J. EDUCATIONAL COMPUTING RESEARCH, Vol. 11(4) 339-346, 1994

PSYCHOLOGICAL ASSESSMENT USING SIMULATIONS WITH UNRESTRICTED NATURAL LANGUAGE INPUT

MICHAEL ANBAR
MICHAEL RAULIN
State University of New York, Buffalo

ABSTRACT

Five computerized role-playing scenarios, which accept unrestricted natural language input, were developed and administered to seventy-two freshman medical students. The scenarios, written in CASIP, measured and automatically scored each response on five psychological dimensions: Social skills, level of frustration, submissiveness, combativeness, and negotiative ability. The programmed scenarios also monitored nonverbal dimensions, which may reflect the emotional state of the testee. These included: The time it took to start an answer; the time spent reviewing the answer; the lengths of answers and of the words used. The testees behaved significantly different in handling the different role-playing scenarios. While no significant correlations were found between the psychological dimensions expressed in the different scenarios, the tests identified individual testees who displayed a pattern of extremes of psychological behavior.

INTRODUCTION

The personalities of physicians probably contributes significantly to their success as health care providers. Possessing maladaptive personality traits, such as combativeness, excessive selfishness, or ineffective communication skills, predicts poor professional performance in clinical practice in many professions [1]. However, the role of personality in predicting behavior and the best procedure to measure personality have been debated extensively in psychology [2, 3]. Personality traits of candidates for medical school are currently evaluated in brief personal interviews plus "reading between the lines" of letters of recommendation. Extensive psychological testing, including interviews with trained psychologists, are too costly to be applied to the hundreds of candidates selected for personal interview by medical schools. Personality assessment during brief personal interviews, generally by untrained interviewers, has severe shortcomings

[4]. In addition to the relative ease of deceiving untrained interviewers, these shortcomings include potential interviewer biases and lack of standardization [1]. Furthermore, an interview assessment of a candidate often correlates poorly with a real-life behavioral assessment [3]. Some of the shortcomings of conventional interviews might be overcome by computerized psychological tests. Although computerized tests cannot evaluate poise, demeanor, composure, and savoir-faire, which usually can be assessed in an interview, they can evaluate many personality traits that are hard to assess in a casual interview and thus minimize the acceptance of unsuitable medical school candidates [5].

For several years, we have been developing computerized psychological assessment tools based on role-playing in a simulated environment [6, 7]. Such role-playing involves verbal interaction with simulated persons who challenge the testee in different situations. The testee is asked to play different roles, ranging from a figure of authority such as a student counselor, to a citizen intimidated by a rude law enforcement officer. These computerized simulations involve an unrestricted natural language interaction, achieved by using CASIP as the authoring tool [8-10]. CASIP parses the testee's input, recognizing *expected* answers by the presence of key words or their contextual synonyms in specified positions in the sentence. The computer's response depends on all the interactions that took place from the start of the session, giving the testee the impression of a dialogue with a live person.

CASIP-authored programs yield a verbatim record of the man-machine dialogue, including several measures of conduct in giving each answer. These include the time it takes a testee to begin to answer and the time spent to review the answer before entering it and waiting for the machine's response. It also measures the number of backstrokes, which may indicate, in addition to the tendency to make typographic errors, a tendency to revise an answer that might be considered unsatisfactory as it was originally typed. CASIP allows the automatic scoring of each of the testee's responses on up to eight different dimensions. These scoring parameters provide a combination of scores that characterizes the testee by a multidimensional behavioral profile. Thus, CASIP's automatic scoring overcomes the major objections to personality assessment by interviews, namely personal bias, lack of norms, and excessive use of professional time.

We developed a scoring manual that scores each potential answer of the testee on a five-point scale for each of five psychological dimensions [11]. The scoring manual gives conceptual criteria for each rating, with several examples to help raters. Interrater reliability of this system ranged from .64 to .89 elevated on pilot data used to develop the scenario programming and the scoring manual [11]. Some scoring dimensions are averaged (e.g., Social Skills), while in others we look for extremes (e.g., combativeness). We look for extremes on dimensions where a single extreme response may indicate critical flaws in testee's personality or control. To validate these computerized tests, we intend to test the entire freshman class in our medical school from the next several years, follow up the

testees through residency, and find to what extent these tests predict professional behavior. The medical school maintains routine evaluations of students' performance in their clinical placements and residency. These records will be our primary data source to evaluate the professional behavior of the students. We are particularly interested in being able to detect the student who shows chronic problems in dealing with the demands of a clinical setting, and therefore is at high risk for dropping out or flunking out of medical training after years of investment. This vast effort is worthwhile if it significantly improves screening of candidates for medical school.

In this article, we describe the findings of our initial evaluation of seventy-two freshman medical students conducted during orientation to medical school. They were randomly selected from a subset of 120 eligible students out of a class of 135. Fifteen students opted not to participate in the study. The objectives of this study included: determining if such a battery of psychological tests can be administered to a large population in a single setting; determining the level of the program's recognition of subject responses in a large, heterogenous population; determining correlation between personality attributes scored in the same scenario, as well as between different scenarios; determining to what extent the different scenarios, which invoke different emotional involvements, spur different modes of behavior of the testees (such as the average length of verbal statements, average length of words used in those statements, the average time of thinking before an answer is given, or the average time spend on reviewing each answer before hitting the *ENTER* key).

TEST SETTINGS AND THE TESTING SCENARIOS

We used the five role-playing scenarios described below. The testing was done on 286 and 386 microcomputers. Each scenario was stopped after fifteen dialogue cycles or after eight minutes, whichever came first. Some scenarios could be terminated earlier if the testee responded in a manner that would normally terminate a similar interaction in real life. These tests were automatically scored, and the results were statistically analyzed to yield the results reported below. The automatic scoring was based on applying the criteria of the personality scoring manual developed for this project [11] to every possible answer recognized by the program. Two raters independently assigned scores on each psychological dimension to each possible answer that was recognized by the programs. Disagreements between raters were resolved through discussion. There were few disagreements on what scores to assign to a particular answer, and none of the disagreements were more than one point on a five-point scale. The five scenarios were pretested on over 135 medical student candidates and undergraduate psychology students to refine the programs and develop the personality scoring features. Those preliminary tests were manually scored to develop the automatic scoring that was used in the full scale test reported here.

The battery of five testing scenarios took up to forty-five minutes to complete. The first scenario required the testee to assume a somewhat higher social status than the computer-emulated person. Here is the introduction to that scenario: "*You are a student peer counselor. Your role is to advise students on subjects concerning their health and welfare. John, a student, is sent to you by the resident advisor because several students have been complaining about his smoking. What will you say to John?*" Testees have handled this situation using several different strategies, ranging from authoritative to friendly. Some subjects try to convince John to quit smoking; others discuss the right to avoid his secondhand smoke. John's behavior was programmed to be inconsistent—oscillating between militancy and compliance—making his handling difficult and frustrating. There is a twist when John mentions, if appropriately interrogated, that the complaint of his mates against him has an ulterior motive—he thinks he is hated because he is a better student and is more popular with girls than his peers. This scenario probes primarily social skills and negotiative power, while frustration might be exhibited here only in individuals with an exceptionally high level of impatience and lack of social skills.

The second scenario puts the testee in a confrontational position with a petty bureaucrat on campus. This is how it opens: "*You try to check out from the library a book that is essential for a term paper due tomorrow, and the librarian will not let you. She insists that you have an overdue fine of sixteen dollars and twenty-five cents because a previous book was returned late. You know that you returned that book on time and that the library is at fault. This library does not issue receipts when books are returned and paying the fine is regarded as a final settlement; usually there is no way to recover a fine once paid. Books in this library are not stamped with dates of check out and check in. What will you say to the librarian at this point?*" The librarian is consistently authoritative and non-yielding to a level that evokes frustration. This scenario probes self-confidence and persistence without over-combativeness. Testees were found to handle this situation in different ways. These include paying the fine right away, seeking help from a higher authority, pleading for understanding, suggesting different possibilities that might have lead to the unjust fine, inventing fake witnesses and receipts, and even becoming abusively combative. We test here negotiative skills in a frustrating situation.

The third scenario puts the testee in an equal status to the emulated person under conditions that may call for suspicion and aggression: "*You are living in a dormitory. You just noticed that your wallet containing your monthly allowance, credit card, and driver's license is missing. There is a young man named Bob in the room. He is a high-school classmate of Pat, your roommate. Bob arrived just yesterday. Pat is going to be in class for the rest of the afternoon. There is a phone on the desk, so you may call Campus Security. What will you say now to Bob?*" This scenario ends with a twist: "*It is 7PM. Pat returns at last and says, while in the doorway: "I found your wallet in the hallway near the elevator. Since I was*

rushing to class I could not get back and tell you. Here it is. You must have been worried. Weren't you? Well let's go and have dinner. By the way, where is Bob?" Like in the former scenarios, testees have handled this situation in a variety of ways. Their strategies range from immediately accusing Bob, who then leaves indignantly, to "beating around the bush" trying to trap Bob as the culprit; some testees call the police, while others ask for Bob's help in finding the missing wallet or borrow money from Bob. The social skills called for in this scenario are very different from those dealing with the bureaucratic librarian, and Bob's behavior is not intended to invoke frustration unless one is unusually egocentric. Becoming combative in this situation is again not uncommon and might indicate some undesirable personality traits.

In the fourth scenario the computer emulates an authoritative bully: "*Your buy a pack of pencils in a local drug store. Just as you are leaving the store, a security guard stops you. He accuses you of stealing a pack of gum. The gum was at your feet, just as you were leaving the checkout counter. It must have fallen out of someone else's bag, but to the guard it appears that it fell out of your bag. What will you say now to the guard in your defense?*" Like in the library scenario, the testee is innocent, but this offense is more serious and the consequences of conceding are much more severe. The guard is assertive and stubborn and cannot be talked out of his accusation. Strategies used by testees in this situation range from aggressive defiance to submissive denial. This scenario often triggers combative behavior, though alternatives of calling a lawyer or the police do exist. The choice of arguments can be used as a measure of negotiative skills.

The fifth and last scenario puts the testee in a situation where the computer emulates an irrational person of equal social status: "*You are at a Delta Tau Delta fraternity party, and one of the brothers, with the smell of beer on his breath, grabs you and says, "Hey... That is my shirt! Give me it! Right Now!" He is very insistent that you have his shirt. Surely you know he is wrong. You are separated from your friends and have no extra clothing. How will you talk your way out of such a situation?*" This again is a frustrating situation, which may have either aggressive or submissive solutions, although certain delay or distraction tactics may also work. This scenario also invokes different reactions in male and female testees; not surprisingly, females feel more threatened. However, only a minority of female testees disclose their gender in the dialogue (e.g., "I am a girl . . . " The program is sensitive to gender differences if the testee discloses it.

PARAMETERS STUDIED

We scored five scenarios (Counselor, Librarian, Lost Wallet, Guard, and Party) on five psychological dimensions (social skills, frustration, submissiveness, combativeness, and negotiative skill). To normalize for variation in the number of answers in a session, we divided the obtained score on each dimension in a

scenario by the number of verbal interactions in that session. We also computed an average across scenarios for each psychological dimension.

We also took advantage of the power of the computer to monitor other *process dimensions*, which may prove to have psychological significance. These included Session Time (time to complete a scenario), Think Time (time from computer's response until the testee starts to respond), Reread Time (time between typing the last character and hitting ***ENTER***), Typing Speed (number of characters typed divided by the time it took to type them), Number of Characters, Number of Backstrokes, Length of Answer (number of words in answer), and Average Length of Words. We computed several composite measures from these process dimensions. We included each of the process dimensions in the statistical analysis. In the following discussion we will refer to the testees' input as *answers* and to the computer's output as *responses*.

RESULTS AND DISCUSSION

The program counts all the answers that were recognized and scored, and compares them with the total number of answers given. The average recognition rate of all sessions was 94 percent, reaching 98 percent in *Librarian* and *Counselor* sessions. This satisfactory level of recognition will be improved in the future by making the program recognize unrecognized or misrecognized answers picked up in this study.

We have statistically analyzed and correlated more than fifty parameters as listed above. Presenting a fifty by fifty correlation matrix is impossible within the space constraints of this article. We found strong correlations between the psychological dimensions we measured within sessions, but with few exceptions, no significant correlations between dimensions across scenarios. For instance, there was a strong positive correlation between *frustration* and *combativeness* and a strong negative correlation between those two dimensions and *social skills* within most of the scenarios. However, each of the psychological dimensions were virtually uncorrelated across dimensions, suggesting that each scenario was tapping different aspects of behavior, and trait-like behavior was not observed across dimensions. We had not expected this finding, although Mischel and his colleagues [12, 13] have argued that this phenomenon is common in studies of human behavior across situations. Moreover, this indicates that the testees get emotionally immersed in each of the simulations to an extent that overshadows behavioral biases from the previous scenarios. This significant change of behavior in different scenarios was corroborated in the process dimensions as well. While there were no significant differences between sessions on typing speed, significant differences in *Think* and *Reread* times were found. The average length of answers and average word lengths also showed significant differences between scenarios. In other words, the tested subjects handled the challenges of the different scenarios differently not only in expressive language but in non-verbal behavior as

well. The importance of these findings, which indicate that computerized simulations can be effective probes of human behavior, cannot be overemphasized.

The absence of significant statistical inter-scenario correlations, which we might have been predicted in a naive model, suggests that characteristics shown in a certain situation may not necessarily come into play in a different scenario. However, some very interesting results are uncovered when we examined the individual performance of subjects. We prepared frequency distributions on each dimension measured (both psychological and process dimensions) in each of the five scenarios. We then identified the two to three outliers in each of these distributions. We found tremendous clustering of outliers (i.e., a few subjects were consistently outliers on many dimensions). For instance, testee #70 showed extreme submissiveness in *Librarian* and in *Counselor*, while showing exceptional low submissiveness in *Guard*, minimal social skills, negotiative power and combativeness combined with an exceptionally high level of frustration in *Librarian*; Tested #51 showed excessive combativeness as a result of extreme behavior in *Party*, but unusually low combativeness in *Counselor*; the same testee also showed high frustration level in *Party*, but an extremely low level of frustration in *Guard* and *Counselor*, exceptional negotiative skills in *Counselor*, and minimal submissiveness in *Guard. Testee #9* showed extraordinarily low level of combativeness in the *Counselor, Guard* and *Librarian* scenarios, and very low frustration level in *Counselor*; the same testee showed extremely high submissiveness in *Librarian* but very low one in *Guard*. Testee #13 showed extremely low combativeness in *Counselor* and in *Librarian*, submissiveness in *Librarian* and in *Party* and extreme frustration in *Librarian*.

These examples suggest that there may be characteristic *patterns* or *profiles* of behavior of individuals when exposed to specific situations, rather than a consistency of trait scores in different scenarios. Fifty-nine of the seventy-one subjects showed no deviant scores in this analysis. When a subject did show a deviant score, they typically showed several deviant scores. The mean number of deviant scores for the thirteen subjects who were deviant at least once was 5.77 (range 3 to 9). Such patterns of behavior in a set of computerized simulations may well be correlated with their real-life behavior. It will take several years to follow-up the testees to establish the relationship between tacking the situations in the role playing scenarios, on one hand, and real life behavior on the other. Still we believe that such an effort will be worthwhile.

REFERENCES

1. N. D. Sundberg, *Assessment of Persons*, Prentice-Hall, Englewood Cliffs, New Jersey, 1977.
2. K. D. Lanning, *Consistency, Scalability, and Personality Assessment*, Springer-Verlag, New York, 1991.
3. W. Mischel, *Personality and Assessment*, Wiley, New York, 1968.

4. K. P. Morganstern, Behavioral Interviewing, in *Behavioral Assessment: A Practical Handbook*, A. S. Bellack and M. Hersen (eds.), Pergamon, New York, 1988.
5. J. N. Butcher (ed.), *Computerized Psychological Assessment: A Practitioner's Guide*, Basic Books, New York, 1987.
6. M. Anbar, *Using CASIP to Assess Aptitudes of Medical Students and of Applicants to Medical School*, Proc. 13th Annual Symposium on Computer Applications in Medical Care, Washington, D.C., pp. 924-927, 1989.
7. M. Anbar, A. Anbar, and M. Raulin, *Natural Language Driven Tests to Assess Knowledge, Personality and Decision Making Ability*, Proc. of the AMIA 1st Annual Educational and Research Conf., Snowbird, p. 49, 1990.
8. M. Anbar, CAI Computer Assisted Instruction: A Way to Avoid the Pitfalls of Multiple-Choice Behavior in Medical Practice, *Medical Electronics, 18*:2, pp. 118-124, 1987.
9. M. Anbar, Computerized Instruction and Testing Based on Conversational Natural Language Input, *IEEE Engineering in Medicine and Biology Magazine, 11*:1, pp. 57-61, 1992.
10. M. Anbar, Use of Natural Language in Interactive Testing of Competence in Medicine, *Journal of Medical Education Technologies, 2*:2, pp. 7-22, 1992.
11. M. L. Raulin and M. Anbar, *Development of an Artificial Intelligence Based on Role-play System for Psychological Assessment*, Poster presented at the Eastern Psychological Association Convention, Boston, April 1992.
12. W. Mischel, Alternatives in the Pursuit of the Predictability and Consistency of Persons: Stable Data that Yield Unstable Interpretations, *Journal of Personality, 51*, pp. 578-604, 1983.
13. W. Mischel, Convergences and Challenges in the Search for Consistency, *American Psychologist, 39*, pp. 351-364.

Direct reprint requests to:

Dr. Michael Anbar
Department of Biophysical Sciences
120 Cary Hall
SUNY School of Medicine
and Biomedical Sciences
Buffalo, NY 14214