ARC Journal of Nursing and Healthcare (AJNH)
Volume 2, Issue 1, 2016, PP 1-8
www.arcjournals.org

Co-Worker Assessment and Physician Multisource Feedback

Gregg Trueman, Jocelyn Lockyer

Mount Royal University

Abstract:

Background: *Multisource feedback (MSF) is increasingly being used as one of the components in revalidation and recertification processes to guide physicians' continuing professional development. Data provided by co-workers (e.g., nurses, pharmacists, technicians) are recognized as integral for assessing a physician's communication, teamwork and interprofessional abilities. The purpose of this study was to examine both the reliability of co-worker scores and the association between co-worker familiarity and physician ratings as both affect perceptions of the quality of feedback and the likelihood that recipients will take their feedback seriously.*

Method: *MSF data from 9674 co-workers of 1341 Alberta physicians across 9 specialty groups were analyzed. Analyses for internal consistency and generalizability theory (G and D-studies) were used to assess reliability. The association between co-worker familiarity and the MSF scores they provided to physicians was assessed using ANOVA.*

Results: *Cronbach's alpha for all co-worker tools was $\geq$ 0.90. Generalizability coefficients (EP^2) varied by specialty and ranged from 0.56 to 0.72. D studies revealed that a minimum of 11 co-workers are necessary to achieve stability (i.e., $EP^2 \geq 0.70$). Co-worker familiarity exerted a significant ($p < .001$) positive main effect on physician performance scores, across all specialty groupings.*

Conclusions *This study confirms the reliability of co-worker scores and provides evidence that co-worker MSF data is stable and consistent for the purposes of providing physicians with feedback for professional development. Attention however needs to be paid to co-worker/physician familiarity as this relationship may favourably bias physician performance scores.*

Keywords: *multisource feedback, workplace-based assessment; collaborative practice; co-worker feedback.*

Word Count: *2503*

1. Co-Worker Familiarity and Physician Multisource Feedback

MSF is a reliable workplace-based assessment strategy[1] used to collect data on the performance of practicing physicians, and has been used extensively to assess physicians and guide continuing professional development in Canada,[2,3] the USA,[4,5] the UK[6,7] and Europe.[8,9] Generally in a MSF assessment, patients, co-workers (e.g., nurses, technicians, pharmacists), and medical colleagues provide data. There may also be a self-assessment questionnaire. In some cases, the co-worker and medical colleague data are combined.[10] In other practice settings, two different questionnaires with different items for medical colleagues and non-physician co-workers are used to provide data.[11] When one questionnaire is used to assess both physician and non physician behaviors, and the feedback report is an aggregate of these data, the unique perspective provided by non-physician co-workers is lost.

The College of Physicians and Surgeons of Alberta (CPSA) Physician Achievement Review (PAR) Program (www.par-program.org) provides a unique opportunity to examine MSF data provided by co-workers across nine specialty groups. The instruments, adapted and implemented over a 12 year period, were first developed for family physicians[12] and later adapted across eight other specialty groups, to inform professional development. In the PAR program, the co-worker is a source distinct from the medical colleague and co-worker assessment questionnaire (CAQ) items are different from items on the medical colleague questionnaire. While each co-worker instrument varies by specialty, their focus has been on interprofessional teamwork (i.e., communication, professionalism and collaboration) and not on medical expertise. This is in direct contrast to UK programs where co-workers complete a colleague instrument that includes both MD and non-MD respondents.[13,14] In part, the CPSA sought to ensure that their tools captured different aspects of the physician's work; different instruments with distinct foci completed by multiple groups would help ensure a broad range

of feedback. Each physician participates in the PAR program every five years and identifies the medical colleagues and co-workers who will provide the data. Reliability speaks to the reducibility of feedback data, something that is particularly relevant when fewer numbers of co-workers are available to provide feedback. Thus, uptake of co-worker data for professional development purposes is predicated on reliable instrumentation.[15]

The PAR instruments were psychometrically assessed with an examination of evidence for validity[16-19] and reliability[20-25] when they were first developed. While aspects of reliability have been re-examined for physician groups that have participated in PAR on more than one occasion,[12] there has not been a comprehensive examination of reliability for the co-worker instruments across all PAR specialties or specifically of reliability, a key component of validity. Of particular concern in the present study is characteristics of MSF design (e.g., number of co-workers providing information) known to influence the stability of MSF feedback.[26] Given the multiple sources of error in MSF data, internal consistency measures - while necessary – are insufficient to establish instrument reliability. Generalizability theory expands on alpha by using two types of studies - generalizability (*G*) and decision (*D*) studies - to quantify the amount of variance associated with different factors or *facets,* and to provide reliability evidence for measurement protocols (i.e., optimal numbers of items and assessors) in fully nested, unbalanced behavioural measures that constitute PAR data.[27]

Also of interest is the influence of familiarity - between the co-worker and index physician - on co-worker performance scores.[11] Initial MSF work did not find a definitive association that would preclude physician selection of their own co-worker.[16,28] More recently however, familiarity has been correlated in UK studies with more favourable feedback.[29] In a study of 68 underperforming physicians referred to the UK's National Clinical Assessment Service, physicians who selected their own co-worker assessors were more likely to obtain higher scores compared to feedback provided by colleagues who were selected for them.[30] Research involving post graduate trainees also found that the length of the working relationship (i.e., familiarity) between assessor and trainee influenced their MSF scores.[13]

This study was undertaken to address the following questions: (1) what is the reliability of the MSF scores provided by PAR co-worker questionnaires?; and (2) what is the association between familiarity and the scores provided by physician selected co-workers? It was believed that this information would inform instrument revision and uptake of co-worker feedback for professional development purposes.

2. METHODS

Pivotal Research Inc, a company that administers the PAR program on behalf of the CPSA, created an anonymous dataset of co-worker data, collected between January 2006 and April 2011, for 150 physicians from each of the nine specialty groups (i.e., anesthesia, diagnostic imaging, episodic medicine, family medicine, laboratory medicine, medical specialists, pediatrics, psychiatry and surgery). The data set represented the most recent assessments of physicians in each specialty grouping. For each physician, responses from up to eight respondents were provided. Depending upon specialty grouping, the co-worker questionnaire contained between 17 and 22 items placed on a five point Likert scale, with a sixth "unable to assess" option. The co-worker's self-reported *familiarity* with the physician was also provided on a five point Likert scale [i.e., 1="not at all"; 2="not well"; 3="somewhat"; 4="well" and 5="very well" familiar]. Data describing the physician included: sex, medical school (Canadian, International), years since graduation, location of practice (urban, regional, or rural), and number of times the physician had participated in PAR, were available for physicians in each specialty group. No data describing the co-workers providing feedback were available for analysis.

3. ANALYSIS

Descriptive statistics were calculated for all physician socio demographic variables. At the instrument level, the number of questionnaires and items including the mean number of co-workers per physician, the mean score with standard deviation and range were calculated.

Cronbach's alpha was calculated to examine each tool's internal consistency. MSF designs can threaten instrument reliability with several sources of error variance in that data are *uncrossed* (i.e., co-workers rate the physician on only one occasion), *unbalanced* (i.e., there are different numbers of co-workers providing input for each physician), and *fully nested* (i.e., co-workers for each physician

are unique to that physician). A*G* study uses repeated-measures analysis of variance to simultaneously quantify the variance embedded in each facet that are interacting to create error but is not captured in the inter-item correlation matrices underpinning internal consistency reliability.[31] *G* studies calculate variance components expressed as a coefficient (EP^2), where $EP^2 \geq 0.70$ is generally considered the minimum threshold suitable for MSF instruments of similar intent.[32] For all co-worker tools, *G* studies were performed to estimate the variance associated with different facets: the *physician*, the *co-worker*, the *questionnaire item*, and *residual error* (i.e., measurement artifact). *D* studies then used the variance components derived from each *G*-study to improve the stability of measurement protocols used to collect physician feedback (e.g., the number of questionnaire items or the number of co-workers).

To assess the influence co-worker familiarity exerted on physician performance scores, a one way analysis of variance (ANOVA) with planned comparisons was used to identify differences between group means within the five levels of familiarity.[33] Effect sizes were calculated to determine the direction and magnitude of familiarity on performance scores to differentiate statistical from clinical significance.[34]

Ethics approval for this research was provided by the University of Calgary Conjoint Health Ethics Research Board.

4. RESULTS

Data from 9674 co-workers, provided to 1341 physicians, were analyzed. Physician socio-demographic characteristics are summarized in Table 1. The majority of physicians were male, practiced in an urban setting, and graduated from Canadian medical schools a mean of 22.6 years ago [range = 8 to 49 years]. With the exception of diagnostic imaging and laboratory medicine, most physicians had received PAR feedback on a previous occasion. At the level of the questionnaire, the mean number of co-workers in each medical specialty ranged from 6.32 for family physicians to 7.48 in diagnostic imaging (Table 2). Overall, co-workers' ratings were negatively skewed leading to a restriction of range at the upper end of the scale and toward favourable views of physician performance (Table 2). Mean item scores and standard deviations ranged from a low of 4.04(.94) [psychiatry] to 4.81(.49) [diagnostic imaging].

Table1. *Descriptive Data (by physician)*

	MD Sex		**School of MD Graduation**		**Years since Graduation**	**Location of MD Practice**			**# of PAR Reports per MD**		**Total # MDs**
	F (%)	M (%)	Canada (%)	IMG (%)	mean (sd)	Urban (%)	Regional (%)	Rural (%)	1 (%)	2 (%)	
Anesthesia	31 (21)	119 (79)	114 (76)	36 (24)	26.1 (8.0)	118 (78)	19 (13)	13 (9)	19 (12)	131(88)	**150**
Diagnostic Imaging	29 (19)	121 (81)	117 (78)	33 (22)	23.1 (10.49)	129 (86)	14 (9)	7 (5)	150 (100)	0	**150**
Episodic Medicine	47 (31)	103 (69)	127 (85)	23 (15)	19 (9.69)	111 (74)	18 (12)	21 (14)	91 (61)	59 (39)	**150**
Family Medicine	63 (42)	87 (58)	81 (54)	69 (46)	18.7 (8.97)	83 (55)	17 (11)	50 (33)	66 (44)	84 (56)	**150**
Laboratory Medicine	47 (31)	94 (63)	73 (52)	68 (48)	27.0 (8.49)	115 (77)	19 (13)	7 (5)	141 (100)	0	**141**
Medical Specialist	47 (31)	103 (69)	114 (76)	36 (24)	20.5 (9.84)	131 (87)	12 (8)	7 (5)	68 (45)	82 (55)	**150**
Pediatrics	61 (41)	89 (59)	92 (61)	58 (39)	24.8 (9.35)	133 (89)	9 (6)	8 (5)	46 (31)	104 (69)	**150**
Psychiatry	53 (35)	97 (65)	100 (67)	50 (33)	26.0 (.85)	122 (81)	9 (6)	19 (13)	29 (19)	121 (81)	**150**
Surgery	37 (25)	113 (75)	118 (79)	32 (21)	18.1 (7.41)	103 (69)	28 (19)	19 (13)	74 (49)	76 (51)	**150**
Total	**415**	**926**	**936**	**405**	**22.58**	**1045**	**145**	**151**	**682**	**659**	**1341**

Table2. *Descriptive Data (by questionnaire)*

	# questionnaires	**# items**	**ave # raters/MD**	**mean(SD) score**	**range (SD)**	**kewness**
Anesthesia	1118	19	7.39	4.31 (.64)	4.51(.75)-4.72(.54)	-0.40
Diagnostic Imaging	1134	20	7.48	4.29 (.66)	4.45(.79)-4.81(.49)	-0.53
Episodic Medicine	1043	20	6.80	4.23 (.65)	4.20(.91)-4.73(.51)	-0.37
Family Medicine	989	17	6.32	4.23 (.67)	4.34(.82)-4.71(.58)	-0.40
Lab Medicine	1069	22	7.52	4.07 (.69)	4.20(.83)-4.71(.48)	-0.44
Medical Specialist	1073	20	7.03	4.18 (.96)	4.68(.59)-4.31(.78)	-0.34
Pediatrics	1080	22	7.09	4.32 (.67)	4.13(.86)-4.75(.49)	-0.60
Psychiatry	1086	22	7.14	4.38 (.67)	4.04(.94)-4.78(.48)	-0.76
Surgery	1082	19	7.10	4.35 (.65)	4.34(.77)-4.68(.55)	-0.60
Total	**9674**					

5. RELIABILITY

Cronbach's alpha across the collection of questionnaires were very high, ranging from 0.90 (episodic medicine) to 0.96 (diagnostic imaging), providing evidence for each tools' internal consistency reliability (Table 3). *G* coefficients ranging from 0.56 to 0.72 indicated a different picture of reliability across the collection of tools, providing new evidence for how behavioural measurement can introduce unique sources of variance influencing MSF reliability (i.e., Table 3). The facet with the greatest variance was *rater by item nested within physician*, ranging from 22.62%: (anesthesia) to 43.41%: (diagnostic imaging) indicating significant variation within the co-worker: physician relationship. Additional error variance across the CAQ, ranging from 19.49%: (anesthesia) to 49.27%: (laboratory medicine), was attributed to *rater nested within physician*. *G* coefficients, and their associated standard error of measurement, ranged from a low of 0.56 (0.020) and 0.56 (0.022) for the diagnostic imaging and episodic medicine instruments respectively, to a high of 0.72 (0.024) for the medical specialist CAQ.

Table3. *Reliability Coefficients and Variance Components*

			Variance Components									
			p		**r:p**		**i**		**pi**		**ri:p**	
PAR Specialty	**Cronbach's α**	**EP²**	σ^2	%	σ^2	%	σ^2	%	σ^2	%	σ^2	%
Anaesthesia	0.94	0.62	0.0447	54.98	0.0183	22.62	0.0038	0.47	0.01972	2.44	0.0156	19.49
Diagnostic Imaging	0.96	0.56	0.02943	8.18	0.1561	43.41	0.0099	2.75	0.0177	4.92	0.0146	40.74
Episodic Medicine	0.90	0.56	0.03338	7.89	0.1586	37.48	0.0144	3.41	0.03286	7.77	0.1839	43.46
Family Medicine	0.95	0.64	0.06341	12.96	0.2085	42.61	0.0091	1.85	0.02125	4.34	0.1872	38.24
Laboratory Medicine	0.92	0.70	0.03983	10.17	0.1108	28.28	0.0194	4.96	0.02869	7.32	0.193	49.27
Medical Specialist	0.92	0.72	0.07168	15.68	0.1744	38.15	0.0099	2.17	0.02492	5.45	0.1762	38.55
Pediatrics	0.94	0.59	0.03357	8.48	0.1513	38.21	0.0149	3.77	0.02167	5.47	0.1745	44.07
Psychiatry	0.93	0.58	0.03336	7.45	0.1527	34.12	0.0201	4.49	0.06047	13.51	0.1809	40.42
Surgery	0.95	0.62	0.04396	10.2	0.1733	40.21	0.0074	1.72	0.01581	3.67	0.1905	44.21
p = physician; r = rater; i = item												
σ^2(p) = variance component for physician												
σ^2(r:p) = variance component for rater within physician												
σ^2(i) = variance component for item												
σ^2(pi) = variance component for physician by item												
σ^2(ri:p) = variance component for rater by item, within physician												

EP^2 coefficient formula: $= \dfrac{\sigma(\lambda)^2}{\sigma(\lambda)^2 + \sigma(\delta)^2}$.

D-studies were conducted to determine if the number of questionnaire items or the number of co-workers were sufficient to produce stable feedback to the individual physician. *G* studies demonstrated a relatively small variance component that could be attributed to CAQ items across specialty grouping. As such, increasing the number of items resulted in little change in the *G* coefficient. However, changing the number of co-workers providing feedback did produce higher *G* coefficients across the entire collection of co-worker tools (Table 4). With the exception of the laboratory medicine tool, our review found that a minimum of 11 co-workers were required to provide stable data with a coefficient $\geq$ 0.70.

Table4. *D-studies by Specialty*

		# of Assessors										
	# items	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**
Anesthesia	19	0.531	0.575	0.611	0.641	0.666	0.688	**0.707**	0.724	0.738	0.752	0.763
Diagnostic Imaging	20	0.467	0.511	0.548	0.724	0.580	0.607	0.631	0.670	0.686	**0.701**	0.714
Episodic Medicine	20	0.487	0.530	0.566	0.596	0.622	0.644	0.664	0.681	0.696	**0.710**	0.722
Family Medicine	17	0.584	0.626	0.660	0.688	**0.712**	0.732	0.749	0.764	0.778	0.789	**0.800**
Lab Medicine	22	0.612	0.652	0.684	**0.710**	0.732	0.750	0.766	0.780	0.791	**0.802**	
Medical Specialist	20	0.654	0.693	**0.723**	0.748	0.768	0.786	**0.800**				
Pediatrics	22	0.506	0.549	0.586	0.616	0.643	0.665	0.685	**0.702**	0.717	0.731	0.743
Psychiatry	22	0.498	0.542	0.578	0.608	0.634	0.656	0.676	0.693	**0.708**	0.721	0.734
Surgery	19	0.540	0.584	0.619	0.649	0.675	0.696	**0.715**	0.732	0.746	0.759	0.771

Threshold for low stakes assessment: $\geq$ 0.70; high stakes assessment: $\geq$ 0.80

6. FAMILIARITY

Overall, 72.4% [range 68% - 78%] of co-workers included their degree of familiarity with the physician ratee as part of their information (Table 5). MSF scores were significantly different between familiarity groups ($p < .001$), producing a moderate to moderately-large effect size across all specialty groupings. These data indicate that as co-worker familiarity with the physician increases, regardless of specialty grouping, so too did the performance scores co-workers assigned. Furthermore, planned comparison analyses showed that co-worker familiarity (i.e., well and very well) was associated with increased PAR scores compared to co-workers who were unfamiliar (i.e., not at all, not well, and somewhat) with the physician they were assessing. Though these findings are inconclusive as to magnitude, they nonetheless identify a direct, linear relationship between co-worker familiarity and the performance scores they assign to the physician.

Table5. *Co-worker Familiarity and PAR Scores (ANOVA)*

Specialty		**Familiarity**						
	n (%) reported	not well	somewhat	well	very well			
		M (SD)	M (SD)	M (SD)	M (SD)	F	*p*	ω^2
Anesthesia	758 (68)	-	69.3 (14.8)	79.4 (12.4)	85.6 (11.1)	57.56	<.001	0.36
Diagnostic Imaging	887 (78)	37 (22.8)	71.4 (18.3)	84.7 (14.0)	91.6 (12.9)	47.55	<.001	0.44
Episodic Medicine	751 (72)	59 (19.3)	76.8 (13.4)	86.1 (13.0)	91.5 (9.8)	31.72	<.001	0.36
Gen Practitioner	719 (73)	17.7 (12.1)	57.9 (17.3)	70.3 (13.0)	77.0 (9.7)	56.03	<.001	0.47
Lab Medicine	794 (74)	48 (24.1)	64.9 (16.3)	78.6 (13.6)	86.1 (11.8)	59.53	<.001	0.47
Medical Specialist	792 (74)	43.3 (25.1)	73.2 (20.4)	90.9 (16.5)	99.3 (14.0)	57.70	<.001	0.48
Pediatrics	785 (73)	58 (23.7)	78.8 (22.9)	93.7 (14.5)	100.4 (11.1)	35.50	<.001	0.43
Psychiatry	768 (71)	52.6 (13.2)	78.6 (21.5)	92.6 (15.6)	101.6 (10.1)	61.30	<.001	0.46
Surgery	754 (70)	42.8 (10.7)	62.9 (19)	77.2 (13.6)	83.7 (11.5)	44.68	<.001	0.42
Effect size: ω = .10 (small); .30 (moderate); .50 (large)								

7. DISCUSSION

This study examined non-physician co-worker feedback provided to 150 physicians across nine medical specialties. The PAR co-worker tools were developed over several years but not reviewed as a unified collection of workplace-based assessments. This study enabled a comparison of co-worker feedback across specialty groupings and allowed for a review of their internal consistency and generalizability coefficients, and the issue of assessor familiarity's influence on performance scores, using data collected over a 5 year period.

All co-worker tools, including their respective sub-scales, demonstrated high internal consistency with alpha scores $\geq$ 0.90, a small standard error of measurement providing evidence for the reliability of the CAQ across specialty grouping. While the range of EP^2 approached 0.70 with eight co-workers, our *D* studies produced lower reliability coefficients across all specialty groupings suggesting that six of nine PAR specialties require a minimum of 11 co-workers for stable data, a finding similar to a UK study.[10] These findings suggest that while current co-worker data is reliable for providing formative professional development information, our generalizability coefficients do not support using co-worker feedback alone for high stakes practice decisions, a finding supported by UK research.[10]

G studies have informed MSF research and influenced data collection procedures concerning the necessary numbers of raters required to provide reliable data or the numbers of items needed on MSF questionnaires. However, *G* studies have not considered how the data collection process itself may influence the feedback provided to individual physicians. In fact, the variance components drawn from MSF *G* studies' are rarely reported or published in research studies[35]

Our study examines variance components associated with multisource feedback [e.g., rater by item nested within the physician (*ri:p) and rater nested within the physician (r:p)]* . In this study, the relationship between the 'rater and physician' and the 'rater crossed by item and nested within physician' contributed the greatest amount of variance. Our study therefore points to the need to consider the influence that data collection processes have on the quality of co-worker ratings in the workplace setting. A review of the quality assurance process associated with MSF may need to include strategies that guide physicians in how and whom they select as co-worker assessors. It may also require a more robust understanding of how co-workers make assessment decisions about

physician practice. Even knowing the value that physicians place on specific items addressed in co-worker questionnaires would be valuable in designing or revising MSF instrumentation.

8. CONCLUSION

This study confirms the reliability of co-worker scores and provides evidence that co-worker MSF data is stable and consistent for the purposes of physicians' continuing professional development. Indeed, co-worker instrumentation that are part of the suite of PAR instruments have been adopted in two other Canadian jurisdictions as part of their process for physician appraisal and professional development.[36] Our study provides further insight regarding the role that co-worker familiarity plays in physician assessment. Unfortunately, the absence of co-worker socio demographic information precluded a more robust exploration of this relationship. Nevertheless our findings suggest a closer look at how physicians select co-workers, and other factors (e.g., how co-workers make decisions related to scoring), that may influence PAR performance is indicated.

8.1. Practice Points

Practice Points
• Physician co-workers provide reliable, formative multisource feedback for the purpose of quality assurance and collaborative interprofessional practice.
• Co-worker familiarity with physician ratees is positively correlated with physician performance scores.
• Routine review of variance components described in MSF *G* and *D*-studies is necessary for ongoing revision of co-worker instrumentation.
• Quality assurance processes associated with MSF may need to include strategies that guide physicians in how and whom they select as co-worker assessors

ACKNOWLEDGEMENTS

I would like to thank **Dr. Pamela Nordstrom, Dr. Tanya Beran** and **Dr Kent Hecker** who supported my doctoral research at the University of Calgary and provided feedback on the development of this manuscript. Special thanks to **Dr. David Keane** at McMaster University who consulted on the ur-Genova software.

Declaration of Interest

The authors report no declarations of interest.

REFERENCES

[1] Bracken D, Timmreck C, Church A. The handbook of multisource feedback: The comprehensive resource for designing and implementing MSF processes. San Francisco, CA Jossey-Bass 2001.

[2] Lockyer J, Clyman S. Multisource feedback (360-degree evaluation). In: Hawkins ESHRE, ed. Practical Guide to Evaluation of Clinical Competence. Philadelphia, PA: Mosby Elsevier; 2008:10.

[3] Sargeant J, Mann K, Sinclair D, van der Vleuten C, Metsemakers J. Challenges in multisource feedback: Intended and unintended outcomes. Medical Education 2007;41:583-91.

[4] Lipner R, Blank L, Leas B, Fortna G. The value of patient and peer ratings in recertification. Academic Medicine 2002;77:S64-S6.

[5] Chesluk B, Bernabeo E, Hess B, Lynn L, Reddy S, Holmboe E. A new tool to give hospitalists feedback to improve interprofessional teamwork and advance patient care. Health Affairs 2012;31:2485-92.

[6] Archer J, McGraw M, Davies H. Assuring validity of multisource feedback in a national programme. Archives of Disease in Childhood 2010;95:330-5.

[7] Comparison of colleague and patient multisource feedback instruments designed for GPs in the UK. Royal College of General Practitioners, 2010. (Accessed at http://www.cfepsurveys.co.uk/documents/RCGP_comparison_of_colleague_and_patient_MSF_instruments_2010.pdf.)

[8] Overeem K, Lombarts M, Arah O, Alazinga N, Grol RPTM, Wollersheim HC. Three methods of multi-source feedback compared: a plea for narrative comments and coworkers' perspectives. Medical Teacher 2010;32:141-7.

[9] Overeem K, Wollersheim H, Arah O, Cruijsberg J, Grol R, Lombarts K. Factors predicting doctors' reporting of performance change in response to multisource feedback. BMC Medical Education 2012;12:52.

[10] Campbell J, Richards S, Dickens A, Greco M, Narayanan A, Brearley S. Assessing the professional performance of UK doctors: An evaluation of the utility of the General Medical Council patient and colleague questionnaires. Quality & Safety in Health Care 2008;17:187-93.

[11] Sargeant J, Mann K, Ferrier S, et al. Responses of rural family physicians and their colleague and coworker raters to a multi-source feedback process: A pilot study. Academic Medicine 2003;78:S42-S4.

[12] Violato C, Lockyer J, Fidler H. Changes in performance: A 5-year longitudinal study of participants in a multi-source feedback programme. Medical Education 2008;42:1007-13.

[13] Archer J, Norcini J, Southgate L, Heard S, Davies HT. mini-PAT (Peer Assessment Tool): A valid component of a national assessment programme in the UK? Advances in Health Sciences Education: Theory and Practice 2008;13:181-92.

[14] Wright C, Richards S, Hill J, et al. Multisource feedback in evaluating the performance of doctors: The example of the UK General Medical Council patient and colleague questionnaires. Academic Medicine 2013;Publish Ahead of Print:10.1097/ACM.0b013e3182724cc0.

[15] Sargeant J, Mann K, Ferrier S. Exploring family physicians' reactions to multisource feedback: Perceptions of credibility and usefulness. Medical Education 2005;39:497-504.

[16] Hall W, Violato C, Lewkonia R, et al. Assessment of physician performance in Alberta: The physician achievement review. CMAJ 1999;161:52-7.

[17] Violato C, Marini A, Toews J, Lockyer J, Fidler H. Feasibility and psychometric properties of using peers, consulting physicians, co-workers, and patients to assess physicians. Academic Medicine 1997;72:S82-S4.

[18] Lockyer J, Violato C, Fidler H. A multi source feedback program for anesthesiologists. Canadian Journal of Anaesthesia 2006;53:33-9.

[19] Lockyer J, Violato C, Fidler H. The assessment of emergency physicians by a regulatory authority. Academic Emergency Medicine 2006;13:1296-303.

[20] Lockyer J, Violato C, Fidler H. Assessment of radiology physicians by a regulatory authority. Radiology 2008;247:771-8.

[21] Lockyer J, Violato C, Fidler H, Alakija P. The assessment of pathologists/laboratory medicine physicians through a multisource feedback tool. Archives Of Pathology & Laboratory Medicine 2009;133:1301-8.

[22] Violato C, Lockyer J, Fidler H. Multisource feedback: A method of assessing surgical practice. BMJ (Clinical Research Ed) 2003;326:546-8.

[23] Violato C, Lockyer J, Fidler H. Assessment of pediatricians by a regulatory authority. Pediatrics 2006;117:796-802.

[24] Violato C, Lockyer J, Fidler H. Assessment of psychiatrists in practice through multisource feedback. Canadian Journal Of Psychiatry 2008;53:525-33.

[25] Violato C, Lockyer J. Self and peer assessment of pediatricians, psychiatrists and medicine specialists: Implications for self-directed learning. Advances in Health Sciences Education: Theory and Practice 2006;11:235-44.

[26] Brennan R, ed. Educational measurement. 2 ed. Westport, CT: American Council on Educaiton and Praeger Publishers; 2006.

[27] Narayanan A, Greco M, Campbell J. Generalisability in unbalanced, uncrossed, and fully nested studies. Medical Education 2010;44:367-78.

[28] Ramsey P, Wenrich M, Carline J, Inui T, Larson E, LoGerfo J. Use of peer ratings to evaluate physician performance. JAMA 1993;269:1655-60.

[29] Campbell J, Roberts M, Wright C, et al. Factors associated with variability in the assessment of UK doctors' professionalism: Analysis of survey results. BMJ 2011;343.

[30] Archer J, McAvoy P. Factors that might undermine the validity of patient and multi-source feedback. Medical Education 2011;45:886-93.

[31] Bloch R, Norman G. Generalizability theory for the perplexed: A practical introduction and guide: AMEE Guide No. 68. Medical Teacher 2012;34:960-92.
[32] Overeem K, Wollersheim H, Arah O, Cruijsberg J, Grol R, Lombarts K. Evaluation of physicians' professional performance: An iterative development and validation study of multisource feedback instruments. BMC Health Services Research 2012;12:1-11.
[33] Tabachnick B, Fidell L, eds. Using multivariate statistics. 5 ed. Phildelphia, PA: Allyn & Bacon; 2006.
[34] Field A. Discovering statistics using SPSS. 3rd ed. London: UK: Sage; 2009.
[35] Hecker K, Adams C, Coe J. Assessment of first-year veterinary students' communication skills using an objective structured clinical examination: The importance of context. Journal of Veterinary Medical Education 2012;Advance Online Article:1-7.
[36] Nova Scotia Physician Achievement Review Pragram (NSPAR). CPSNS, 2005. (Accessed July 25, 2012, at http://www.nspar.ca/.)

AUTHORS' BIOGRAPHY

Gregg C. Trueman, PhD, NP is a nurse practitioner and associate professor in the School of Nursing and Midwifery at Mount Royal University in the Faculty of Community and Health Studies: Y348, 4825 Mount Royal Gate SW, Calgary, Alberta Canada, This research was conducted as part of his doctoral program under the supervision of Dr Jocelyn Lockyer at the University of Calgary.

Jocelyn M. Lockyer, PhD is professor and Senior Associate Dean-Education in the Faculty of Medicine at University of Calgary.