Multicenter Validation of a Customizable Scoring Tool for Selection of Trainees for a Residency or Fellowship Program

The EAST-IST Study

Gabriel T. Bosslet[1], W. Graham Carlos III[1], David J. Tybor[2], Jennifer McCallister[3], Candace Huebert[4], Ashley Henderson[5], Matthew C. Miles[6], Homer Twigg III[1], Catherine R. Sears[1], Cynthia Brown[1], Mark O. Farber[1], Tim Lahm[1], and John D. Buckley[1]

[1]Division of Pulmonary, Critical Care, Sleep and Occupational Medicine, Indiana University School of Medicine, Indianapolis, Indiana; [2]Tufts University School of Medicine, Boston, Massachusetts; [3]Division of Pulmonary, Critical Care, and Sleep Medicine, The Ohio State University Wexner Medical Center, Columbus, Ohio; [4]Division of Pulmonary, Critical Care, Sleep, and Allergy, University of Nebraska Medical Center, Omaha, Nebraska; [5]Division of Pulmonary and Critical Care, Marsico Lung Institute, University of North Carolina Healthcare, Chapel Hill, North Carolina; and [6]Division of Pulmonary, Critical Care, Allergy, and Immunologic Disease, Wake Forest School of Medicine, Winston-Salem, North Carolina

ORCID ID: 0000-0002-7841-6298 (G.T.B.).

Abstract

Rationale: Few data have been published regarding scoring tools for selection of postgraduate medical trainee candidates that have wide applicability.

Objectives: The authors present a novel scoring tool developed to assist postgraduate programs in generating an institution-specific rank list derived from selected elements of the U.S. Electronic Residency Application System (ERAS) application.

Methods: The authors developed and validated an ERAS and interview day scoring tool at five pulmonary and critical care fellowship programs: the ERAS Application Scoring Tool–Interview Scoring Tool. This scoring tool was then tested for intrarater correlation versus subjective rankings of ERAS applications. The process for development of the tool was performed at four other institutions, and it was performed alongside and compared with the "traditional" ranking methods at the five programs and compared with the submitted National Residency Match Program rank list.

Results: The ERAS Application Scoring Tool correlated highly with subjective faculty rankings at the primary institution (average Spearman's $r = 0.77$). The ERAS Application Scoring Tool–Interview Scoring Tool method correlated well with traditional ranking methodology at all five institutions (Spearman's $r = 0.54$, 0.65, 0.72, 0.77, and 0.84).

Conclusions: This study validates a process for selecting and weighting components of the ERAS application and interview day to create a customizable, institution-specific tool for ranking candidates to postgraduate medical education programs. This scoring system can be used in future studies to compare the outcomes of fellowship training.

Keywords: education; internship and residency; job application

(Received in original form November 29, 2016; accepted in final form January 30, 2017)

Supported by the Association of Pulmonary and Critical Care Medicine Program Directors, American College of Chest Physicians, and an American Thoracic Society Education Research Award.

Author Contributions: G.T.B., W.G.C., D.J.T., J.M., C.H., A.H., M.C.M., H.T., C.R.S., C.B., M.O.F., T.L., and J.D.B. gave substantial contributions to the conception and design of the work, the acquisition, analysis, and interpretation of data. G.T.B drafted the work and all authors revised it critically for important intellectual content. All authors gave final approval of the version to be published, and all agree to be accountable for all aspects of the work in ensuring that questions related to the accuracy and integrity of any part of the work are appropriately investigated and resolved.

Correspondence and requests for reprints should be addressed to Gabriel T. Bosslet, M.D., M.A., Assistant Professor of Clinical Medicine, Fellowship Director, Division of Pulmonary, Critical Care, Sleep, and Occupational Medicine, Indiana University School of Medicine, 1120 West Michigan Street, CL 285, Indianapolis, IN 46202. E-mail: gbosslet@iu.edu

This article has an online supplement, which is accessible from this issue's table of contents at www.atsjournals.org

Ann Am Thorac Soc Vol 14, No 4, pp 517–523, Apr 2017

DOI: 10.1513/AnnalsATS.201611-938OC
Internet address: www.atsjournals.org

In 2016, over 35,000 medical school graduates applied to over 4,800 residency programs in the United States (1). Graduate medical training programs are tasked each year with selecting candidates for interview and ranking them through the National Residency Match Program (NRMP). The database used by these programs, the Electronic Residency Application Service (ERAS), gathers all trainee-specific data in a single place. Programs filter through these data to determine which candidates they believe are worthy of interview and potential ranking through the NRMP.

Training programs may receive ERAS applications for hundreds to thousands of candidates each year. Selecting from among this many applications is a laborious process. Much of the selection process entails considerable subjectivity regarding which applicants programs choose for interview or ranking. Many training programs have developed internal scoring tools for evaluation of applications; however, many of these tools lack objectivity. In addition, little has been published about how these tools were developed or validated (2, 3).

Because the ERAS application is standardized and contains the same variables for each candidate, it provides an opportunity for the development of an internally consistent process for objectively selecting among candidates for ranking. If the individuals selecting applicants for a given program can agree on the important variables and a process can fairly weight each of these variables, a standardized process for selection can be developed, which may more closely resemble the global assessment of those involved in the selection process and mitigate bias.

Due to the disparate nature of medical specialties and the various missions of individual training programs, a universal, single scoring system will not work for all residencies or fellowship training programs. Each scoring system would need to be tailored to the mission and training culture of the individual training program. As a result, the *process* by which the scoring tool is developed and internally validated becomes the generalizable element of the tool and of applicable utility for a wider audience.

We developed and validated a process for selecting and weighting components of both the ERAS application and interview day for use in ranking applicants for potential selection to graduate medical education training programs. This process creates a repeatable and evaluable scoring method for evaluating the ERAS application, the ERAS Application Scoring Tool, or ERAS Application Scoring Tool (EAST). Combining this with the similarly created Interview Scoring Tool (IST) creates a reliable scoring tool for applications and interviews (the EAST-IST). This tool represents a data collection process by which programs can more rigorously and reliably evaluate their selection criteria over time. It would also allow for quantitatively measuring fellowship inputs (candidates) to compare with the outputs of fellowship (graduating fellows) in future studies.

This article describes the EAST-IST development and weighting process, and validates the resulting tool at five pulmonary and critical care medicine training programs to compare its performance over an interview cycle with a "usual method" of ranking candidates.

Methods

This study was approved by the institutional review board at Indiana University (Indianapolis, IN; primary institution protocol no. 1405123506) and considered exempt at the University of Nebraska (Omaha, NE; institutional review board no. 562-15-EX). An institutional agreement was obtained with The Ohio State University (Columbus, OH), Wake Forest School of Medicine (Winston-Salem, NC), and the University of North Carolina (Chapel Hill, NC).

Process Description

(*See* Appendix E1 in the online supplement for a "how-to" document that outlines in more detail the steps of this process.)

The ERAS application is comprised of 26 distinct variables for all applicants. These variables can be hierarchically categorized into five groups: Academic History; Written Lauds; Leadership and Awards; Research and Publications; and Miscellaneous (Figure 1). The 15 members of the primary institution's fellowship committee iteratively agreed on which components of the ERAS application and interview day factors were important in the ranking process for their institution (those selected are highlighted in *gray* in Figure 1). These variables were used as the basis of our weighting process.

In the variable weighting process, each member of the selection committee individually allotted 100 "priority points" to each level of the hierarchy (Appendix E1b is a sample-weighting tool, and Figure 2 demonstrates how the weighting points were used to calculate the variable weights). In other words, each member allotted a total of 100 points between the EAST and the IST, to represent the perceived relative importance of each of these two components in the process. They then allotted 100 points between Academic History, Written Lauds, Leadership and Awards, Research and Publications, and Miscellaneous. Then, each individual allotted 100 points to each of the variables contained within each of these categories. All 15 members of the selection committee completed the weighting tool independently, thereby reducing undue influence on each other's selected weights. The structure of the weighting tool is such that individual members do not know their exact weights for each variable until the calculation is completed.

After the weighting tools were completed, author G.T.B. collected the forms and calculated the variable weights. These weights were unique to each individual rater (*see* Figure 2 for an explanation and sample calculation). The final program EAST-IST weights were determined by calculating the mean of each of the variables among all of the members of the weighting group. These program weights were combined with a weight multiplier to develop the score for each variable response option (*see* Appendix E1c for response options and multiplier values).

Evaluation of Intrarater Agreement for the EAST Tool

To see how the EAST tool compared with the traditional evaluation process for ERAS applications, the 15 members of the primary institution's fellowship committee ranked 10 random, deidentified ERAS applications from the 2012 interview cycle from 1 to 10 in descending order, with 1 being the most desirable (henceforth referred to as the "gestalt ranking"). This was done by each

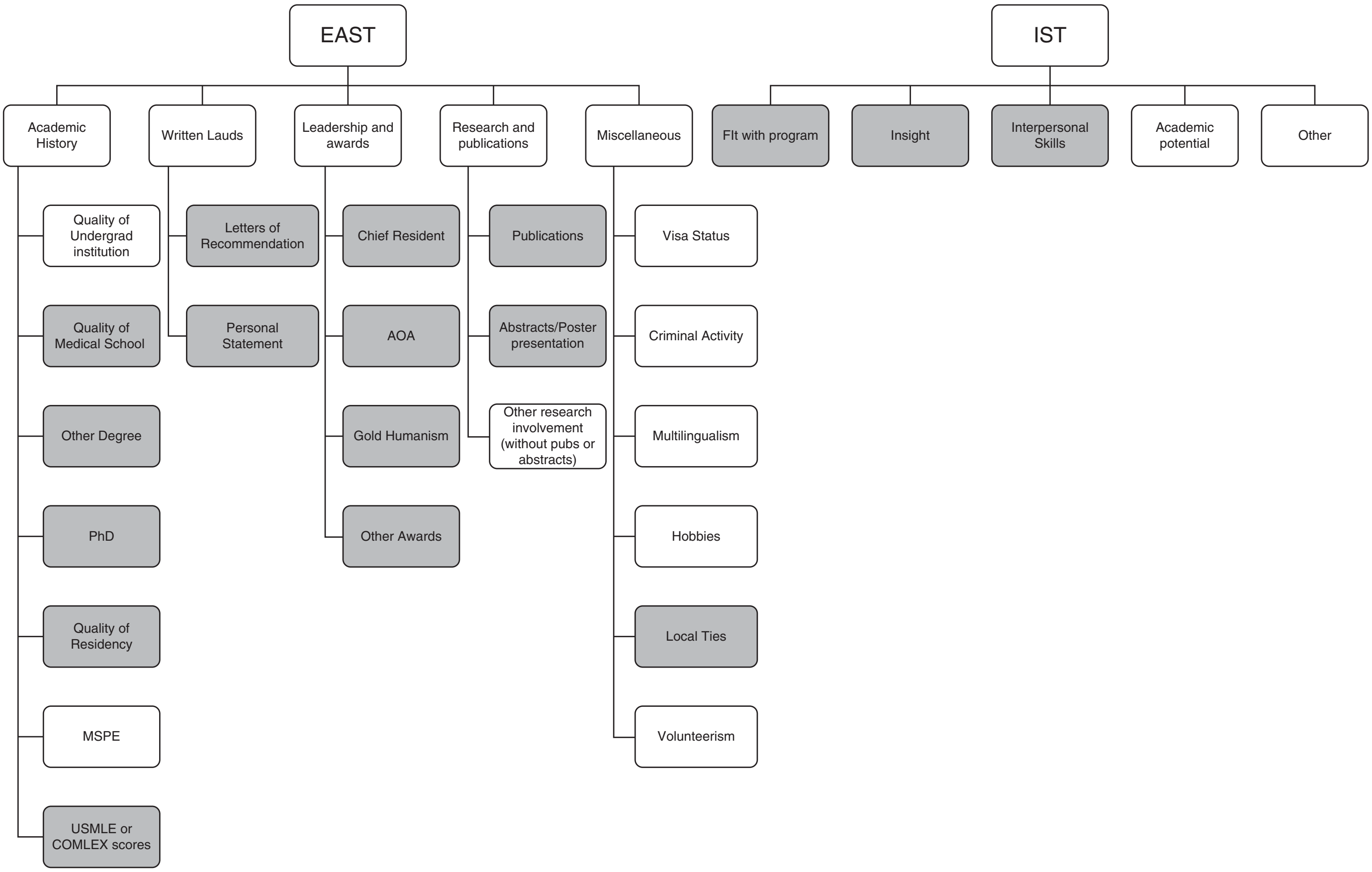

Figure 1. The U.S. Electronic Residency Application System (ERAS) application and interview day variables, categorized into a weighting hierarchy. The ERAS Application Scoring Tool (EAST) variables are fixed on the ERAS application and are not changeable. The Interview Scoring Tool (IST) variables are at the discretion of the training program (those listed here are done so as an example). The variables *shaded gray* are the ones selected by the primary institution for their EAST-IST. AOA = Alpha Omega Alpha; COMLEX = Comprehensive Osteopathic Medical Licensing Examination; MSPE = Medical Student Performance Evaluation; PhD = doctor of philosophy; USMLE = U.S. Medical Licensing Exam.

weighting group member individually, generating 15 separate ranked lists.

After 4 weeks (enough time for memory washout), each member individually scored each of these same applications again, this time using their weighted EAST tool (*see* Appendix E2 for a sample EAST form). The weights used in this process were the individual's weights only, to account only for *their* preferences in ranking. Members were unaware of their variable weights and were blinded to the scoring. The results of the EAST tool process were another 15 separate 1–10 rankings of the 10 applications. For each member of the committee, we then assessed intrarater reliability by Spearman's rank correlation coefficient, comparing the applications' ranks from the gestalt system to the ranks derived from each reviewer's EAST weights.

Evaluation of the Program EAST-IST Rank List versus Usual Method Rank List

Next, the EAST-IST method was validated by comparing it to the traditional method of applicant ranking in the 2014 interview season at the primary institution. Each candidate was interviewed by six faculty members who varied based upon the interview day. In 2014, we interviewed 32 candidates for 6 fellowship positions. The traditional method included a five-point Likert scale with the following four domains: Application and Academic Record, Interpersonal Skills, Fit with the Program, and Insight (*see* Appendix E3 for the traditional method scoring tools for each participating institution). This tool was completed by all interviewers immediately after the interview, and generated a score (sum of the average Likert responses) for each candidate. This score was used to compile an ordered list that was used as the starting point for our committee's rank list meeting, during which the final rank list was agreed upon.

In addition to our traditional method of rank list development (described previously here), interviewers also completed the EAST (before the interview day) and ISTs (immediately after the interview) for each candidate. The EAST-IST data were collected electronically and the EAST-IST scores were calculated by G.T.B. *after* submission of the rank order list to the NRMP. No committee members, including the study authors, were aware of the EAST-IST variable weights or scores until they were tabulated after the rank list submission. We then compared the

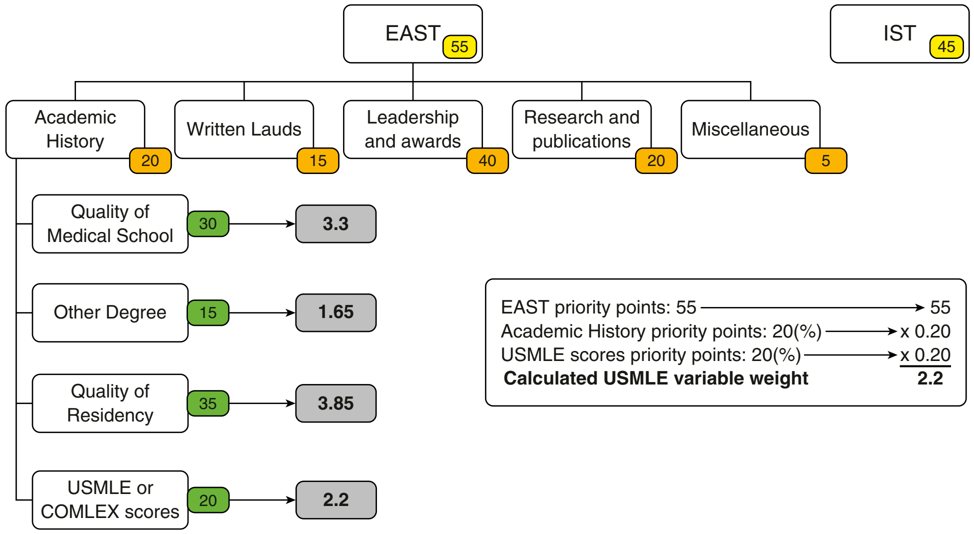
EAST 55
IST 45
Academic History 20
Written Lauds 15
Leadership and awards 40
Research and publications 20
Miscellaneous 5
Quality of Medical School 30
3.3
Other Degree 15
1.65
Quality of Residency 35
3.85
USMLE or COMLEX scores 20
2.2
EAST priority points: 55 → 55
Academic History priority points: 20(%) → x 0.20
USMLE scores priority points: 20(%) → x 0.20
Calculated USMLE variable weight 2.2

Figure 2. Sample calculation of the U.S. Electronic Residency Application System (ERAS) Application Scoring Tool–Interview Scoring Tool (EAST-IST) weighting tool. This figure represents one person's hypothetical calculated variable weights. Each level of the hierarchy is represented by the different *colored boxes*. In the weighting process, each member of the committee is asked to allocate 100 "priority points" to the components of each level of the hierarchy. Committee members may use any numbers, so long as each hierarchy level adds up to 100. For example, in the first level (*yellow boxes*), this committee member has allotted 55 points to the EAST tool and 45 points to the IST. The variable weight calculation is shown in the *box*. The levels of the hierarchy are multiplied to create a single weight for each variable (calculated weights for the variables are in the *gray boxes*). Please note that not all variables within the EAST or IST are shown for space reasons. COMLEX = Comprehensive Osteopathic Medical Licensing Examination; USMLE = U.S. Medical Licensing Exam.

submitted NRMP rank order list to the rank list generated by the EAST-IST using Spearman's rank correlation coefficient.

Multicenter Validation

Finally, we performed the EAST-IST development and validation at four other pulmonary and critical care programs during the 2015 interview season. The characteristics of each program are shown in Table 1. Using the same iterative process described previously here, each program selected their own EAST-IST variables to be included, and then invited selected faculty members (often the fellowship committee) to perform the weighting of these variables.

Program directors were allowed to select the faculty members they thought most appropriate to participate in the EAST-IST variable selection and weighting. Author G.T.B. calculated the programs' weights. Each program completed the EAST-IST scoring in addition to their traditional method of applicant evaluation (each traditional method can be found in Appendix E3). Program directors and faculty members were blinded to the EAST-IST weights. After each programs' NRMP rank lists were submitted, the mean EAST-IST scores for each applicant were calculated and candidates were ranked according to the EAST-IST score. We then compared each program's submitted NRMP rank order list to the rank list generated by the EAST-IST using Spearman's rank correlation coefficient.

Data Collection and Statistical Analysis

EAST and IST data were collected and managed using Research Electronic Data Capture (REDCap) hosted at Indiana University (4). We used Stata v13 (StataCorp, College Station, TX) for all statistical analysis, using two-sided statistical tests with an α level of 0.05.

Table 1. Characteristics of the Participating Pulmonary and Critical Care Training Programs, including the Primary Development Program (Program 1)

Program	No. of Fellows Matched (in the Year of Study)	No. of Faculty Participating in the Study	No. of Candidates Ranked in 2015	EAST Evaluations per Ranked Candidate	IST Evaluations per Ranked Candidate
1	6	15	26	6.0	6.0
2	4	7	32	2.2	1.9
3	4	9	43	3.3	3.3
4	6	9	39	3.9	4.5
5	4	7	36	2.6	2.4

Definition of abbreviations: EAST = U.S. Electronic Residency Application System Application Scoring Tool; IST = Interview Scoring Tool.

Results

Primary Program EAST Tool Intrarater Agreement

We first compared each of the 15 faculty members' gestalt rankings with the EAST rankings on the 10 randomly chosen applications (using only their own variable weights). For each individual faculty member, Spearman's rank correlation coefficient ranged from 0.59 to 0.93, with a mean of 0.77.

Primary Program Comparison of the EAST-IST with the Traditional Method of Candidate Selection

Figure 3 demonstrates the scatter plot of the EAST-IST ranking process versus the submitted NRMP rank list, derived via the traditional method (program 1 is the primary program's development results). The Spearman's rank correlation coefficient was 0.77. The EAST-IST was able to select 10 of the top 11 candidates selected via the traditional method. Of candidates ranked in the top tertile by the traditional method, 10 of 11 fell into the top tertile by the EAST-IST ranking process, a sensitivity of 91%; the corresponding specificity was 95%.

Multicenter Validation—EAST-IST versus Traditional Methods

Table 1 demonstrates the characteristics of the participating programs. Figure 3 demonstrates the scatter plots for all participating programs. The Spearman's rank correlation coefficients were 0.54, 0.72, 0.84, and 0.65. The ability of the EAST-IST to correctly identify which candidates were ranked in the top tertile by the traditional methods was different in each program, with sensitivities of 69, 47, 92, and 58%, and specificities of 84, 73, 96, and 79%, respectively.

Discussion

The process of selecting candidates to a postgraduate medical education program is fraught with subjectivity (5). To our knowledge, this is the first study that validates a process for weighting and scoring applicants to postgraduate medical education programs with external validity. The EAST tool demonstrated high intrarater correlation with gestalt application rankings, and the EAST-IST correlated well with the NRMP rank lists generated by traditional methods at five pulmonary and critical care medicine fellowship programs.

Little has been published regarding the process of developing postgraduate medical application scoring tools. In 1989, an application scoring form and interview scoring form were described from George Washington University Medical Center (2), demonstrating the utility of a scoring system for uniformity and objectivity at a radiology program at single institution. In 2006, a scoring tool was developed at a single orthopedic surgery residency, and was shown to be more effective at predicting in-training exam scores, board scores, and appointment to chief resident

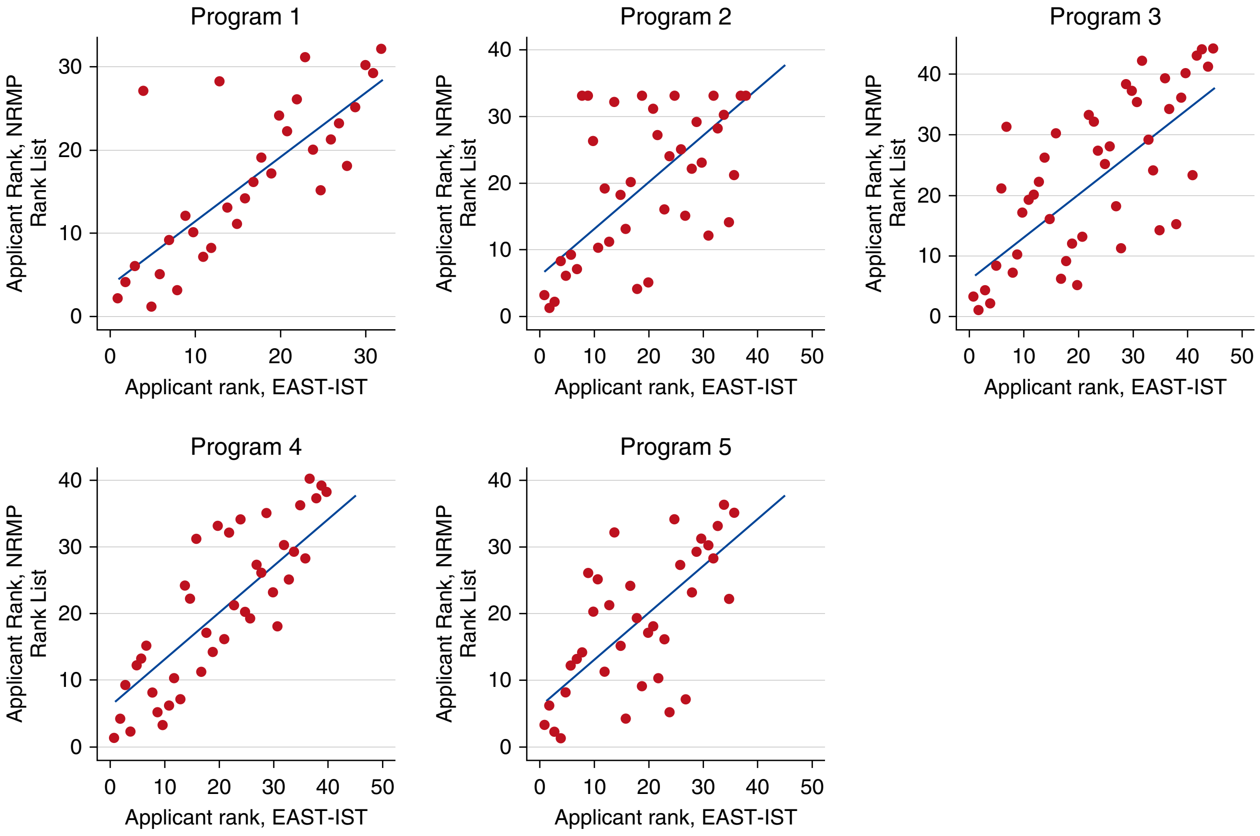

Figure 3. U.S. Electronic Residency Application System (ERAS) Application Scoring Tool–Interview Scoring Tool (EAST-IST) versus submitted National Residency Match Program (NRMP) rank lists using traditional methods at four validation programs. Spearman's *r*: program 1, $r = 0.77$, $P < 0.01$; program 2, $r = 0.54$, $P < 0.01$; program 3, $r = 0.72$, $P < 0.01$; program 4, $r = 0.84$, $P < 0.01$; program 5, $r = 0.65$, $P < 0.01$.

than individual predictors (U.S. Medical Licensing Exam I score, Alpha Omega Alpha status, junior year clinical clerkship scores) (3). The contents of these scoring forms were not designed for external generalizability, and the process by which they were developed or weighted was described in a way that limited their more generalized use.

There is considerable literature in human resource management and psychology journals that outlines best practices for reducing bias in the candidate selection process (6). There are few data that postgraduate medical education selection programs have integrated these practices (7). Recent studies have demonstrated considerable and seemingly arbitrary bias in even the most important, complex, person-related decisions (8)—one recent study suggested that irrelevant extraneous variables (e.g., time of last meal) can have considerable effect on judicial rulings regarding parole (9).

The EAST-IST process outlined in this report attempts to minimize some types of bias as much as possible by using what James Suroweicki has termed the "wisdom of crowds" to assign a strict weight to each potential variable, rather than relying on a largely subjective evaluation of a candidate's application and interview skills (10). This collective wisdom requires that the crowd adhere to four conditions: (*1*) a true diversity of opinions; (*2*) independence of those opinions; (*3*) decentralization of experience; and (*4*) a suitable mechanism for aggregation of opinions.

The EAST-IST weighting process aggregates opinions from a variety of diverse, independent, and decentralized individuals (faculty members) to develop weights that represent the desired candidate traits of the group as a whole. This requires that those who participate in the process come from varied backgrounds, work in diverse areas within the institution, and are able to project their own independent voice for what is important in selecting candidates for training. The EAST-IST process attempts to protect this independence by having each individual perform the weighting tool independent of the input of others. Diversity and decentralization are accomplished through the makeup of the committee.

Although our study did not attempt to prove that this process mitigates bias, an adherence to the EAST-IST evaluation at least allows for a standardized evaluation of each candidate with a wider input than was previously available. Because this study evaluated the process for development of a tool, it allows for applicability across a variety of different programs and specialties. It is natural and reasonable that different programs will, by design, select different components of both the EAST and IST portions of the tool to reflect their unique goals and priorities in selecting training candidates. Only a *process* for variable selection and weighting would be applicable to a heterogeneous population of programs.

Strengths and Limitations

The tool created by the process described here allows for several advantages over typical traditional ranking systems. First, because the evaluator does not have to directly compare one application to the others, it does not require the evaluator to have knowledge of the other applicants' quality to effectively evaluate the quality of the candidate. Because the application and interview days are split into discrete, weighted data points, an evaluator could even be interrupted in the evaluation of a single candidate and be able to return to that evaluation with relative ease. Several participants self-reported that completion of this tool took no more time than the traditional methods of candidate evaluation. Second, as the same weights are used over time, a program can compare the quality of candidates from year to year reliably. Alternatively, a program could use the same variable weighting process to reformulate the EAST and IST variables and weights from year to year to capture a dynamic mission, training culture, or selection committee. Third, this weighting process values equally the input of all participants, allowing all voices to be equally valuable. Finally, using a weighting process allows the members of a program to know precisely how its committee values the variables, allowing for discussion and transparency to the program—and potentially even to applicants.

It also has advantages over a more traditional, unweighted individual scoring method, as the weights are able to represent the input of a larger number of individuals than can generally interview a candidate. This can help to include the input of more individuals into the scoring process, even if they are unable to participate in the interview process. In addition, a weighted process can help programs more precisely quantify their values and focus on those values as they rank candidates.

Our study has several limitations. First and foremost, comparison of the tool to gestalt rankings and a "traditional method" are likely not the most ideal outcomes with which to compare this process. An ideal tool would be able to predict which candidates best fit the values and goals of the fellowship—this can only be determined after the candidate has begun the process of training. Ultimately, programs want to be able to predict the quality of candidates that finish their program; however, this would reflect both the quality of the candidate and the effectiveness of the training program. Future research should address how this tool is able to predict both the newly arrived candidate (reflecting only the fit of the candidate with the program) and graduates of the program (reflecting fit and program quality). Our study simply demonstrates this process to be comparable to a usual method. Second, we only validated this process within pulmonary and critical care fellowship training programs, which are generally considerably smaller than residency programs. Finally, not all geographic regions of the country were represented by our sampling of programs.

Future Research

This tool has potential value to any postgraduate training program that uses ERAS. This methodology allows the quantitation of the *inputs* of fellowship. Further studies are planned to apply this methodology for quantitating the *outputs* of fellowship (fellow graduates) based upon program-specific criteria, using the same process for selection of variables and weighting. Once this is complete, the inputs can be compared with the outputs, and a more rigorous evaluation of those factors that select high performers can be carried out.

Conclusions

This study validates a process for selecting and weighting components of the ERAS application and interview day to aggregate them into a customizable single tool for scoring candidates to postgraduate medical education programs. The scoring tool correlates highly with individuals' gestalt rankings of selected ERAS applications, and also correlates in a moderate to high fashion when compared with NRMP rank lists at five pulmonary and critical care training programs. This process and subsequent tool needs to be validated in a larger cohort of diverse residencies and fellowships. It can then be used for comparison to trainee performance during and at the conclusion of training. ■

Author disclosures are available with the text of this article at www.atsjournals.org.

Acknowledgment: The authors thank all of the faculty members who participated in the U.S. Electronic Residency Application System Application Scoring Tool–Interview Scoring Tool process at each participating institution. G.T.B. also thanks his wife, Sarah Spitznagle Bosslet, for helping with manuscript and life clarity.

References

1 National Resident Matching Program. Press release: Results of 2016 NRMP main residency match largest on record as match continues to grow. Washington, D.C.: NRMP; 2016 Mar 18 [accessed 2017 Feb 17]. Available from: http://www.nrmp.org/press-release-results-of-2016-nrmp-main-residency-match-largest-on-record-as-match-continues-to-grow/
2 Curtis DJ, Riordan DD, Cruess DF, Brower AC. Selecting radiology resident candidates. *Invest Radiol* 1989;24:324–330.
3 Turner NS, Shaughnessy WJ, Berg EJ, Larson DR, Hanssen AD. A quantitative composite scoring tool for orthopaedic residency screening and selection. *Clin Orthop Relat Res* 2006;449:50–55.
4 Harris PA, Taylor R, Thielke R, Payne J, Gonzalez N, Conde JG. Research electronic data capture (REDCap)—a metadata-driven methodology and workflow process for providing translational research informatics support. *J Biomed Inform* 2009;42:377–381.
5 Bass A, Wu C, Schaefer JP, Wright B, McLaughlin K. In-group bias in residency selection. *Med Teach* 2013;35:747–751.
6 Huffcutt AI. From science to practice: seven principles for conducting employment interviews. *Appl HRM Res* 2010;12: 121–136.
7 Kim RH, Gilbert T, Suh S, Miller JK, Eggerstedt JM. General surgery residency interviews: are we following best practices? *Am J Surg* 2016;211:476–481.e3.
8 Vohs KD, Baumeister RF, Schmeichel BJ, Twenge JM, Nelson NM, Tice DM. Making choices impairs subsequent self-control: a limited-resource account of decision making, self-regulation, and active initiative. *J Pers Soc Psychol* 2008; 94:883–898.
9 Danziger S, Levav J, Avnaim-Pesso L. Extraneous factors in judicial decisions. *Proc Natl Acad Sci USA* 2011;108:6889–6892.
10 Surowiecki J. The wisdom of crowds, 1st Anchor Books ed. New York: Anchor Books; 2005.