BRIEF REPORTS

A New Instrument for Residency Program Evaluation

Lisa Bellini, MD, Judy A. Shea, PhD, David A. Asch, MD, MBA

The purpose of the study was to develop a comprehensive program evaluation instrument. Following pilot work with residents, a 69-item instrument consisting of statements with 5-point strongly agree to strongly disagree response options was distributed to 107 residents; 104 responded. Psychometric analyses revealed no ceiling or floor effects; 9 items were deleted. There were three subscales: workload (19 items; $\alpha = .76$); educational environment (29 items; $\alpha = .72$) and lifestyle issues (12 items; $\alpha = .62$). Mean item scores were significantly higher for educational environment (3.53; SD 0.28) than for workload (2.78; SD 0.39) and lifestyle (2.96; SD 0.42). Items with the lowest scores reflected issues that were largely fixable. The 60-item instrument appears to be psychometrically sound, comprehensive, and exportable.

KEY WORDS: residency; program evaluation.

J GEN INTERN MED 1997;12:707-710.

Resident-generated program evaluation can identify strengths and weaknesses of a training program through residents' eyes, document the educational climate of the program, and give residents a voice in program structure.¹⁻⁶ Despite the importance of evaluation, comprehensive instruments are difficult to find. Many of those that exist are directed at specific aspects of residency training such as its stress, and the impact of introducing night-float systems or other changes.^{3,7-11} We describe the development of a program evaluation instrument that expands on three previously identified domains—workload, education, and lifestyle—to include specific factors that might be targeted for intervention.^{12,13}

METHODS

A MEDLINE search using the search words "residency" and "program evaluation" from 1991 to 1996 identified 145 citations. One evaluation instrument developed by Seelig was found among these citations.^{12,13} The 33-item questionnaire assessing resident satisfaction with

workload, learning environment, and stress served as the foundation of our evaluation instrument. We supplemented this initial list with other questions based on our clinical and educational experiences.

The initial draft instrument contained 63 items. All questions were written in statement form. Response options were formatted as a 5-point Likert-type scale, with strongly disagree to strongly agree scored from 1 to 5, respectively.

The questionnaire was pilot tested on 20 residents and faculty members to judge completeness, readability, and accuracy in reflecting the residents' perceptions. On the basis of the pilot test responses, 6 questions were added. The final instrument contained 69 evaluation items and 11 demographic questions (Appendix A). It was distributed in May 1995 to all 107 residents in one university program; 104 residents responded.

Psychometric analyses focused on examining the frequency distributions of items to identify those with large amounts of missing data, ensure that distributions were interpretable, identify items that were "reverse" scored, and look for items with ceiling effects. Following Seelig's conception of subscales, the workload subscale had 19 items, the educational environment subscale had 29 items, and the lifestyle subscale had 12 items. For each subscale we calculated an internal consistency reliability coefficient (Cronbach's α) and disattenuated correlations among the subscale scores.¹⁴ When analyzing subscale scores, we substituted item mean values based on the responses of all who answered the item for missing data.

Exploratory analyses consisted of comparing the three subscale mean scores, shown as the percentage of total available points and mean item scores, to observe which aspects of the program were most problematic to the residents, and looking at variation in scale scores according to 10 items describing demographics and career plans. In addition, we examined the content of the 10 items with the most and the least favorable responses.

RESULTS

Psychometric analyses revealed three items with more than 33% missing data. Because the content applied to only a subset of residents in our program, these items were deleted from further analyses. The amount of missing data on other items ranged from 0% to 19%, with a mean of 4.3%. No items had floor or ceiling effects. The internal consistency reliability coefficients (Cronbach's α)

Received from the Department of Medicine, University of Pennsylvania School of Medicine (LB, JAS, DAA), and Veterans Affairs Medical Center (LB, DAA), Philadelphia, Pa.

Dr. Asch is the recipient of a Health Services Research and Development career development award from the Department of Veterans Affairs.

Address correspondence and reprint requests to Dr. Bellini: Department of Medicine, Hospital of the University of Pennsylvania, 3400 Spruce St., Philadelphia, PA 19104.

Table 1. Summary of Subscale Scores

Subscale	Percent of Total (SD)	Mean Item Score (SD)
Workload	0.56 (0.08)	2.78 (0.39)
Educational environment	0.71 (0.06)	3.53 (0.28)
Lifestyle	0.59 (0.08)	2.96 (0.42)

were .76 for workload, .72 for educational environment, and .62 for lifestyle issues. Disattenuated correlations among the three subscale scores were .39 between workload and educational environment, .70 between workload and lifestyle, and .39 between educational environment and lifestyle issues, suggesting that at least the lifestyle and workload subscales measure something different from the educational environment subscale.

Each item's polarity was examined. Forty-six items were worded so that "disagree" or "strongly disagree" indicated favorable opinions. Scoring for these items was changed so that 5 (the highest item score) consistently indicated a favorable response. Six items were deleted because we could not agree on an unambiguous favorable or unfavorable interpretation.

The results of the exploratory analyses include a summary of subscale scores (Table 1) and the 10 items with the most and the least favorable responses (Table 2). Scores were significantly different between the workload and lifestyle subscales ($F = 20.96, p < .0001$), but the differences were more dramatic between the educational environment and lifestyle subscales ($F = 174.08, p < .0001$)

and between the educational environment and workload subscales ($F = 347.28, p < .0001$). Only 2 of 30 analyses examining score variation according to demographics and career plans were statistically significant. The 10 items with the most and least favorable responses revealed that many highly endorsed items are related to the educational environment or have content about colleagues, while items with which residents are least satisfied are largely fixable.

DISCUSSION

The most useful aspect of this project for those outside our institution is probably the instrument itself. Although the developmental effort will certainly benefit us in the future, viewing the items with the most and the least favorable responses was very instructive. The content of the favorable responses showed that the housestaff were generally satisfied with the educational program and that our best learning venues were rotations in the intensive care unit (ICU) and emergency room (ER), resident report, and our weekly management conference series. The items with the least favorable responses reflected both institutional and program issues. On the basis of these responses, we have made several changes. For example, we reorganized the phlebotomy and transport services, reassessed the responsibilities of the inpatient clerical staff, and implemented a strict days-off and work-hour policy.

This study has several limitations. First, it represents the experience of only one university-based training program. Still, we believe that the instrument addresses issues

Table 2. Ten Items with the Highest and Lowest Levels of Endorsement

Component	Item Content	Mean
Highest levels of endorsement		
Education	I learn a lot on ICU.	4.62
Education	Resident report has contributed to my learning.	4.46
Education	My colleagues contribute to my learning on the wards.	4.39
Education	Management conference has contributed to my learning.	4.16
Lifestyle	The personal support I receive from my colleagues is adequate.	4.13
Lifestyle	The personal support I receive from my family and friends is adequate.	4.12
Education	I learn a lot on ER rotation.	3.97
Education	I have too much responsibility for my ward patients.	3.92
Education	I have too much responsibility for my ICU patients.	3.88
Education	Firm Chief conference has contributed to my learning.	3.87
Ten items with the lowest item endorsements*		
Workload	The computer system meets the demands of care.	1.28
Workload	I spend too much time trying to access test results (inpatient).	1.46
Education	A formal mentoring program beginning in internship would be helpful.	1.95
Education	A more structured approach to career planning would be helpful.	2.01
Lifestyle	The amount of time I have to recover between ICU call days is adequate.	2.22
Lifestyle	The amount of sleep I get while on call is adequate.	2.22
Lifestyle	The amount of time I have to read medicine is adequate.	2.27
Workload	I spend too much time ordering tests (inpatient).	2.28
Workload	I spend too much time drawing blood (inpatient).	2.31
Workload	Clerical support is adequate.	2.45

*Residents were also dissatisfied with the following item, which was omitted because of large amounts of missing data: Residency training compromises my family responsibilities (mean 2.26, SD 0.94).

that are relevant to most training programs and thus is easily exportable. Second, the study contains no intervention. We cannot assess the impact of the instrument on the program this time, although the weaknesses we identified have allowed us to implement changes in the program. Third, despite the extensive developmental process, we deleted nine questions from the final instrument. This finding, which highlights the difficulty of developing useful evaluation instruments, is important because many program directors do not have the time or local resources to develop and pilot an evaluation instrument and even fewer will subject an evaluation instrument to psychometric analysis.

Resident-generated program evaluation is essential to assess residents' satisfaction with their training. We believe we have developed a comprehensive program evaluation instrument that expands on three previously identified domains—workload, education, and lifestyle. The instrument is generic and thus exportable to other programs in need of such a tool. Areas of future research include refinement of the instrument, yearly distribution to monitor our progress with ongoing issues, as well as continued monitoring to identify new areas of concern.

The authors thank Christine Weeks for her help with data collection, entry, and analysis.

REFERENCES

- 1996–1977 Graduate Medical Education Directory. Chicago, Ill: American Medical Association; 1996.
- Gray JD. Global rating scales in residency education. *Acad Med.* 1996;71(suppl):S55–63.
- Butterfield PS. The stress of residency: a review of literature. *Arch Intern Med.* 1988;148:1428–35.
- Norcini JJ, Grosso LJ, Shea JA, Webster GD. The relationship between features of residency training and ABIM certifying examination performance. *J Gen Intern Med.* 1987;2:330–6.
- Norcini JJ, Webster GD, Grosso LJ, Blank LL, Benson JA. Ratings of residents' clinical competence and performance on certifying examination. *J Med Educ.* 1987;62:457–62.
- Grosso L, Norcini JJ. Does feedback on examination performance help directors of internal medicine residencies evaluate the medical knowledge of their residents against national norms? *Acad Med.* 1990;69:979–82.
- McCue JD. The distress of internship: causes and prevention. *N Engl J Med.* 1985;312:449–52.
- Small GW. House officer stress syndrome. *Psychosomatics.* 1981;22:860–9.
- Residency Services Committee, Association of Program Directors in Internal Medicine. Stress and impairment during residency training: strategies for reduction, identification, and management. *Ann Intern Med.* 1988;148:1428–35.
- Schwartz AJ, Black ER, Goldstein MG, Jozefowicz RF, Emmings FG. Levels and causes of stress among residents. *J Med Educ.* 1987;62:744–53.
- Trontell MC, Carson JL, Taragin MI, Duff A. The impact of the night float system on internal medicine residency programs. *J Gen Intern Med.* 1991;6:445–9.
- Seelig CB. Quantifying qualitative issues in residency training: development and testing of a scaled program evaluation questionnaire. *J Gen Intern Med.* 1993;8:610–3.
- Seelig CB, DuPre CT, Adelman HM. Development and validation of a scaled questionnaire for evaluation of residency programs. *S Med J.* 1995;88:745–9.
- Guilford JP, Fruchter B. *Fundamental Statistics in Psychology and Education.* New York, NY: McGraw-Hill; 1978.

APPENDIX A

Program Evaluation Instrument for Residents

Each of the original 69 items was presented as a statement, and respondents were asked to indicate their agreement on a 5-point scale: strongly disagree (1), disagree (2), neutral (3), agree (4), or strongly agree (5). Items in italics at the end of each section were deleted after the initial administration. Items with an asterisk (*) are reverse scored.

Workload

1. Clerical support on ward medicine is reasonable.
 2. I have too many patients on my ward service.*
 3. The average number of admissions on call is too high.*
 4. The workload on the wards is too heavy.*
 5. I spend too much time drawing blood.*
 6. I spend too much time trying to access test results.*
 7. I spend too much time ordering tests.*
 8. Medical records are generally available for the patients I evaluate.
 9. Clerical support in my continuity clinic is reasonable.
 10. I have too many patients in my continuity clinic.*
 11. I spend too much time in clinic.*
 12. I see too many patients during a given continuity clinic session.*
 13. Medical records are generally available for the patients I evaluate in clinic.
 14. The computer system meets the demands of care.
 15. Over the last year, I have had too many ward rotations.*
 16. Over the last year, I have had too many ICU rotations.*
 17. Over the last year, I have had too many ER rotations.*
 18. The amount of time I have to recover between ward call days is adequate.
 19. The amount of time I have to recover between ICU call days is adequate.
- ♦ *The addition of subspecialty preceptors in my continuity clinic would reduce my need for subspecialty consultation.**
- ♦ *I spend too much time in the outpatient setting compared to the inpatient setting.*
-

APPENDIX A
Continued

Educational Environment

1. I get useful feedback from my colleagues on the wards.
2. I get useful feedback from faculty on the wards.
3. I learn a lot on my ward rotations.
4. I learn a lot on my ICU rotations.
5. I learn a lot on my ER rotations.
6. I have too much attending supervision on my ward rotations.*
7. I have too much attending supervision on my ICU rotations.*
8. I have too much attending supervision on my ER rotations.*
9. I have too much responsibility for my ward patients.*
10. I have too much responsibility for my ICU patients.*
11. The diagnoses represented on my ward service are diverse.
12. Faculty contribute to my learning on the wards.
13. My colleagues contribute to my learning on the wards.
14. I get useful feedback from faculty in my continuity clinic.
15. My Firm Chiefs contribute to my learning in continuity clinic.
16. I learn a lot in continuity clinic.
17. The diagnoses represented in my continuity clinic are diverse.
18. I have too much responsibility for my clinic patients.*
19. I have the opportunity to follow my clinic patients as inpatients.
20. I have too much supervision in my continuity clinic.*
21. I have the opportunity to follow my inpatients in my continuity clinic.
22. The instruction I receive on what is expected of me is adequate.
23. My understanding of the benefits of health prevention in primary care is adequate.
24. The instruction I receive on providing feedback is adequate.
25. The instruction I receive on critically reviewing the literature is adequate.
26. The instruction I receive on DNR status is adequate.
27. Resident report has contributed to my learning.
28. Management conference has contributed to my learning.
29. Firm Chief conference has contributed to my learning.
- ♦ *Clinical competency criteria should be met prior to advancement to the next PGY level.*
- ♦ *I will be competent to practice general internal medicine upon completion of this training program.*
- ♦ *My continuity clinic is a valuable learning experience compared to ward medicine.*

Lifestyle Issues

1. The amount of leisure time I have is adequate.
 2. The amount of time I have off from direct patient responsibilities is reasonable.
 3. The amount of time I have to read medicine is adequate.
 4. The personal support I receive from family and friends is adequate.
 5. The personal support I receive from my colleagues is adequate.
 6. My benefits package is adequate.
 7. The creation of peer support groups would be helpful.*
 8. Sexual harassment is a problem at this institution.*
 9. The on-call facilities are adequate.
 10. The amount of sleep I get while on call is adequate.
 11. A more structured approach to career planning would be helpful.*
 12. A formal mentoring program beginning in internship would be helpful.*
 - ♦ *Residency training compromises my family responsibilities.**
 - ♦ *I moonlight to make ends meet.*
 - ♦ *I moonlight for discretionary income.*
 - ♦ *My total indebtedness related to medical training has impacted my anticipated career path.**
-