Psychological Assessment Using Simulations with Unrestricted Natural Language Input

Michael Anbar and Michael Raulin

Department of Biophysical Sciences, School of Medicine & Biomedical Sciences, and Department of Psychology, Faculty of Social Sciences, State Univ. of New York, Buffalo, NY.

In an attempt to develop a computerized methodology to screen medical school candidates for undesirable personality traits, a battery of five computerized role-playing psychological tests that accept unrestricted natural language input were developed and administered to a group of 72 Freshman medical students. The tests measured five psychological traits: Social skills, level of frustration, submissiveness, combativeness, and negotiative ability. The tests, written in CASIP, automatically scored each response given on those five psychological dimensions. In addition, the tests monitored a series of non-verbal parameters, such as the time it took to start an answer, the time spent reviewing the answer, the length of answers and of the words used, all of which may reflect the emotional state of the testee. It was found that the testees behaved significantly differently in handling the different role-playing scenarios. While no significant correlations were found between the psychological traits expressed in the different scenarios, the tests identified individual testees who displayed a pattern of extremes of psychological behavior.

1. Introduction

The personality of physicians is probably the most important factor in their success as providers of adequate health care. Possessing maladaptive personality traits, such as combativeness, excessive selfishness or ineffective communication skills, predict poor professional performance in clinical practice. 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 tests, 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. In addition to the relative ease of deceiving untrained interviewers, these shortcomings include bias by the personality match of the interviewer and interviewee, lack of standardization, and the lack of opportunity to assess the candidate's behavior in a real-life-like environment, when the interviewee is less self-conscious and relatively free of stress. 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 can be assessed only in a person-to-person interaction, they can evaluate many personality traits that are hard to assess in a casual interview and thus minimize the acceptance of misfits.

For several years, we have been developing computerized psychological assessment tools based on role-playing in a simulated environment.[1] Such role-playing involves verbal interaction with simulated persons who challenge the testee in different situations. In such a test the testee may 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[2] as the authoring tool. CASIP parses the testee's input, recognizing *expected* answers by the presence of keywords 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 dialog with a live person.

CASIP-authored programs yield a verbatim record of the man-machine dialog, including several measures of conduct in giving each answer. These include the time it takes a testee to come up with 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 indicate, in addition to correcting of typographic errors, a tendency to revise an answer that might be considered unsatisfactory as it stands. 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, often in conjunction with any of the answers previously given. Certain scoring dimensions are averaged, while in others we look for extremes, since a single critical answer, or a sequence of two answers, can unambiguously reveal a certain attribute of the testee's personality. To validate these computerized tests, we intend to test the entire freshman class in our medical school several years in a row, follow up the testees through residency, and find to what extent these tests predict professional behavior. This vast effort is worthwhile if it significantly improves screening of candidates for medical school.

In this paper we describe the findings of a large pilot experiment in which we tested in a *single* session 72 Freshman medical students on the day of orientation. The were randomly selected from a subset of 120 eligible students out of a class of 135. The objectives of this study include: Determine 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 spent on reviewing each answer before hitting the <ENTER> key).

2. Test settings and the testing scenarios

In this personality test we used five different role-playing scenarios that will be described. The tests were carried out on 72 microcomputers, 286 and 386 IBM-AT's or IBM-AT clones. The test was administered through a batch file. Each scenario was stopped after 15 dialog cycles or after 8 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 5 scenarios were pretested on more than 70 medical student candidates, more than fifty undergraduate psychology students and subsequently on 15 Freshmen medical students from the same class of 135. These 15 students participated in a summer preparatory course and were then excluded from the large scale test. Those preliminary tests were manually scored to develop the automatic scoring that was used in the full scale test reported here.

The battery of 5 testing scenarios took up to 45 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, which, by the way, is not called for in this case. Other testees discuss the rights of others to avoid his secondhand smoke. John has an inconsistent personality that oscillates between militancy and compliance, making his handling intentionally 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 led to the unjust fine, inventing fake witnesses and receipts, and even becoming abusively combative. We test here negotiative skills in an callous frustrating framework.

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 common and might indicate some undesirable personality traits.

In the fourth scenario the computer emulates an authoritative bully: "*You 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 tests the point where the testee breaks down and becomes combative, though alternatives of calling a lawyer or the police do exist. The choice of arguments can be used as a measure of negotiative skills, while admission of guilt may point to a very strange personality indeed.

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; females feel more threatened. However, only a minority of female testees disclose their gender in the dialog, e.g., "I am a girl..."; the program is sensitive to gender differences if the testee discloses it.

In this pilot study we examined the differences in human behavior expressed in this diversity of scenarios and found that each of them measures *independently* different attributes of the testees.

3. Parameters Studied

The scenarios, or sessions, used in this study were: Counselor (***CR***), Librarian (***LB***), Lost Wallet (***LW***), Guard (***GD***) and Party (***PY***). The scores of each answer were summed for each session to give the following cumulative session scores for social skills (***SS***), frustration (***FR***), submissiveness (***SB***), combativeness (***CM***), and negotiative power (***NG***). The parameters monitored were, therefore, ***SSCR*** = social skill in the Counselor session, ***FRLW*** = frustration score in the Lost Wallet session, ***CMLB*** = combativeness in the Librarian session, ***SBPY*** = submissiveness in the party session, etc. To normalize for variation in the number of answers in a session we used in the statistical analysis the *average* scores (cumulative scores of a session divided by the number of verbal interactions in that session). For instance, ***ACMLB*** = ***ΣCMLB*** / (Number of answers in the Librarian session). In our evaluations, we also used the average score of a trait in all five scenarios, e.g., ***AVSS*** = ***(ASSLB*** + ***ASSCR*** + ***ASSGD*** + ***ASSPY*** + ***ASSLW***)/5.

Other parameters of interest that were measured or calculated are: Session Time (***ST***) = the time it took a testee to complete an interactive scenario; Think Time (***TT***) = the time it takes from the appearance of the computer's response on the screen until the testee starts to type the answer; Fraction of Think Time (***FTT***) = ***ΣTT/ST***; Reread Time (***RR***) = the time elapsed between typing the last character of the answer and hitting the <ENTER> key; Fraction of Reread Time (***FRR***) = ***ΣRR/ST***; Typing Speed (***TYS***) = the number of characters in the answer divided by the time it took to type them; number of characters (***CH***) and number of backstrokes (***BS***) from which the Backstroke ratio (***BSR*** = ***BS/CH***) was derived; Length of Answer (***AANL***) = number of words in answer; Length of Words (***AWDL***) = the average number of characters per word in an answer. Each of the latter parameters was evaluated for each session and used in the statistical analysis. For instance, ***FRRPY*** is the fractional reread time in the Party session, ***AWDLLB*** is the average word length in the Librarian session, etc. In the following discussion we will refer to the testees' input as *answers* and to the computer's output as *responses.*

4. 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 94.1 ± 6.8% reaching 98% 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 50 parameters as listed above. Presenting a 50 by 50 correlation matrix is impossible within the space constraints of this paper. Table 1 presents, therefore, a small sample of the correlation matrix and additional findings are presented in a narrative form. Table 1 presents correlation coefficients (their p values are in *italics*) between the scores of four personality traits of testees within three of the scenarios.

As shown in Table 1, there are strong positive or negative correlations between the personality attributes within each session but, with a few exception, there are no significant correlations between the different attributes scored in different scenarios. For instance, there is a highly significant correlation between *frustration* and *combativeness* while a highly significant negative correlation is seen between those two attributes and *social skills.* The lack of significant correlation between the same attributes exhibited in different scenarios suggests that the testee's responses depend on the specific situation encountered. Moreover, this indicates that the testees get emotionally immersed in each of the simulations to an extent that overshadows any significant behavioral bias from the previous experience that occurred just a few minutes before. This significant change of behavior in different scenarios was corroborated also in the non-verbal parameters. While there were no significant differences between ***TYS*** of individuals in the 5 sessions, there were significant differences in *Think* and *Reread* times: ***FTTPY*** > ***FTTLW*** > ***FTTLB*** , ***FRRLB*** > ***FRRLW***, ***FRRGD*** > ***FRRPY***, ***FRRCR*** > ***FRRPY*** and ***FRRGD*** > ***FRRLW*** were statistically significant. The average answer and word lengths also showed significant differences: ***AANLLB*** > ***AANLPY***, ***AANLLB*** > ***AANLLW***, ***AANLGD*** > ***AANLPY***, ***AANLCR*** > ***AANLLW*** > ***AANLPY*** and ***AWDLLB*** > ***AWDLPY***, ***AWDLLB*** > ***AWDLLW***, ***AWDLCR*** > ***AWDLPY***, ***AWDLCR*** > ***AWDLLW***. The importance of these findings, which indicate that computerized simulations can be effective probes of human behavior, cannot be overemphasized.

Table 1
A sample of the correlation matrix of scores of psychological attributes

	ASSCR	***AFRCR***	***ASBCR***	***ACMCR***	***ASSGD***	***AFRGD***	***ASBGD***	***ACMGD***	***ASSPY***	***ACMPY***
ASSCR	1.0000	-.6767	-.5010	-.6099	-.1287	.0145	-.0176	.1528	-.0270	.0190
	.0000	*.0000*	*.0000*	*.0000*	*.2848*	*.9045*	*.1202*	*.2649*	*.8218*	*.8743*
AFRCR	-.6767	1.0000	.4217	.6930	.1551	-.0816	.1861	-.1341	.0052	-.0391
	.0000	*.0000*	*.0002*	*.0000*	*.1966*	*.4985*	*.1202*	*.2649*	*.9656*	*.7445*
ASBCR	-.5014	.4217	1.0000	.4988	.2120	-.0541	-.0838	-.1318	.1843	-.0426
	.0000	*.0002*	*.0000*	*.0000*	*.0759*	*.6543*	*.4872*	*.2734*	*.1211*	*.7224*
ACMCR	-.6099	.6930	.4988	1.0000	.2598	-.0447	.1076	-.1385	.0122	.0044
	.0000	*.0000*	*.0000*	*.0000*	*.0287*	*.7113*	*.3716*	*.2493*	*.9189*	*.9708*
ASSGD	-.1287	.1551	.2120	.2598	1.0000	-.4600	-.0269	-.3266	.2081	-.0750
	.2848	*.1966*	*.0759*	*.0287*	*.0000*	*.0001*	*.8235*	*.0054*	*.0816*	*.5344*
AFRGD	.0145	-.0164	-.0541	-.0447	-.4600	1.0000	-.0033	.1980	.0021	-.1265
	.9045	*.4985*	*.6543*	*.7113*	*.0001*	*.0000*	*.9784*	*.0979*	*.9860*	*.2930*
ASBGD	-.0176	.1861	-.0838	.1076	-.0269	-.0033	1.0000	-.0121	-.0660	-.0145
	.8844	*.1202*	*.4872*	*.3716*	*.8235*	*.9784*	*.0000*	*.9199*	*.5843*	*.9048*
ACMGD	.1528	-.1341	-.1318	-.1385	-.3266	.1980	-.0121	1.0000	-.1675	.1282
	.2034	*.2649*	*.2734*	*.2493*	*.0054*	*.0979*	*.9199*	*.0000*	*.1628*	*.2868*
ASSPY	-.0270	.0052	.1843	.0122	.2081	.0021	-.0660	-.1675	1.0000	-.8149
	.8218	*.9656*	*.1211*	*.9189*	*.0816*	*.9860*	*.5843*	*.1628*	*.0000*	*.0000*
ACMPY	.0190	-.0391	-0426	.0044	-.0750	-.1265	-.0145	.1282	-.8149	1.0000
	.8743	*.7445*	*.7224*	*.9708*	*.5344*	*.2930*	*.9048*	*.2868*	*.0000*	*.0000*

The absence of significant statistical inter session correlation, which we might have been predicted in a naive model, suggests that a trait revealed in a certain situation may not necessarily come into play in a different scenario. On the other hand, when we examined individually the scores on the five dimensions in each of the scenarios as well as the average score of each trait, e.g., ***AVSS*** or ***AVCM***, we identified individual testees with consistent ***extreme*** behavior characteristics. 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***; Testee #51 showed excessive combativeness in ***AVCM*** 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***, extreme 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. Such patterns of behavior in a set of computerized simulations must be correlated with the real-life behavior of the same individuals. It will take thousands of tests and several years of follow-up of the testees to establish reliable correlations between tackling 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.

[1] Anbar M., 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, DC; pp. 924-927; 1989.
Anbar M., Anbar A. and Raulin M. 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.

[2] Anbar M. CAI Computer Assisted Instruction: A way to avoid the pitfalls of multiple-choice behavior in medical practice. *Medical Electronics,* 1987, 18(2):118-124.