Information within residency monthly evaluation forms at two institutions

Christopher J. Cooper, Paulette Wehner, Cindy Dailey, Nanette O’Connor, James Kleshinski and Joseph I. Shapiro

ABSTRACT
Periodic review of resident performance is an important aspect of residency training. Amongst allopathic residency programs, it is expected that the performance of resident physicians which can be grouped based on the ACGME core competencies, be assessed so as to allow for effective feedback and continuous improvement. Review of monthly evaluation forms for residents in the core ACGME programs at Marshall University and the University of Toledo demonstrated a wide spread in the number of Likert questions that faculty were asked to complete. This number ranged from a low of 7 in Surgery to a high of 65 in Psychiatry (both Marshall Programs). Correlation and network analysis were performed on these data. High degrees of correlations were noted between answers to questions (controlled for each resident) on these forms at both institutions. In other words, although evaluation scores varied tremendously amongst the different residents in the programs studied, scores addressing different competencies tended to be very similar for the same resident, especially in some of the programs which were studied. Network analysis suggested that there were clusters of questions that produced essentially the same answer for a given resident, and these clusters were bigger in some of the different residency program assessment forms. This seemed to be more the rule in the residency programs with large numbers of Likert questions. The authors suggest that reducing the number of monthly questions used to address the core competencies in some programs may be possible without substantial loss of information.

Introduction
Evaluation of postgraduate or resident performance has been occurring since the development of such programs in the early 20th century following the famous ‘Flexner’ report calling out for improvements in the education of physicians [1]. However, the modern assessment of resident performance across the spectrum of competencies [2] has become much more complex [3]. Although the six basic competencies (1. medical knowledge, 2. patient care and procedural skills, 3. interpersonal and communication skills, 4. professionalism, 5. system-based practice and 6. practice-based learning and improvement) form the root of most evaluation systems [2,4,5], specialty evaluations have been allowed to, if not encouraged to, expand to query along far more axes of evaluation. This is not to say that American Board of Medical Specialties (ABMS) residency review committees (RRC) have insisted on longer evaluation forms; in fact they have not. However, virtually all of these RRCs have increased the granularity of competency assessment, and incorporation into monthly evaluation forms would seem to logically follow. For example, on the Nephrology program website, more than 50 items of assessment have been detailed under the heading of ACGME core competencies [6]. As most residency evaluation forms have incorporated the concepts of milestones and PGY-specific evaluations, these forms which are filled out by faculty (as well as other learners and, in some cases, other members of the health care team) have become more complex.

In the current study, we attempted to evaluate the information present in the monthly residency evaluation forms used by core ACGME programs at the University of Toledo and Marshall University. Hypothesizing that there might be redundancy in some of these forms, especially those that asked what superficially seemed to be a large number of duplicative questions, we decided to perform analysis examining how much correlation the answer to the different queries (controlled for the specific resident) had with each other.

Methods
Records of the monthly evaluations of all residents from the academic year 2015–16 from the core residency programs of pediatrics (PED), internal medicine (IM), Obstetrics and Gynecology (OB), family and community medicine (FM), psychiatry (Psych) and Surgery (Surg) were queried from a data base created and maintained with New Innovations™ software. These monthly
evaluations of all residents were obtained in a de-
deidentified manner (CD for Marshall, JK for University of Toledo) and shared as Excel worksheets. These work-
sheets were converted into csv files for importation into
the open source analysis package R. Both the University of Toledo and Marshall University IRB categorized the
study as IRB exempt.

Correlation and network analysis were performed on
only the Likert scale questions from each institution.
As obstetrics and gynecology at UT used many
different forms for different rotations, we did not
attempt correlation or network analysis with this
residency. In addition to the Likert scale questions,
all of the residency evaluation forms had summative
questions and allowed for the entry of free text com-
ments. These fields were not analyzed in this paper.

Data analysis of these Likert questions was per-
formed with R, [7] primarily using statistics functions
present in the base package as well as a heatmap/net-
work plot called heatmap2 in the gplots package, [8]
which allowed for Pearson correlations (r values) to be
performed on a matrix of numerical data derived from
answers to questions on the monthly evaluation form.
In addition to restricting ourselves to Likert style ques-
tions, questions that were often unanswered (i.e. > 50%)
on the forms were deleted from subsequent analysis.
With some residency programs with high percentages
of complete records (i.e., no missing values), we also
performed principal component analysis using the
princomp function of the stats package within R [7].

Some of the residency programs provided data that
was suitable for principal component analysis. Rather
than try to substitute average values into records to
replace missing values, we restricted this analysis to
those programs where a large percentage of resident
evaluations were complete for every question. These
were restricted to Marshall where there was a paucity
of resident-level specific Likert questions. As the
Marshall Psychiatry program had very few complete
records, we restricted the principal component analysis
to Marshall IM, OB, PEDS, Surg and FM.

Results
Details for each residency program are provided in
Table 1. We noted that variation in scores between
residents varied considerably with all programs but
that variations in the scores for specific residents on
the different questions were quite similar. This is
illustrated by the histogram detailing the responses
to the third Marshall IM question that shows values
across the answer scale. In contrast, looking at the
scores on the 29 questions for the third resident in
our list (no identification inherent in this), all the
scores given were exactly 4 (Figure 1). The range of
scores per de-identified resident are shown in Figure
2 and presented as boxplots.

Much of the Likert data correlated well within the
resident evaluations, especially at the Marshall pro-
grams (Table 2, Figure 3). This is apparent even for
those programs with the smallest number of ques-
tions such as Marshall FM and Surg. This can be seen
on the correlation matrices shown in Table 2 where
average Pearson correlation coefficients ranged from
a low of 0.44 in the UT_Psych data set to average
values exceeding 0.9 in Marshall Surg and IM.

From these correlation values, network analysis
was also performed. For Marshall PED, there were
essentially three groups of highly correlated
answers. For Marshall IM, the degree of correla-
tion was so high as to make specific networks
somewhat speculative. For Marshall OB, there were
also three large grouping. In contrast, the
answers to questions were only modestly corre-
lated. Marshall Psych, like IM, had such highly
correlated answers that groupings were difficult
to visualize. Although Marshall Surgery had only
7 questions, the answers to all correlated quite well
with each other. For the UT programs, the corre-
lations were less high as previously noted. UT_PED showed a very highly correlated group
of 4 questions with much less correlation noted
between the remaining questions. UT_IM had
a number of pockets of highly correlated answers
as did UT_FM and UT_Psychiatry. UT Surgery
had a large group of highly correlated questions
with a subset (approximately 15%) of questions
that correlated poorly with all others.

When we performed principal component analy-
asis, we found that > 95% of the variance could be
explained by the first 5 principal components in all
programs. In all but FM, almost all of the variance

| Program     | Marshall University | University of Toledo |
|-------------|---------------------|-----------------------|
|             | Evaluations (N)     | Likert Questions (N)  |
| Psychiatry  | 64                  | 65                    |
| Surgery     | 491                 | 7                     |
| Internal Medicine | 728 | 29                  |
| Pediatrics  | 453                 | 22                    |
| Obstetrics  | 742                 | 16                    |
| Family Medicine | 372  | 9                    |

University of Toledo Obstetrics used a number of different evaluation forms for different rotations and evaluators. Correlation
analysis was not felt possible with the limited sample size for each form.
was contained within the first 2 principal components (Figure 4).

Although it was tempting to speculate that the longer questionnaires had greater correlation values, this was not necessarily the case. In fact, the Marshall Surgery program with only 7 Likert questions had the highest average correlations between the answers. Next highest at Marshall was the Psych questionnaire with 65 Likert questions. Conversely at UT, Psych had the lowest average correlation between answers to Likert questions while having the highest number of Likert questions within that institution.

Discussion

The evaluation of post-graduate physician trainees in ACGME programs has undergone an infusion of rigor and structure during the past decades. Some of this can be attributed to the introduction of the six core competencies approximately 20 years ago [9].

While increased granularity of evaluation is clearly desirable, our data would suggest that we see diminishing returns. As demonstrated in Figure 3, one must speculate that if the answers to Likert questions correlate so highly with one another, there must be some degree of redundancy. Without proof, we would conjecture that the thoughtfulness of response might be compromised by a task perceived as onerous by the responding faculty member [10]. An alternative is that faculty, when evaluating the clinical performance of residents, simply don’t or aren’t able to discriminate between performance in core competencies. As an example, if the resident is perceived to be high performing, that may be generalized to all
domains of the evaluation. In contrast, if the resident is perceived as poorly performing, that may also be generalized to all components of their evaluation. Certainly, the high degree of correlation between questions suggests that residents receive modest benefit in the identification of specific areas of competence that could be improved.

When we examined our data between the different institutions, it was apparent that the correlation coefficients varied more within the UT than Marshall programs, suggesting that not only were the correlations between answers to questions lower but the variability of said correlations was greater. This suggests that more information might be present in the UT evaluations; unfortunately, without a true gold standard it is difficult to confirm this. As we examine the specific Likert scale responses between departments, it was also clear that the scores given to the Marshall Psych residents were on average lower than that for the other core residency programs. This discrepancy is almost certainly due to the Psych residency being new and our cadre of Psych residents limited to the PGY1 and 2 years for the academic year under consideration. It seems that the Likert questions alone for the Psychiatry form at Marshall would require in excess of one hour to complete. That said, it would be unlikely that faculty would

\[
\begin{array}{|c|c|c|c|}
\hline
\text{Program} & \text{Marshall University} & \text{University of Toledo} \\
& \text{Mean} & \text{SD} & \text{Mean} & \text{SD} \\
\hline
\text{Psychiatry} & 0.88 & 0.04 & 0.45 & 0.34 \\
\text{Surgery} & 0.91 & 0.05 & 0.60 & 0.30 \\
\text{Internal Medicine} & 0.91 & 0.03 & 0.66 & 0.22 \\
\text{Pediatrics} & 0.83 & 0.08 & 0.46 & 0.37 \\
\text{Obstetrics} & 0.84 & 0.07 & 0.59 & 0.16 \\
\text{Family Medicine} & 0.74 & 0.12 & 0.59 & 0.16 \\
\hline
\end{array}
\]
Figure 3. Correlation heat maps with network groupings shown for Marshall (top Panel) and UT (bottom Panel) residency programs. Color code for correlations shown on graphs. Network groupings fixed by R software using the command heatmap.2 within the package gplots. The results depict correlation between items with highly correlated items coded light green, as demonstrated by the line of identity for identical items that are displayed from bottom left to top right, to less well correlated items that are depicted in in darker shades.

Figure 4. Principal component analysis of Marshall Residency programs. Note that for all programs, the first principal component contained the vast majority of information. In fact, for all programs, more than 95% of variance was contained within first 5 principal components. For FM, the distribution of significant variance went up to the 5th component whereas for other programs, it was limited to the first 1 or 2.
uniformly fill such a form out with diligence and thoughtfulness. The IM questionnaires from both institutions as well as Psychiatry, FM and Surgery from UT each had nearly ½ that number of Likert questions and would also appear to be perceived as onerous. In contrast, Marshall Surgery and Marshall Family Medicine managed to evaluate their residents with much briefer forms.

Following the analysis presented in this manuscript, the authors have concluded that residency evaluation forms should be simplified. It would seem that an evaluation form focused on the six basic competencies along with an additional question or two regarding other relevant activities (e.g., teaching medical students [11]) would represent an improvement over the larger, unwieldy forms employed by some of our residency programs. We would also suggest that such forms be evaluated on an ongoing basis; although comparison over time is an appealing advantage for maintaining consistency, it may be more important to modify these forms to reflect the actual axes of evaluation that are performed by evaluators which may, in fact, change year to year.

Disclosure statement

No potential conflict of interest was reported by the authors.

References

[1] Flynn JT. Legacy of the Flexner Report. Science. 1965;150:554.

[2] Swing SR. Assessing the ACGME general competencies: general considerations and assessment methods. Acad Emerg Med. 2002;9:1278–1288.

[3] Carraccio C, Englelander R. Evaluating competence using a portfolio: a literature review and web-based application to the ACGME competencies. Teach Learn Med. 2004;16:381–387.

[4] Choe JH, Knight CL, Stiling R, et al. Shortening the miles to the milestones: connecting EPA-based evaluations to ACGME milestone reports for internal medicine residency programs. Acad Med. 2016;91:943–950.

[5] Meier AH, Gruessner A, Cooney RN. Using the ACGME milestones for resident self-evaluation and faculty engagement. J Surg Educ. 2016;73:e150–e7.

[6] ACGME Program Requirements for Graduate Medical Education in Nephrology (Internal Medicine). 2017. Available from: https://www.acgme.org/Portals/0/PFAssets/ProgramRequirements/148_nephrology_2017-07-01.pdf

[7] R: A language and environment for statistical computing. 2018. Available from: https://www.R-project.org/

[8] gplots: various R programming tools for plotting data. 2019. Available from: https://CRAN.R-project.org/package=gplots

[9] Barone JE. The ACGME competencies. Curr Surg. 2002;59:349–350.

[10] Blankush JM, Shah BJ, Barnett SH, et al. What are the associations between the quantity of faculty evaluations and residents’ perception of quality feedback? Ann Med Surg (Lond). 2017;16:40–43.

[11] Weissman MA, Bensinger I, Koestler JL. Resident as teacher: educating the educators. Mt Sinai J Med. 2006;73:1165–1169.