# Quantification of the Effectiveness of a Residency Program Using the Resident In-Service Examination

**Claudiu V. Cotta, MD, PhD[1], Deborah J. Chute, MD[1], and Karl S. Theil, MD[1]**

## Abstract

This study describes a quantitative tool in the assessment of residency programs, in which national ranking of residents after the resident in-service examination in postgraduate year 4 is compared to that in postgraduate year 1. The relationship between training and changes in ranking, resident in-service examination results before and after training in specific areas are also compared. To illustrate the use of this novel approach, data from a large residency program were analyzed. The 70 residents were ranked as a postgraduate year 1 group at the 50th national percentile. As postgraduate year 4 residents, they were ranked at the 59th percentile, a significant ($P < .003$) improvement. There was moderate correlation between performance in postgraduate year 1 and that in postgraduate year 4 (0.61); however, initial ranking was no indication of the final ($R^2 = .34$), with the exception of high performers. Training in specific areas improved ranking, demonstrating association between training and performance. In conclusion, the effectiveness of training provided by a residency program can be quantified using the resident in-service examination. This should provide a quantitative tool in the assessment of postgraduate programs.

## Introduction

The standards of the Accreditation Council for Graduate Medical Education (ACGME) and its accreditation system insure that most residency programs provide their trainees with the skills necessary to practice and to pass certification examinations.[1] However, in addition to personal or geographic preferences, there is significant variability in program size, curriculum, patient volume, faculty number, and so on, which makes assessing the effectiveness of residency programs difficult. To guide candidates and accreditation organizations, indices such as percentage of trainees passing specialty examinations,[2] employment placement, publications and even resident surveys have been used, but assessing a training program remains a challenge.

In contrast to the limited means to assess programs, there are multiple methods to monitor individual trainee progression. Instituted by ACGME, the Milestones program[3] guides and follows the professional development of trainees. Resident in-service examinations (RISEs) predate the Milestones and are standardized tests aiming to quantify the accumulation of theoretical and practical knowledge.[4] For example, the RISE developed by the American Society for Clinical Pathology has been administered to pathology residents since 1993.[5] The questions are generated by experts, are

---

[1] R.J. Tomsich Pathology and Laboratory Medicine Institute, Cleveland Clinic, Cleveland, OH, USA

**Corresponding Author:**
Claudiu V. Cotta, R.J. Tomsich Pathology and Laboratory Medicine Institute, 9500 Euclid Avenue/L-30, Cleveland, OH 44122, USA.
Email: cottac@ccf.org

updated, and at least partially emulate the certification examination administrated by the American Board of Pathology (ABP). The RISE results are reported both as absolute numbers and as percentiles ranking the test taker in his national postgraduate year (PGY) peer group.[6] The consistent test format throughout training allows trainees and program directors to monitor the progression of individuals. The Milestones incorporate performance on standardized tests as a reliable (and recommended) method of assessment, especially as senior resident performance on the RISE correlates with outcomes of ABP examinations.[6,7] However, the question remains: While almost all residents progress during training (in absolute scores on standardized tests,[5] Milestones and from a subjective point of view), how to quantify the contribution of the program?

We use the RISE to measure the effectiveness of a specific residency program. This is a shift from its use in the assessment of individuals to providing quantitative information on the program. Professional growth requires an individual effort and effective training. Averaging the results of a large number of trainees reduces the variability induced by differences in drive, test-taking ability, and previous training. In consequence, comparing the national peer group ranking of a group of residents at the end of their training to that at the beginning should quantify the impact of the program. We apply this to a large residency program hoping to answer a few questions. First, can we detect changes in peer group ranking after training? Second, are changes an exclusive function of the initial ability of the resident? Third, is specific training associated with changes in ranking? To answer these questions, results of the RISE taken by 70 residents as PGY4 were compared to those in PGY1. Absolute numerical scores were ignored, focus was exclusively on the national peer group ranking of residents, with the idea that training in an ineffective program should lead to lower ranking as PGY4 than as PGY1 and the reverse, better ranking should be achieved in an effective training environment. To investigate the link between training and changes in ranking, we took advantage of a particularity of the program: training in transfusion medicine (TM) and in hematopathology (HP) was provided during PGY2. In consequence, ranking in these fields as PGY1 was used as baseline, while the changes in PGY2 were associated with training. This largely ruled out the possibility that the changes at the end of PGY4 were exclusively due to individual preparation, without a significant contribution of the training program. Overall, we illustrate the notion that changes in aggregate percentile ranking from PGY1 to PGY4 of all the residents as a group can measure the effectiveness of the program.

## Methods

### Participant Selection

After internal review board approval in 2017, 70 residents training in the same Anatomic and Clinical Pathology (AP/CP) program between 2006 and 2017 were identified. The
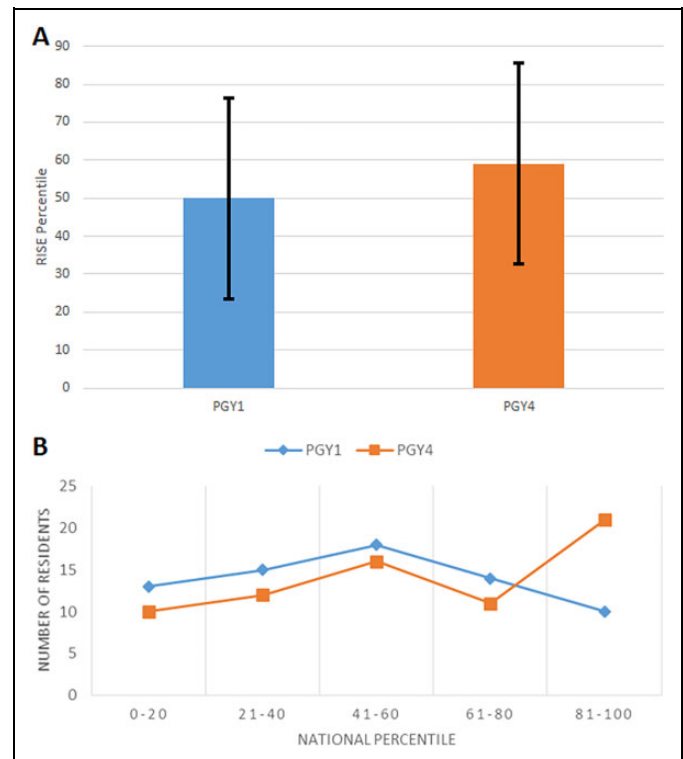


**Figure 1.** A, Average resident in-service examination (RISE) percentile. The increase of 9.3% in resident rank between postgraduate year (PGY1 and 4) is statistically significant ($P = .003$). B, Number of residents by quintile ranking. Postgraduate year 1 had a normal distribution, while PGY4 has a skew toward high performers.

AP- or CP-only trainees were not included. The results on the RISE were anonymized. Analyzed were the overall national percentile ranking on the tests taken during the PGY1 and 4 and the percentile ranking in TM and HP as PGY1 and PGY2, between 2006 and 2017.

### Statistical Analysis

Anonymized data were stored, analyzed, and visually represented using Microsoft Excel (Microsoft, Seattle, Washington). Statistical tests used included average, standard deviation (SD), and paired *t* test. The difference in the performance on RISE for each resident was calculated by deducting the percentile ranking as PGY1 from the percentile as PGY4.

## Results

### National Ranking as Postgraduate Year 1

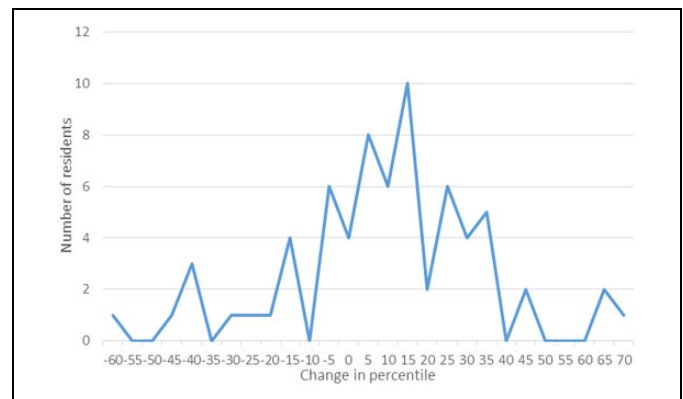The mean national peer group percentile ranking of PGY1 residents was 50 (26.3; Figure 1A). With 70 individuals enrolled, it is not surprising that the performance of the group was no different from that of the national reference group of PGY1 residents. The SD is a reflection of the wide variation in performance, 13 residents ranking in the bottom 20%, while 16 in the top 80 (Figure 1B). The distribution of percentile
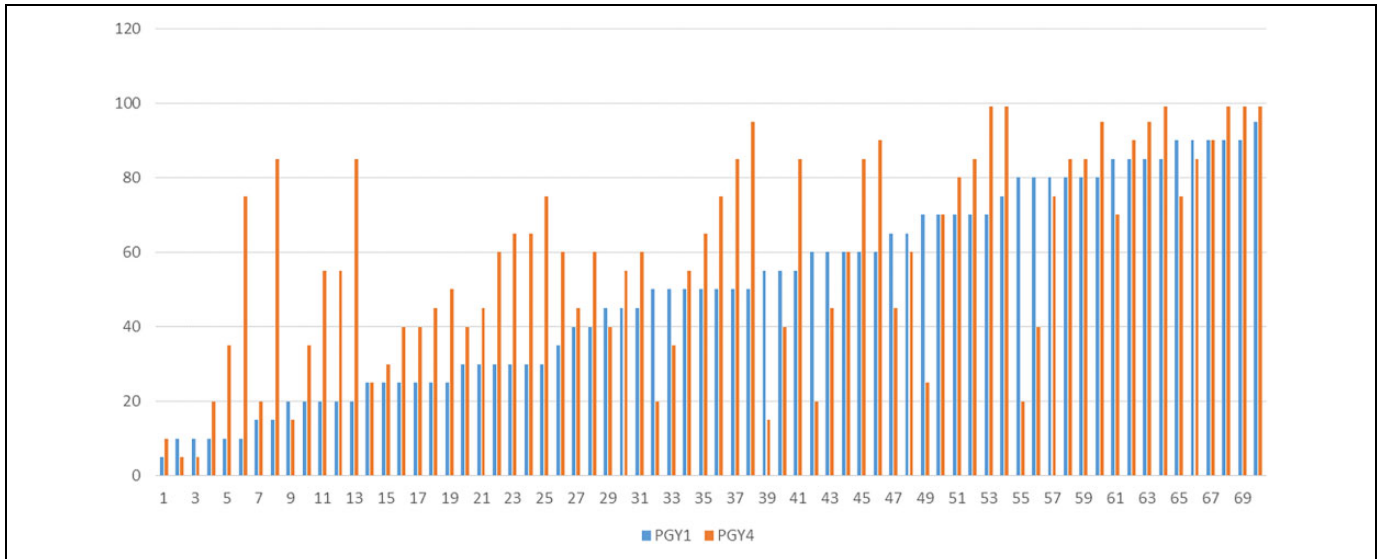
**Figure 2.** Individual resident in-service examination (RISE) percentile in postgraduate year (PGY) 1 and 4.
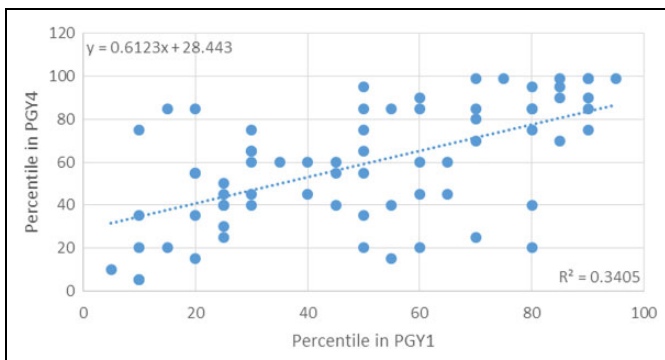


**Figure 3.** Correlation between ranking as postgraduate year (PGY) 1 and PGY4. There is a moderate correlation between ranking as PGY1 and PGY4 (slope 0.61); however, the coefficient of determination is low at 0.34. Dotted line = line of best fit.

rankings was normal, symmetrical, and roughly approximating that of a bell-shaped curve (Figure 1B).

## National Ranking as Postgraduate Year 4

The mean percentile ranking of PGY4 residents was 59 (27.6; Figure 1A). Like for PGY1, ranking ranged from very low to very high, but the distribution curve was not symmetrical, this time there was a significant skew toward high performers (Figure 1B), defined as residents ranked in the upper national quintile (80%-100%).

## Changes in Ranking

The individual percentile ranking as PGY1 and PGY4 of participants is displayed in Figures 2 and 3. The average change in ranking was 9.27 percentiles, but the SD was very large at 24.7, indicating variability in performance, even if the distribution of the values was normal, with a bell-shaped curve centered



**Figure 4.** Changes in percentile ranking between postgraduate year (PGY) 4 and PGY1. Overall, a normal distribution "bell curve."

around the value of 9 (Figure 4). The difference between the ranking as a PGY4 and that as a PGY1 was statistically significant ($P < .003$; Figure 1A). If any change in performance is taken into consideration, 48 residents improved their performance, 4 had a similar performance, while 18 performed worse as PGY4 than as PGY1, a resident was 2.7 times more likely to improve than to fall in ranking. If only changes larger than 5% are considered, the differences between improving and declining performance residents are more significant: 40 improved, 18 stayed the same while 12 declined in ranking, indicating that a resident was 3.3 times more likely to improve than to decline.

## Correlation Between Performance as Postgraduate Year 1 and Postgraduate Year 4

Initial performance was not a strong predictor of final ranking, as shown in Figures 2 and 3, significant improvement in ranking being achieved across all quintiles. The lower quintile (0%-20%) had a significant number of PGY1 residents who
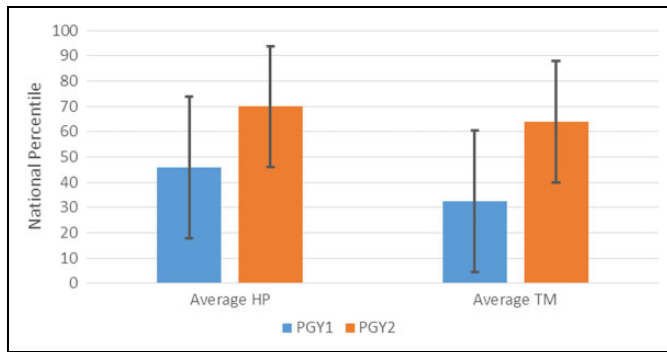
**Figure 5.** The impact of training in specific areas on national ranking. Training in these disciplines was offered at the beginning of postgraduate year (PGY) 2. Ranking was performed based on the performance on the resident in-service examination (RISE) administered in the second half of PGY2. The differences in ranking before and after training are significant in both disciplines ($P < .001$ for both). HP indicates hematopathology; TM, transfusion medicine.

improved as PGY4s, the proportion of improved-no change-declined being 10-0-3, with at least 2 residents becoming high performers. In fact, this quintile registered most significant gains, but performance was very heterogeneous, a few residents (6 of 13) remaining low performers even as PGY4, registering no or minimal improvements. Most impressive was the second quintile, with 14-1-0, no resident declining in performance, one staying the same, while 14 improved. The middle quintile (41%-60%) had a mixed performance, 10-1-7, while the worst performing fourth quintile (61%-80%) had an even number of improvers over nonchanging or worsening performers 7-1-6, with losses in ranking more severe than the gains registered (Figures 2 and 3). As expected, the high performers continued to be ranked highly and when losses in ranking were registered, they were not severe. Overall, the slope of the linear regression equation was 0.61, indicating a moderate correlation between the rankings as PGY1 and PGY4, but the coefficient of determination was low at 0.34, indicating that the performance as a PGY4 of a particular PGY1 resident was difficult to predict.

### Impact of Specific Training

No experience in HP and TM was correlated with below average ranking in these disciplines (Figure 5) on the test administrated in the second half of PGY1. Rotations at the beginning of PGY2 resulted in improvements in national ranking in HP (30 percentiles) and TM (25 percentiles) on the test administrated at the end of PGY2 (Figure 5), clearly indicating the association between training and performance.

## Discussion

To quantify the effectiveness of a program, we compared the national peer ranking of a group of PGY4 residents to that in PGY1. The idea was that residents in effective programs should improve their ranking, while those with ineffective training

should decline. In other words, differences in resident ranking as a group are at least partially dependent on program effectiveness. Individual differences in drive, test-taking ability, or personal histories were counterbalanced by the large number of residents involved (70).

We detected a difference of over 9 percentiles in ranking, but smaller changes may not be detectable in smaller programs. This could be circumvented by multigenerational data. Sure, programs change over the periods necessary to acquire data, but some changes may impact the residents at national level and certain parameters with a major impact on training may change very slowly: number of patients/cases/procedures, ratio faculty/trainee, location, patient population, affiliation, and so on. In addition, if differences are too small to detect with data from 20 to 30 residents, maybe the program has an average impact on resident training, neither beneficial nor detrimental.

We also investigated the correlation between individual ranking as PGY1 and as PGY4. The conclusions are mixed: Moderate correlation exists, but predicting individual PGY4 performance based on ranking as PGY1 is impossible, with the exception of very high performers. This is further argument for the role of the training program. An absent correlation would have been in direct contradiction with intuitive and statistical observations showing that individuals with strong performance tend to continue to perform well. A very strong correlation (basically preserved ranking from PGY1 to PGY4) would have shown unexpected uniformity in the effectiveness of training and total absence of individual factors, casting doubts over the accuracy of the data. Overall, the mixed results are not only realistic but also encouraging: Initial lackluster performance can be significantly enhanced, while strong performance can be maintained.

The possibility that progress was exclusively consequence of individual efforts (even when the cohort was sufficiently large to make this unlikely) was investigated. The impact of specific training was clearly demonstrated by the significant upgrade in ranking. These changes were more significant than those in overall ranking as PGY4, probably due to lack of standardization of the curriculum at national level: Some residents have HP or TM training in PGY1, some in PGY3, or maybe the time between training and testing allowed for information to be forgotten. Regardless, it is obvious that significant improvement in specific areas is linked to training, supporting the notion that changes in overall ranking after residency are impacted by program effectiveness.

The main limitation of the study is the possibility that the RISE may not cover relevant or current information and that important aspects of training are not addressed in this test. In the absence of an alternative method of standardized assessment of residents and with experience indicating that RISE performance correlates with that on the ABP certification examination,[6] we feel that the data generated through the RISE should be taken into consideration.

The main finding is that candidates and regulatory agencies can obtain quantitative information on the effectiveness of a specific program. Residents gaining in national ranking (becoming better trained than their peers) is a good indicator that the program

has dedicated faculty and resources, regardless of the subjective impression of inspectors, faculty, or trainees. The opposite is also obvious: When trainees become less competitive in spite of the programs' stated goals, impressive appearances on paper, or high morale, alarm bells should sound. A search of the literature shows that this type of quantitative assessment has not been used before by residency programs in any specialty, a somewhat surprising finding, as in-service examinations of varying types are widespread and as data are easily analyzed and interpreted. Implementation of this type of analysis in a consistent and transparent manner could have a significant impact on how residency programs are accredited and funded. One could imagine accreditation agencies withdrawing support for programs who repeatedly fail to quantitatively demonstrate their effectiveness and consistently lag behind the other programs in that specialty. Hospitals may choose to divert scant resources across specialties, encouraging effective programs, and decreasing the resources allocated to ineffective ones. For candidates, ranking programs on the match rank order list would become a more objective endeavor with the quantitative knowledge described earlier. These data should allow effective programs to answer the question every recruit should ask: Why would I train in this program and not in a competing one?

## Conclusion

Changes from PGY1 to PGY4 in aggregate national percentile ranking of residents as a group can measure the effectiveness of the training.

### Authors' Note

C.V.C. and K.S.T. contributed equally to this work.

### Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

### Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

### References

1. Accreditation Council for Graduate Medical Education. http://www.acgme.org. Accessed May 30, 2018.
2. The American Board of Pathology Newsletters. http://www.abpath.org/index.php/newsletters. Accessed May 30, 2018.
3. Accreditation Council for Graduate Medical Education. http://www.acgme.org/What-We-Do/Accreditation/Milestones/Overview. Accessed May 30, 2018.
4. American Society for Clinical Pathology. http://pathologylearning.org/residents/rise. Accessed May 30, 2018.
5. McKenna BJ. The American Society for Clinical Pathology resident in-service examination: does resident performance provide insight into the effectiveness of clinical pathology education? *Clin Lab Med*. 2007;27:283-291.
6. Rinder HM, Grimes MM, Wagner J, et al. Senior pathology resident in-service examination scores correlate with outcomes of the American Board of Pathology certifying examinations. *Am J Clin Pathol*. 2011;136:499-506.
7. Accreditation Council for Graduate Medical Education. The pathology milestone project. http://www.acgme.org/Portals/0/PDFs/Milestones/PathologyMilestones.pdf. Accessed May 30, 2018.