted of the same instructions that were given to Mechanical Turk Workers; however, due to cost, consisted of only 36 questions that were also used on Mechanical Turk. The health professional group was not given a qualification test as both doctors and nurses should have recieved some form of empathy training during their education.

	Patient Name	Empathetic Opportunity
1	Jackie Dauer	I am really worried about the swallowing problems being a symptom of my cancer coming back. I've heard that if you get cancer a second time it's almost always fatal. Do you think this is cancer?
2	Vinny Devito	Can you tell me if I can ever start eating like before? I'm a food guy and love eating. It sucks that all that I've been able to eat is pureed food for the last year. I just want to be able to eat like before and be normal at our family get-togethers. I feel so out-of-place now that I can't eat what I want. Please tell me if I would ever get back to my original diet.
3	Vinny Devito	I am worried if I will ever be able to lead a normal life again.
4	Marty Graw	Doctor, imagine you being sick all the time. How would you feel about being sick and coughing while talking to your patients? My condition is the same. I am a chef but cannot even taste any of the food I'm cooking.
5	Marty Graw	This health problem could not have come at a more worse time. I'm already stressed about financial problems and my daughter's depression and now I have these issues as well. Can you give me at least some temporary relief so that I can handle my other issues first? Then I don't mind falling sick again.

Table 1: Empathetic Opportunities from Virtual Patients

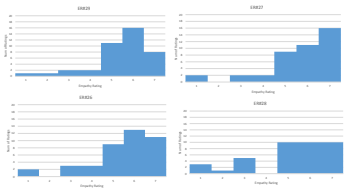

Figure 3: Example of Layperson Consensus on High Empathetic Responses

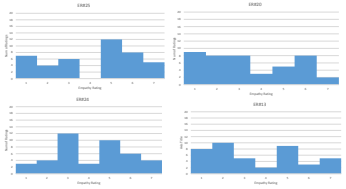

Figure 4: Example of Layperson Low Consensus

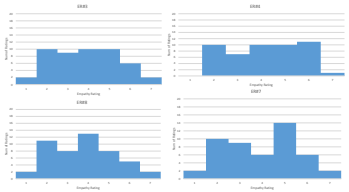

Figure 5: Example of Health Professional Low Consensus

Results - Rating Empathy

The ratings of each empathetic response were averaged and Mann-Whitley U-tests were performed on the empathy ratings. Histograms of ratings for each empathetic response were created to analyze the distribution of ratings per response. All analysis was performed on only the same 36 empathetic responses that both the health professionals and laypeople rated.

Based on the results, the layperson group was found to come to a consensus on a majority of the empathy ratings. Of the 36 empathetic responses, there was a consensus for 29 of these responses (both high (Figure 3) and low empathy). Of those 29, 7 were high empathy ratings and 22 were low empathy ratings. A third category (Figure 4) of ratings (7/36) emerge that reveals empathy can be interpreted differently by different people.

For the health professional group, there was less consensus. Of the 36 empathetic responses, there was only a consensus on 15 of the responses(both high and low empathy). Of those 15, 9 were high empathy ratings and 6 were low empathy ratings. Again, there was a third category (Figure 5) in which there was little consensus.

These categories were determined by analyzing the skewness values of the distributions. The three categories were created on the following criteria:

- High Consensus (High Empathy): Skewness values less than -0.5
- High Consensus (Low Empathy): Skewness values greater than 0.5
- Low Consensus: Skewness values between -0.5 and 0.5

Discussion

The first two categories are promising since a large crowd can come to a general consensus on how empathetic a response is. While there may be outliers present in the distributions, there are still clear trends to one side of the scale. These outliers suggest that some people have different standards as to what constitutes high or low empathy. These different standards are reflected even more so in the third category of empathy ratings.

This third category does not mean necessarily mean that the crowd is unable to rate empathy; rather, in some cases, perception of empathy can be different to many people. This category is valuable as it still provides insight on the perception of a response.

This "eye of the beholder" effect reinforces the need to analyze the distributions rather than the mean. The mean for the response "I cannot tell you if your cancer is coming back however, I can tell you about your swallow and let your physician know if anything looks abnormal" was 4.19 which would suggest that this response was viewed as relatively neutral and a safe response for a doctor to say. The histogram (top left of Figure 4) reveals the opposite: the response was not viewed as neutral. The response is actually risky. Depending on the person, the response may be interpreted as either empathetic or not empathetic.

Results - Empathy Perception

The results showed that out of 36 ratings, approximately 61% were statistically different between health professionals and laypeople (refer to Table 2 for a breakdown of consensus).

Discussion

There were certain categories of responses that caused the high number of statistical differences. One category of

responses was medical relevant responses. Another category was ambiguous or vague responses. Despite the 61% that were statistically different, 39% of the ratings were consistent between both groups. Again, there were certain categories of responses that led to these ratings being similar. These categories include responses that were easily identifiable as high empathy or low empathy responses. An example of a high empathy response is "I can't promise you anything, but we wil do everything we can to find out what's going on..." (Figure 6), and a low empathy response is "I think you should ask your primary doctor that question" (Figure 7).

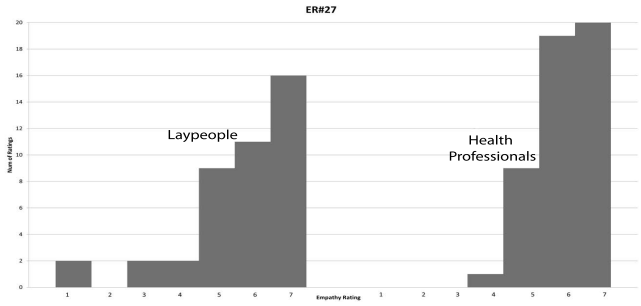

Figure 6: Distribution of Ratings for a High Empathy Response

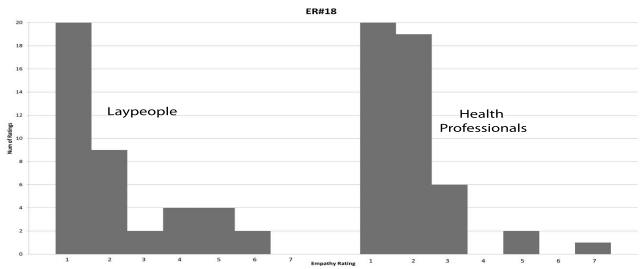

Figure 7: Distribution of Ratings for a Low Empathy Response

Ambiguous or Vague Responses

Medically relevant responses are responses that, although ignore the virtual patient's empathetic opportunity, ask a question relevant to the symptoms of the patient. For example, if a student responded to E0#2 with the question "When did the swallowing problems start?," laypeople view this as a low empathy response while health professionals view this as more empathetic (although no consensus). Figure 8 visually shows that the two groups have a different perception of empathy.

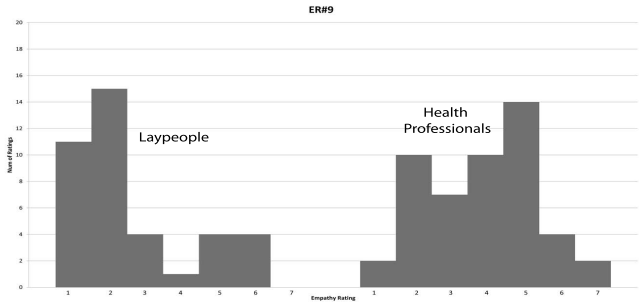

Figure 8: Distribution of Ratings for a Medically Relevant Response

Medically Relevant responses are the majority of the responses that resulted in statistically different ratings. The health professional group most likely views these responses as positive because as doctors/nurses, they ask medically relevant questions every day. While these questions might ignore the patient's concerns, doctors may view them as more empathetic because medically relevant questions help diagnose the patients. To the laypeople, medically relevant questions are not empathetic because the response fails to address the actual concerns of the patients. Clearly, the health professionals and laypeople have a different perception on empathy.

The different perception of empathy does not necessarily imply that either group is terrible at rating empathy; however, the differences reinforce that individuals have their own interpretation of what is and isn't empathetic. These differences are also useful to understand how people can perceive the same response differently. For medical students, understanding that there are different perceptions is useful to make them more aware of what they say to their patients.

Medically Relevant Responses
Ambiguous responses are responses which don't clearly identify if the doctor was trying to be empathetic. Some parts of their responses might seem empathetic, while other parts may not.

When looking at the ambiguous response "No I don't" (Figure 9), the ratings are inconsistent. This mixed rating reveals the complex nature of a person's view on what is empathetic. While the ambiguous response quite clearly states the doctor does not think the patient has cancer, this is not necessarily a satisfactory response in the eye of some raters.

In advocating the need to analyze empathy visually, looking at the mean empathy ratings for the response "No I don't", the mean is approximately the same in both groups (laypeople - 3.55, health professionals - 3.41, means were not significantly different). This would suggest that both groups think this is a slightly less empathetic response; however, looking at the rating distribution reveals that in the health professional group, 55% of ratings were on the low empathy side (< 4) and 27% were in the the high empathy side (> 4), and in the layperson group, 55% were on the low empathy side and 40% were on the high empathy side.

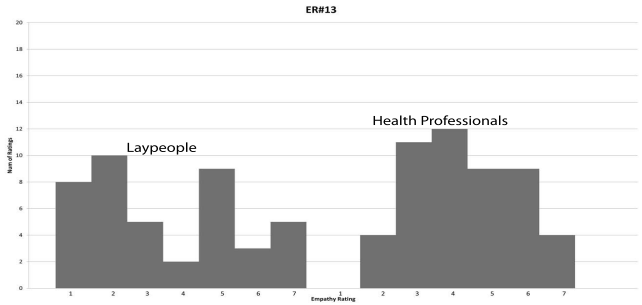

Figure 9: Distribution of Ratings for the response "No I don't"

General Discussion

The results reveal the difficult nature of subjectively rating empathy; however, by analyzing the histograms for every empathetic opportunity response, the response's reception by the rater becomes apparent.

Since we consider empathy rating as a histogram rather than a mean, a prediction can be made as to how a response will be received. These visualizations could be valuable in the classroom to give an idea as to what doctors should and shouldn't say in a medical interview. The distributions that show a mix of ratings are also valuable because the distribution reveals that while maybe 50% of people find the response to be highly empathetic, the other 50% may find it to have no empathy whatsoever. These distributions reveal responses that have "high risk" because the interpretation depends on the patient as to how well they will receive your response. Low empathy responses are also high risk responses. "Low risk" distributions are ones in which the ratings are on the high empathy side of the scale.

	Consensus	Consensus (High Empathy)	Consensus (Low Empathy)	No Consensus
Laypeople	29	7	22	7
Health Professionals	15	9	6	21
Between Groups	14	7	7	22

Table 2: Consensus Within and Between Groups

Limitations

While the results revealed that empathy can be rated by the crowd, these ratings are only in the context of one point in the conversation with a virtual patient. The student could have been empathetic for the overall conversation; however, raters were only asked to consider the response to the empathetic opportunity.

Detecting tone in the conversation is also difficult. While a student may have had good intentions with a response, lack of tone while reading a text-based conversation could lead to some misinterpretations.

Because the interaction was text-based, empathy could not be looked at in the context of non-verbal communication skills. In addition, prior research has shown that while medical students do respond empathetically to virtual patients, they do so with less frequency and quality [3]. This leads to data that includes responses that ignore the empathetic opportunity and move on to the next question. Our data contained many of these topic changes which results in ratings to responses that are off-topic and do not really address the empathetic opportunity.

Conclusion and Future Work

By using Mechanical Turk to obtain ratings for empathy, we found that, for the large majority of the virtual human interactions, the crowd can rate empathy. The health professional group was able to rate empathy; however, they had more ratings that were low consensus. There were three categories of ratings distributions:

- Crowd agrees on a high empathetic response
- Crowd agrees on a low empathetic response
- Crowd is undecided on whether a response is empathetic or not

This third category is interesting because although the crowd could not come to a consensus on every rating, the distributions suggest there are some responses that can be interpreted differently by everyone. This category reinforces that, in some cases, empathy is not black and white. This distribution also strengthens the argument to analyze the histograms rather than the means.

By analyzing the histograms of empathy ratings, we also found that the layperson and health professional groups, in many cases, have a different perception as to what defines empathy. The differences also correlate with previous literature on perception of doctors' communication skills [6].

Because empathy is complex, we found it effective to analyze the histograms rather than the means. The means do not necessarily reflect how a response was perceived. Some distributions have no general consensus; however, this lack of consensus is powerful since it indicates that some responses can be open to interpretation.

Empathy ratings could be shown to the medical student to see if there is an improvement in their empathy. This is a novel approach to improving a medical student's empathy because the students are given a rating from someone who has no affiliation with their education.

In addition to showing empathy ratings to the student, future work should go into investigating the effectiveness of leveraging the crowd to rate other communication skills in virtual human interactions. One example would be to obtain ratings on the ability of the student to clearly explain a diagnosis or procedure. Ratings would be useful to students, and again, the histograms of the ratings may reveal when an explanation was poor or unclear.

Acknowledgements

We thank members of the Virtual Experiences Research Group at the University of Florida for their help during the study as well as during the writing of this paper.

References

[1] Anderson, B. Learning objectives for medical student education–guidelines for medical schools: report I of the Medical School Objectives Project. *Academic medicine : journal of the Association of American Medical Colleges 74*, 1 (Jan. 1999), 13–8.

[2] Bylund, C. L., and Makoul, G. Empathic communication and gender in the physician-patient encounter. *Patient education and counseling 48*, 3 (Dec. 2002), 207–16.

[3] Deladisma, A. M., Cohen, M., Stevens, A., Wagner, P., Lok, B., Bernard, T., Oxendine, C., Schumacher, L., Johnsen, K., Dickerson, R., Raij, A., Wells, R., Duerson, M., Harper, J. G., and Lind, D. S. Do medical students respond empathetically to a virtual patient? *The American Journal of Surgery 193*, 6 (June 2007), 756–760.

[4] Hojat, M. A Definition and Key Features of Empathy in Patient Care. In *Empathy in Patient Care*. 2007, 77–85.

[5] Hojat, M., Mangione, S., Nasca, T. J., Cohen, M. J. M., Gonnella, J. S., Erdmann, J. B., Veloski, J., and Magee, M. The Jefferson Scale of Physician Empathy: Development and Preliminary Psychometric Data. *Educational and Psychological Measurement 61*, 2 (Apr. 2001), 349–365.

[6] Kenny, D. a., Veldhuijzen, W., Weijden, T. V. D., Leblanc, A., Lockyer, J., Légaré, F., and Campbell, C. Interpersonal perception in the context of doctor-patient relationships: a dyadic analysis of doctor-patient communication. *Social science & medicine (1982) 70*, 5 (Mar. 2010), 763–8.

[7] Kim, S. S., Kaplowitz, S., and Johnston, M. V. The Effects of Physician Empathy on Patient Satisfaction and Compliance. *Evaluation & the health professions 27*, 3 (Sept. 2004), 237–51.

[8] Kotranza, A., Cendan, J. C., Johnsen, K., and Lok, B. Virtual Patient with Cranial Nerve Injury Augments Physician-Learner Concern for Patient Safety. *Journal of Bio-algorithms and Med-systems 6*, 11 (2010), 25–34.

[9] Morris, R. R., and Picard, R. Crowdsourcing Collective Emotional Intelligence. *Proc. of Collective Intelligence* (Apr. 2012).

[10] von Ahn, L., and Dabbish, L. Labeling images with a computer game. *Proc. of the 2004 conference on Human factors in computing systems - CHI '04* (2004), 319–326.