# Reliability of Trainees' Endorsements on Standardized Psychiatric Interviews

Thomas E. Uttley, M.D.

This study investigated the utilization of standardized psychiatric interviews (SPI's) in psychiatric training programs. As the field of psychiatry, and the training of its' new members moves to conform its' principles to scientific models, the reliable use of SPI's is useful in reaching that goal, in both research and diagnostic applications. This investigation contains two studies. In Study 1, a random survey of 20 per cent of all psychiatric residency training programs was conducted to determine the prevailing level of training devoted to SPI's. The resulting findings are referred to as "training as usual" (TAU). Study 1 shows that residents are not sufficiently trained in the use of SPI's insomuch as more than 85 per cent of training programs offered no training in their administration. Study 2 tested residents' inter-judge reliabity upon administration of the Psychiatric Status Schedule (PSS) both before and after they received an intensive training intervention. The purpose of the training intervention was to increase the skills necessary for residents to improve their inter-judge reliability in administering the PSS. Results of Study 2 show a highly significant increase in the residents' inter-judge reliability from before to after training ($p < .005$). All seven residents in the study had total agreement on an average of 64 per cent of the PSS items before training (that is, when they received the prevailing amount of training (TAU) as found in Study 1) and on 90 per cent of the PSS items after the training intervention. This investigation was useful in showing that psychiatry can further its goal of conforming to scientific models by providing the type of training necessary to yield the high inter-judge reliability levels needed to achieve those goals.

## INTRODUCTION

In psychiatry's pursuit of moving itself toward a scientific model, standardization has become a central theme. Standardization is the force behind the progression of Diagnostic Statistical Manuals (DSM-III-R), behind the development of an ever increasing number of standardized interviews, such as the Schedule for Affective Disorders (SADS) and the Psychiatric Status Schedule

Special thanks to Arturo Rio, Ph.D., Jose Szapocznik, Ph.D., and Daniel Santisteban, ABD for assistance in carrying out this project.
This project was partially supported by grant #DA5334 from the National Institute on Drug Abuse.

(PSS) and finally behind the delineation of "therapeutic windows" for various psychoactive medications. For a phenomena to conform to a scientific model it must have certain measurable, reproducible and consistent features over time. For the broader scientific community to accept a particular fields phenomenae as scientific, they look to their "proofs" through research. One of the key psychiatric research tools to have developed over the last two or three decades has been the standardized interview schedule. The clinical delivery of psychiatric care has also benefited from the development of such "intruments," especially in light of the increased demands for standard diagnoses by third-party interests.

In the training of new psychiatrists, where the goal of the training program is to help its residents develop such previously mentioned standardized skills, there are significant efforts in the areas of nosology, psychotherapeutic techniques and pharmacological interventions. Training in the use of standardized interview schedules has not received as much effort in either training programs or the psychiatric literature (1).

Although standarized psychiatric interviews (SPI's) are frequently cited in research articles, their validity and reliability remain a source of some debate. Among other topics of debate concerning SPI's, a principle area of discussion focuses on the nature of the training required to insure that the instruments' rater can endorse a consistent item-response when it is applicable to a clinically elicited stimulus (2).

In this article, the notion that you can give a mental health professional, be he/she a resident or experienced psychologist, an SPI without adequate training, and expect reliable and valid performance, is challenged. It is further suggested that specific training in SPI's would significantly enhance inter-judge reliability (3). In an attempt to address the previously cited concerns, the author conducted two studies. The first of these explored the prevailing level of residency training, in a nationwide sample of training programs, of specific training in the administration of SPI's. The second investigated the before to after effect of a specific training intervention, measuring inter-judge reliabilities on the PSS, with a sample of psychiatric fellows and psychologists.

## STUDY 1

### *Introduction*

As previously mentioned, SPI's have become an important vehicle on the road to the alignment of psychiatry to a scientific model. Residency training efforts on SPI's have not enjoyed the same emphasis as have training efforts in psychiatric diagnosis and management (4,5). A review of the literature did not reveal a systematic effort to provide SPI training in residency programs (6). In keeping with the notion that some level of training is required in the administration of SPI's so as to benefit from their usage, Study 1 surveyed the current level of training in American psychiatric training programs. The results of this survey

will be referred to as "training as usual" (TAU), that is, the average amount of training effort found as a result of the survey. As will become evident in the results section of Study 1 and as it is used in Study 2, TAU was a somewhat arbitrarily arrived at "average" of the prevailing training efforts as described by the various program directors responding to the survey. Additionally, TAU is to be held in contrast to the more intensive training intervention described in Study 2.

*Methods*

Sample

The study population were the accredited American Psychiatric Training programs as they appear in the 1988–89 *DIRECTORY OF GRADUATE MEDICAL EDUCATION PROGRAMS* as accredited by the Accreditation Council for Graduate Medical Education. From these programs, a table of random numbers was used to select a sample of approximately 20 per cent of the total population of programs (41 of 206).

Questionnaire

A five part questionnaire was developed to determine present training efforts for SPI's at psychiatry residency programs. The general concepts addressed by the questionnaire concerned; how integral a part of the overall training experience/clinical evaluation process are SPI's in any particular program, [having to choose from MANDATORY program usage and importance (that is, in some form or another, a trainee in that particular program would be 100% expected to use an SPI in the course of his/her evaluation routine), FREQUENT (probably a 50–95% expectation that a trainee would use an SPI in the course of his/her evaluation routine), INFREQUENT (probably a 5–45% expectation of usage and NOT AT ALL (meaning that, in that particular program a trainee would not be likely, at all, to encounter an SPI as a routine part of training in said institution)], which particular SPI's are routinely used and how many hours of formal training is devoted to their use. The actual questions were as follows:

1. Do your trainees utilize SPI's as part of their patient evaluations?
2. What are the three most common SPI's used by your trainees?
3. Is there scheduled DIDACTIC training time allotted to these SPI's? (hours per academic year)
4. Does your department use actual interviews and/or videotapes to provide "hands-on" training EXPERIENCE on specific SPI's?
5. Is time devoted to ANALYZING and discussing with your trainees the results they obtained during these "hands-on" experiences?

Procedure

The questionnaires were mailed to the 41 program directors in the sample. If at the end of three weeks no response was obtained, a follow-up questionnaire, along with a personalized cover letter, was remailed. The remaining non-

responders, after an additional three weeks, were contacted by telephone with a request for the information as written on the questionnaire.

*Results*

Overall, 35 of 41 (85.4%) training directors in the sample responded. Of this total, 21 responded to the first mailing, 11 responded to the second mailing and an additional 3 responded to the direct phone call.

Results obtained on the questionnaire were as follows:

| | | | |
|---|---|---|---|
| QUESTION 1: | relative importance program placed on SPI's | | |
| | MANDATORY | 5.7% | 2 of 35 |
| | FREQUENT | 5.7% | 2 of 35 |
| | INFREQUENT | 28.6% | 10 of 35 |
| | NOT AT ALL | 51.4% | 18 of 35 |
| QUESTION 2: | particular SPI's in use | | |
| | SADS/KSADS | 6 programs | |
| | DIS/DISC | 5 programs | |
| | SKIDS | 2 programs | |
| | SCID | 1 program | |
| | GAS | 1 program | |
| | BPRS | 1 program | |
| | ISC | 1 program | |
| | PDI | 1 program | |
| | NONE | 25 programs | |
| QUESTION 3: | didactic training time (hours per academic year) | | |
| | no training | 85.7% | 30 of 35 |
| | 0 to 2 hours | 2.9% | 1 of 35 |
| | 2 to 4 hours | 2.9% | 1 of 35 |
| | >4 hours | 8.5% | 3 of 35 |
| QUESTION 4: | experiential training time | | |
| | no training | 88.5% | 31 of 35 |
| | 0 to 2 hours | 2.9% | 1 of 35 |
| | 2 to 4 hours | 2.9% | 1 of 35 |
| | >4 hours | 5.7% | 2 of 35 |
| QUESTION 5: | analytical training time | | |
| | no training | 94.2% | 33 of 35 |
| | 0 to 2 hours | 2.9% | 1 of 35 |
| | 2 to 4 hours | 2.9% | 1 of 35 |
| | >4 hours | .0% | 0 of 35 |

*Discussion*

The aim of Study 1 was to determine the extent to which psychiatric residents receive specific training in the rating of widely used SPI's. Results

indicate that first, slightly less than 12 per cent of programs provided for appreciable use of SPI's by their trainees. Second, the most commonly used SPI's in this sample were the SADS/KSADS and DIS/DISC. These two instruments were mentioned by 61 per cent of the training directors who specified a training instrument of choice. Lastly, with regards to actual hours per academic year of training, there was NO training at the didactic, experiential and discussion phase in 85.7, 88.5 and 94.2 per cent, respectively. Of significance, the sites that reported high levels of training tended to self-describe as research oriented programs. It is therefore concluded that generally, residents are not appreciably trained in the use of SPI's and that training programs are not utilizing these potentially valuable resources. As such an opportunity is lost to further psychiatry's endeavors at aligning itself as a scientific model.

## STUDY 2

### *Introduction*

As the first study in this investigation focused on the amount of actual training of residents (TAU) on SPI's, this second section built upon those results by testing a group of residents for their inter-judge reliability when trained at that TAU level. Afterwards, this group of trainees' inter-judge reliabilities were determined. The group was then subjected to a more comprehensive SPI training intervention and their subsequent inter-judge reliabilities were determined. It is the notion of Study 2 that the existing training effort, as revealed in Study 1, is inadequate and that it is possible to improve residents inter-judge reliabilities when using SPI's by intervening to give them more adequate and thorough training.

Study 2 used a group of seven mental health professionals (also referred to as the RATERS) with no prior experience on the SPI used here, the Psychiatric Status Schedule (PSS), to test the above hypothesis that more adequate training will improve inter-judge reliabilities beyond those that one might expect to obtain from residents who receive only the limited, or non-existent training as is typical of residency training programs today. This group of seven was trained to an extent that hopefully was representative of the TAU that a typical resident would receive in a typical residency program today (also referred to as BASELINE training or TAU), they were tested to see how well they performed, they were then trained more thoroughly and then retested to look for performance improvements in their reliabilities.

In choosing the subjects (the patients) on which the raters administered their PSS's, there was a general attempt to assure that their demographics and pathologies were similar but more importantly, a particular statistical analysis was used that operated independently of pathology level. In other words, the pathology of the various subjects used in the study was not being tested. The study was testing how well a group of trainees could interpret the same

pathology stimulus received from that subject. As an analogy, this process would be similar to training a group of rookie baseball umpires to call balls or strikes. If the same group of umpires saw the same pitch (the stimulus), you would expect them to all interpret it the same way, independent of whether it was a curve ball or a fast ball being thrown by the pitcher (the pathology being presented by the subject). An umpire should be able to call balls vs strikes whether he is umpiring in a little-league game or the World Series.

The particular type of statistical test used to allow for this type of analysis was McNemar's Test of Correlated Proportions. An indepth discussion of the test is beyond the scope of this paper and will be dealt with in another publication by this author.

*Methods*

Research Design


The research design was a one group pre-test—post-test design. The intervention consisted of an intensive training procedure. The pre-training prior to the pre-testing represents an effort to standardize the baseline to "training as usual" levels established in Study 1.

Instrument The Psychiatric Status Schedule (PSS) (8) was used as the SPI to be investigated for inter-judge reliabilities in this study. The PSS is an instrument designed to improve the research value of clinical judgements of psychopathology and role functioning. It includes sections to detect the usual mental status type of signs and symptoms of psychiatric disorder plus sections on 1) impairment of formal role functioning; 2) impairment in efficiency and conduct of leisure time activities and daily routine; 3) impairment in interpersonal relationships; and 4) the use of drugs and alcohol and illegal or other antisocial activity.

The PSS booklet is a standardized interview format of 321 precoded items. Evaluations usually take 30–50 minutes. The scoring system involves 17 symptom and 6 role scales. Four of these role scales did not apply to our adolescent study population (e.g.—parent role, wage earner role) and were therefore not included in the analyses. Additionally the instrument specifies a number of items that apply only to certain conditions not applicable to this outpatient sample. These items were deleted from the analyses to avoid artificially inflating agreement ratings. These exclusions therefore reduced the number of analyzable items from 321 to 232 for the present study.

Raters There were seven raters, all mental health professionals with no prior experience in administration of the PSS. Specifically, the population

included 3 psychologists at the PhD level and 4 psychiatrists, all at the fellowship level in child and adolescent training.

Procedures

1. "Training as Usual." Training as usual consisted of a 20 minute introduction to the use of the PSS, including a lecture and the distribution of a copy of the published summary article on the PSS (8). The didactic lecture briefly presented the general features of the instrument and its applications. The trainees were then informed that additional information was available in a central office file.

2. Pre-testing. A senior doctoral level psychologist with 10 years of experience in the administration of the PSS (over 300 administrations) administered the PSS to two randomly selected adolescents referred to an urban public mental health evaluation center. Each of these two interviews were videotaped and later independently rated by the seven raters in this study (9). The raters did not have access to the administrator's ratings. The results of these 14 data sets comprised the pre-test data.

3. Experimental Training Intervention. The seven raters were then subjected to an intensive three-phase training protocol which consisted of (a) didactic, (b) "hands on" experience and (c) follow-up discussion phases.

a.) *Didactic Phase* Didactic training was a 2½ hour training session led by an experienced clinical researcher. During these two and a half hours this instructor discussed the Manual of Instructions for the PSS (10) in a detailed fashion. Attention was paid to such aspects of the instrument as specific wording of questions, probing, follow-up questions and the interpretation of stimulus provided by the subject (11).

b.) *Hands-on Phase* The "hands on" experience phase of training was comprised of listening to a PSS training-case audio tape provided by the PSS authors. This tape included two complete PSS administrations. The trainees completed the PSS rating forms on each of these two standardized sample cases. In addition, each trainee conducted one PSS administration on an adolescent subject on his/her own. Therefore, a total of three "hands-on" experiences were obtained by each trainee.

c.) *Follow-up discussion phase* The discussion phase of the intervention consisted of an extensive open group discussion, led by the clinical researcher, of the results of the endorsements of each rater, on an item-by-item basis, on each of the two audiotaped interview ratings. Careful attention was devoted to resolving disagreements in individual raters' endorsements.

The three-phase training approach to the above model procedural intervention required a total of 5½ hours.

4. Post-testing. To obtain post-training test data, the same procedure was followed at post as at pre-testing, but with two new interview subjects. The same psychologist conducted the interviews which were videotaped and later indepen-

dently rated by the 7 raters. The interview subjects were from the same urban catchment area, of the same age and accessed through the same public mental health referral procedure.

*Data Preparation Procedures*

The data consisted of the four videotaped PSS administrations—rated by the seven judges. Of the 321 items, 89 were excluded due to inapplicability to the subjects, as mentioned above. Therefore, 28 data sets were available for 232 items. These were entered into Fortran coding sheets, along with other pertinent variables such as subject ID, rater ID, time of testing (pre or post training), etc. Item responses were coded true, false or blank (when no answer was given). Analyses were conducted using the SPSS-X (release 3.1) program on an IBM-3090 200E computer. The List Cases By Variable feature was performed to double-check the accuracy of the data file. Frequencies and condescriptive statistics by subjects and raters were obtained for the purpose of item-by-item analyses.

*Results*

Table 1 presents the total number of items (and percentage) of inter-rater agreement at different levels of agreement per item. For example, on subject #1 (pre-training case), seven out of seven judges agreed on a total of 150 out of the 232 possible items (65%). On subject #3 (post-training case) seven out of seven judges agreed on 210 of the 232 items (90%).

McNemar's (12) Test of correlated proportions was applied to determine statistical probabilities associated with *across* time differences between pre and post-subject (1 vs 3, 1 vs 4, 2 vs 3, 2 vs 4) and *within* time [pre (1 vs 2) and post (3 vs 4)].

The results are presented in Table 2. As can be seen, there were no

**TABLE 1.**

**Total Number of Items of Inter-rater Agreement by Number of Raters Agreeing per Item. Total Items = 232.**

| | Pre Training | | Post Training | |
|---|---|---|---|---|
| Number of raters agreeing | Subject 1 | Subject 2 | Subject 3 | Subject 4 |
| 7 of 7 | 150 (65%) | 147 (63%) | 210 (90%) | 209 (90%) |
| 6+ of 7 | 186 (80%) | 189 (81%) | 224 (97%) | 226 (97%) |
| 5+ of 7 | 214 (92%) | 205 (88%) | 229 (99%) | 229 (99%) |

**TABLE 2.**

**McNemar's $x^2$ of Correlated Proportions for Inter-rater Agreement *Within* Time and *Across* Time**

| | Within Time | | Across Time | | | |
|---|---|---|---|---|---|---|
| $x^2$, df = 1 (P) | (Pre) Subjects 1 × 2(( | (Post) Subjects 3 × 4 | Subjects 1 × 3 | (Pre Subjects 1 × 4 | to Subjects 2 × 3 | Post) Subjects 2 × 4 |
| 7 of 7 | .46 | .00 | 59.21 | 55.15 | 68.91 | 61.19 |
| | (n.s.) | (n.s.) | (<.001) | (<.001) | (<.001) | (<.001) |
| 6+ of 7 | .14 | .27 | 34.91 | 43.35 | 39.51 | 34.69 |
| | (n.s.) | (n.s.) | (<.001) | (<.001) | (<.001) | (<.001) |
| 5+ of 7 | .74 | .13 | 9.38 | 17.63 | 24.30 | 14.69 |
| | (n.s.) | (n.s.) | (<.005) | (<.001) | (<.001) | (<.001) |

significant differences in the six comparisons *within* time [analyses within pre cases and within post cases, e.g. subject 1 vs 2, at 7 of 7 agreement level per item—$X^2 = .46$ (n.s.)]. This suggests that the rate of agreement between raters remained consistent. Highly significant differences, however, were obtained for the 12 analyses *across* time (from untrained to trained, e.g. subject 1 vs 3 at 7 of 7 agreement level per item—$x^2 = 59.21$, $p < .001$) at the three different levels of agreement (7/7, 6+/7, 5+/7). These represent highly significant net gains in agreement from before to after training.

*Discussion*

A modified 232 item PSS was rated by a group of 7 mental health professionals (3 junior level PhD psychologists and 4 child psychiatry fellows) on 4 adolescent male patients sampled from an urban mental health referral system. The group of raters completed the PSS questionnaire from videotaped interviews of the 4 patients. These raters underwent a comprehensive 3 phase training protocol on the standardized procedure for translating patient responses into item endorsements by the raters.

Two subjects were rated pre-training, and two subjects were rated after the training intervention.

Using McNemar's Test of Correlated Proportions, there were highly significant differences in inter-judge reliability in all combinations of across intervention analyses (subjects 1 vs 3, 1 vs 4, 2 vs 3 and 2 vs 4), and non-significant differences between subjects *within* training conditions (subjects 1 vs 2 and 3 vs 4).

It should be noted that interrater reliabilities have been frequently computed in terms of item aggregates reflected in scale scores (13,14). In this study, however, a more rigorous interrater reliability procedure was chosen that

investigates item-by-item agreement rather than merely looking at agreement on item aggregates.

## SUMMARY

The first part of the present research attempted to determine what constitutes a "training-as-usual" protocol directed at the reliable administration of SPI's such as it might exist among accredited American Psychiatric residency training programs. In a nationwide survey of one in five (41 of the 206) training programs, a better than 85% survey response rate was obtained. Results revealed that less than 12% of programs provided substantial training on an SPI. There was no training effort in standardizing residents procedural skills in applying SPI's, (indicated as hours/academic year in didactic, "hands-on" experience or follow-up discussion phase training) in more than 85% of programs sampled. The small percentage of program directors who did indicate SPI skill standardization tended to self-describe as research oriented programs.

These results of Study 1, if generalized from this survey, reflect that relatively little training is offered to trainees on the proper administration of SPI's. Although low levels of training were anticipated by the author, the survey revealed surprisingly little training effort.

Study 2 suggests that raters who are not trained in a procedure to reliably interpret a standardized psychiatric interview (SPI) will not significantly agree among themselves when they attempt to rate the various items, on the same interview, from the same patient stimulus, at the same time.

Limitations of the present study included: a 20% sample of training programs to establish training as usual may not have been a sufficiently large proportion to adequately sample the general population of American training programs. Nevertheless, the sample was randomly selected and the response rate was high. A further limitation is the seemingly arbitrary criterion for the "training as usual" intervention (20 minutes), however, the author did base this intervention on the average training effort reported by program directors.

Perhaps a more serious limitation, however, was that there was a failure to counterbalance raters by subject. In the present research design, all raters used the same two subjects at pretraining and then used a second set of subjects at post training. As such, the competing hypothesis that significant differences from pre to post may be due to subject difference rather than intervention effects—cannot be completely ruled out.

With significant concern existing within the psychiatric community relating to psychiatry's image as an inexact science, a model procedural format to standardize psychiatrists' skills in endorsing SPI's could be an important step toward a standardized skill and a concommitant image change. A standardized skill such as this has a potential role in private and institutional use of SPI's, their research use, and alternatively, being able to critically review research literature where SPI's are employed.

## REFERENCES


1. Sanson-Fisher RW, Martin J: Standardized interviews of children. Br. J. Psychiatry 139:138–143, 1981
2. Maguire P, Fairbairn S, Fletcher C: Consultation skills of young doctors: I—benefits of feedback training in interviewing as students persist. Br Med J (Clin Res) 14: 292(6535):1573–6, 1986
3. Herjanic B: Systematic diagnostic interviewing of children: Present state and future possibilities. Psychiat Day, 2(2):115–30, 1984
4. Fernando T, Mellsop G, Nelson K, Peace K, Wilson J: The reliability of axis V of DSM-III. Am J Psychiatry 143(6):752–5, 1986
5. Hyler SE, Williams JB, Spitzer RL: Reliability in the DSM-III field trails: Interview V case summary. Arch Gen Psychiatry 39(11):1275–8, 1982
6. Keller MB, Lavori PW, Andreasen NC, Grove WM, Shapiro RW, Scheftner W: Test-retest reliability of assessing psychiatrically ill patients in a multi-center design. J Psychiatric Res 16(4):213–27, 1981
7. Monti PM: The social skills intake interview: Reliability and convergent validity assessment. J Behav Ther Exp Psychiatry 14(4):305–10, 1983
8. Spitzer RL, Endicott J, Fleiss JL, Cohen J: The Psychiatric Status Schedule—A technique for evaluating psychopathology and impairment in role functioning. Arch Gen Psychiatry 23(7):41–55, 1970
9. Tyrer P, Cicchetti DV, Casey PR, Fitzpatrick K, Oliver R, Balter A, Giller E: Gross national reliability study of a schedule for assessing personality disorders. J Nerv Ment Dis 172(12):718–21, 1984
10. Spitzer RL, Endicott J, Cohen GM: Manual of Instructions Psychiatric Status Schedule, (2nd edition). New York, NY: New York State Department of Mental Hygiene, Biometrics Research, 1968
11. Semler G, Wittchen HR, Joschke K, Zaudiz M, von Geiso T, Kaiser S, von Cranach M, Pfister H: Test-retest reliability of a standardized psychiatric interview (DIS/CIDI0). Eur Arch Psychiatry Neurol Sci 236(4):214–22, 1987
12. Rosner B: Fundamentals of Biostatistics, (2nd Edition). Boston, MA: Duxbury Press, 1986
13. Tyrer P, Strauss J, Cicchetti D: Temporal reliability of personality in psychiatric patients. Psychol Med 13(2):393–8, 1983
14. Riskind JH, Beck AT, Berchick RJ, Brown G, Steer RA: Reliability of DSM-III diagnoses for major depression and generalized anxiety disorder using the structured clinical interview for DSM-III. Arch Gen Psychiatry 44(9):817–20, 1987