RESIDENT & FELLOW SECTION

Section Editor
Mitchell S.V. Elkind, MD, MS

Education Research: Unsatisfactory NEX rating correlations

Searching for the reasons

Zachary London, MD
Lori Schuh, MD
Douglas J. Gelb, MD, PhD
Lonni Schultz, PhD

Correspondence to Dr. London: zlondon@med.umich.edu

ABSTRACT

Objectives: To determine whether the previously demonstrated poor correlation between local faculty and external American Board of Psychiatry and Neurology (ABPN) examiners evaluating the Neurology Evaluation Exercise (NEX) is attributable to a difference between raters who know the residents and raters who do not, a difference between raters with ABPN experience and raters without it, or some other factor.

Methods: Deidentified NEX encounters were videotaped at 2 neurology residency programs. Each video was graded by 1 local faculty examiner, 1 external faculty examiner with ABPN experience, and 1 external faculty examiner without ABPN experience, using the ABPN-sanctioned form. Acceptable/unacceptable rates were compared using Cohen κ statistic.

Results: Fifty-eight videotaped NEX encounters involving 20 residents were evaluated by 12 local faculty examiners, 13 ABPN examiners, and 10 external non-ABPN examiners. The level of agreement between groups failed to meet our target κ of 0.7 (ABPN vs non-ABPN external examiners: κ = 0.47 [95% confidence interval 0.21–0.73]; local vs non-ABPN external examiners: κ = 0.37 [95% confidence interval 0.08–0.66]; local vs ABPN external examiners: κ = 0.40 [95% confidence interval 0.14–0.67]). Local, non-ABPN, and ABPN examiners assigned a failing grade to 13 (22%), 11 (19%), and 16 (28%) of the NEX encounters, respectively.

Conclusions: The disappointing correlation between local examiners, non-ABPN external examiners, and ABPN external examiners is not solely attributable to bias toward familiar residents. Inadequate training in NEX administration and scoring could be a factor. It is also possible that the NEX is not a valid tool. Further study is necessary. *Neurology*® 2013;80:e142–e145

GLOSSARY

ABPN = American Board of Psychiatry and Neurology; **CI** = confidence interval; **ICC** = intraclass correlation coefficient; **NEX** = Neurology Evaluation Exercise.

Oral examinations are used in neurology training programs worldwide as a means of assessing the clinical skills of resident physicians. The American Board of Psychiatry and Neurology (ABPN) replaced the centralized oral examination with a series of clinical skill examinations administered by individual programs (usually to their own residents), the Neurology Evaluation Exercise (NEX). The NEX was adopted before any validation studies were performed.

In a previous study by our group, local faculty and unaffiliated ABPN examiners evaluating videotaped NEX encounters had poor interrater reliability for a determination of pass or fail (κ = 0.32; 95% confidence interval [CI] = 0.11–0.53).[1] Of 98 total encounters, the ABPN examiners assigned a failing grade to 27, but local faculty recommended a failing grade for only 11 (40.7%) of these. Local faculty were twice as likely to agree with the ABPN examiners who assigned a passing grade to a given encounter than those who assigned a failing grade.

We discussed several possible explanations for the disparity between ABPN and local faculty in the grading of NEX encounters, one of which was that local faculty members might be biased in favor of passing their own residents. It is easy to conceive of reasons why such a bias might occur. Local evaluators might be more willing to forgive a single unsatisfactory performance by a resident who is well-

From the Department of Neurology (Z.L., D.J.G.), University of Michigan, Ann Arbor; and Departments of Neurology (Dr. L. Schuh) and Public Health Sciences (Dr. L. Schultz), Henry Ford Hospital, Detroit, MI.

Go to Neurology.org for full disclosures. Funding information and disclosures deemed relevant by the authors, if any, are provided at the end of the article.

© 2013 American Academy of Neurology

© 2013 American Academy of Neurology. Unauthorized reproduction of this article is prohibited.

known to them, interpreting the encounter in the context of the resident's overall performance. Faculty may be less willing to fail a resident with whom they work because of the potential repercussions (such as strained interpersonal relationships, the additional work generated by repeat NEX encounters, or even the possibility of dismissal of a resident from the program).

Regardless of the cause, if local bias is common then the elimination of the centralized oral examination could result in insufficient scrutiny of neurology trainees' clinical skills. This study was designed to determine whether bias on the part of local evaluators is the primary reason for the disparity in NEX grading between local faculty and unaffiliated ABPN examiners. The study was also designed to determine the extent to which the disparity related to a difference between evaluators who had previously served as ABPN examiners and those who had not.

METHODS **Standard protocol approvals, registrations, and patient consents.** The Institutional Review Board at each site approved the study, and residents gave written informed consent. Patients gave written authorization for videotaping. ABPN board examiners were recruited through a mail posting describing the study and were paid a small honorarium. External and local faculty were recruited by site investigators and also received a small honorarium. Residents and patients were recruited by site investigators. Patients received compensation in the form of parking reimbursement and a small gift card. Residents received no compensation.

All encounters could be categorized as ambulatory, neurodegenerative, or neuromuscular. Critical care and pediatric neurology encounters were excluded from the study because of challenges related to obtaining authorized videotaping.

Although most encounters were true clinical skills examinations with real patients, some encounters were between residents and standardized patients. In these encounters, residents were instructed to perform at a failing level intentionally. Evaluators were informed that there might be intentionally poor performances among the encounters that they were evaluating.

Investigators developed written instructions for performing and evaluating the NEX and distributed these to all participants. To minimize potential evaluator bias and protect confidentiality, patients and residents used aliases and physician identification badges were removed. Residents were asked to avoid divulging the location of the encounter during the interview.

A coinvestigator or designee videotaped the clinical encounter between the resident and the patient. Encounters were transferred to DVD. Each DVD was viewed and evaluated by 3 different people: an ABPN examiner (all were active at the time), a board-certified neurologist working at an academic institution other than the one at which the resident trains, and a board-certified neurologist working at the same institution as the resident. Local and non-ABPN external faculty were matched for academic rank. Evaluators completed the ABPN-sanctioned NEX 1 form, which uses an 8-point grading system. To receive an overall passing grade, residents must pass each section: medical interviewing skills, examination skills, and humanistic/professionalism skills.

Statistical analysis. The overall pass/fail response along with 8-point grades for skills listed on the NEX 1 form were determined by the ABPN, local faculty, and external faculty examiners for each resident encounter. Cohen κ statistics were computed to assess the interrater agreement between the ABPN examiners and the local and external faculty for the pass/fail responses. Kappa statistics >0.75 represent excellent agreement, 0.4 to 0.75 moderate agreement, and less than 0.4 poor agreement. The predetermined target κ was ≥ 0.7 to reflect very good to excellent agreement.

Intraclass correlation coefficients (ICCs) were computed to assess the agreement between ABPN examiners and local and external faculty for each performance-rating component or skill on the NEX 1 form. ICC can be interpreted the same way as κ statistics, with a target of ≥ 0.7 suggesting very good to excellent correlation.

RESULTS Fifty-eight NEX encounters from 21 residents were evaluated by 13 ABPN examiners, 12 local faculty examiners, and 10 external faculty examiners. The breakdown of the case scenarios was 32 ambulatory, 15 neuromuscular, and 11 neurodegenerative. Of the types of encounters, 50 were real examinations, whereas 8 were intentionally unsatisfactory examinations. The median number of years teaching was 3.5 with a range from 0 to 32 years for local faculty and 4.5 with a range from 2 to 15 for external faculty. For ABPN examiners, the median was 20 years with a range from 6 to 41.

Local faculty, external faculty, and ABPN examiners assigned a failing grade to 13 (22%), 11 (19%), and 16 (27%) of the NEX encounters, respectively.

Table 1 compares ABPN examiners with local faculty examiners. Of the 42 encounters considered acceptable by ABPN examiners, 37 (88.1%) were also rated as acceptable by the local faculty examiners. However, of the 16 encounters considered unacceptable by ABPN examiners, only 8 (50%) were deemed unacceptable by the local faculty examiners. The κ statistic assessing the agreement between the 2 types of examiners was 0.40 (95% CI = 0.14–0.67), which is lower than the predetermined cutoff of 0.7.

Table 2 compares ABPN examiners with non-ABPN external faculty examiners. Of the 42 encounters considered acceptable by ABPN examiners, 39 (92.9%) were also rated as acceptable by the non-ABPN external faculty examiners. However, of the 16 encounters considered unacceptable by ABPN examiners, only 8 (50%) were evaluated as unacceptable by the non-ABPN external faculty examiners. The κ statistic assessing the agreement between the

Table 1 Comparison of ABPN and local faculty examiners

	Local faculty acceptable	Local faculty unacceptable	Total
ABPN acceptable	37	5	42
ABPN unacceptable	8	8	16
Total	45	13	58

Abbreviation: ABPN = American Board of Psychiatry and Neurology.

© 2013 American Academy of Neurology. Unauthorized reproduction of this article is prohibited.

Table 2 Comparison of ABPN and non-ABPN external faculty examiners

	Non-ABPN external faculty acceptable	Non-ABPN external faculty unacceptable	Total
ABPN acceptable	39	3	42
ABPN unacceptable	8	8	16
Total	47	11	58

Abbreviation: ABPN = American Board of Psychiatry and Neurology.

2 types of examiners was 0.47 (95% CI = 0.21–0.73), lower than the designated cutoff of 0.7.

Table 3 compares local with non-ABPN external faculty examiners. Of the 45 encounters considered acceptable by local faculty examiners, 40 (88.9%) were also rated as acceptable by the non-ABPN external faculty examiners. However, of the 13 encounters considered unacceptable by local faculty examiners, only 6 (46.2%) were evaluated as unacceptable by the non-ABPN external faculty examiners. The κ statistic assessing the agreement between the 2 types of examiners was 0.37 (95% CI = 0.08–0.66), once again lower than the predetermined cutoff of 0.7.

Table 4 shows how the 2 other groups of evaluators compare with ABPN evaluators in the rating of the component performance categories of the NEX. With the ABPN evaluators serving as the comparison group, the ICC measures for the local and non-ABPN external faculty were similar for most components. None of the components met the target ICC of 0.7.

DISCUSSION For each pair of evaluator groups, interrater reliability (both for the determination of acceptable vs unacceptable performance and for the individual component scores) was lower than the predetermined cutoff. The comparison between ABPN and non-ABPN external evaluators is particularly compelling. The disappointing interrater reliability between these 2 groups cannot be blamed on bias in favor of familiar residents, because neither group knew the residents on the videotapes. Further evidence against preferential treatment of familiar residents is the fact that local evaluators rated more encounters unacceptable than the non-ABPN external evaluators did. This suggests that if there was local bias, it was as likely to work against familiar residents as in their favor.

Table 3 Comparison of local and non-ABPN external faculty examiners

	Non-ABPN external faculty acceptable	Non-ABPN external faculty unacceptable	Total
Local acceptable	40	5	45
Local unacceptable	7	6	13
Total	47	11	58

Abbreviation: ABPN = American Board of Psychiatry and Neurology.

It could be that ABPN examiners are more reliable evaluators than the other 2 groups—either because of their ABPN experience and training, or because of some inherent quality that led to their selection as ABPN examiners in the first place. The average ABPN examiner in this study had been teaching for many more years than the average non-ABPN examiner. Given their seniority, ABPN examiners might have a substantially different perspective on history-taking skills and examination techniques relative to the other 2 groups of examiners. This would not explain the poor correlation between local and non-ABPN external evaluators, however.

It is conceivable that some combination of these 2 explanations could account for all of our results. The discrepancy between ABPN and non-ABPN evaluators could be attributable to the superior rating skills of the ABPN group, the discrepancy between local and non-ABPN external evaluators could be attributable to local bias (although, as noted above, this bias would have to be more complex than a unidirectional preference in favor of familiar residents), and the discrepancy between local and ABPN external evaluators could be attributable to either factor (or both). This seems unlikely, however. A much more straightforward interpretation of our results is that the current implementation of the NEX is inadequate. This could be either because evaluators have been inadequately trained or because of inherent flaws in the instrument.

The value of training clinical skills evaluators is unclear.[2,3] In the hope that a uniform training program could lead to increased interrater reliability in NEX grading, the American Academy of Neurology has developed the Clinical Skills Evaluation Training program, a moderator-led training session that is available for download on the American Academy of Neurology Web site. The program is based on written vignettes rather than live or videotaped encounters. Aggregate improvement, as defined by a short-term change in the evaluators' determinations of pass vs fail in the desired direction, was 10.3% (95% CI 5.8%–14.9%, $p < 0.001$). Larger studies demonstrating improvement in skill assessment with long-term outcomes are needed to validate this program.[4]

It remains possible that the NEX itself is simply not a valid assessment tool. The NEX evaluation form has no behavioral anchors or milestones for the evaluation scale, and the form is not differentiated by type of patient encounter. If the NEX evaluation forms do not accurately address the key facets of clinical skills evaluation, no amount of training would improve correlation among examiners. It is therefore important to study the effectiveness of the NEX, both short and long term, to prevent wasted effort by program

© 2013 American Academy of Neurology. Unauthorized reproduction of this article is prohibited.

Table 4 Comparison of the other 2 groups of evaluators with ABPN evaluators in NEX component performance categories

	Local faculty		Non-ABPN external faculty	
Performance category	No.	ICC	No.	ICC
Medical interviewing skills	56	0.331	57	0.481
Neurologic examination skills				
Mental status	53	0.399	54	0.353
Cranial nerves	56	0.031	57	0.204
Sensory	52	0.275	52	0.268
Motor examination	57	0.107	57	−0.023
Reflexes	56	0.388	55	0.302
Cerebellar	56	0.269	57	0.316
Station and gait	53	0.311	54	0.210
Humanistic qualities	53	0.392	57	0.451

Abbreviations: ABPN = American Board of Psychiatry and Neurology; ICC = intraclass correlation coefficient; NEX = Neurology Evaluation Exercise.

directors, faculty, and residents, and to maintain high standards for neurology board certification. Perhaps training programs outside of the United States can draw on the experience of the ABPN when designing their own clinical skills examinations.

Additional investigation is necessary. For example, a study of interrater reliability among former ABPN examiners rating the NEX could be informative. A high interrater reliability would suggest that with appropriate training or selection of evaluators, the NEX can be a valid instrument. A low interrater reliability would be subject to several possible interpretations, but it would further highlight concerns regarding the NEX. Until high interrater reliability between local and external evaluators can be achieved, NEX results must be interpreted with caution.

AUTHOR CONTRIBUTIONS

Dr. London and Dr. Schuh: study concept and design, acquisition of data, critical revision of the manuscript for important intellectual content. Dr. Gelb: acquisition of data, critical revision of the manuscript for important intellectual content. Dr. Schultz: statistical analysis and interpretation.

STUDY FUNDING

This study was funded by an American Academy of Neurology Education Research Grant.

DISCLOSURE

Z. London serves as an editor for *MedLink Neurology* online. L. Schuh reports no disclosures. D. Gelb has served as an ABPN examiner; receives royalties from the publication of *Introduction to Clinical Neurology* (Oxford University Press, 2012), *UpToDate* (chapter author, 2001–2012), and *MedLink Neurology* (chapter author, 1998–2012); and has received honoraria for writing questions for Continuum. L. Schultz reports no disclosures. Go to Neurology.org for full disclosures.

REFERENCES

1. Schuh LA, London Z, Neel R, et al. Education research: bias and poor interrater reliability in evaluating the neurology clinical skills examination. Neurology 2009;73:904–908.
2. Holmboe ES, Hawkins RE, Huot SJ. Effects of training in direct observation of medical residents' clinical competence: a randomized trial. Ann Intern Med 2004;140:874–881.
3. Cook DA, Dupras DM, Beckman TJ, Thomas KG, Pankratz VS. Effect of rater training on reliability and accuracy of mini-CEX scores: a randomized, controlled trial. J Gen Intern Med 2009;24:74–79.
4. Shanker V, Kass J, Potrebic S, Abramowitz M, London Z. The Clinical Skills Examination Training Program (CSET): an interactive training exercise for evaluators. Presented at American Academy of Neurology Annual Meeting; 2012; New Orleans.

© 2013 American Academy of Neurology. Unauthorized reproduction of this article is prohibited.

Neurology®

Education Research: Unsatisfactory NEX rating correlations: Searching for the reasons

Zachary London, Lori Schuh, Douglas J. Gelb, et al.

Neurology 2013;80;e142-e145
DOI 10.1212/WNL.0b013e318289702a

This information is current as of March 25, 2013

Updated Information & Services	including high resolution figures, can be found at: http://n.neurology.org/content/80/13/e142.full
References	This article cites 3 articles, 1 of which you can access for free at: http://n.neurology.org/content/80/13/e142.full#ref-list-1
Subspecialty Collections	This article, along with others on similar topics, appears in the following collection(s): **All Education** http://n.neurology.org/cgi/collection/all_education **Methods of education** http://n.neurology.org/cgi/collection/methods_of_education
Permissions & Licensing	Information about reproducing this article in parts (figures,tables) or in its entirety can be found online at: http://www.neurology.org/about/about_the_journal#permissions
Reprints	Information about ordering reprints can be found online: http://n.neurology.org/subscribers/advertise

Neurology ® is the official journal of the American Academy of Neurology. Published continuously since 1951, it is now a weekly with 48 issues per year. Copyright © 2013 American Academy of Neurology. All rights reserved. Print ISSN: 0028-3878. Online ISSN: 1526-632X.

AMERICAN ACADEMY OF NEUROLOGY®