Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Assessing Physician Performance Using Multisource Surveys: Do Biases Exist Due to Gender, Country of Training, Native Language, and Age?

Julie J Lanz
University of Nebraska Kearney https://orcid.org/0000-0003-2090-5475
Paul Gregory (paul@pdpflorida.com)
https://orcid.org/0000-0003-0476-2611
Larry Harmon
Physicians Development Program Inc / PULSE 360 Program

Research article

Keywords: 360-degree performance appraisal, 360-degree survey, multi-source rater feedback, interpersonal and communication skills, emotional intelligence

Posted Date: December 15th, 2020

DOI: https://doi.org/10.21203/rs.3.rs-69954/v2

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Abstract

Background: With a growing number of foreign-trained physicians joining the United States workforce, there is a need to assess their job performance fairly. The purpose of this study was to explore the fairness of a multisource competency assessment on U.S.- and non-U.S.-trained physicians.

Methods: We conducted a non-experimental retrospective analysis on physicians working in the United States (n=258) who participated in a physician assessment and education program.

Results: There were no significant differences in performance outcomes at the scale level of teamwork, motivating or discouraging behaviors, technical practice, and patient interactions, nor the item level based on demographic differences.

Conclusions: The PULSE 360 is a powerful tool that can be used to evaluate physician performance without bias due demographic differences including: gender, country of physician medical training, physician native language, or age.

Background

In 2017, foreign-trained physicians made up over 25% of practicing physicians in the United States.[1] Given their substantial contributions to the U.S. healthcare system,[2,3] fairly assessing their performance is critical. Policy suggestions have been developed recently to promote the benefits of employing international health care workers, such as treating them transparently and fairly.[4] However, comparing the performance of U.S. trained physicians (USTP) to non-U.S. trained physicians (NUSTPs) is not well-understood.[5] Exploring soft skills like professionalism, interpersonal and communication skills, teamwork, and patient interactions is critical, and physician competency assessments should be fair regardless of demographic differences like age, sex, and nationality. The purpose of this paper is to explore the fairness of a multisource competency assessment program when evaluating U.S. and non-U.S. trained physicians working in the United States. While previous studies have explored where physicians were educated or trained (international medical graduates), in this study, NUSTP refers to physicians who completed their residencies outside the United States.

Assessing Physician Competence

Fair assessment of physician competencies is a key step in improving job performance. Reasons for evaluating physician performance range from appraisal to recertification, identifying high-risk physicians, and remediating those with a previous history of poor performance.[6] Maintaining quality patient care is important because 6-12% of physicians are referred to remediation for poor clinical skills.[7] The Institute of Medicine estimates that physician dyscompetency is one contributor to preventable medical errors at an estimated cost of $17 billion.[7,8] To determine which physicians should be referred for dyscompetency, one model for performance remediation starts with an assessment of the physician's competence.[9] A

common framework for maintaining physician competency is the American Board of Medical Specialties, which developed the Maintenance of Certification (ABMS MOC). This four-part framework includes: maintaining licensure, lifelong learning, cognitive expertise, and quality improvement.[10]

Multisource Feedback

Multisource feedback is the use of physicians' team members (other physicians, nurses, and staff) to evaluate job performance.[11,12] The scope and depth of multisource feedback is valuable given the argument that patient evaluations of physician performance are subjective at best.[13] For example, patient evaluations have been found to be influenced by the race and gender of the physician such that only physicians who were white and male benefited from a customer satisfaction judgment, even after controlling for objective measures of performance.[14] Beyond clinical skills, physician performance is based on a combination of individual differences including specialty area, gender, and age.[15,16] Evidence suggests that biases against international medical graduates may lead to more complaints against physicians and disciplinary outcomes,[17] but findings on biased physician performance evaluations are mixed.[18] Given the inconclusive evidence, having two examiners appears to mitigate potential sex or ethnic biases against physicians who are being evaluated based on their clinical performance.[19]

The PULSE 360 Program has been conducting multisource feedback assessments for physicians since 2002 with over 15,000 unique healthcare professionals participating in the program receiving over a million completed surveys of feedback. The original PULSE 360 Survey was developed with the help of subject matter experts (SMEs) including senior physician leaders and other healthcare administrators to determine the behaviors that they believed were most relevant to physicians being able to succeed as part of an interdisciplinary healthcare team. This led to the creation of over 100 behavioral rating items, which have been revised through years of item analyses and outcome studies to the most commonly used survey today, which consists of 25 behavioral items. PULSE conducts assessments for academic medical centers, community hospitals, and other healthcare organizations throughout the US and Canada. The PULSE 360 family of surveys have been used in several scholarly research projects showing strong validity and reliability in evaluating physicians' performance as part of a healthcare team.

Some research has explored the use of multirater assessments on international medical graduates and found them reliable,[20] but little research has examined bias in physician assessment as a function of training country (i.e., USTPs versus NUSTPs). Of the research on assessing physician performance, one experiment found that after holding education, experience, and personality consistent, international medical graduates were rated more poorly than those who had born in the prospective patients' home country. However, physicians who had been trained in an industrialized and high-income country benefited on their evaluations.[21] There are no significant differences in mortality rates for international versus national practitioners, but differences may exist in regard to the soft skills of communication, teamwork, and ethical issues.[22-24] Part of this bias may be a function of the examiners themselves.[25] In one study, international medical graduates had lower mortality rates than U.S. medical graduates.[26]

Further, there is evidence that in Canada, international medical graduates are disciplined for misconduct more frequently than North American medical graduates.[27] In Australia, international medical graduates receive more complaints and disciplinary adverse findings.[17] Thus, there is a critical need for fair tools to evaluate USTPs and NUSTPs on their job performance. Given the mixed findings on the effects of gender, country of training, language, and age on performance, the following is predicted:

Hypothesis 1: There will be significant differences in PULSE 360 physician performance (as rated by colleagues) based on gender such that females will have higher scores.

Hypothesis 2: There will be significant differences in PULSE 360 physician performance based on country where training occurred such that USTPs will have higher scores.

Hypothesis 3: There will be significant differences in PULSE 360 physician performance based on first language spoken such that native English speakers will have higher scores.

Hypothesis 4: There will be significant differences in PULSE 360 physician performance based on age such that younger physicians will have higher scores.

Methods

Design

A non-experimental retrospective analysis of data was conducted for two hundred and fifty-eight physicians (n=258) who participated in a physician education program focused on fitness for duty.

Statistical Analyses

For hypotheses 1-4, we conducted a post-hoc power analysis using G*Power v. 3.0.10 to determine the sample size needed to detect significant effects.[28] Given a one-tailed independent samples t-test with a large effect size (d = 0.5, α = .05), the preferred sample size is 176 participants. Independent samples t-tests using a Bonferonni adjusted alpha of 0.0125 (.05/4) were conducted in order to evaluate potential biases in PULSE 360 scale scores due to demographic differences including: gender, country in which training occurred, first language spoken, and age; see Tables 1-4.

Participants

Seventy percent (80%) of the physicians were males (n=206), 62% were trained in the United States (n=157), 66% had English as their first language (n=166), the average physician age was 61 (range 41-84), and 78% were board certified (n=202). Age was median split into two groups: 1) 62 or younger (n=126, 49%), 2) 63 or older (n=125, 48%), n=7 missing data (3%). There were thirty (n=30) different specialties represented within the sample of physicians, including internal medicine, obstetrics and gynecology, anesthesiology etc.). All physicians in the sample were participating in an assessment conducted through a physician education program at an academic medical center in the United States.

Physicians had been referred to this program by a healthcare organization for a fitness for duty evaluation.

PULSE 360 Survey

The PULSE 360 Survey is an assessment of leadership, teamwork, communication, professionalism, and other physician behaviors based on multisource feedback from other physicians, advanced practice providers, clinical staff, and administrative staff members who interact with a physician. The survey used with the current sample included n=25 items and is made up of 5 performance domains, including a total composite performance score known as the Teamwork Index (TI) Score. All items are scored using a 5-point Likert type scale regarding the extent to which raters perceive a physician engages in a target behavior. The internal consistency reliability estimates for all performance domains are as follows: TI Score (25 items) α = .92, Motivating Behavior Score (9 items) α = .85, Discouraging Behavior Score (7 items) α = .84, Technical Practice Score (5 items) α = .82, and Patient Interaction Score (4 items) α =.79. TI scores typically range from 0 to 100 with a national mean score of 68.9 for physicians. The other scale scores range from 20 to 100. Prior research has demonstrated both the internal and external validity of PULSE 360 scores in relation to important physician outcomes such as malpractice risk and patient satisfaction.[11,29-34]

Data Analyses

The PULSE 360 survey item data is collected at the ordinal level of measurement while scale scores created from this data are interval level data. We opted to perform parametric analyses (independent sample t-tests) because the observed data demonstrated an approximately normal distribution. However, we also conducted non-parametric chi-square comparisons of expected distribution of scores for each hypothesis given the ordinal nature of the item level data. We report the results of the parametric analyses only because the non-parametric analyses produced the same results/conclusions at both the scale and item level.

Results

Hypothesis 1 was not supported. There were non-significant differences in the mean PULSE 360 scores for male vs. female physicians (see Tables 1-4 for independent samples t-test results). Hypothesis 2 was not supported; there were non-significant differences in the mean PULSE 360 scores for US-trained vs. foreign-trained physicians. Hypothesis 3 was not supported; there were non-significant differences in the mean PULSE 360 scores for native English speakers vs. non-native English speakers. Hypothesis 4 was not supported; there were non-significant differences in the mean PULSE 360 scores between age groups (see Table 4). Additionally, all post hoc comparisons of item level differences yielded non-significant differences in mean scores on all PULSE 360 Scale scores for all comparison groups mentioned in Hypotheses 1-4.

Discussion

This study compared the performance of USTPs and NUSTPs using a multisource assessment tool. Given the growing demand for NUSTPs in the United States and potential bias in how their performance is assessed, there is a need to fairly select, train, and support this diverse group of international physicians.[34] Physicians in this study were evaluated on their performance using the PULSE 360 Survey and were compared across gender, country of training, native language, and age. There were no significant differences in their reported performance on professionalism, teamwork, motivating behaviors, discouraging behaviors, technical practice style, or patient interactions. These findings suggest that there are valid and reliable tools established to fairly evaluate the performance of both U.S.-trained physicians and foreign-trained physicians. This is valuable given that previous research has found that some measures discriminate against some protected classes.[17]

The physicians in our current sample were recruited to the physician assessment and education program for a variety of reasons that may not be representative of practicing physicians in the United States. However, the use of multisource data allows us to move beyond self-reported performance to a more comprehensive view of physicians' performance on the job. Further research will be needed to more thoroughly explore these findings, but at least within our sample, there were no significant variations in feedback scoring patterns attributable to protected class membership.

Conclusions

The use of multisource feedback can provide a more comprehensive and unbiased assessment of others' perceptions of physician behavior and performance within the healthcare team than traditional single source methods of feedback.

List Of Abbreviations

United States trained physicians (USTP), non-U.S. trained physicians (NUSTP).

Declarations

Ethics approval and consent to participate: Because this study used de-identified archival data, it was deemed unnecessary to ask participants for informed consent. This study was approved by the UNK IRB #041320-2.

Consent for publication: Not applicable.

Availability of data and materials: The dataset used and analyzed during the current study are available from the corresponding author on reasonable request.

Competing interests: Two of the authors are employees of the Physicians Development Program Inc / PULSE 360 Program, Miami, FL, USA.

Funding: Not applicable.

Authors' contributions: Study concept and design: PG and LH. Analysis and interpretation of data: PG, JL, and LH. Drafting and revising of the manuscript: JL, PG, and LH. All authors read and approved the final version.

Acknowledgements: Not applicable.

References

[1]American Immigration Council. Foreign-trained doctors are critical to serving many U.S. communities. https://www.americanimmigrationcouncil.org/research/foreign-trained-doctors-are-critical-serving-many-us-communities. January 17, 2018. Accessed May 4, 2020.

[2]Jolly P, Boulet J, Garrison G, Signer MM. Participation in US graduate medical education by graduates of international medical schools. *Acad Med* 2011;*86*(5):559-564. doi:10.1097/ACM.0b013e318212de4d

[3]Pinsky WW. The importance of international medical graduates in the United States. *Ann Intern Med* 2017;166(11):840-841. doi:10.7326/M17-0505

[4]Chen PG, Auerbach DI, Muench U, Curry LA, Bradley EH. Policy solutions to address the foreign-educated and foreign-born health care workforce in the United States. *Health Affair* 2013;32(11):1906-1913. doi:10.1377/hlthaff.2013.0576

[5]Rao NR, Yager J. Acculturation, education, training, and workforce issues of IMGs: Current status and future directions. *Acad Psychiatr* 2012;36(4):268-270. doi:10.1176/appi.ap.12050095

[6]Humphrey C. Assessment and remediation for physicians with suspected performance problems: An international survey. *J Contin Educ Health* 2010;30(1):26-36. doi:10.1002/chp.20053

[7]Williams BW. The prevalence and special educational requirements of dyscompetent physicians. *J Contin Educ Health Prof* 2006;26:173–191. doi:10.1002/chp.6

[8]Kohn LT, Corrigan J, Donaldson MS. Committee on Quality of Health Care in America, Institute of Medicine. *To err is human: Building a safer health system*. Washington, DC: National Academy Press, 1999

[9]Hauer KE, Ciccone A, Henzel TR, Katsufrakis P, Miller SH, Norcross WA, Papadakis MA, Irby DM. Remediation of the deficiencies of physicians across the continuum from medical school to practice: a

thematic review of the literature. *Acad Med* 2009;84(12):1822-1832. doi:10.1097/ACM.0b013e3181bf3170

[10]Hawkins RE, Lipner RS, Ham HP, Wagner R, Holmboe ES. American Board of Medical Specialties Maintenance of Certification: Theory and evidence regarding the current framework. *J Contin Educ Health* 2013;33(S1):S7-S19. doi:10.1002/chp.21201

[11]Lanz JJ, Gregory PJ, Menendez ME, Harmon L. Dr. Congeniality: Understanding the importance of surgeons' nontechnical skills through 360° feedback. *J Surg Educ* 2018;75(4):984-992. doi:10.1016/j.jsurg.2017.12.006

[12]Castanelli D, Kitto S. Perceptions, attitudes, and beliefs of staff anaesthetists related to multi-source feedback used for their performance appraisal. *Brit J Anaesth* 2011;107(3):372-377. doi:10.1093/bja/aer152

[13]Glickman SW, Schulman KA. The mis-measure of physician performance. *Am J Manag Care* 2013;19(10):782-785. PMID:24304157

[14]Hekman DR, Aquino K, Owens BP, Mitchell TR, Schilpzand P, Leavitt K. An examination of whether and how racial and gender biases influence customer satisfaction. *Acad Manage J* 2010;53(2):238-264. doi:10.5465/amj.2010.49388763

[15]Grace ES, Wenghofer EF, Korinek EJ. Predictors of physician performance on competence assessment: Findings from CPEP, the Center for Personalized Education for Physicians. *Acad Med* 2014;89(6):912-919. doi:10.1097/ACM.0000000000000248

[16]Wenghofer EF, Williams AP, Klass DJ. Factors affecting physician performance: Implications for performance improvement and governance. *Healthc Policy* 2009;5(2):e141-e160. PMID:21037818

[17]Elkin K, Spittal MJ, Studdert DM. Risks of complaints and adverse disciplinary findings against international medical graduates in Victoria and Western Australia. *Med J Australia*, 2012;197(8):448-452. doi:10.5694/mja12.11702

[18]Denney ML, Freeman A, Wakeford R. MRCGP CSA: are the examiners biased, favouring their own by sex, ethnicity, and degree source?. *Br J Gen Pract* 2013;63(616):e718-e725. doi:10.3399/bjgp13X674396

[19]McManus IC, Elder AT, Dacre J. Investigating possible ethnicity and sex bias in clinical examiners: an analysis of data from the MRCP (UK) PACES and nPACES examinations. *BMC Med Educ* 2013;13(103):1-11. doi:10.1186/1472-6920-13-103

[20]Nair BR, Alexander HG, McGrath BP, Parvathy MS, Kilsby EC, Wenzel J, Frank IB, Pachev GS, Page GG. The mini clinical evaluation exercise (mini-CEX) for assessing clinical performance of international medical graduates. *Med J Australia* 2008;189(3):159-161. PMID:18673104

[21]Louis WR, Lalonde RN, Esses VM. Bias against foreign-born or foreign-trained doctors: Experimental evidence. *Med Educ* 2010;44(12):1241-1247. doi:10.1111/j.1365-2923.2010.03769.x

[22]Norcini JJ, Boulet JR, Dauphinee WD, Opalek A, Krantz ID, Anderson ST. Evaluating the quality of care provided by graduates of international medical schools. *Health Aff (Millwood).* 2010;29:1461-1468. doi:10.1377/hlthaff.2009.0222

[23]Slowther A, Lewando Hundt GA, Purkis J, Taylor R. Experiences of non-UK-qualified doctors working within the UK regulatory framework: A qualitative study. *J R Soc Med* 2012;105:156-67. doi:10.1258/jrsm.2011.110256

[24]Ko DT, Austin PC, Chan BTB, Tu JV. Quality of care of international and Canadian medical graduates in acute myocardial infarction. *Arch Intern Med* 2005;165:458-463. doi:10.1001/archinte.165.4.458

[25]Yeates P, Woolf K, Benbow E, Davies B, Boohan M, Eva K. A randomised trial of the influence of racial stereotype bias on examiners' scores, feedback and recollections in undergraduate clinical exams. *BMC Med* 2017;15(179):1-11. doi:10.1186/s12916-017-0943-0

[26]Tsugawa Y, Jena AB, Orav EJ, Jha AK. Quality of care delivered by general internists in US hospitals who graduated from foreign versus US medical schools: Observational study. *BMJ* 2017;356(j273):1-8. doi:10.1136/bmj.j273

[27]Alam A, Matelski JJ, Goldberg HR, Liu JJ, Klemensberg J, Bell CM. The characteristics of international medical graduates who have been disciplined by professional regulatory colleges in Canada: A retrospective cohort study. *Acad Med* 2017;92(2):244-249. doi:10.1097/ACM.0000000000001356

[28] Faul F, Erdfelder E, Lang AG, Buchner A. G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behav Res Methods.* 2007:39:175-191.

[29] Gregory PJ, Robbins B, Schwaitzberg SD, Harmon L. Leadership development in a professional medical society using 360-degree survey feedback to assess emotional intelligence. *Surg Endosc* 2017;31(9):3565-3573. doi:10.1007/s00464-016-5386-8

[30] Gregory PJ, Ring D, Rubash H, Harmon L. Use of 360° feedback to develop physician leaders in orthopaedic surgery. *J Surg Orthop Adv* 2018;27(2):85-91. PMID:30084813

[31]Hageman MG, Ring DC, Gregory PJ, Rubash HE, Harmon L. Do 360-degree feedback survey results relate to patient satisfaction measures? *Clin Orthop Relat R* 2015;473(5):1590-1597. doi:10.1007/s11999-014-3981-3

[32]Hammerly ME, Harmon L, Schwaitzberg SD. Good to great: Using 360-degree feedback to improve physician emotional intelligence. *J Healthc Manag* 2014;59(5):354-366. PMID:25647955

[33]Lagoo J, Berry WR, Miller K, Neal BJ, Sato L, Lillemoe KD, Doherty GM, Kasser JR, Chaikof EL, Gawande AA, Haynes AB. Multisource evaluation of surgeon behavior is associated with malpractice claims. *Ann Surg* 2019;270(1):84–90. doi:10.1097/SLA.0000000000002742

[34]Peile E. Selecting an internationally diverse medical workforce. *BMJ* 2014;348(g2696):1-2. doi:10.1136/bmj.g2696

Tables

Table 1. t-Test Comparison of Mean PULSE 360 Scale Scores by Gender.

	Gender					
	Male (n=206)		Female (n=52)		Mean Comparison (df=256)	
PULSE 360 Scale Score	m	sd	m	sd	t	p
1. Teamwork Index Score	66.5	24.0	65.0	23.4	.401	.69^{ns}
2. Motivating Behavior Score	81.4	10.2	81.2	10.2	.153	.88^{ns}
3. Discouraging Behavior Score	28.2	9.9	29.2	9.4	-.625	.53^{ns}
4. Technical Practice Score	86.2	10.3	87.9	8.3	-1.14	.25^{ns}
5. Patient Interaction Score	87.1	9.7	87.5	9.1	-.225	.82^{ns}

**All post hoc comparisons for between group differences were non-significant for all scale scores*

Table 2. t-Test Comparison of Mean PULSE 360 Scale Scores by Trained in US Status.

	Physician trained in the US?					
	Yes (n=157)		No (n=97)		Mean Comparison (df=252)	
PULSE 360 Scale Score	m	sd	m	sd	t	p
1. Teamwork Index Score	65.3	22.9	67.5	25.6	-.712	.48^{ns}
2. Motivating Behavior Score	81.1	9.6	81.7	11.1	-.433	.67^{ns}
3. Discouraging Behavior Score	28.9	9.6	27.7	10.2	.946	.35^{ns}
4. Technical Practice Score	86.0	9.7	87.3	10.5	-.952	.34^{ns}
5. Patient Interaction Score	87.0	9.1	87.4	10.3	-.324	.75^{ns}

missing n=4

**All post hoc comparisons for between group differences were non-significant for all scale scores*

Table 3. t-Test Comparison of Mean PULSE 360 Scale Scores by Native English Speaker Status.

	Native English Speaker?					
	Yes (n=166)		No (n=84)		Mean Comparison (df=248)	
PULSE 360 Scale Score	M	sd	m	sd	t	p
1. Teamwork Index Score	66.4	23.3	66.8	24.8	-.122	.90^{ns}
2. Motivating Behavior Score	81.4	9.9	81.5	10.7	-.032	.98^{ns}
3. Discouraging Behavior Score	28.3	9.6	28.1	9.8	.207	.84^{ns}
4. Technical Practice Score	86.6	9.5	86.1	10.9	.359	.72^{ns}
5. Patient Interaction Score	87.3	9.3	87.4	9.9	-.031	.98^{ns}

missing n=8

**All post hoc comparisons for between group differences were non-significant for all scale scores*

Table 4. t-Test comparison of Mean PULSE 360 Scale Scores by Age Range.

	Age Range of Physicians					
	62 or younger (n=126)		63 or older (n=125)		Mean Comparison (df=249)	
PULSE 360 Scale Score	M	sd	m	sd	t	p
1. Teamwork Index Score	67.7	22.6	64.5	24.5	1.06	.290^{ns}
2. Motivating Behavior Score	82.0	9.9	80.6	10.2	1.10	.272^{ns}
3. Discouraging Behavior Score	27.8	9.0	28.9	10.3	-0.92	.357^{ns}
4. Technical Practice Score	86.4	8.9	87.8	9.1	.275	.783^{ns}
5. Patient Interaction Score	87.8	9.1	86.5	9.8	1.13	.257^{ns}

missing n=7

**All post hoc comparisons for between group differences were non-significant for all scale scores*