Article

Testing Psychodynamic Psychotherapy Skills Among Psychiatric Residents: The Psychodynamic Psychotherapy Competency Test

Linda S. Mullen, M.D.

Ronald O. Rieder, M.D.

Robert A. Glick, M.D.

Bruce Luber, Ph.D.

Paul J. Rosen, B.A.

Objective: Training in psychodynamic psychotherapy remains a core requirement of psychiatric residency training programs, yet no standard measures of competency exist to document residents' knowledge and skills in this area. To address this issue, the authors developed a written test of applied knowledge of psychodynamic psychotherapy technique and theory, the Psychodynamic Psychotherapy Competency Test. Their goal in this article was to evaluate the validity of this test.

Method: The test was given to a group of 36 psychoanalytic experts and 206 residents in their second, third, and fourth psychiatric postgraduate years from 10 programs located in different parts of the United States. Program directors provided information on the number of hours of psychodynamic didactic teaching, supervision, and resident-conducted psychodynamic psychotherapy and rated the psychodynamic psychotherapy skills of residents in their fourth postgraduate year on the basis of cumulative supervisor reports.

Results: There were significant differences in test performance between residents and faculty experts and between residents in their second and fourth postgraduate years: more advanced residents and experts had progressively better scores. The mean scores of fourth-year residents in different programs differed significantly, but the scores of second-year residents did not. Higher test scores were positively associated with both number of hours of resident-conducted psychotherapy and number of hours of supervision. Among fourth-year residents, test scores correlated significantly with program director evaluations.

Conclusions: This initial study supports the validity of the Psychodynamic Psychotherapy Competency Test as well as the feasibility of testing psychotherapy skills in a standardized fashion.

(Am J Psychiatry 2004; 161:1658–1664)

Training in psychodynamic psychotherapy in psychiatric residency programs remains a fundamental educational objective in the face of a burgeoning number of new psychiatric treatment paradigms. A survey of the field conducted jointly by the American Association of Directors of Psychiatric Residency Training and the Association for Academic Psychiatry (1) found that psychodynamic understanding and technique remain fundamental to both training and the practice of psychiatry. Understanding of patients' psychological and emotional communications and proficiency in their management are requisite capabilities for psychiatrists, and the opportunity to conduct supervised psychotherapy may be crucial to the acquisition of "the psychological knowledge and observational skills that are psychiatry's unique contribution to all of medical practice" (2).

At present, psychiatric residency training programs must provide psychodynamic psychotherapy training to meet regulations specified by the Psychiatric Residency Review Committee. These regulations, which took effect on Jan. 1, 2001, state,

> The program must demonstrate that residents have achieved *competency* [italics added] in at least the following forms of treatment: brief therapy, cognitive-behavioral therapy, combined psychotherapy and psychopharmacology, psychodynamic therapy, and supportive therapy. (3)

It is impossible to meet the requirement of assessing competency in psychodynamic psychotherapy and these other areas without reliable, valid, and cost-effective measures. In addition, a valid measure of psychodynamic psychotherapy skill would allow for the study of the necessary experiences and time spent in psychotherapy training to develop such competency.

The primary method of evaluation of psychodynamic psychotherapy abilities in most residencies has been supervisory reports of trainees' performance. When supervisors have few supervisees, and when they are primarily in a teacher/mentor role, it is inevitable that such evaluations are subject to bias and inconsistency. Numerous investigators and educators have attempted to improve such ratings by developing instruments designed to measure the acqui-

sition of psychotherapy skills. In 1981, Liston et al. (4) developed the Psychotherapy Competence Assessment Schedule, an 85-item scale designed to assess videotaped recordings of residents conducting psychodynamic psychotherapy. They found that residents acquired skills over the course of training, but interrater agreement was low, particularly for judgments about "technique skills" compared with "communication skills." Additionally, the time and effort required to perform such evaluations made them unfeasible, and the scale is not now in use.

A number of other authors have attempted variations on this method, either by rating therapists' performance in videotaped or audiotaped psychotherapy sessions directly or by collecting self, peer, or supervisor rating questionnaires based on a general sense of performance (5–11). Some studies using these instruments have claimed to show acquisition of psychotherapy skills by residents over time, but levels of interrater reliability were uniformly low and training and manpower issues made such measures characteristically cumbersome and difficult to apply to a large group of subjects. Additionally, since the raters in these studies were usually not blind to the level of training, these results may reflect supervisor and trainee expectations of the growth of skills during residency training.

In 1985, Moline and Winer (12) attempted to create a more objective measure of psychotherapy skills with the use of written clinical vignettes that required trainees to choose among intervention options. Moline and Winer found that beginning residents performed better on the test than did more advanced residents; however, their study was limited by its small sample size, its limited number of questions, its lack of conceptual breadth, and its lack of validation against other measures.

Our overall goal was to evaluate the validity of a newly developed written test, the Psychodynamic Psychotherapy Competency Test. Specifically, we had the following goals:

1. Determine the reliability of expert clinicians' responses to the test items as the criterion standard. Our hypothesis was that there would be an acceptably high level of consensus among expert clinicians on the psychodynamic psychotherapy concepts and techniques underlying the correct responses presented in the test.
2. Determine the discriminant validity of the test. Our hypothesis was that the test results would differ across groups known to have varying levels of experience and training in psychodynamic psychotherapy.
3. Determine the criterion validity of the test. Our hypothesis was that the test scores would show a correlation with supervisory evaluations, the standard measure of capability in this area.
4. Determine the feasibility and applicability of the test. Our belief was that psychiatry residency program directors would be willing participants in this type of evaluation of the psychotherapy skills of their residents. We also believed that the test could be applied widely across programs differing in size, location, educational resources, and investment in the teaching of psychodynamic psychotherapy.

Method

Test Development

We developed a multiple-choice, written test of psychodynamic psychotherapy skills. The test was designed to assess knowledge of psychodynamic theory and technique as applied to clinical situations, represented by actual psychotherapy sessions, briefly described in vignettes. These vignettes describing situations experienced by residents are similar to case presentations in classes teaching psychodynamic psychotherapy theory, technique, and process. Our goals were to develop a test that could validly measure the acquisition of core psychodynamic psychotherapeutic concepts and skills during residency training and to correlate the measured acquisition of knowledge with components of education (i.e., classroom teaching, supervision, and hours of patient contact).

The Psychodynamic Psychotherapy Competency Test consists of eight psychotherapy vignettes, each presenting historical and current information about a patient engaged in a psychodynamic treatment, and one or more specific psychotherapy sessions at various time points during that treatment, for a total of 17 clinical encounters. There are 57 five-option, multiple-choice questions related to the written clinical material. The vignettes are designed to assess the following core areas of psychodynamic psychotherapy knowledge and skills: 1) establishing a therapeutic framework and alliance, 2) recognizing and managing transference, countertransference, and resistance, 3) recognizing defensive organization and therapeutic change, and 4) assessing and recommending several psychodynamic interventions. Most items encompass multiple concepts in psychodynamic psychotherapy, allowing us to test a broad range of knowledge with a relatively small number of questions. We accomplished this by designing a variety of wrong answers that reflected different types of common errors. Vignettes are written in a way that protects patient confidentiality.

The questions were developed by two of us (L.S.M. and R.O.R.) and reviewed by three collaborating psychoanalysts. These five analysts constituted the initial group of expert consultants. The level of agreement we required for inclusion of any given item was 80%, or four of five experts in this group. Questions with less than 80% agreement were rewritten to attain the 80% standard. An educational consultant then reviewed the questions and provided advice regarding National Board of Medical Examiner guidelines for test writing. The test was designed to take approximately 2 hours to administer. Scoring is done by calculating the percentage of questions answered correctly out of a total of 57. A correct answer was defined as one that was endorsed by the highest percentage of experts in the study.

We have included a sample question with brief answers in Appendix 1. The entire test can be reviewed on our web site at http://psychotherapy.columbia.cursum.net.

Subjects

The group of 50 faculty experts approached for participation in this study consisted of academically active psychoanalysts. They ranged in age from approximately 36 to 72 years. Two-thirds of the group came from Columbia University, and the rest came from academic centers in different parts of the United States. The 36 experts (72%) who agreed to participate were demographically representative of the group as a whole but differed in their avail-

TABLE 1. Characteristics of 36 Experts and 206 Residents Participating in a Study of the Psychodynamic Psychotherapy Competency Test

	Female Gender		Age (years)	
Group	N	%	Mean	SD
Experts (N=36)	12	33	—[a]	
Residents				
Postgraduate year 2 (N=76)	35	46	32.4	5.1
Postgraduate year 3 (N=69)	32	46	34.0	6.8
Postgraduate year 4 (N=61)	27	44	35.0	7.2

[a] Information on age was not obtained for experts.

ability to participate in research; time pressures were the most common reason given for refusal to participate.

Ten psychiatric residency training directors with an interest in psychodynamic psychotherapy (three of whom were in the expert sample) were asked to administer the test to their residents in postgraduate years 2, 3, and 4. Five agreed to do so. An open invitation to participate in the pilot study of the test was extended to residency directors attending a workshop where the project was presented during the March 1999 Annual Meeting of the American Association of Directors of Psychiatric Residency Training. Directors of five additional programs volunteered and were able to meet the time requirements for inclusion in the pilot study for 1999. This procedure yielded a study group of programs varying in location, size, and level of commitment to the training of psychodynamic psychotherapy. These programs were at Columbia University, Cornell University, Creedmoor Psychiatric Center, the Karl Menninger School of Psychiatry, New England Medical Center, New York University, Northwestern University, Shepard Pratt at the University of Maryland, the University of New Mexico, and the University of North Carolina in Chapel Hill.

Test Procedure and Additional Data Collected

Psychoanalytic faculty members were asked to take the test at their convenience before a deadline. They were instructed to take the test in a 2-hour block, without interruptions and without discussing the material with other participants, including residents. Residents were tested over 2 hours, under standardized testing conditions, in the late spring of the 1999 academic year, reflecting maximal training at their respective postgraduate levels.

Program directors who agreed to participate were asked to provide information regarding their programs as well as evaluations of the psychodynamic psychotherapy skills of their fourth-year residents. The program directors completed a form describing their programs' hours of psychodynamic didactic teaching, supervision, and clinical work for each class of residents. Program directors also rated fourth-year residents for their psychodynamic psychotherapy skills on the basis of cumulative supervisory evaluations over the course of their training. The ratings made were for overall skill plus six specific skills: ability to recognize and manage transference, ability to use one's own emotional responses, knowledge of psychological terms, knowledge of psychodynamic concepts and literature, ability to use the patient's emotional responses, and ability to use the supervision. Each of these items was rated on a 7-point Likert scale with anchor points that we defined; a higher rating indicated better performance. Demographic data on residents, including age and gender, were also collected. For experts, data were collected on gender but not on age.

Validation Strategy and Data Analysis

Our validation strategy was to 1) obtain a group of items for which we had expert consensus, 2) determine if residents with ascending levels of training had escalating scores approaching those of experts, 3) correlate residents' scores with specific aspects of psychodynamic training.

We administered the test to the expert group and to the residents simultaneously. This method precluded determining whether the experts agreed with the a priori defined correct answers and which items had an acceptable level of expert consensus before the administration of the test to the residents. However, it allowed us to determine the level of agreement among a larger group of experts subsequently and, thus, to examine important characteristics of test items, such as the discriminating ability of items with varying levels of consensus among experts. We planned to eliminate from scoring any item where the plurality of experts did not agree with our a priori defined correct answer. On the advice of our educational consultant, we established, a priori, an expert consensus level of 70% or more for an item to be acceptable, which would be low for "fact"-based information but not for measures of clinical judgment. Items with less than 70% consensus would be reviewed to examine whether they had other valuable qualities.

We studied discriminant validity by examining the extent to which residents' scores on the test were associated with their level of training. Data for residents and experts were compared by analysis of variance (ANOVA). Residents were compared by ANOVA across postgraduate year levels and across programs. With four groups (three groups of residents and one group of experts) in an ANOVA, a sample size of 209 was necessary to achieve a conventional power level of 0.8, given an effect size of 5% variance accounted for and a criterion alpha of 0.05. For an ANOVA on scores across programs, with 10 institutions, 146 examinees were needed to ensure a power level of 0.8 with an alpha of 0.05 for an effect size of 10% variance accounted for (13). We also examined criterion validity by correlating test scores for all fourth-year residents with supervisor evaluation scores.

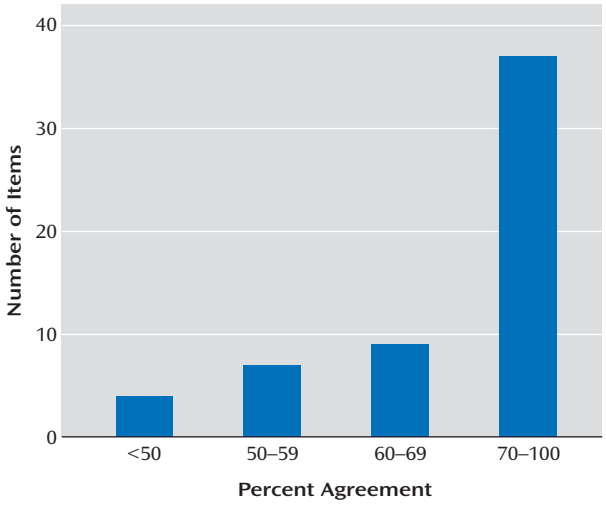
Residents' test scores were correlated with program characteristics (i.e., number of hours of didactic teaching, supervision, and clinical work in psychodynamic psychotherapy) to explore which educational components had the strongest relationship. We also examined the relationship between these program characteristics and mean scores of residents from all programs, using a multiple regression.

Individual test items were evaluated in terms of expert consensus as well as two other measures—item difficulty factor (number of subjects who failed the item/number of subjects who responded to the item) and discrimination index (correct response frequency of upper quartile minus correct response frequency of lower quartile/number of respondents in upper quartile). A difficulty factor between 0.2 and 0.8 is usually considered acceptable; in a reasonably heterogeneous sample, item difficulty should approach 0.5 for best discrimination. For the discrimination index, a value at or above 0.2 is considered a high level of discrimination (14).

Results

Ten psychiatric residency training programs administered the test to most of their residents in postgraduate years 2, 3, and 4, for a total of 206 residents. Residents were approximately equally distributed across postgraduate years, and although men slightly outnumbered women at all postgraduate year levels, this did not reach statistical significance. Likewise, residents' mean ages reflected their postgraduate year level and were similar across years. In the group of 36 psychiatric/psychoanalytic expert faculty members who took the test, men outnumbered women by about two to one (Table 1).

FIGURE 1. Consensus Among 36 Experts on Answers to 57 Questions on the Psychodynamic Psychotherapy Competency Test[a]



[a] Thirty-seven of the 57 items had ≥70% agreement among the expert group.

Experts' Results

Thirty-seven (65%) of 57 questions had a level of expert consensus of at least 70% (Figure 1), in agreement with the a priori defined answers. In the other 20 items, the answer most frequently chosen by the expert group was without exception the a priori defined correct answer. Consensus ranged from 48% to 69% (Figure 1). Thus, the agreement on all items was well above the chance agreement level of 20% for a five-option multiple-choice item, and all a priori correct answers chosen by the authors became the correct answer most chosen by experts. Because of substantial consensus on items with less than 70% expert consensus, and because these items added value to the test because of their other qualities, including a high discrimination index, we decided against deleting any items at this stage of development of the instrument. Thus, all items were included in the total score. The mean score for the group of experts was 43.2 (SD=3.5), or 75.8% (Table 2). Gender was not a significant determinant of mean score in this group.

Residents' Results

The ANOVA comparing mean total scores of experts and residents at all levels was highly significant (Table 2); more advanced residents and experts scored progressively better. A post hoc comparison of means (Tukey-Kramer, $p<0.05$) revealed that experts were significantly different from residents at all levels and that second-year residents differed significantly from fourth-year residents.

Residents' scores were significantly correlated with specific training measures. Total scores of all residents correlated significantly with cumulative number of hours of didactic instruction in psychodynamic theory and technique, individual supervision, and conducting psychodynamic psychotherapy; residents having a greater number of hours of training had higher scores. Scores of residents in their second postgraduate year correlated only with hours of didactic instruction, and scores of residents in their fourth postgraduate year correlated only with hours of psychodynamic therapy experience and supervision (Table 3). Number of hours of conducting psychotherapy and number of hours of supervision were found to be highly correlated with each other ($r=0.84$, $N=26$, $p=0.0001$).

TABLE 2. Experts' and Residents' Scores on the Psychodynamic Psychotherapy Competency Test[a]

Group	Number Correct of 57 Questions		Percent Correct of 57 Questions
	Mean	SD	
Experts (N=36)	43.2	3.5	75.8
Residents			
Postgraduate year 2 (N=76)	34.4	6.2	60.4
Postgraduate year 3 (N=69)	36.3	5.4	63.7
Postgraduate year 4 (N=61)	37.4	5.5	65.6

[a] Significant difference among groups ($F=22.1$, $df=3, 238$, $p<0.0001$).

Test scores of residents in their fourth postgraduate year correlated significantly with all measures of resident psychodynamic psychotherapy skill evaluated by using the 7-point supervisor ratings (Table 4). Multiple regression across all supervisor ratings was significant (Table 4). The total score on the test correlated moderately with the supervisors' overall skill rating (Table 4). The highest test score correlation was with the rating for knowledge of psychodynamic concepts and literature (Table 4).

Program Results

Figure 2 shows the mean total scores on the Psychodynamic Psychotherapy Competency Test for the 10 programs. The mean scores ranged from 28.8 to 38.8. An ANOVA comparing total scores of residents across programs revealed significant differences (Figure 2). Post hoc testing (Tukey-Kramer honestly significant difference) showed the two lowest-scoring programs (1 and 5 in Figure 2) to be significantly different from the three highest-scoring programs, with the third-lowest-scoring program also significantly different from the highest-scoring one.

The ANOVA comparing mean total scores for residents in their second postgraduate year across the six programs that had at least five residents in each postgraduate year did not show significant differences (Figure 3), whereas an ANOVA comparing mean total scores of residents in their fourth postgraduate year in the same programs showed significant differences among programs (Figure 4).

Test Characteristics

According to the level of consensus of experts on items (Figure 1), 37 (65%) of the 57 items were highly discriminatory (discrimination index ≥0.20). Of the 20 questions that had a level of expert consensus less than 70%, 16 (80%) had a discrimination index of at least 0.20 (Table 5). Forty-five (79%) of the 57 items had a difficulty factor between 0.2 and 0.8. All 20 questions that had a level of expert con-

TABLE 3. Correlations Between Scores on the Psychodynamic Psychotherapy Competency Test and Cumulative Hours of Psychotherapy Training of 199 Residents[a]

Type of Training	Correlation (r) All Residents (N=199)	Postgraduate Year 2 (N=72)	Postgraduate Year 3 (N=65)	Postgraduate Year 4 (N=62)
Instruction	0.18**	0.23*	0.10	–0.05
Therapy	0.26**	–0.03	0.24	0.33**
Supervision	0.27**	0.19	0.14	0.38**

[a] One program did not provide information on type of psychotherapy training.
*p<0.05. **p<0.01.

TABLE 4. Correlations Between Scores on the Psychodynamic Psychotherapy Competency Test With Supervisor Evaluations of 61 Residents in Their Fourth Postgraduate Year[a]

Psychodynamic Skill	Correlation[b] r	p
Overall skill	0.40	<0.001
Ability to recognize and manage transference	0.39	<0.001
Knowledge of psychological terms	0.44	<0.001
Knowledge of psychodynamic concepts and literature	0.46	<0.001
Ability to use patient's emotional responses	0.29	<0.004
Ability to use own emotional responses	0.29	<0.004
Ability to use supervision	0.22	<0.04

[a] Supervisors rated residents on a 7-point scale; a higher rating indicated better performance.
[b] r^2=0.26, N=7, p<0.001.

sensus less than 70% had a difficulty factor in this range. Forty-four (77%) of the 57 items had at least two of the desirable test qualities described earlier in this article. The majority of questions not having at least two of these qualities tended to be "easy" and were basic questions that did not distinguish between high and low scorers.

Discussion

Our study and results lead us to conclude that this first version of the Psychodynamic Psychotherapy Competency Test shows considerable evidence of being a valid instrument, demonstrating both discriminant and criterion validity. The results also indicate that certain elements of residency education—clinical experience and supervision—contribute to the acquisition of knowledge about psychodynamic theory and technique.

Our data demonstrate that recruitment of expert faculty to participate in the project is feasible and that adequate levels of expert consensus on clinical judgments in psychodynamic psychotherapy can be obtained. These are extremely important factors in the continued development of the test. In this first version, we explored the level of agreement among experts. The findings that 65% of items showed ≥70% agreement and that all but five of the 57 items had >50% agreement demonstrate the existence of a set of shared concepts among the experts as well as substantial reliability of the criteria (i.e., correct answers) used in our instrument.

As stated earlier, we established the 70% goal for expert agreement rather than a higher level a priori because we believe that such a level of agreement would be appropriate given that we are attempting to measure clinical judgments and decision making, not factual knowledge. Similar levels of agreement are considered acceptable in establishing the reliability of clinical diagnoses among experts using diagnostic criteria and rating scales.

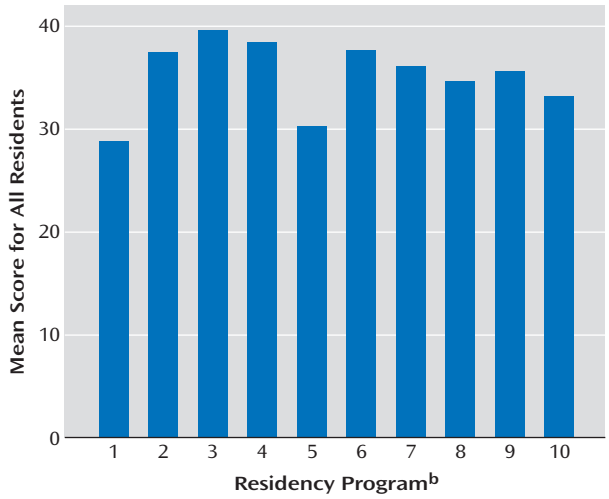
Giving the same version of the test to both experts and residents enabled us to measure important test item characteristics independent of expert consensus. We found that all of the items having less than 70% agreement had other valuable test characteristics, especially an ability to discriminate (80%) among high and low scorers, both faculty and residents. This indicates that even though there is substantial consensus among experts, there are some areas of clinical judgment in which there is somewhat less agreement, and it is in these areas that more accomplished clinical thinking can be tested. In other words, ignoring these areas by eliminating items of this sort might lead to ignoring important areas that should be tested. Another possibility is that a lack of consensus on an item may reflect poor item construction rather than a more stringent test of clinical thinking. As we accumulate more expert data, we should be able to answer this question.

We examined the criterion validity of the test, but we recognize that there is no established standard in this area. We were able to show a substantial correlation with supervisor evaluations, the most widely used current measure of psychotherapy skills. More importantly, we showed discriminant validity of the test across groups differing in level of training, experience, and expertise. Advanced residents and experts performed progressively better on the test. However, we have not yet measured the progression of scores in the same group of residents over the course of their training. In such a study we would want to have data on residents before training or very early in training, whereas our current data begin with residents in the late spring of their second postgraduate year.

In addition to showing differences across postgraduate year level, we found that these differences correlated with the amount of experience in conducting psychodynamic psychotherapy and the amount of supervision provided to residents, further evidence of discriminant validity as well as an indication of a measurable teaching effect on these clinical skills.

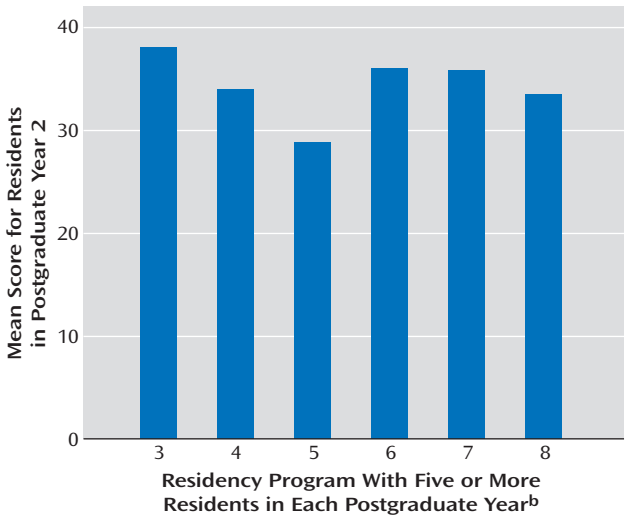
In addition to analyzing test performance across postgraduate year levels, we compared test performance across programs and found that programs varied in terms of the performance of advanced residents but not beginning residents. Although we found very substantial correlations of the mean differences with program educational characteristics in residents in their fourth postgraduate year across programs, we did not have enough residency programs participating in this study to meaningfully ex-

FIGURE 2. Mean Scores on the Psychodynamic Psychotherapy Competency Test of 206 Residents in 10 Residency Programs[a]



[a] Mean score=number of items answered correctly of 57.
[b] Significant difference among programs (F=6.1, df=9, 196, p<0.0001).

FIGURE 3. Mean Scores on the Psychodynamic Psychotherapy Competency Test of 63 Beginning Residents in Six Residency Programs[a]



[a] Mean score=number of items answered correctly of 57.
[b] No significant difference among programs (F=2.10, df=5, 57, n.s.).

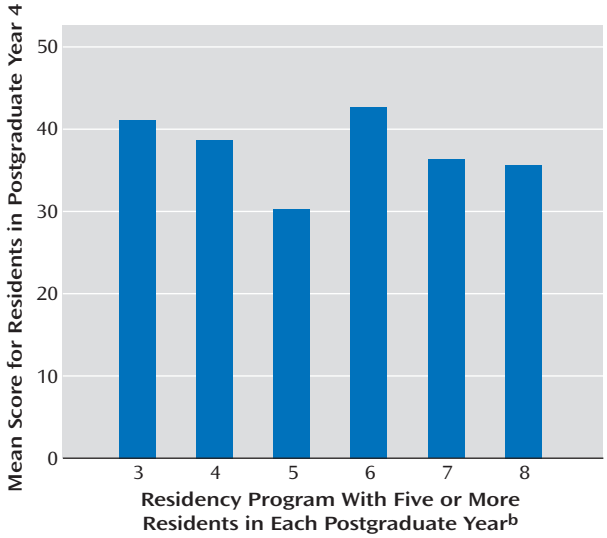
amine this issue. This finding suggests that the test may be sensitive to the degree to which programs emphasize or effectively teach psychodynamic psychotherapy. Differences may also reflect recruitment patterns or the preference of residents interested in psychodynamic psychotherapy for one program over another.

The test represents an attempt to create a valid instrument in an area where no gold standard exists. We do not know whether test performance reflects actual practice or, even if it does, whether these practices predict good patient outcome. Although the number of participants was substantial, it borders on being inadequate for certain analyses. Finally, at the time of this writing, we have not yet studied test-retest reliability and are thus unable to look at variance of the same residents over time. Future studies should include following a group of residents over the course of their training, in addition to measuring test-retest reliability in the group of faculty experts, because residents' scores would be expected to change over a time that would allow for the forgetting of previous answers.

We regard the high level of agreement among experts as an indication of reliability. It would also be useful to convene the experts to determine reasons for lack of consensus on some items. At present, the test provides a distribution of proficiency of residents at each level of training, both within individual programs and across a group of programs with varying teaching and practice patterns.

Our pilot data suggest that this testing method could provide one way to measure the competency of resident psychiatrists in psychodynamic psychotherapy. We believe the test should be integrated with other observations to establish competency and progression. It has the

FIGURE 4. Mean Scores for 54 Advanced Residents in Six Residency Programs[a]



[a] Mean score=number of items answered correctly of 57.
[b] Significant difference among programs (F=5.24, df=5, 48, p<0.0006).

advantage of being more objective and of having national norms. It could identify "outlying" residents for special remediation, and it could also track residents' progress if used on an annual basis. The test also provides substantial data for programs to evaluate themselves, allowing for a more objective and valid comparison of resident groups within a program as well as comparisons with other programs. Programs will have to interpret any deficiency in their residents' performance as being related to resident aptitude or quality of education and training.

TABLE 5. Levels of Agreement Among 36 Experts Regarding Answers to Questions on the Psychodynamic Psychotherapy Competency Test With Different Discrimination Indices

	Level of Agreement Among Experts[b]											
	90%–100% (N=20)		80%–89% (N=7)		70%–79% (N=10)		60%–69% (N=9)		50%–59% (N=7)		<50% (N=4)	
Item[a]	N	%	N	%	N	%	N	%	N	%	N	%
Questions with discrimination index ≥0.20 (N=34)	10	50	4	57	7	70	7	78	5	71	1	25
Questions with discrimination index <0.20 (N=23)	10	50	3	43	3	30	2	22	2	29	3	75

[a] N=number of questions with given discrimination index.
[b] N=number of questions with given level of agreement.

APPENDIX 1. Example Question[a]

The resident should have (choose the best response):
- Been more flexible in changing the time of the session so as to support the patient's developing assertiveness in the therapy. [Incorrect—unnecessary gratification of patient or appeasement]
- Done nothing differently, as the process appears to be evolving at a reasonable rate and intensity. [Correct—patient is increasingly self reflective and tolerant of the treatment frame]
- Refused to change the appointment as the patient is treating the resident as if she is the patient's husband (i.e., "acting-in"). [Incorrect—misinterpretation of patient's transference and arbitrary withholding as a counterresponse]
- Handled the request for the change similarly, while being more explicit about the patient's need to avoid sexual material. [Incorrect—patient is manifesting greater self awareness and self esteem, not currently avoiding sexual material]
- Handled the request for the change similarly, while interpreting the patient's sadomasochistic transference, as it is now more fully apparent in the treatment. [Incorrect—misinterpretation of patient's healthy self assertion and capacity to tolerate frustration as sadomasochism]

[a] Includes rationale for correct and incorrect responses. The test may be viewed in its entirety on our web site at http://psychotherapy.columbia.cursum.net.

Presented at the annual meetings of the American Association of Directors of Psychiatric Residency Training, San Juan, Puerto Rico, March 9–12, 2000, and the American Psychiatric Association, Chicago, May 13–18, 2000. Received May 16, 2002; revision received Nov. 4, 2003; accepted Nov. 24, 2003. From the Department of Psychiatry, Department of Residency Education, and Center for Psychoanalytic Training and Research, Columbia University College of Physicians & Surgeons and New York State Psychiatric Institute. Address reprint requests to Dr. Mullen, Columbia University/New York State Psychiatric Institute, 1051 Riverside Dr., Unit 63, New York, NY 10032; lsm23@columbia.edu (e-mail).

Supported, in part, by the Irville and Helen MacKinnon Graduate Education Fund.

The authors thank Drs. Elizabeth L. Auchincloss, Carol Bernstein, Stanley Bone, Peter Buckley, Lisa Dixon, Ronald Krasner, Alan Maltbie, Lisa Mellman, Paul Mohl, Nancy Morrison, Steven P. Roose, Hillary Schmidt, Marc Sorenson, Deborah Spitz, George Thompson, and Joel Yager.

References

1. Mohl PC, Lomax J, Tasman A, Chan C, Sledge W, Summergrad P, Notman M: Psychotherapy training for the psychiatrist of the future. Am J Psychiatry 1990; 147:7–13
2. Nemiah JC: Supervision: teaching or psychotherapy? Canada's Ment Health Suppl 1971; 66:3–63
3. Accreditation Council for Graduate Medical Education: Program Requirements for Residency Training in Psychiatry, Sept 2000. http://www.acgme.org
4. Liston EH, Yager J, Strauss GD: Assessment of psychotherapy skills: the problem of interrater agreement. Am J Psychiatry 1981; 138:1069–1074
5. Buckley P, Conte HR, Plutchik R, Karasu TB, Wild KV: Learning dynamic psychotherapy: a longitudinal study. Am J Psychiatry 1982; 139:1607–1610
6. Fuqua D, Newman J, Scott T, Gade E: Variability across sources of performance ratings: further evidence. J Couns Psychol 1986; 33:353–356
7. Winer JA, Mostert M: Evaluation of residents' dynamic psychotherapy skills. J Psychiatr Educ 1988; 12:329–337
8. Robiner W, Fuhrman M, Bobbit B: Supervision in the practice of psychology: toward the development of a supervisory instrument. Psychotherapy in Private Practice 1990; 8:87–98
9. Alberts G, Edelstein B: Therapist training: a critical review of skill training studies. Clin Psychol 1990; 10:497–511
10. Jones S, Krasner R, Howard K: Components of supervisors=ratings of therapists=skillfulness. Acad Psychiatry 1992; 16:29–36
11. Beitman BD, Yue D: A time-efficient, research-based, outcomes-measured psychotherapy training program. Acad Psychiatry 1999; 23:95–102
12. Moline R, Winer JA: Assessment of residents=ability to do psychotherapy. J Psychiatr Educ 1985; 9:329–337
13. Murphy KR, Myors B: Statistical Power Analysis. Mahwah, NJ, Lawrence Erlbaum Associates, 1998, p 56
14. Golden CJ, Sawicki RF, Franzen MD: Test construction, in The Handbook of Psychological Assessment. Edited by Goldstein G, Hersen M. New York, Pergamon Press, 1984, pp 19–37