Transforming Resident Assessment: An Analysis Using Deming's System of Profound Knowledge

Eric J. Warm, MD, Benjamin Kinnear, MD, Matthew Kelleher, MD, MEd, Dana Sall, MD, MEd, and Eric Holmboe, MD

Abstract

W. Edwards Deming, in his System of Profound Knowledge, asserts that leaders who wish to transform a system should understand four essential elements: appreciation for a system, theory of knowledge, knowledge about variation, and psychology. The Accreditation Council for Graduate Medical Education (ACGME) introduced the milestones program as a part of the Next Accreditation System to create developmental language for the six core competencies and facilitate programmatic assessment within graduate medical education systems. Viewed through Deming's lens, the ACGME can be seen as the steward of a large system, with everyone who provides assessment data as workers in that system. The authors use Deming's framework to illustrate the working components of the assessment system of the University of Cincinnati College of Medicine's internal medicine residency program and draw parallels to the macrocosm of graduate medical education. Successes and failures in transforming resident assessment can be understood and predicted by identifying the system and its aims, turning information into knowledge, developing an understanding of variation, and appreciating the psychology of motivation of participants. The authors offer insights from their experience for educational leaders who wish to apply Deming's elements to their own assessment systems, with questions to explore, pitfalls to avoid, and practical approaches in doing this type of work.

In *The New Economics for Industry, Government, Education*, W. Edwards Deming[1] introduced the System of Profound Knowledge, asserting that leaders who wish to *transform* a system should understand four essential elements: appreciation for a system, theory of knowledge, knowledge about variation, and psychology.

The Accreditation Council for Graduate Medical Education (ACGME) created the milestones program as a part of the Next Accreditation System.[2] Part of the rationale for introducing milestones was to create developmental language for the six core competencies and to facilitate programmatic assessment within residencies and fellowships.[3–5] This was a significant change, in essence asking residency programs to *transform* assessment.[2] Viewed through Deming's[1] lens, the ACGME can be seen as the steward of a large system, with everyone who provides assessment data as workers in that system. Transformation has not come easily, however. These efforts have met with variable resistance, prompting some to challenge the validity of the competency-based frameworks on which this work relies.[6–8]

In this article, we use Deming's System of Profound Knowledge to analyze the components of the well-developed assessment system of the University of Cincinnati College of Medicine's internal medicine residency program.[9,10] Although we were not aware of Deming's work when we started implementing our new system in 2011, it is clear in retrospect that our successes and failures could have been predicted had we applied his model at the outset. We hope that by sharing Deming's insights and drawing parallels to the larger macrocosm of graduate medical education (GME), we can assist others who wish to maximize the success of their own assessment systems.

Case Study

We will use the case of Resident N to frame our discussion of Deming's four elements: Resident N is in month 31 of our three-year internal medicine residency program. He has already matched into a fellowship program. He is on schedule to graduate in five months, but his performance has been inconsistent to date. (Identifying details of this case have been changed.)

Deming's System of Profound Knowledge

Appreciation for a system

Deming defined a *system* as "a network of interdependent components that work together to try to accomplish the aim of the system."[1] In 2011, we created and implemented a system of assessment based on entrustment of observable practice activities (OPAs).[9,10] OPAs are discrete workplace-based assessment elements rated on a five-level entrustment scale (1 = critical deficiency, 2 = direct supervision, 3 = indirect supervision, 4 = no supervision, 5 = aspirational performance). OPAs can be content specific and vary from rotation to rotation (e.g., manage pancreatitis), or they can be process related and be conserved over rotations (e.g., manage an interdisciplinary team). As of September 2018, we have created more than 450 OPAs, each of which is mapped to ACGME subcompetencies.[11] Faculty members, peers, and allied health professionals independently provide thousands of entrustment ratings of OPAs for residents over the course of their residencies, and these data are

Please see the end of this article for information about the authors.

Correspondence should be addressed to Eric J. Warm, University of Cincinnati College of Medicine, 231 Albert Sabin Way M.L. 0557, Cincinnati, OH 45267-0557; e-mail: warmej@ucmail.uc.edu; Twitter @CincyIM.

Acad Med. 2019;94:195–201.
First published online October 16, 2018
doi: 10.1097/ACM.0000000000002499

tracked over time.[9,10] Resident N, for example, has accumulated 3,156 faculty subcompetency assessments in his first 31 months of residency. Our aim for this system is to use subcompetency entrustment data for formative feedback, summative decisions, and regulatory reporting. Program directors and the Clinical Competency Committee (CCC) monitor these data in real time and meet with residents periodically to review their progress and develop interventions and learning plans as necessary.

Deming[1] felt that a system must be managed, as it will not manage itself, and the bigger the system is, the more difficult it is to manage. Since 2011, our OPA system has collected hundreds of thousands of data points from nearly 1,000 assessors at multiple training sites for 200 residents; as such, it represents a very large system. Deming[1] also believed that everyone in a system should share a clear understanding and commitment to the aim of the system. When we began collecting the OPA data, faculty members, residents, and even the CCC had little appreciation for the aim of the system—or for the system itself. Faculty members viewed their assessment duties as summative in nature, and they often felt like they were passing a grade on to a resident rather than collecting data for formative feedback. Residents—despite having access to data for thousands of possible assessment points—continued to score our program low on the annual ACGME Resident/Fellow Survey question "Are you satisfied with feedback after assignments?"[12] In addition, the program director and CCC gathered and reported OPA data, but they lacked sophisticated ways of interpreting these data for the purpose of feedback and professional development. Failure of the people in our system to understand the interconnectedness of its components put our system at risk. The same could be said of other training programs that collect and report data to the ACGME. How many of these programs understand the aim of the milestones system or view it as a system of interdependent parts? Without such clarity, Deming[1] believed, the components of a system will act in their self-interests and destroy the system.

Deming also held that a system cannot understand itself and needs guidance from the outside. It is unlikely that any resident, faculty member, CCC, or program director would change their behavior without outside forces weighing in.[13] In the microcosm of our residency, the "outside forces" of program leadership needed to create clear communication strategies with all actors in the system, using knowledge, variation, and psychology (examples described below) as the basis of this messaging. To address similar issues in the GME macrocosm, national leaders should recognize the disconnects present in the current milestones system and ensure that all participants understand and move toward clearly defined, shared aims of accrediting bodies.[2]

Theory of knowledge

Deming[1] cautioned that information is not knowledge: Rules of interpretation or theory must be applied to draw meaningful conclusions about a system. *Operational definitions* must be created to apply theory to a data set. These are explicit procedures with which measurements are taken, such that the individuals taking the measurements have a shared mental model of goals and objectives for data collection.[1] Once meaningful data are generated following the operational definitions, rules of interpretation can be applied.

Previous studies have shown that supervisors have "built-in" entrustment scales, often using themselves or other context-specific elements as their default assessment framework.[14–16] These built-in scales can be highly variable among faculty members because they commonly use "self" as the frame of reference (i.e., "How I would perform this task?").[15] Operational definitions to promote shared mental models of essential tasks are therefore necessary for building a validity argument for learner assessments. A validity framework put forth by Kane uses a series of inferences to connect frontline assessment (scoring) to data use (implications).[17] Frontline assessors must understand the operational definitions of the constructs being assessed, as well as how assessment tools relate to these constructs (scoring inference).[18] CCCs and residency program leaders need to have a shared mental model of how to interpret scores to relate them to real-world performance (extrapolation inference) and make summative decisions (implications inference).[17]

To further explore the concepts put forth by Kane, consider assessment systems that evaluate core competencies using a norm-referenced Likert scale that is the same for residents in all postgraduate year (PGY) classes, such as the nine-point scale in Chart 1 where ratings of 1–3 = "unsatisfactory," 4–6 = "satisfactory," and 7–9 = "superior." Use of this type of assessment scale, although common, can be challenging for assessors. If, for example, a resident performs an accurate medical interview but performs worse on the physical exam, how should the assessor judge the resident's overall performance on the patient care competency?

If core competencies for residents were to be evaluated by the CCC in this manner, residents' scores might be reported as they are in Chart 1. At first glance, it appears from these data that all residents have been rated at least satisfactory and most have been rated superior, so there should be no cause for concern with respect to promotion or graduation decisions. Taking a closer look at Chart 1 highlights several difficulties, however. First, anchors such as "superior" may be interpreted by assessors as either norm *or* criterion referenced. In the case of norm referencing, it is not possible that most residents are superior because they are being compared with one another. In the case of criterion referencing, programs may have residents who are truly superior performers, but how do the assessors know how well their residents are performing compared with residents in other programs?[19,20] Second, the data do not discriminate between residents across PGY classes, despite likely progression of resident skills over time. Third, this system does not clearly identify struggling learners. For example, Resident 2's rating on each competency is at least "satisfactory," yet this resident's overall average score is one standard deviation (SD) below the other PGY-1 residents' scores. Whether this is significant depends on which theoretical lens educators use to interpret the scores. As these examples show, information (data) is not necessarily knowledge. To generate meaningful assessment data when using such a scale, one has to change what one asks of assessors (operational definitions), the rules applied to interpret assessment data (theory), or both.

The type of scale used may be less important than the operational definitions applied to the scale, however.

Chart 1

Sample Rating Scale and Representative Ratings of Residents Across Three PGY Classes for the Six ACGME Core Competencies

Rating Scale

Unsatisfactory			Satisfactory				Superior	
1	2	3	4	5	6	7	8	9

Residents by class	Representative ratings by competency[a]						Average overall score[a]
	PC	MK	PBLI	ICS	PROF	SBP	
PGY-1 class							
Resident 1	7.0	6.8	6.5	7.3	7.9	7.1	7.1
Resident 2	6.2	6	5.9	5.8	6.3	6.4	6.1
Resident 3	8.0	7.2	6.9	7.4	7.1	7.0	7.3
Class average	7.1	6.7	6.4	6.8	7.1	6.8	6.8
PGY-2 class							
Resident 4	8.1	8.2	7.2	7.5	7.9	7.8	7.8
Resident 5	8.2	8.0	8.1	7.3	7.6	7.9	7.9
Resident 6	7.9	7.5	7.1	7.3	7.5	7.6	7.5
Class average	8.1	7.9	7.5	7.4	7.7	7.8	7.7
PGY-3 class							
Resident 7	7.5	7.2	7.2	7.1	7.8	7.3	7.4
Resident 8	7.1	7.3	7.4	8.2	8.1	8.2	7.7
Resident 9	8.1	8.0	7.2	7.5	7.9	7.8	7.8
Class average	7.6	7.5	7.3	7.6	7.9	7.8	7.6

Abbreviations: ACGME indicates Accreditation Council for Graduate Medical Education; PGY, postgraduate year; PC, patient care; MK, medical knowledge; PBLI, practice-based learning and improvement; ICS, interpersonal and communication skills; PROF, professionalism; SBP, systems-based practice.
[a]Rating scale, as illustrated above: 1–3 = unsatisfactory, 4–6 = satisfactory, 7–9 = superior.

Ordinal scales using quality-based adjectival anchors such as "satisfactory" or "superior" require several layers of translation, as shown in the above example, and these scales are not well aligned with the rating task being asked of the faculty members or other assessors. The bottom line for any scale is whether its users share a clear understanding of the operational definitions (i.e., have a shared mental model) for that scale.

In our residency program, we chose a different approach to generate more meaningful data and create a higher degree of construct alignment between assessors and the assessment tool.[19] To do this, we abandoned the norm-referenced Likert scales we used for broad themes (e.g., patient care) in favor of a competency-based entrustment framework of discrete, observable skills (i.e., OPAs).[9] In our OPA system, assessors rate skills using the five-level entrustment scale (described above) and a simple criterion-based question: At what level does the assessor trust the learner to perform the skill? Entrustment levels for each skill are directly linked to mapped ACGME subcompetencies and collected over time. (It should be noted, though, that use of entrustment scales does not relieve the need for faculty training.)

For example, Figure 1 displays Resident N's aggregate assessment data over the first 31 months of his residency. During the first 7 months, Resident N was not progressing to higher levels of entrustment (time A in Figure 1). When the CCC met, its investigation into narrative data (submitted by raters along with entrustment-level data) suggested that Resident N was struggling across multiple competency domains. The program director and faculty created a significant and direct intervention for Resident N, and he began to show progressive entrustment to the level of indirect supervision (time B). The CCC used these data to predict that Resident N would be able to perform the supervisory role of a senior resident and promoted him to PGY-2. Resident N continued to show progressive entrustment until an acute drop occurred at months 22–24 (time C). Review by the CCC showed that Resident N was on a basic science rotation and needed direct supervision to carry out lab-related tasks with which he had minimal experience. This was felt to be an appropriate entrustment level for this rotation. When Resident N returned to clinical rotations in month 25 (time D), he continued to show progressive entrustment over time, and the CCC concluded that he was on track toward an on-time graduation.

As this case illustrates, by creating operational definitions (criterion-based OPAs) and applying rules of interpretation—that is, aggregate values on OPA assessments should show progressive entrustment over time, and narrative comments should justify scores—the data in Figure 1 become knowledge. We also collect similar information by subcompetency and OPA. Developing a shared understanding of what constructs are represented when data are viewed in different ways (e.g., by OPA, by subcompetency, or in aggregate) is important for turning information into knowledge, and for determining how CCCs should aggregate and view data.

Knowledge about variation

Deming[1] suggested that converting information into knowledge can be difficult without understanding variation. Generally speaking, there are two broad categories of variation within a system. *Common cause variation* has no assignable source and is inherent to the system itself.[1,21,22] An example would be obtaining slightly differing weights when weighing a person daily on a given scale. A slight variation in the readings may be caused by measurement error inherent in the scale and by subtle expected differences in a person's intake and volume status over time. *Special cause variation* is due to an assignable source that has an effect on the system.[21,22] For example, if a person being weighed daily holds a brick one day and steps onto the scale, the reading that day will likely increase similar to the weight of the brick. This variation has

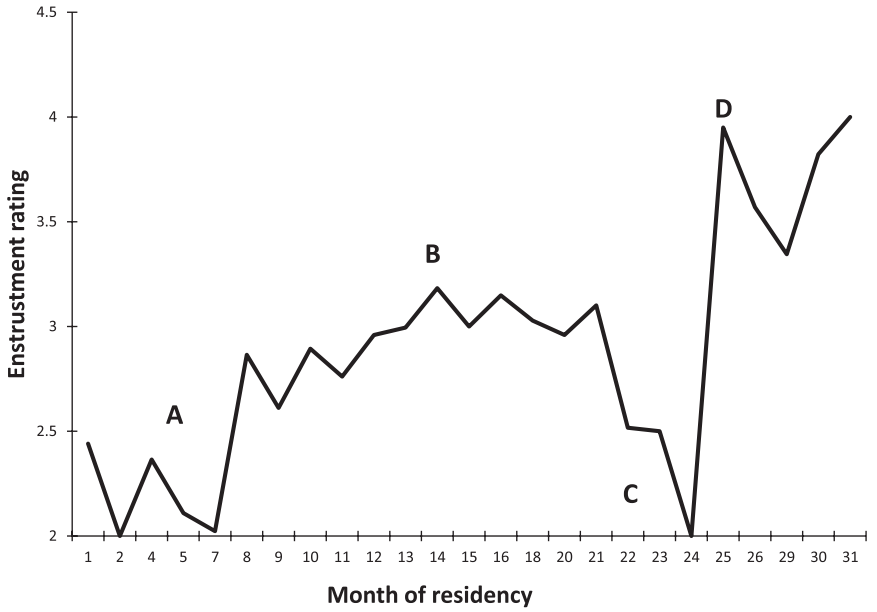

Figure 1 Aggregate entrustment scores over time for Resident N. This graph displays 3,156 faculty assessment data points across all ACGME subcompetencies for Resident N during the first 31 months of the 3-year internal medicine residency program at the University of Cincinnati College of Medicine. To reveal trends in the data, time is measured in months on the x-axis (month 1 = July PGY-1, month 13 = July PGY-2, month 25 = July PGY-3; some months are not represented because of vacation or failure of assessors to complete evaluations). Assessment scores are reported on the y-axis as entrustment ratings using a five-level scale, where 1 = critical deficiency, 2 = direct supervision, 3 = indirect supervision, 4 = no supervision, 5 = aspirational performance. Time points: A = failure to progress at the start of residency due to resident-specific issues, B = stable rise in entrustment after intervention, C = acute decline in entrustment associated with a basic science research rotation (minimal experience with lab-related tasks), D = high performance associated with area of interest/fellowship after graduation. Abbreviations: ACGME indicates Accreditation Council for Graduate Medical Education; PGY, postgraduate year.

an assignable cause—the brick. Special cause variation is neither inherently good nor bad, but it represents change to a system that demands investigation. One might think of special cause variation as *signal* and common cause variation as *noise*. A common struggle for educators within a system is to identify which assessment data represent signals (e.g., scores indicating a resident who is struggling) and which represent noise (e.g., scores reflecting nonsignificant ups and downs).[23]

Multiple techniques can be used to identify variation in the data. First, data are easier to interpret in graphical rather than tabular form.[24] For example, imagine the interpretation challenges if we had presented Resident N's 3,156 data points as a spreadsheet instead of a graph (Figure 1). Second, placing time on the x-axis of a graph can reveal trends in the data that may otherwise be hidden. Third, statistical rules can help distinguish the signal of special cause variation (when a system has truly changed) from the noise of usual or common cause variation.[22,24–27] Simple run charts and more sophisticated control charts (also called "process behavior charts"[22,24–27]) combine these three concepts by showing time on the x-axis, the measure of choice on the y-axis, and a mean or median line to help track progress over time.

To measure and track variation, our residency program uses regression modeling with over 600,000 historical assessment data points to produce entrustment expected scores.[28] This process calculates the level at which a typical resident would be entrusted under a given set of circumstances (rotation, assessor, time of year, etc.), and we plot this expected score against the resident's raw entrustment score. Because we anticipate that most learners will gain competence, and assessment scores will rise over time (rather than reach or remain at a stable mean or median), the statistical rules we use in creating our run charts and control charts will not help us accurately identify special cause variation. Therefore, we also convert the raw entrustment score into a standard score (z score), which shows the number of SDs away from the expected score that the resident's raw score falls. We then add the z score, which sometimes can be negative, to an arbitrary positive integer (3 in this example) for the sole purpose of eliminating negative numbers. These standard scores and the expected scores are plotted on control charts (see Figure 2), with standard scores outside the control limits (3 SDs above or below the mean) representing special cause variation, which must be investigated to understand the reason. Typical explanations include learner performance that is significantly different from expected performance (higher or lower), poor assessor performance (e.g., grade inflation, misunderstanding the scale criterion), or low numbers of data points collected over a certain time period. Returning to the theory of knowledge,[1] it is important that CCC members share an understanding of what constructs are represented by these control charts and how to interpret them. Faculty development to build expertise may be necessary prior to incorporating tools such as control charts into a GME system.

Plotting Resident N's standard score (z score + 3) on a control chart (Figure 2) shows multiple areas in which his data fall outside the control limits, indicating special cause variation. As described earlier, the CCC investigated these points as part of its usual workflow by reviewing frontline assessment data, narrative comments, and other measures of assessment (e.g., in-training exam scores and ambulatory evaluations) to determine reasons for the special cause variation. Resident N's struggles at time A were true signal rather than noise (i.e., common cause variation). His improvement with respect to the expected score is visible around time B. The drop at time C was also true signal but, as noted above, was due to his inexperience with basic science research. At time D, Resident N was performing at a level significantly above expected, with the majority of these assessments being made in the same clinical area as the fellowship he would be joining after graduation. In this example, understanding variation helped the CCC hone in on which time points to investigate. When combined with the theory of knowledge, an approach that applies knowledge about variation

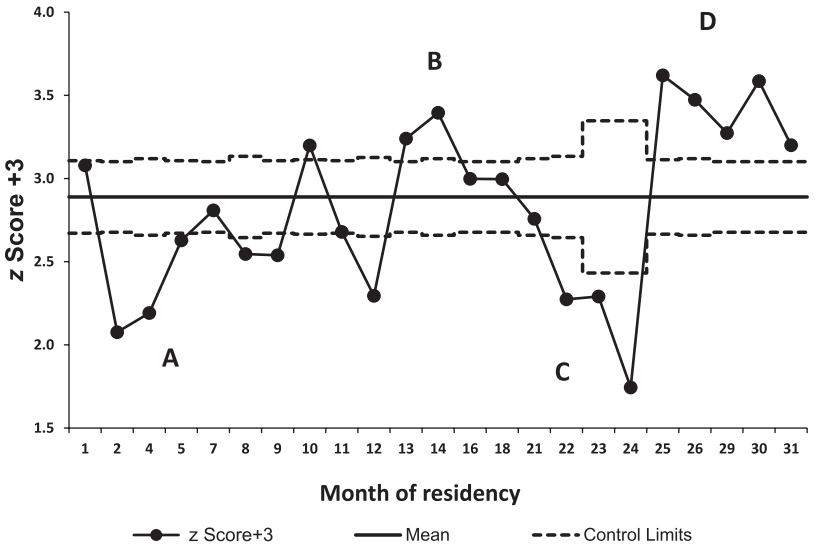

Figure 2 Resident N's standard scores over time displayed on a control chart. These data show the first 31 months of faculty assessments (n = 3,156) of Resident N aggregated across all ACGME subcompetencies, University of Cincinnati College of Medicine internal medicine residency program. The standard score, or *z* score, is the number of standard deviations away from the expected score that a resident's raw entrustment score falls. A random integer—3 in this example—is added to the *z* score to avoid negative numbers. The solid center line is the mean; standard scores above and below this line represent raw scores above and below the expected score. The dotted lines represent the upper and lower control limits (by definition, three standard deviations above or below the mean). To reveal trends in the data, the x-axis shows time in months (month 1 = July PGY-1, month 13 = July PGY-2, month 25 = July PGY-3; some months are not represented because of vacation or failure of assessors to complete evaluations). Assessment scores (as *z* score + 3) are reported on the y-axis. Time points: A = failure to progress at the start of residency due to resident-specific issues, B = stable rise in entrustment after intervention, C = acute decline in entrustment associated with a research rotation (minimal experience with lab-related tasks), D = high performance associated with area of interest/fellowship after graduation. Abbreviations: ACGME indicates Accreditation Council for Graduate Medical Education; PGY, postgraduate year.

can be important for managers of any GME assessment system where data are numerous and faculty/CCC time is limited.

Psychology

Deming[1] explained that all systems that include human interactions require an understanding of psychology. He suggested that people desire to learn, develop, and be connected to others. A primary responsibility of those managing a system is to align the core motivations of individuals with the aims of the system. Managers often resort to extrinsic motivators such as rewards, mandates, or punishments to encourage and guide behavior. These techniques work to a degree, increasing compliance and generating short-term results.[23] However, when people become dependent on extrinsic motivation, their intrinsic motivation is diminished.[29] Behavior guided by intrinsic motivation, or driven by internal rewards, is generally associated with better outcomes.[30] Self-determination theory (SDT), an approach to the psychology of motivation, outlines three innate psychological needs for development of intrinsic motivation: autonomy (control of one's own behavior), competence (feeling of mastery for a specific action), and sense of relatedness (feeling connected to others).[31]

We have used both extrinsic and intrinsic motivators in our assessment system. When our new assessment system was introduced in 2011, only about 60% of faculty members completed evaluations on time. We successfully lobbied the Department of Internal Medicine to withhold teaching practice payments from assessors who were not compliant, following a well-established practice at our institution,[32] and on-time faculty member evaluations increased to greater than 90%. However, these short-term wins came at a cost as some faculty members were disgruntled by the consequences imposed. Although most evaluations were completed on time, the quality varied widely.

Since then, we have made changes to develop the intrinsic motivation of our assessors. To foster a sense of autonomy, we engaged each division in the creation of OPAs, giving ownership of each rotation's assessment form to the faculty assessors.[9] We also incorporated a "not observed" option on the form, so individual faculty assessors never feel forced to assess skills they did not observe.[9,10] To increase assessor competence, we organized a system to evaluate all assessments completed by our faculty.[33] We call this "feedback on the feedback," and we use it to identify those who struggle to understand assessment best practices. To build a sense of relatedness, we have shown faculty members how their assessment data fit into the global picture and are used to help residents improve. Figure 3 illustrates these three psychological needs from the SDT literature,[31] with some key points on how each relates to our assessment system.

As our case study illustrates, high-quality assessment data provided by motivated assessors can lead to a formative plan for improvement. Resident N's struggles were identified early in PGY-1 (time A) by motivated faculty who understood our assessment system and were invested enough to compile rich and nuanced numerical and narrative data about their observations. Together, faculty and Resident N used this information to create an improvement plan tailored to the resident's needs. We believe that faculty members who are intrinsically motivated to gather and deliver assessment data are more likely to provide valuable information than are faculty members who are extrinsically motivated.

Conclusion

Our assessment system has gone through many iterations since its inception, driven by the desire to improve assessment for our learners. However, our initial efforts were not guided by systems theory. As a result, our learning curve was steep, and our failures were significant. Now, we see our efforts as part of a system for which the aims are clear; we look for ways to turn information into knowledge; we have a deeper and more nuanced understanding

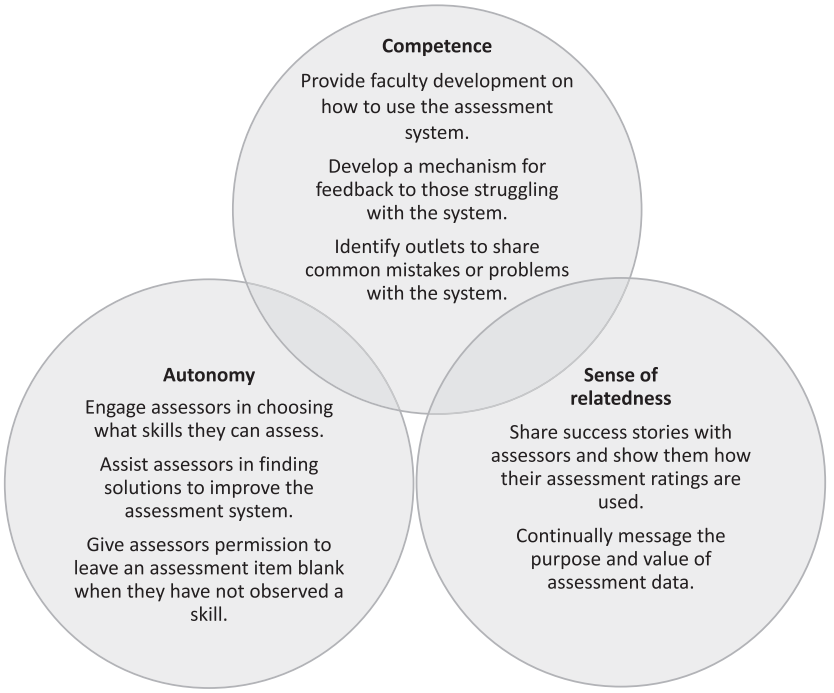

Figure 3 Self-determination theory's[30] three innate psychological needs for the development of intrinsic motivation—autonomy, competence, and sense of relatedness—as they relate to the assessment system of the University of Cincinnati College of Medicine internal medicine residency program.

regarding variation; and we try to appreciate the psychology of motivation of assessors in the system to maximize value.[1] We have had some success with our assessment approach,[9,10] although more study is needed to understand how best to optimize formative and summative value for learners. Appendix 1 summarizes the key points for educational leaders who wish to apply Deming's framework to their own assessment systems, with questions to explore, potential pitfalls to avoid, and practical approaches in doing this type of work. We hope our experience, viewed through the lens of Deming's framework, will help others as they begin or continue their assessment journey.

Funding/Support: None reported.

Other disclosures: None reported.

Ethical approval: The use of milestones data was approved by the Institutional Review Board of the University of Cincinnati, study no. 2014-2042.

Previous presentations: These concepts were presented as part of a mini-course at the Accreditation Council for Graduate Medical Education meeting, Orlando, Florida, March 11, 2017.

E.J. Warm is professor of medicine and program director, Department of Internal Medicine, University of Cincinnati College of Medicine, Cincinnati, Ohio; ORCID: https://orcid.org/0000-0002-6088-2434.

B. Kinnear is assistant professor of medicine and pediatrics and associate program director, Department of Internal Medicine, University of Cincinnati College of Medicine, Cincinnati, Ohio.

M. Kelleher is assistant professor of medicine and pediatrics and associate program director, Department of Internal Medicine, University of Cincinnati College of Medicine, Cincinnati, Ohio.

D. Sall is assistant professor of medicine and associate program director, Department of Internal Medicine, University of Cincinnati College of Medicine, Cincinnati, Ohio.

E. Holmboe is senior vice president, Milestones Development and Evaluation, Accreditation Council for Graduate Medical Education, Chicago, Illinois; adjunct professor of medicine, Yale University, New Haven, Connecticut; and adjunct professor, Feinberg School of Medicine at Northwestern University, Chicago, Illinois.

References

1 Deming WE. The New Economics for Industry, Government, Education. 2nd ed. Cambridge, MA: MIT Press; 2000.
2 Nasca TJ, Philibert I, Brigham T, Flynn TC. The next GME accreditation system—Rationale and benefits. N Engl J Med. 2012;366:1051–1056.
3 Batalden P, Leach D, Swing S, Dreyfus H, Dreyfus S. General competencies and accreditation in graduate medical education. Health Aff (Millwood). 2002;21:103–111.
4 Van Der Vleuten CPM, Schuwirth LWT, Driessen EW, Govaerts MJB, Heeneman S. Twelve tips for programmatic assessment. Med Teach. 2015;37:641–646.
5 Holmboe ES, Rodak W, Mills G, McFarlane MJ, Schultz HJ. Outcomes-based evaluation in resident education: Creating systems and structured portfolios. Am J Med. 2006;119:708–714.
6 Witteles RM, Verghese A. Accreditation Council for Graduate Medical Education (ACGME) milestones—Time for a revolt? JAMA Intern Med. 2016;176:1599–1600.
7 Boyd VA, Whitehead CR, Thille P, Ginsburg S, Brydges R, Kuper A. Competency-based medical education: The discourse of infallibility. Med Educ. 2018;52:45–57.
8 Hong R. Observations: We need to stop drowning—A proposal for change in the evaluation process and the role of the clinical competency committee. J Grad Med Educ. 2015;7:496–497.
9 Warm EJ, Mathis BR, Held JD, et al. Entrustment and mapping of observable practice activities for resident assessment. J Gen Intern Med. 2014;29:1177–1182.
10 Warm EJ, Held JD, Hellmann M, et al. Entrusting observable practice activities and milestones over the 36 months of an internal medicine residency. Acad Med. 2016;91:1398–1405.
11 University of Cincinnati Internal Medicine. Resident education curriculum 2017–2018. http://med.uc.edu/docs/default-source/default-document-library/curriculum-2017-18-final.pdf?sfvrsn=0. Accessed September 10, 2018.
12 Accreditation Council for Graduate Medical Education. Resident/Fellow and Faculty surveys. https://www.acgme.org/Data-Collection-Systems/Resident-Fellow-and-Faculty-Surveys. Accessed September 23, 2018.
13 Ashforth BE, Anand V. The normalization of corruption in organizations. Res Organ Behav. 2003;25:1–52.
14 Kennedy TJ, Regehr G, Baker GR, Lingard L. Point-of-care assessment of medical trainee competence for independent clinical work. Acad Med. 2008;83(10 suppl):S89–S92.
15 Kogan JR, Hess BJ, Conforti LN, Holmboe ES. What drives faculty ratings of residents' clinical skills? The impact of faculty's own clinical skills. Acad Med. 2010;85(10 suppl):S25–S28.
16 Kogan JR, Conforti LN, Iobst WF, Holmboe ES. Reconceptualizing variable rater assessments as both an educational and clinical care problem. Acad Med. 2014;89:721–727.
17 Cook DA, Brydges R, Ginsburg S, Hatala R. A contemporary approach to validity arguments: A practical guide to Kane's framework. Med Educ. 2015;49:560–575.
18 Cook DA, Beckman TJ. Current concepts in validity and reliability for psychometric instruments: Theory and application. Am J Med. 2006;119(2):166.e7–e16.
19 Crossley J, Johnson G, Booth J, Wade W. Good questions, good answers: Construct alignment improves the performance of workplace-based assessment scales. Med Educ. 2011;45:560–569.
20 Rekman J, Gofton W, Dudek N, Gofton T, Hamstra SJ. Entrustability scales: Outlining their usefulness for competency-based clinical assessment. Acad Med. 2016;91:186–190.
21 Deming WE. Out of the Crisis. Cambridge, MA: MIT Press; 2000.
22 Langley GJ, Moen RD, Nolan KM, Nolan TW, Norman CL, Provost LP. The Improvement

Guide: A Practical Approach to Enhancing Organizational Performance. 2nd ed. San Francisco, CA: Jossey-Bass; 2009.
23 Baker K. Determining resident clinical performance: Getting beyond the noise. Anesthesiology. 2011;115:862–878.
24 Wheeler D. Understanding Variation: The Key to Managing Chaos. 2nd ed. Knoxville, TN: SPC Press; 2000.
25 Guthrie B, Love T, Fahey T, Morris A, Sullivan F. Control, compare and communicate: Designing control charts to summarise efficiently data from multiple quality indicators. Qual Saf Health Care. 2005;14:450–454.
26 Noyez L. Control charts, Cusum techniques and funnel plots. A review of methods for monitoring performance in healthcare. Interact Cardiovasc Thorac Surg. 2009;9:494–499.
27 Benneyan JC, Lloyd RC, Plsek PE. Statistical process control as a tool for research and healthcare improvement. Qual Saf Health Care. 2003;12:458–464.
28 Schauer D, Warm EJ. Measurable bias in resident assessment: The development of an expected entrustment score. Unpublished manuscript. 2018.
29 Deci EL, Koestner R, Ryan RM. A meta-analytic review of experiments examining the effects of extrinsic rewards on intrinsic motivation. Psychol Bull. 1999;125:627–668.
30 Wrzesniewski A, Schwartz B, Cong X, Kane M, Omar A, Kolditz T. Multiple types of motives don't multiply the motivation of West Point cadets. Proc Natl Acad Sci U S A. 2014;111(30):10990–10995.
31 Ryan RM, Deci EL. Self-determination theory and the facilitation of intrinsic motivation, social development, and well-being. Am Psychol. 2000;55:68–78.
32 Rouan GW, Wones RG, Tsevat J, Galla JH, Dorfmeister JW, Luke RG. Rewarding teaching faculty with a reimbursement plan. J Gen Intern Med. 1999;14:327–332.
33 Warm E, Kelleher M, Kinnear B, Sall D. Feedback on feedback as a faculty development tool. J Grad Med Educ. 2018;10:354–355.

Appendix 1

Practical Considerations for Applying Deming's[1] System of Profound Knowledge to a Resident Assessment System

Deming element	Summary[a]	Questions to explore	Pitfalls and considerations	Practical approaches
Appreciation for a system	A system is a network of interdependent parts working together toward a common aim.	• What is the assessment system's aim? • What components make up the assessment system? • How are these components interconnected? • Who is managing the assessment system?	• Lack of communication and of shared mental models/ operational definitions can hinder cooperation. • A system cannot understand itself—a team must be in charge of managing it.	• Prior to beginning the work, gather all members of the assessment team to plan the system as a whole. • Include systems thinking as part of faculty development efforts as the system rolls out. • Continuously review how each component of assessment fits into the work as a whole.
Theory of knowledge	Meaningful interpretations about a system can only be drawn when viewed through the lens of a theory. Not all information is knowledge.	• What theory informs the interpretation of assessment data? • What operational definitions are being used in the assessment system? • Has a shared mental model of data interpretation been developed?	• If learner successes or failures are not predicted by the information collected, then change the way data are collected, interpreted, or both.	• Develop a clear set of operational definitions during the system planning phase. • Share these definitions frequently with all assessors and learners. • Continuously measure and refine assessment methods and/or data interpretation to improve system performance.
Knowledge about variation	All systems contain variation that is inherent (common cause) and variation that has an assignable source (special cause).	• How can longitudinal assessment data be viewed to distinguish between types of variation? • What faculty development exists to teach about types of variation seen in the data? • What mechanisms exist for the education team to investigate sources of special cause variation?	• If the education team does not understand variation, there is risk of both overreacting and undercorrecting. • Identifying variation in assessment data does not improve the quality of the data. Robust programs of assessment are necessary to collect quality data.	• Identify or develop expertise in assessing, interpreting, and accounting for variation within the system. • Develop techniques (e.g., rater training) to reduce unnecessary variation in the system.
Psychology	Goals of the system should be aligned with motivation of people working within that system.	• What motivates people in the assessment system? • How can goals of the assessment system align with motivations of the people using it? • Which aspects of self-determination theory[30] would have the largest impact on increasing intrinsic motivation for people in the system?	• Relying solely on external motivators for people using the system can hinder their intrinsic motivation.	• Meet with assessors and learners to delineate the mix of internal and external drivers present in the system. • Be deliberate in choosing which drivers to emphasize, with particular attention to optimizing conditions where internal drivers can flourish.

[a]Summaries are quoted or adapted from Deming.[1]