Education Research: Bias and poor interrater reliability in evaluating the neurology clinical skills examination

L.A. Schuh, MD
Z. London, MD
R. Neel, MD
C. Brock, MD
B.M. Kissela, MD
L. Schultz, PhD
D.J. Gelb, MD

Address correspondence and reprint requests to Dr. Lori Schuh, Department of Neurology, Henry Ford Hospital, 2799 West Grand Blvd., Detroit, MI 48202
lschuh@neuro.hfh.edu

ABSTRACT

Objective: The American Board of Psychiatry and Neurology (ABPN) has recently replaced the traditional, centralized oral examination with the locally administered Neurology Clinical Skills Examination (NEX). The ABPN postulated the experience with the NEX would be similar to the Mini-Clinical Evaluation Exercise, a reliable and valid assessment tool. The reliability and validity of the NEX has not been established.

Methods: NEX encounters were videotaped at 4 neurology programs. Local faculty and ABPN examiners graded the encounters using 2 different evaluation forms: an ABPN form and one with a contracted rating scale. Some NEX encounters were purposely failed by residents. Cohen's kappa and intraclass correlation coefficients (ICC) were calculated for local vs ABPN examiners.

Results: Ninety-eight videotaped NEX encounters of 32 residents were evaluated by 20 local faculty evaluators and 18 ABPN examiners. The interrater reliability for a determination of pass vs fail for each encounter was poor (kappa 0.32; 95% confidence interval [CI] = 0.11, 0.53). ICC between local faculty and ABPN examiners for each performance rating on the ABPN NEX form was poor to moderate (ICC range 0.14–0.44), and did not improve with the contracted rating form (ICC range 0.09–0.36). ABPN examiners were more likely than local examiners to fail residents.

Conclusions: There is poor interrater reliability between local faculty and American Board of Psychiatry and Neurology examiners. A bias was detected for favorable assessment locally, which is concerning for the validity of the examination. Further study is needed to assess whether training can improve interrater reliability and offset bias. ***Neurology***® **2009;73:904–908**

GLOSSARY

ABIM = American Board of Internal Medicine; **ABPN** = American Board of Psychiatry and Neurology; **CI** = confidence interval; **HFH** = Henry Ford Hospital; **ICC** = intraclass correlation coefficients; **IM** = internal medicine; **mini-CEX** = Mini-Clinical Evaluation Exercise; **NEX** = Neurology Clinical Skills Examination; **RITE** = residency inservice training examination; **UC** = University of Cincinnati; **UM** = University of Michigan; **USF** = University of South Florida.

The American Board of Psychiatry and Neurology (ABPN) eliminated its centralized oral examination for several reasons. First, the encounters were not standardized. Second, a single high-stakes, high-stress encounter is not necessarily representative of a candidate's performance. Third, deficiencies were difficult to remediate for failed candidates as they had already completed training. To replace the centralized oral examination, the ABPN Neurology Council voted to require residency programs to administer a set of clinical skills examinations called the Neurology Evaluation Exercise (NEX) to all residents entering training as of July 1, 2005, and document every graduating resident's satisfactory performance on these. The ABPN proposed this approach after reviewing the American Board of Internal Medicine (ABIM) experience with clinical skills evaluation in internal medicine (IM) residency using the Mini-Clinical

Supplemental data at www.neurology.org

Editorial, page 826

e-Pub ahead of print on July 15, 2009, at www.neurology.org.

From the Departments of Neurology (L.A.S.) and Biostatistics and Research Epidemiology (L.S.), Henry Ford Hospital, Detroit, MI; Department of Neurology (Z.L., D.J.G.), University of Michigan Medical School, Ann Arbor; Department of Neurology (R.N., B.M.K.), University of Cincinnati, OH; and Department of Neurology (C.B.), University of South Florida, Tampa.

This study was funded by an AAN Education Research Grant.

Disclosure: Author disclosures are provided at the end of the article.

Evaluation Exercise (mini-CEX), which had also been developed to replace a single bedside oral examination that was felt to be poorly generalizable.[1-3]

The mini-CEX was designed to assess residents with respect to clinical skills, attitudes, and behaviors that are essential in providing high-quality patient care. In a study of 421 PGY-1 IM residents from 21 programs, evaluated by 316 physicians, the 7 components of competence were highly correlated and reliable.[4] Four encounters per resident produced acceptable confidence intervals (CIs) for residents with aggregate scores at the midpoint (marginal/satisfactory performance) or higher on the evaluation scale. Other investigators, using scripted videotapes of standardized patients and standardized residents, demonstrated construct validity (degree to which a test measures the theoretical concept it intends to measure) of the mini-CEX, with faculty evaluators demonstrating discrimination among unsatisfactory, satisfactory, and superior clinical skills by standardized residents.[5] Concurrent validity, or the degree to which a measurement instrument produces the same results as another proven instrument measuring the same variable, has also been demonstrated with the mini-CEX, but only in one residency program with correlations between monthly evaluation forms and inservice training examination scores.[6]

In contrast to the mini-CEX, no validation studies have been performed on the NEX. The ABPN selected 5 clinical evaluations as the minimum needed based on extrapolation from the mini-CEX studies, but the neurologic and general medical examinations differ, so it is not clear how many examinations are needed for reliable results. The need for evaluation and validation of the NEX is urgent. If the process if flawed, this must be learned quickly to prevent wasted effort from program directors, minimize inconvenience to residents, and protect the public by maintaining high standards for neurology board certification. Furthermore, the mini-CEX studies did not address the potential for bias from local evaluating faculty. The reliability and validity studies of the mini-CEX were performed long after the live patient examination had been eliminated from the ABIM certifying examination, so comparison to gold standard ABIM board examiners was not possible. Neurology has a unique opportunity to address how presumptive gold standard ABPN examiners would grade NEX encounters.

The primary aim of this study was to study interrater reliability and bias in the NEX comparing untrained local faculty to trained, unaffiliated ABPN examiners. The secondary aims were to study the concurrent validity of resident NEX evaluations with neurology residency inservice training examination (RITE) scores and to study the number of examinations needed for reliable testing and a passing score.

METHODS Residency program directors were offered an opportunity to participate in multicenter educational research at the 2006 Consortium of Neurology Program Director's Meeting. Our goal was to enroll a variety of programs from across the United States, with differing levels of academic achievement as demonstrated by 2006 RITE scores of PGY-2 residents. Ultimately, 4 programs participated, 3 in the Midwest and 1 in the Southeast: Henry Ford Hospital (HFH), University of Michigan (UM), University of Cincinnati (UC), and University of South Florida (USF). Three of these programs are university-based, and one is an urban community-based residency program.

Current ABPN board examiners were recruited through a mail posting and announcement describing the study at an examination. Local faculty evaluators were voluntarily recruited by study co-investigators at each site.

Demographic data were collected on each participating resident, ABPN evaluator, and faculty evaluator. Demographic data on residents included months of training in all residencies, months of training in neurology, age, and sex. Demographic data on ABPN and faculty evaluators included age, sex, years of teaching experience, academic rank, number of ABPN examinations service, and status as program director or associate program director.

Patients were offered compensation with parking reimbursement and a small gift card. ABPN examiners were offered compensation with a small honorarium. Residents and local faculty were not compensated.

Each site's local Institutional Review Board approved the study. Resident informed consent was obtained prior to participation in this study. Patient consent for videotaping was obtained. Local faculty evaluators and ABPN examiners were exempted from informed consent. We preferentially recruited PGY-2 residents, given the greater likelihood of variable and unsatisfactory performances in less senior trainees. Of the 5 required settings for NEX encounters (child neurology, ambulatory/episodic disorders, neuromuscular disorders, critical care, and neurodegenerative/movement disorders), we excluded critical care and pediatric encounters for technical reasons. At 2 sites (HFH and UC), residents knew which faculty were evaluating them. Lack of anonymity could conceivably bias local faculty in favor of passing their residents. We attempted to minimize this ef-

fect by informing faculty that residents would sometimes be intentionally performing at a failing level. A standardized patient encounter was developed in which residents were coached to perform at a failing level in a specific manner. Residents were limited to 45 minutes for each encounter, an ABPN requirement. Two evaluation forms were used in each encounter: the ABPN sanctioned NEX 1 form with an 8-point grading scale,[7] and a modified form with a contracted 3-point rating scale (appendix e-1 on the *Neurology*® Web site at www.neurology.org). Residents must pass each section (medical interviewing skills, examination skills, and humanistic/professionalism skills) for an overall passing score.

To standardize encounters, written instructions for performing and evaluating the NEX were developed and distributed to all participants (appendix e-2). ABPN grading instructions were distributed to local faculty and ABPN examiners.[8] Each local faculty and ABPN examiner viewed no more than one purposely failed encounter. Encounters were de-identified as much as possible to minimize potential evaluator bias and protect confidentiality. Aliases were used by residents and patients. Residents were instructed to avoid questions that might identify the location of the encounter, but this was divulged on rare occasion.

The NEX encounters were videotaped by the co-investigator or designee. One local faculty evaluator and ABPN examiner reviewed each DVD NEX encounter and completed the NEX 1 and modified form. All data were entered into a database by the statistician investigator.

Descriptive statistics were computed for the demographic information for residents, local faculty examiners, and ABPN examiners. Kappa statistics were computed to assess the interrater agreement between faculty and ABPN examiners for pass/fail responses. Kappa statistics greater than 0.75 represent excellent agreement, 0.4 to 0.75 good agreement, and less than 0.4 poor agreement.[9] Intraclass correlation coefficients (ICC) were computed to assess the agreement between faculty and ABPN examiners for each performance rating component on the NEX form.

Wilcoxon 2-sample *t* tests were used to evaluate the relationship between RITE scores and the individual resident's performance. Residents were separated into 2 groups for this analysis, based on whether they passed their first NEX encounter.

Generalized estimating equations methods were used to assess the passing evaluation rate with increasing number of encounters. This method was selected to take into account multiple encounters from the same resident. In addition, the number of encounters was grouped into first encounter, second encounter, and 3 or more encounters. For these analyses, only residents with encounters performed over time and provided feedback between encounters were included.

RESULTS Thirty-two residents (21 men, 11 women) participated. The majority (63%) were PGY-2 residents. Pre-study performance on the 2006 RITE was variable among these programs with a mean of 31.5% (range 0%–60%) of PGY-2 residents at these programs scoring above the 75th percentile among residents in the same year of training. Twenty-two local faculty with a mean of 7.9 ± 8.5 years of teaching experience participated. Most (68%) were assistant professor rank or lower and only 2 local faculty evaluators had ever served as ABPN examiners. Individuals participated as a local faculty evaluator or ABPN examiner, not both. Eighteen ABPN examiners were recruited with a mean of 21.1 ± 10.6 years of teaching. The majority (78%) were associate professor rank or higher. A total of 98 NEX encounters (20 neurodegenerative, 29 neuromuscular, and 49 ambulatory) were reviewed by both local faculty and ABPN examiners. Twelve encounters were intentional failures. Thirty-three encounters were from HFH, 25 from UM, 20 from UC, and 20 from USF.

Table 1 ABPN and local examiners encounter results

		Local examiners		
		Pass	Fail	No.
All encounters				
ABPN	Pass	63	8	71
Examiners	Fail	16	11	27
	No.	79	19	98
True encounters				
ABPN	Pass	62	7	69
Examiners	Fail	14	3	17
	No.	76	10	86
Intentional failure encounters				
ABPN	Pass	1	1	2
Examiners	Fail	2	8	10
	No.	3	9	12

ABPN = American Board of Psychiatry and Neurology.

Table 1 presents the concordant and discordant results between ABPN and local faculty examiners for all encounters, as well as for true encounters and intentional failure encounters, separately. The interrater reliability for a determination of pass vs fail for each encounter was poor (kappa = 0.32; 95% CI = 0.11, 0.53). For all encounters in which ABPN examiners failed a performance, local faculty agreed 40.7% of the time (individual program agreement ranged from 14.3%–60%). When ABPN examiners passed a performance, local faculty agreed 88.7% of the time for all encounters (individual program agreement ranged from 76.9% to 100%). There was no relationship between local faculty years of teaching experience and grading correlation with ABPN examiners. No difference was seen when the 2 local faculty who had served as ABPN examiners were excluded from the analysis (kappa = 0.32; 95% CI = 0.10, 0.54). Faculty at the sites where residents were blinded to the identity of their evaluators did not have a higher concordance with ABPN examiners than faculty at the non-anonymous sites. In fact, there was a trend for faculty from the non-anonymous sites to agree more closely with ABPN examiners (non-anonymous kappa = 0.49 vs anonymous kappa = 0.12, $p = 0.07$).

Table 2 ABPN pass rates by number of encounters performed (n = 20 residents)

No. of encounters	No.	No. pass (%)
1	20	13 (65)
2	18	16 (88.9)
3 or more	12	11 (91.7)

ABPN = American Board of Psychiatry and Neurology.

ICCs between local faculty and ABPN examiners for each performance rating on the NEX 1 form were poor to moderate, ranging 0.14–0.44 for each rating. The ICC did not improve with the use of a contracted 3-point rating form (ICC range 0.09–0.36).

RITE examination scores from 2008 were available for 28 residents. One resident completed only a single intentional failure and was not included in the RITE analysis. Of the remaining 27, 20 passed their first encounter. This group of passing residents had higher median RITE scores than the 7 residents with a failed first encounter, but none of these differences was significant (median percentile RITE score vs same year in training 46 vs 31 and median percent correct 55.5 vs 48).

Residents at UM and HFH performed examinations over a period of time and were given feedback following each examination. There were 20 residents with 50 true encounters. There was a positive trend in the ABPN examiner assessed rate of passing NEX evaluations with later encounters (table 2: 65% for first encounter, 88.9% for second encounters, and 91.7% for 3 or more encounters). There was improvement between the first and second encounters ($p = 0.033$) and a strong trend between the first encounter and 3 or more encounters ($p = 0.05$). No improvement was seen between the second encounter and 3 or more encounters ($p = 0.442$).

DISCUSSION If the ABPN examiner's evaluations are a gold standard, these results suggest local faculty may be biased in assessing their own resident's NEX performance. Local faculty were twice as likely to concur with ABPN examiners on a passing vs failing grade. This result did not depend on whether residents knew who was evaluating them. Some of the difference in grading may be due to stratification of grading by local faculty. Local faculty knew the training level of the resident and may have assessed performance based on the expected performance for that year of training (e.g., PGY-2), rather than on the expected performance of a neurology residency graduate, as ABPN examiners have been trained to do. Local examiners might be more likely to forgive unsatisfactory performance among residents they know well, while ABPN examiners might consider these behaviors grounds for failure. Local faculty may consider the daily clinical performance of a resident, and not just the performance in this staged situation. Therefore, even if bias exists among local faculty, it does not necessarily follow that this bias invalidates the results of determining clinical competence. In fact, local results may be more valid than ABPN-based results, where truly competent neurologists could have failed for performance errors under pressure in an artificial situation.

Even though this study demonstrated poor interrater reliability between local faculty and ABPN examiners, there was agreement between both about 75% of the time on the essential issue of overall pass/fail performance. With multiple resident observations, this may still represent an improvement over the single high stakes observation of the traditional centralized live patient examination. The ABPN maintains that the hurdle to neurology board certification is not passing the oral examination, but the written examination. The NEX therefore may be sufficient to the needs of board certification even if it is not a strongly reliable tool.

It might be possible to design a faculty development course to improve assessment of neurology resident clinical skills. One group of investigators found that 11 of 40 faculty inaccurately rated taped mini-CEX performances as satisfactory when they were not.[5] In a follow-up study, the investigators used videotapes in a 4-day faculty development course with lectures and a variety of interactive evaluation exercises with standardized residents and patients.[10] Eight months later, those randomized to the educational intervention rated videotaped encounters between standardized residents and patients more stringently than control faculty who did not take the course. It is not known whether the same result could be accomplished with less training. Earlier studies with short duration faculty training did not improve faculty rating skills.[1] It would be very difficult for programs to find and fund faculty to attend a 4-day-long course to become NEX evaluators. Any requirement of prior faculty training to serve as a NEX evaluator could pose logistical difficulties, and could negate one of the express purposes of the NEX: to have multiple observers evaluate each resident. From a logistic standpoint, a Web-based training and assessment program would be preferred, but any faculty development tool must be studied for effectiveness and ability to mitigate bias.

Concern about interrater reliability is mitigated by improvement in ABPN examiners' ratings of resident performance with increasing number of examinations. Reliable resident performance was obtained

after only 2 NEX encounters. Reducing the number of completed NEX encounters from 5 to a lower number may reduce stress on residents and programs without reducing the reliability of the NEX, although one reason to retain all 5 is that they each require demonstration of unique history and examination skills.

One limitation of this study is the limited number of participating neurology programs, residents, and NEX encounters. Another limitation is that each NEX encounter was evaluated by only one ABPN examiner, so the interrater reliability among ABPN examiners was not assessed. Before making decisions based on the gold standard of ABPN examiners, their interrater reliability should be evaluated. If their interrater reliability is good, then the NEX process might be strengthened by routinely sending videotapes to ABPN examiners for review. Given that the supply of ABPN examiners will diminish with time, an alternative would be to send the videotapes to faculty outside the local program. Thus, another possible follow-up study may be to evaluate the interrater reliability between outside faculty and ABPN examiners.

Follow-up studies are essential. We must first address the interrater reliability between 2 groups of ABPN examiners to determine if the NEX tool is reliable. If there is good interrater reliability between ABPN examiners, we should address the interrater reliability between unaffiliated faculty and ABPN examiners. If a faculty development tool is developed for assessing resident skills in the NEX, it is essential that the effectiveness of that tool be rigorously studied to make certain it is accomplishing its stated goals. We cannot assume that training local faculty will eliminate bias.

AUTHOR CONTRIBUTIONS

Statistical analysis was completed by Lonni Schultz, PhD, Henry Ford Hospital.

ACKNOWLEDGMENT

The NEX Study Group thanks the neurology residents and faculty from Henry Ford Hospital, University of Michigan, University of Cincinnati, and University of South Florida, and the ABPN Examiners who contributed to this work.

DISCLOSURE

This study was funded by an American Academy of Neurology Education Research Grant. Dr. Schuh has received funding for travel from the ACGME and the AAN and has received an honorarium from the ACGME (Parker Palmer Award). Dr. London serves as an editor for *Medlink Neurology* online. Dr. Neel reports no disclosures. Dr. Brock serves as Book Review Editor for the *Journal of Neuroimaging* and serves on a speakers' bureau for Serono. Dr. Kissela has served as an ABPN examiner and is a current member of the ACGME Neurology Review Committee; has served on a scientific advisory board for Northstar Neuroscience and a speakers' bureau for Boehringer-Ingelheim; serves on the Neurology Residency Review Committee; has served as an expert witness in medicolegal cases related to stroke (has performed chart reviews and been deposed once); and receives research support from the NIH (Principal Investigator: NIH-NINDS R-01 NS30678, Executive Committee or Site PI: NIH-NINDS R-01 NS039987, NIH-NINDS U-01 NS041588). Dr. Schultz reports no disclosures. Dr. Gelb has served as an ABPN examiner; serves as an editor for *Current Opinion in Internal Medicine*; receives royalties from the publication of *Introduction to Clinical Neurology* (Elsevier 1995), *UpToDate* (chapter author, 2001–2009), and *MedLink Neurology* (chapter author, 1998–2009); and has received speakers honoraria from Washington University, Mayo Clinic, the American University of the Caribbean, and the University of Pennsylvania.

Received January 21, 2009. Accepted in final form April 24, 2009.

REFERENCES

1. Noel GL, Herbers JE, Jr., Caplow MP, et al. How well do internal medicine faculty members evaluate the clinical skills of residents? Ann Intern Med 1992;117:757–765.
2. Kroboth FJ, Hanusa BH, Parker S, et al. The inter-rater reliability and internal consistency of a clinical evaluation exercise. J Gen Intern Med 1992;7:174–179.
3. Woolliscroft JO, Stross JK, Silva J, Jr. Clinical competence certification: a critical appraisal. J Med Educ 1984;59:799–805.
4. Norcini JJ, Blank LL, Duffy FD, Fortna GS. The Mini-CEX: a method for assessing clinical skills. Ann Intern Med 2003;138:476–481.
5. Holmboe ES, Huot S, Chung J, Norcini H, Hawkins RE. Construct validity of the MiniClinical Evaluation Exercise (MiniCEX). Acad Med 2003;78:826–830.
6. Durning SJ, Cation LJ, Markert RJ, Pangaro LN. Assessing the reliability and validity of the mini-clinical evaluation exercise for internal medicine resident training. Acad Med 2002;77:900–904.
7. American Board of Psychiatry and Neurology. NEX 1 form. Available at: http://www.abpn.com/downloads/forms/ABPN_NEX_form_v1.pdf. Accessed January 15, 2009.
8. American Board of Psychiatry and Neurology. Clinical skills evaluation of residents in neurology and child neurology. Available at: http://www.abpn.com/downloads/forms/NEX%20Clinical%20Skills%20Evaluation%20Instructions.pdf. Accessed November 10, 2008.
9. Fleiss JL, Levin B, Paik MC. Statistical Methods for Rates and Proportions: Third Edition. John Wiley & Sons: New Jersey; 2003:604.
10. Holmboe ES, Hawkins RE, Huot SJ. Effects of training in direct observation of medical residents' clinical competence. Ann Intern Med 2004;140:874–881.

Education Research: Bias and poor interrater reliability in evaluating the neurology clinical skills examination

L. A. Schuh, Z. London, R. Neel, et al.

Neurology 2009;73;904-908 Published Online before print July 15, 2009
DOI 10.1212/WNL.0b013e3181b35212