Does Formal Instruction About the BI-RADS Ultrasound Lexicon Result in Improved Appropriate Use of the Lexicon?

Tamara Ortiz-Perez[1], **Eric J. Trevino**[2], **Karla A. Sepulveda**[1], **Susan G. Hilsenbeck**[3], **Tao Wang**[3], **Emily L. Sedgwick**[1]

[1]Department of Radiology, Baylor College of Medicine, One Baylor Plaza, MS 360, Houston, TX 77030

[2]Pearland, TX

[3]Duncan Cancer Center, Baylor College of Medicine, Houston, TX

Abstract

OBJECTIVE.—The purpose of this article is to determine whether formal instruction regarding the BI-RADS ultrasound lexicon results in improved appropriate use of the lexicon.

SUBJECTS AND METHODS.—Ninety test questions depicting the features outlined by the 2003 BI-RADS lexicon were identified in our PACS. Informed consent was obtained from 34 radiology residents. The participants took the preinstruction test and then had 1 hour of formal instruction regarding the BI-RADS ultrasound lexicon, which included images depicting the different sonographic features and final assessment (including subcategories 4a, 4b, and 4c). The participants then completed the postinstruction test, which examined the same content. Test scores were calculated for both the pre- and postinstruction tests and then were compared by a linear mixed model and Wilcoxon signed rank tests.

RESULTS.—The participants' postinstruction test scores showed significant improvement in the overall use of the BI-RADS ultrasound lexicon ($p < 0.0001$). There was also significant improvement in the following specific areas: final assessment ($p = 0.0005$), margin ($p = 0.0003$), orientation ($p = 0.0104$), and lesion boundary ($p = 0.0050$). The categories for which test scores did not show significant improvement were echo pattern ($p = 0.07$), posterior acoustic features ($p = 0.50$), shape ($p = 0.98$), and subset of the final assessment ($p = 0.24$).

CONCLUSION.—Formal instruction regarding the BI-RADS ultrasound lexicon results in improved lesion characterization and final assessment.

Keywords

BI-RADS ultrasound lexicon

In 2003, the American College of Radiology (ACR) created the first edition of the BI-RADS lexicon for breast ultrasound in an effort to standardize image interpretation and to

---

improve communication among health care providers [1]. The BI-RADS breast ultrasound lexicon has also been used to develop standardized management strategies based on lesion characteristics.

Previous research has found variability in the use of the BI-RADS ultrasound lexicon. Specifically, studies have identified significant user variation for lesion descriptors, final assessment categories, and BI-RADS category 4 subcategories (4a, 4b, and 4c). Prior work has shown fair interobserver agreement for echo pattern [2, 3]. Other studies revealed fair-to-moderate agreement for the margin descriptor [3, 4]. Moderate agreement for the final assessment has been seen; however, low interobserver agreement was noted for the BI-RADS category 4 subcategories [2, 4].

Sources of variability in interpretation of breast ultrasound have been identified. Variability in operator technique is inherent to ultrasound as a modality. Interpretations may also be influenced by the amount of training or experience of the radiologist. This source of variability may be affected by formal training [2]. Research has provided evidence that the appropriate use of the BI-RADS mammography lexicon improves with formal training [2]. To our knowledge, no study has been performed to document the utility of such training in the BI-RADS ultrasound lexicon.

The objective of this study is to determine whether formal instruction about the BI-RADS ultrasound lexicon results in improved appropriate use of the lexicon. To our knowledge, this is the first study to assess the effect of formal instruction on correct use of the BI-RADS ultrasound lexicon.

Subjects and Methods

This prospective study was approved by the institutional review board and was HIPAA compliant. Informed consent was obtained from study participants. Authors included two breast imaging fellows (< 1 year of experience) and two breast imaging fellowship-trained attending physicians (6 and 9 years of experience).

Queries using the ultrasound BI-RADS descriptors were performed in our PACS (ISite, Philips Healthcare) using the Primordial Search Tool (Primordial Customized Radiology Solutions) and our PenRad system (PenRad Technologies). A total of 1515 patient images were retrieved from January 2009 to June 2011. These images were reviewed by two breast imaging fellows (< 1 year of experience) and subsequently by one fellowship-trained breast imager (9 years of experience). A final test set of 90 questions was developed by author consensus from patients with ultrasound examinations that showed findings described by the BI-RADS ultrasound lexicon. These findings were categorized retrospectively using the lexicon for ultrasound established by the ACR in 2003 [1]. Histologic diagnosis was available via ultrasound-guided biopsy for lesions that were originally categorized as BI-RADS category 4 or 5. For masses representative of the category 4 subcategories, masses with imaging features depicted in a previous publication were used [1, 5]. Use of the subcategories is intended to further help with risk stratification and therefore to help both the physicians and the patients in guiding further management.

Subcategory 4a denotes lesions with a low suspicion for malignancy. For this subcategory, a benign pathologic profile is expected and, therefore, considered concordant. A 6-month or routine follow-up after a benign concordant biopsy or cytology result is appropriate. This group includes solid masses without all the benign features (such as fibroadenomas), a palpable complicated cyst, and a probable abscess [1].

Subcategory 4b denotes lesions with an intermediate suspicion for malignancy. Imaging follow-up and radiologic-pathologic correlation are integral components in using this subcategory appropriately because the pathologic results may be benign or malignant. Therefore, the next step in management will be determined according to the concordance of the result. For example, a pathologic result that yields a fibroadenoma or fat necrosis is considered concordant for a partially circumscribed partially obscured mass. Follow-up imaging after a benign biopsy result depends on the radiologic-pathologic correlation. However, biopsy results yielding a papilloma may warrant excisional biopsy [1].

Subcategory 4c denotes lesions with moderate suspicion for malignancy that lack the classic characteristics to be assessed as category 5 lesions. For this subcategory, a malignant pathologic profile is expected and, therefore, considered concordant. An example is an irregular solid mass [1].

Category 5 is used for masses that have a 95% or higher likelihood of malignancy. This category should be reserved for findings that are classic representations of breast cancer. An example is a spiculated irregular high-density mass [1].

Most masses originally categorized as probably benign had subsequent follow-up imaging confirming the benign cause. Masses with no follow-up imaging available were reviewed and were found by author consensus to fulfill the BI-RADS category 3 criteria. For the masses originally categorized as benign, no further workup was done.

Selected images from each study were copied into a Microsoft Office PowerPoint presentation as JPEG images after the patient identifiers were removed. In some cases, images were cropped to focus attention on the main finding. Each image was then categorized using the BI-RADS ultrasound lexicon with agreement reached with all authors. Question stems were developed to accompany each image, with one image and one question per slide. A total of 90 questions were presented for both the pre- and postinstruction tests (Table 1; the PowerPoint presentation with the questions, Fig. S1, can be seen in the *AJR* electronic supplement to this article, available at www.ajronline.org). Descriptors included echo pattern, lesion boundary, orientation, margin, shape, posterior acoustic features, BI-RADS assessment, and use of BI-RADS 4 subcategories.

Thirty-four radiology residents representing all levels of training were the study participants (Table 2). Each participant was given a unique identifier and was asked to record his or her answer for the preinstruction test set on the standard form (Scantron). The answer forms were immediately collected after the last question was completed. The participants then received formal training in the use of ultrasound descriptors and mass assessment according to the BI-RADS ultrasound lexicon. The didactic training lasted for 1 hour. Immediately after this training, the participants were asked to complete the postinstruction test. Every

participant was able to complete all three components of the assessment (preinstruction test, lecture, and postinstruction test). Performance was measured both before and after formal training was provided.

The numbers and percentages of correct answers were added for all 90 questions and for the sets of questions in the different feature analysis categories (shape, echo pattern, lesion boundary, margin, lesion orientation, acoustic features, final BI-RADS assessment, and BI-RADS 4 subcategories final assessment). The scores for the participants in the pre- and postinstruction tests were defined as the percentage of correct answers of the total 90 questions. Test results were described with median, interquartile range, minimum, and maximum values. A linear mixed model was used to examine the differences between the pre- and posttest scores and among the years of radiology residents, with test scores (percentage correct) as the repeated measurement. Differences between the pre- and postinstruction test results for each category of questions were compared by nonparametric Wilcoxon signed rank tests. All analyses were conducted using SAS software (version 9.2 for Windows, SAS Institute); p values of 0.05 or less were considered statistically significant.

Results

The participants' postinstruction test scores showed significant improvement in overall use of the BI-RADS ultrasound lexicon ($p < 0.0001$). There was also significant improvement in the following categories: final assessment ($p = 0.0005$), margin ($p = 0.0003$), orientation ($p = 0.0104$), and lesion boundary ($p = 0.0050$) (Figs. 1–4). Categories that did not show significant improvement included echo pattern ($p = 0.07$), posterior acoustic features ($p = 0.50$), shape ($p = 0.98$), and subset of the final assessment ($p = 0.24$) (Figs. 5–8 and Table 3). The test scores were significantly different among the residents at different levels of radiology training, with an average increased score of 2.7% per year of training ($p = 0.0002$). However, all training groups derived a statistically significant benefit, and there was no interaction between experience and benefit from training ($p = 0.79$) (Fig. 9).

Discussion

Breast cancer continues to be an important cause of morbidity and mortality in the United States. An estimated 226,870 women were diagnosed with invasive breast cancer in the United States in 2012, and many of those women underwent an ultrasound-guided core needle biopsy [6]. Ultrasound is a cost-effective modality that does not use ionizing radiation, which makes it an ideal modality for use in large populations. Its utility, however, is dependent on the operator for both mass detection and mass assessment. Efforts have been made to minimize ultrasound operator variability in mass detection (e.g., automated breast ultrasound). New technologies have been developed to improve mass assessment (e.g., elastography). Our study shows that formal instruction in the appropriate use of the BI-RADS ultrasound lexicon is an effective method for improved mass characterization and final assessment.

This study found statistically significant improvements in lesion boundary, margin, orientation, and final assessment categories. Improvement was noted in all of the categories analyzed, including the feature analysis of echo pattern, shape, posterior acoustic features, and BI-RADS 4 subcategories, although this was not statistically significant. Of note, one of the greatest improvements was in the category of final assessment. Because final assessment ultimately dictates management, this category has the greatest effect on patient care. Our study results are concordant with the results of another study assessing the impact of formal instruction on the appropriate use of the BI-RADS mammography lexicon [2].

The final assessment also serves as the main form of communication with the referring clinicians. Although there was significant improvement in the overall final assessment categories, there was little improvement in performance in the category 4 subset of final assessment. The subdivision of BI-RADS category 4 into the subcategories 4a, 4b, and 4c was created to better inform clinicians and pathologists of the radiologist's degree of concern about the ultrasound finding. The subcategories were also intended to assist in internal audits and radiologic-pathologic correlation. However, our study found that, despite formal teaching about the subcategories, participants continued to have a poor understanding regarding the appropriate use of the BI-RADS category 4 subcategories. Indeed, scarce literature exists showing examples of the subcategories to improve understanding of their appropriate use. There is also a paucity of evidence to accurately categorize the expected number of malignancies in each of the subcategories. The new edition of the ACR BI-RADS lexicon is expected shortly. Once those categories are better understood, more studies could be performed to evaluate the incidence of malignancy in each group and the interobserver variation in the use of the BI-RADS 4a, 4b, and 4c subcategories.

The median postinstruction test score (80%) was a significant improvement when compared with the median preinstruction test score (72.8%). It is unknown whether the degree of improvement would continue with more formal instruction. The pre- and postinstruction test scores were higher in residents with more radiology experience; however, all participants derived the statistically significant benefit from formal instruction about the ultrasound lexicon (Fig. 9).

Although test scores improved in the areas of shape, echo pattern, and posterior acoustic features, the improvements were small and not statistically significant. Of note, the individual observer performance regarding the correct use of these descriptors was high (> 80%) before formal training, and there was limited room for improvement. On the basis of these data, we concluded that the participants had a solid understanding of the appropriate use of these descriptors before formal training was provided; therefore, formal training on these descriptors did not have a statistically significant effect on their performance.

For meaningful ultrasound research to be performed, all participants must understand the BI-RADS ultrasound lexicon and use it correctly. The preponderance of ultrasound research relies on the physician's ability to correctly assess feature analysis and come to the appropriate final assessment. If a radiologist's ability to correctly assess an ultrasound finding is inconsistent, research regarding management strategies may not be useful. A relevant example would be the management of a probable fibroadenoma. Multiple articles

advocating imaging follow-up of this common finding rely heavily on the radiologist's ability to accurately assess features and come to an appropriate final assessment [7, 8]. This research is meaningful because of the radiologists' abilities to consistently perform feature analysis and final assessment.

Our study has weaknesses. First, the number of the residents who participated in the study ($n = 34$) was relatively small. Second, the study participants in this group were radiology residents. Board-certified radiologists may have a better understanding of the BI-RADS ultrasound lexicon, although other studies indicate a wide interobserver variation. Our study did, however, show that all participants, regardless of level of training, derived the same benefit from formal instruction. We also did not study whether the formal training had a lasting effect on the participants. Ongoing education about the BI-RADS ultrasound lexicon could be beneficial. Online material has been developed for continued instruction on the BI-RADS ultrasound lexicon [9]. Reinforcement of the lexicon could be performed during ACR ultrasound accreditation (via image selection) and during the American Board of Radiology Maintenance of Certification examinations.

In summary, formal training regarding the BI-RADS ultrasound lexicon improves observer performance in ultrasound feature analysis of masses (particularly lesion boundary and margin) and final assessment categories, which improves the quality of the study interpretation. After training, the participants improved their recognition of suspicious features, which should ultimately result in more appropriate management. Consistent and appropriate use of the lexicon may improve physician communication and serve as a constant for future ultrasound research. Although the study was performed with radiology residents, training could also be useful for radiologists with more experience. ACR ultrasound accreditations and American Board of Radiology Maintenance of Certification examinations could be tools through which the appropriate use of the lexicon could be reinforced.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

References

1. Mendelson EB, Baum JK, Berg WA, et al. BI-RADS: ultrasound, In: D'Orsi CJ, Mendelson EB, Ikeda DM, et al. Breast Imaging Reporting and Data System: ACR BI-RADS—breast imaging atlas. Reston, VA: American College of Radiology, 2003

2. Berg WA, Blume JD, Cormack JB, Mendelson EB. Operator dependence of physician-performed whole-breast US: lesion detection and characterization. Radiology 2006; 241:355–365 [PubMed: 17057064]

3. Lazarus E, Mainiero MB, Schepps B, Koelliker SL, Livingston LS. BI-RADS lexicon for US and mammography: interobserver variability and positive predictive value. Radiology 2006; 239:385–391 [PubMed: 16569780]

4. Abdullah N, Mesurolle B, El-Khoury M, Kao E. Breast imaging reporting and data system lexicon for US: interobserver agreement for assessment of breast masses. Radiology 2009; 252:665–672 [PubMed: 19567644]

5. Raza S, Goldkamp AL, Chikarmane SA, Birdwell RL. US of breast masses categorized as BI-RADS 3, 4, and 5: pictorial review of factors influencing clinical management. RadioGraphics 2010; 30: 1199–1213 [PubMed: 20833845]

6. American Cancer Society. What are the key statistics about breast cancer? American Cancer Society web site. www.cancer.org/Cancer/BreastCancer/DetailedGuide/breast-cancer-key-statistics. Published August 23, 2012. Updated February 26, 2013. Accessed April 17, 2013

7. Graf O, Helbich TH, Hopf G, Graf C, Sickles EA. Probably benign breast masses at US: is follow-up an acceptable alternative to biopsy? Radiology 2007; 244:87–93 [PubMed: 17581897]

8. Harvey JA, Nicholson BT, Lorusso AP, Cohen MA, Bovbjerg VE. Short-term follow-up of palpable breast lesions with benign imaging features: evaluation of 375 lesions in 320 women. AJR 2009; 193:1723–1730 [PubMed: 19933671]

9. University of California School of Medicine. BI-RADS tutor: the electronic teaching file. BI-RADS Tutor website. biradstutor.com. Published January 1, 2008. Accessed April 13, 2013
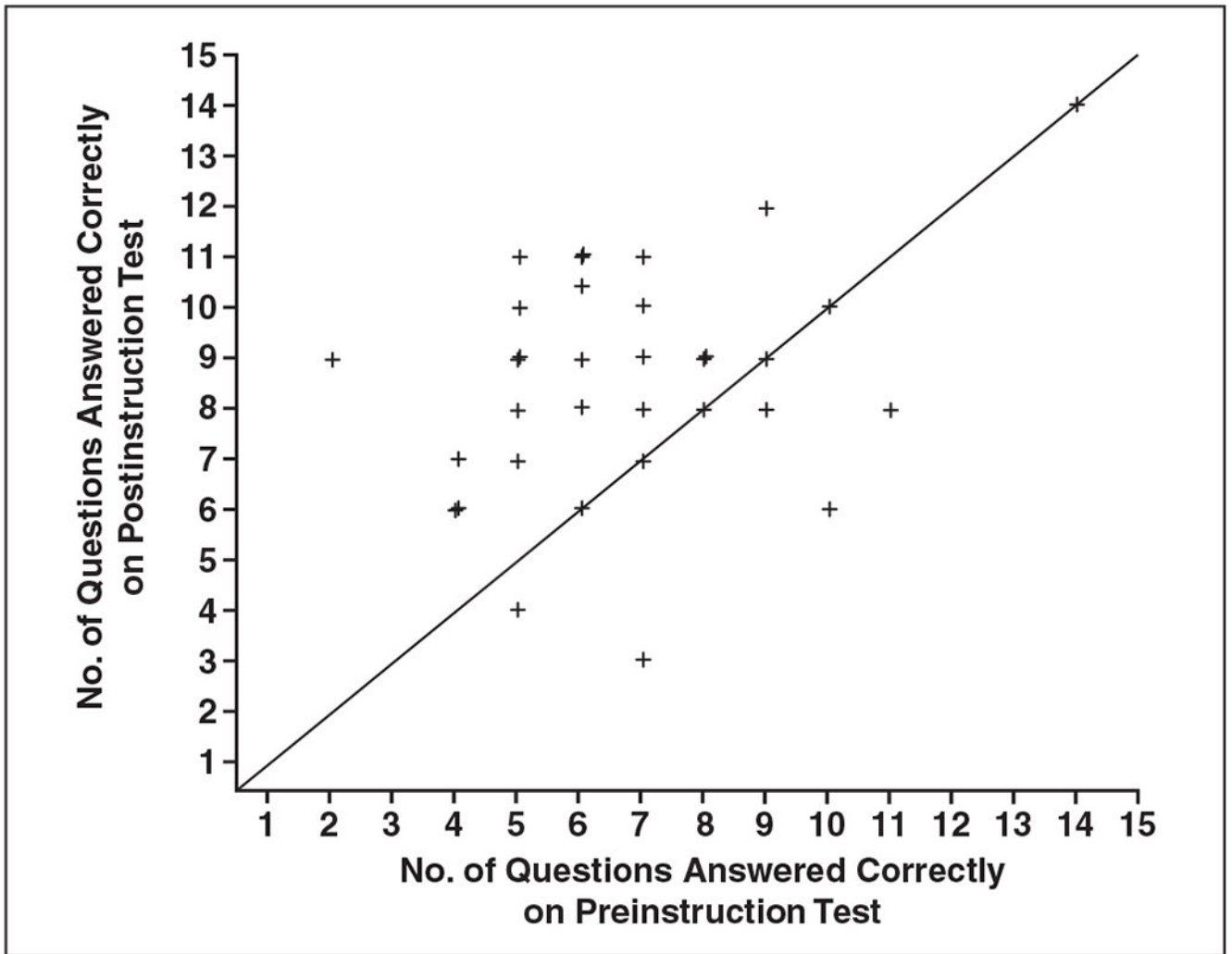
Fig. 1—.
Pre- and postinstruction test performance for final assessment of BI-RADS category. Data points are number of questions answered correctly on each test. Diagonal line indicates median number of questions answered correctly.
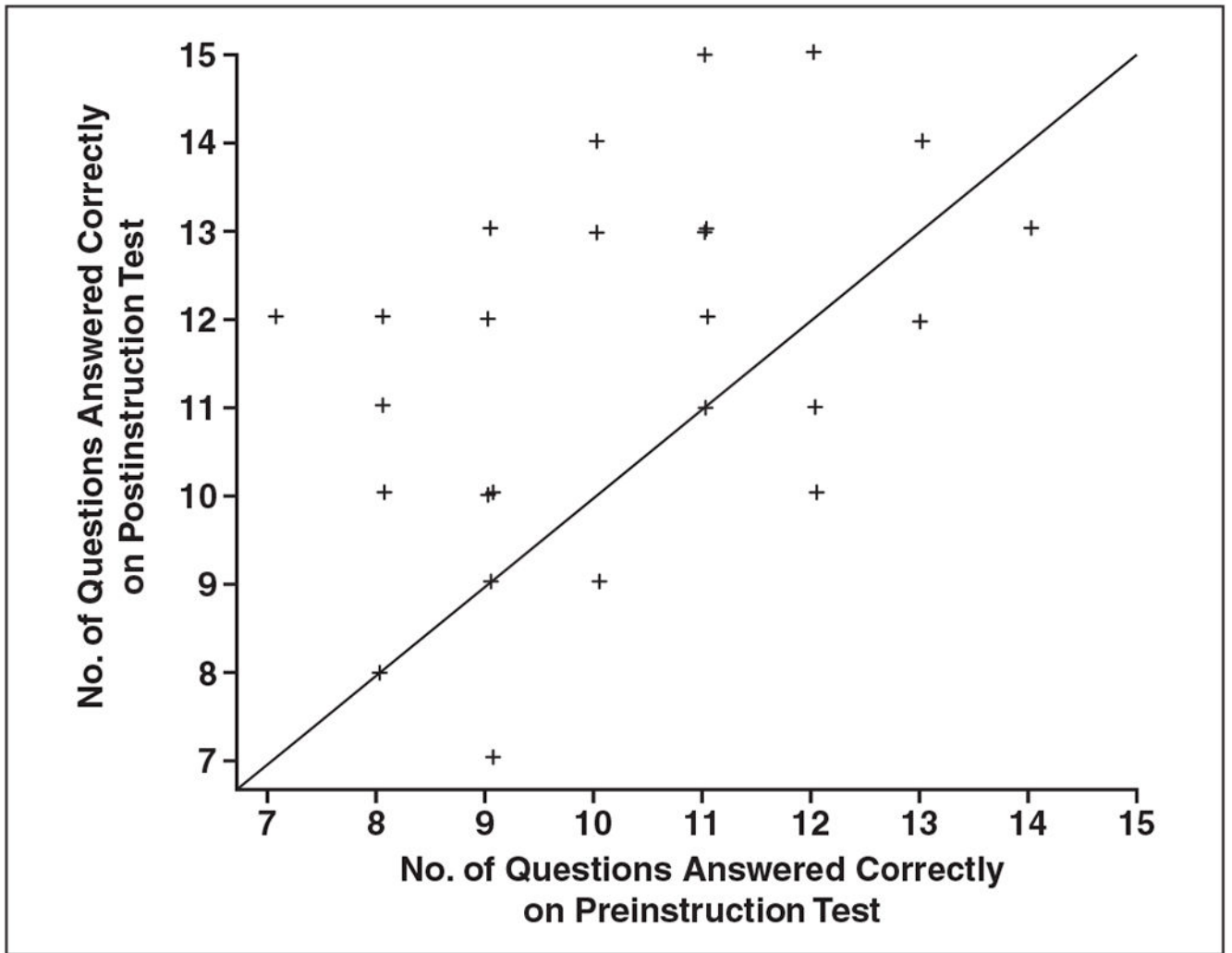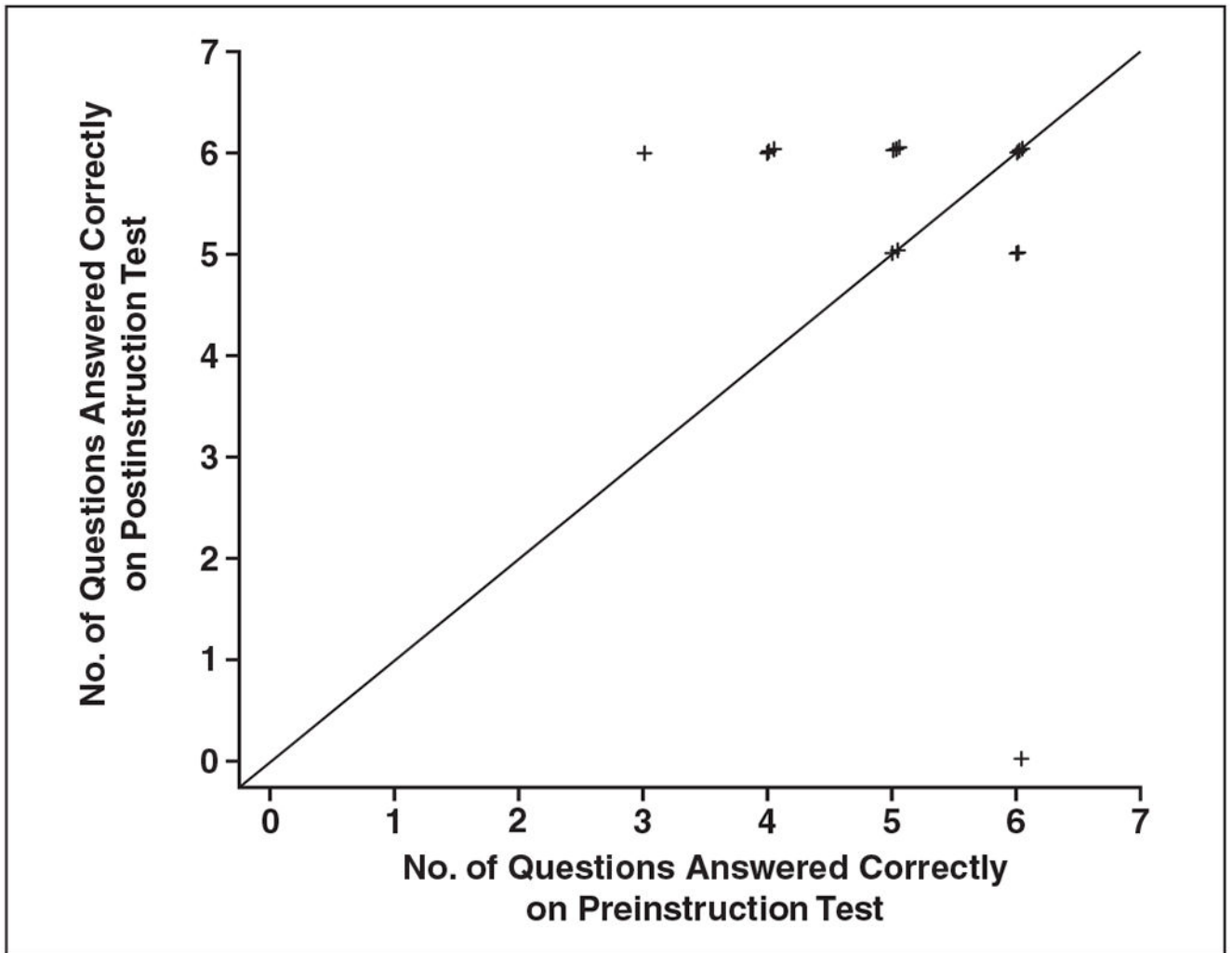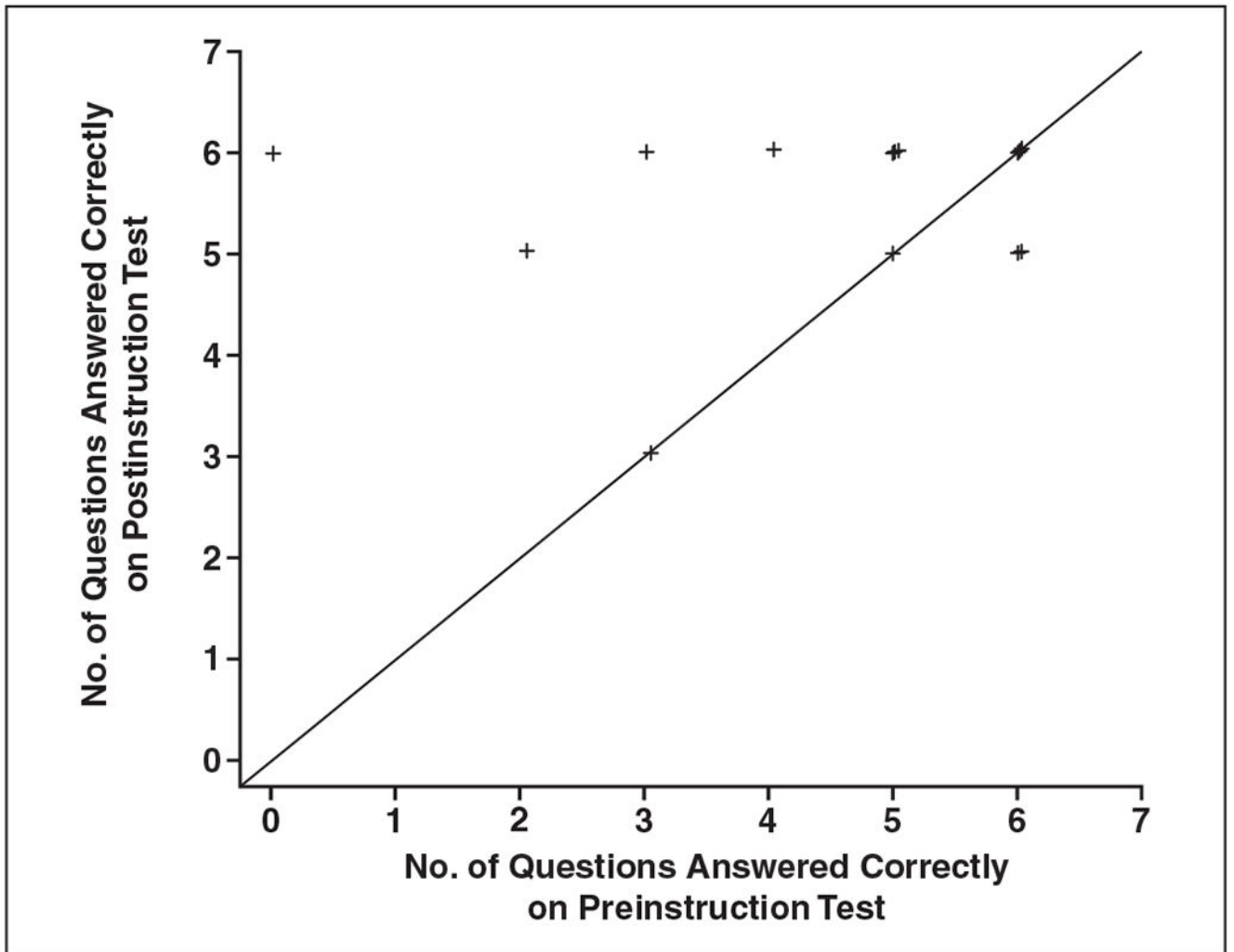
Fig. 2—.
Pre- and postinstruction test performance for lesion margin. Data points are number of questions answered correctly on each test. Diagonal line indicates median number of questions answered correctly.

Fig. 3—.

Pre- and postinstruction test performance for lesion orientation. Data points are number of questions answered correctly on each test. Diagonal line indicates median number of questions answered correctly.

Fig. 4—.
Pre- and postinstruction test performance for lesion boundary. Data points are number of questions answered correctly on each test. Diagonal line indicates median number of questions answered correctly.
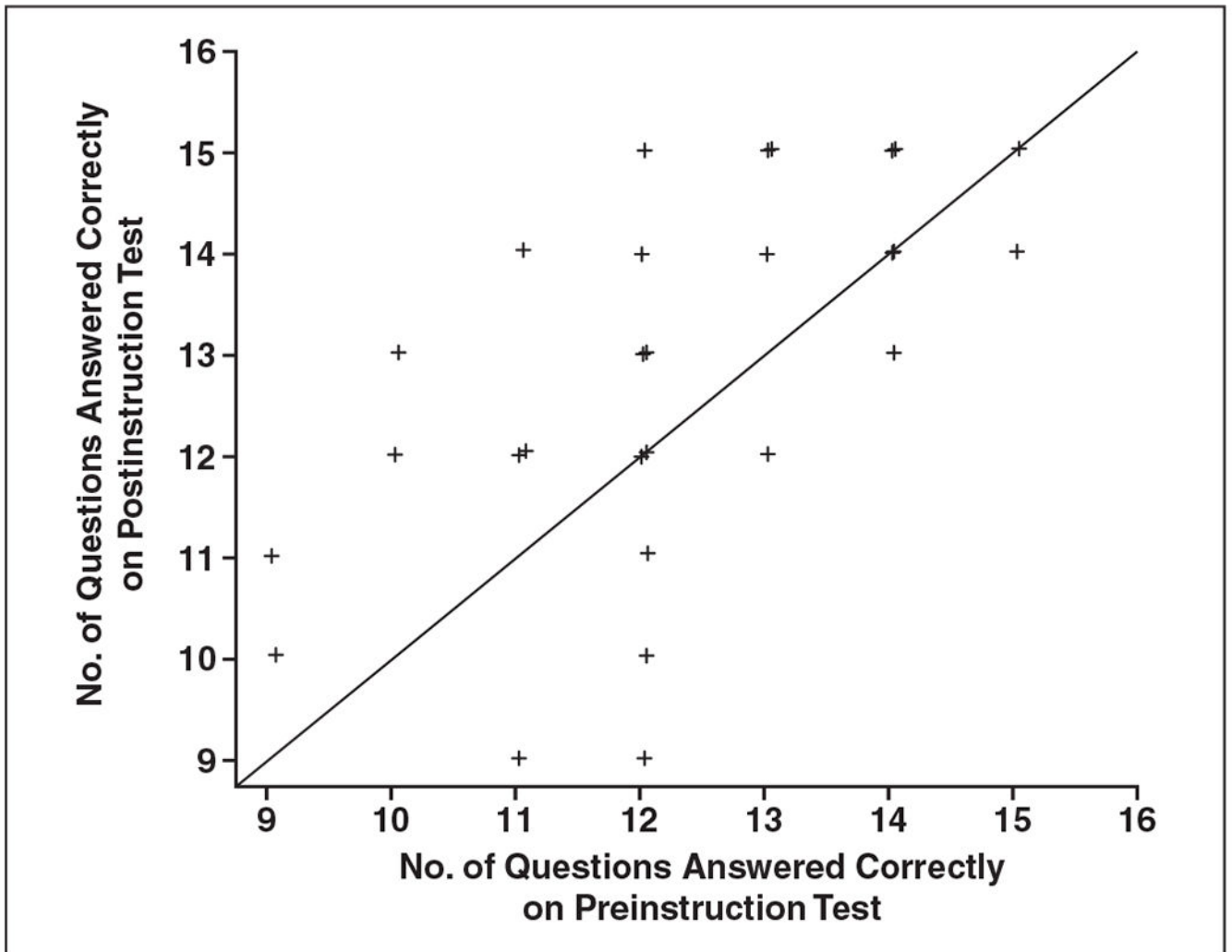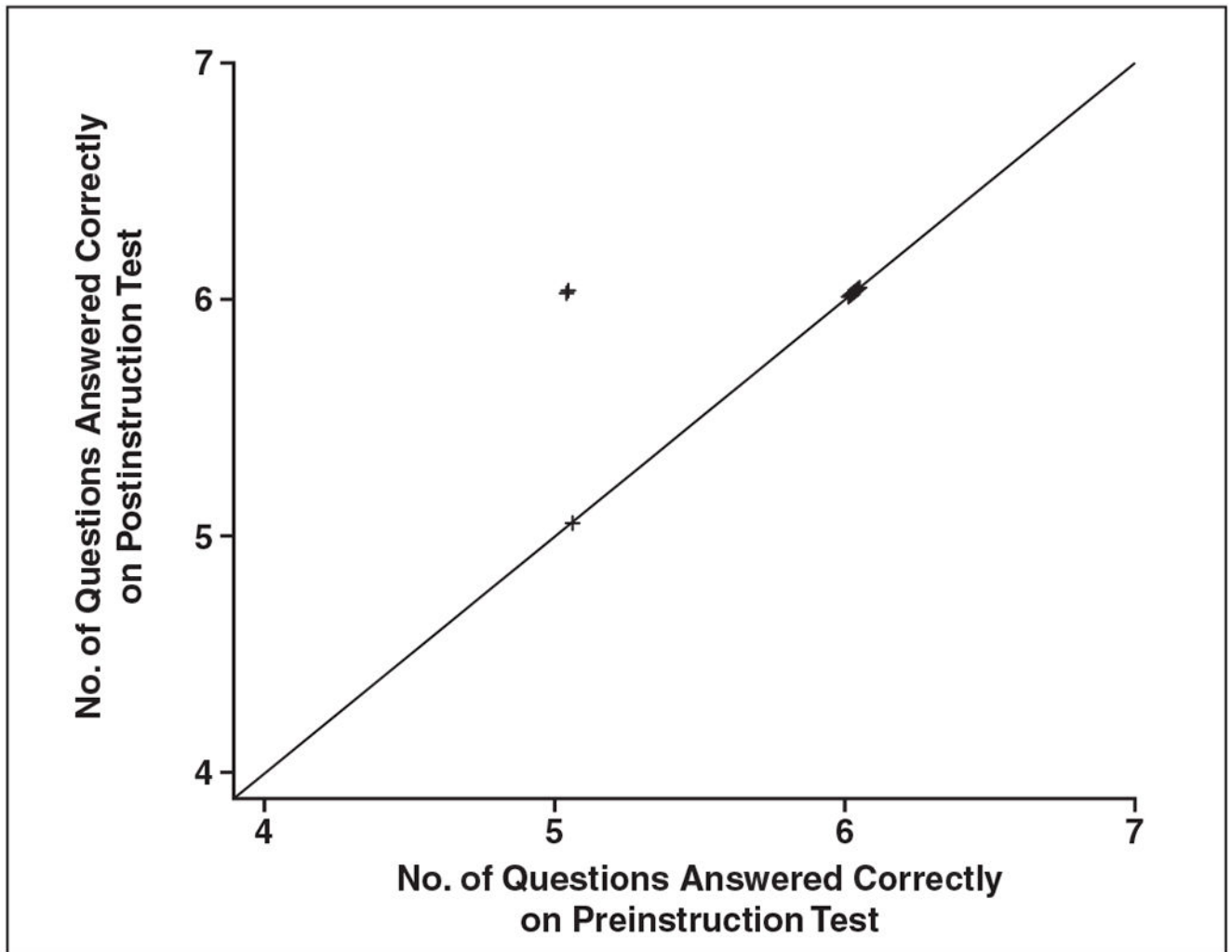
Fig. 5—.

Pre- and postinstruction test performance for lesion echo pattern. Data points are number of questions answered correctly on each test. Diagonal line indicates median number of questions answered correctly.

Fig. 6—.

Pre- and postinstruction test performance for lesion posterior acoustic features. Data points are number of questions answered correctly on each test. Diagonal line indicates median number of questions answered correctly.
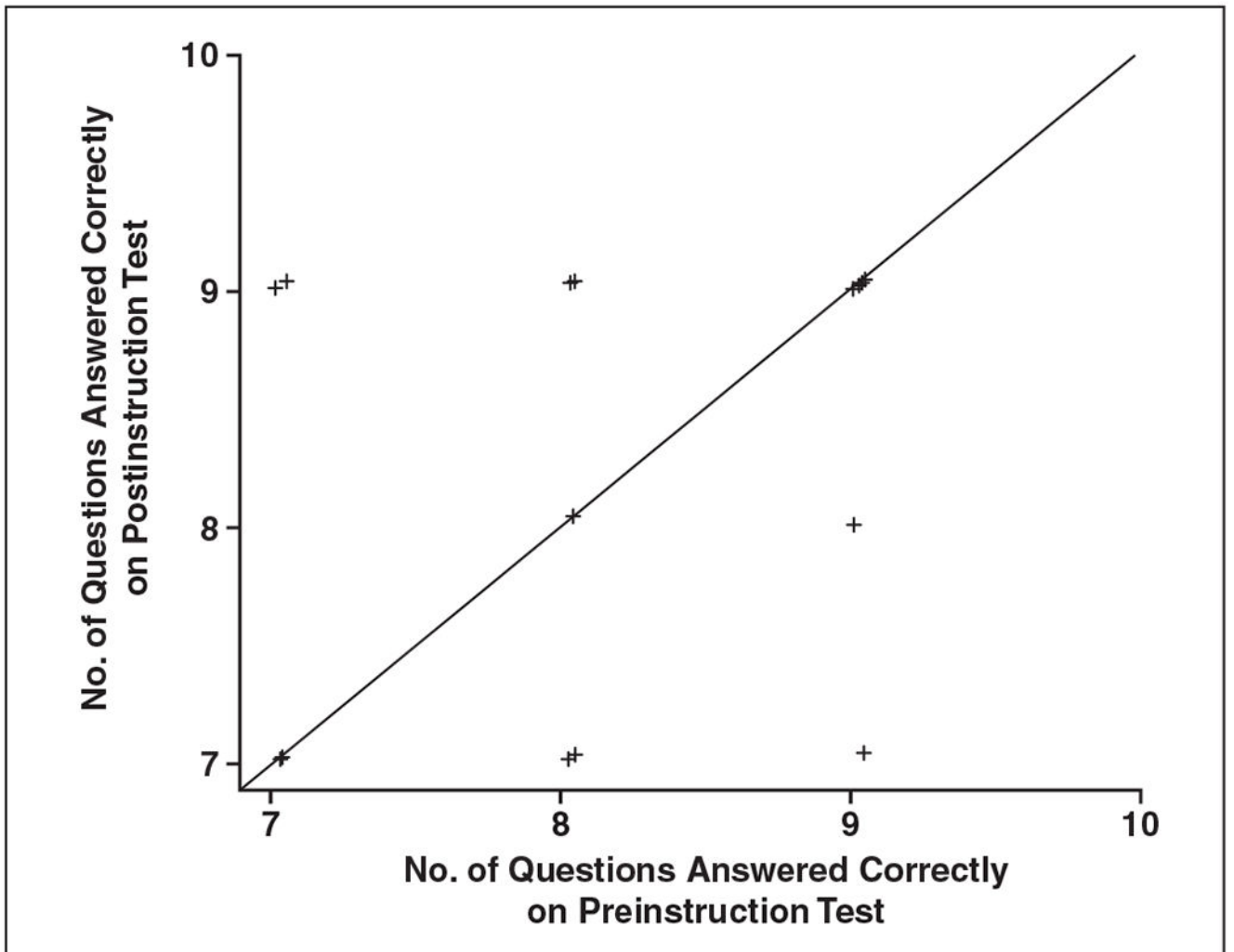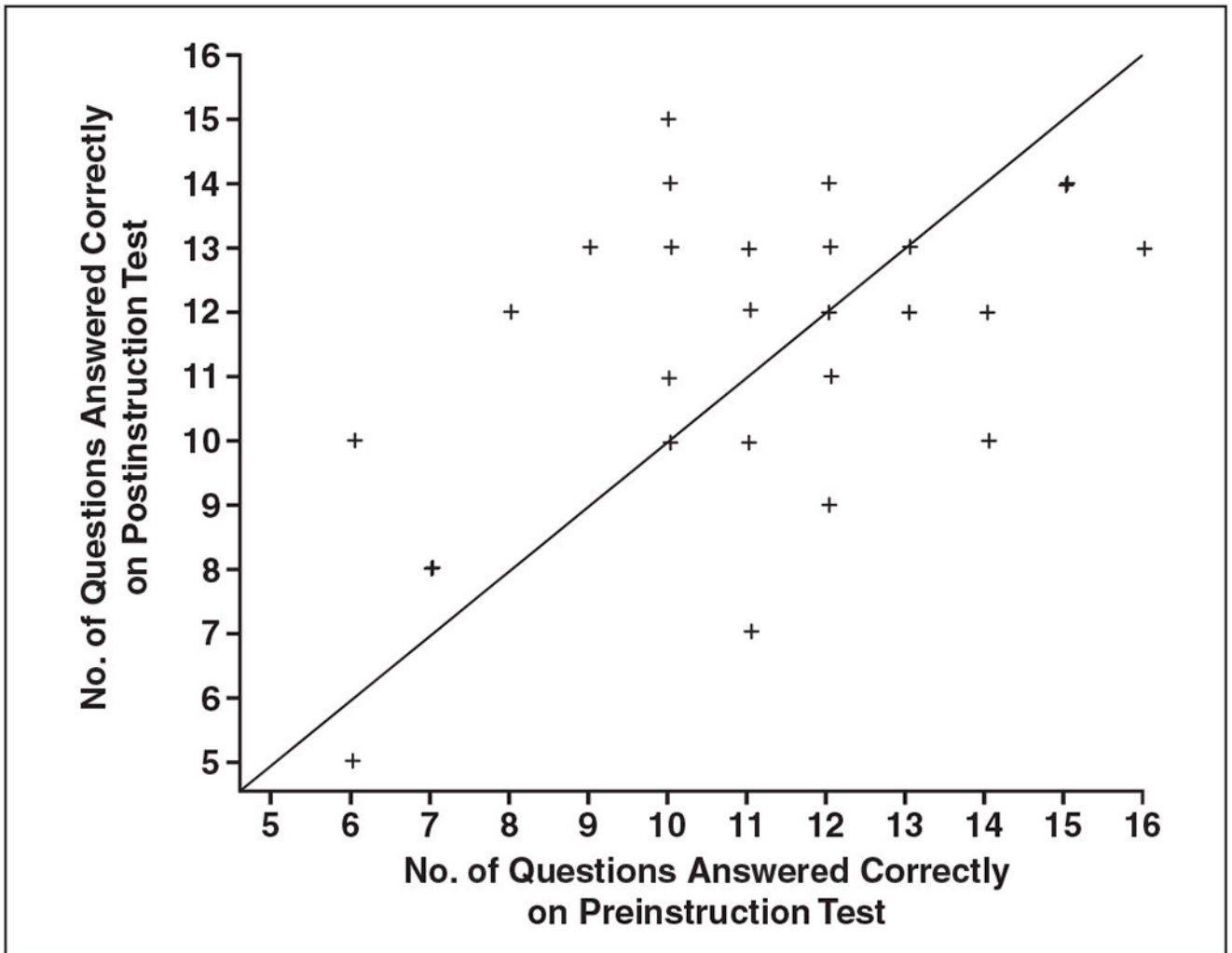
Fig. 7—.
Pre- and postinstruction test performance for lesion shape. Data points are number of questions answered correctly on each test. Diagonal line indicates median number of questions answered correctly.

Fig. 8—.

Pre- and postinstruction test performance for use of BI-RADS subcategories 4a, 4b, and 4c. Data points are number of questions answered correctly on each test. Diagonal line indicates median number of questions answered correctly.
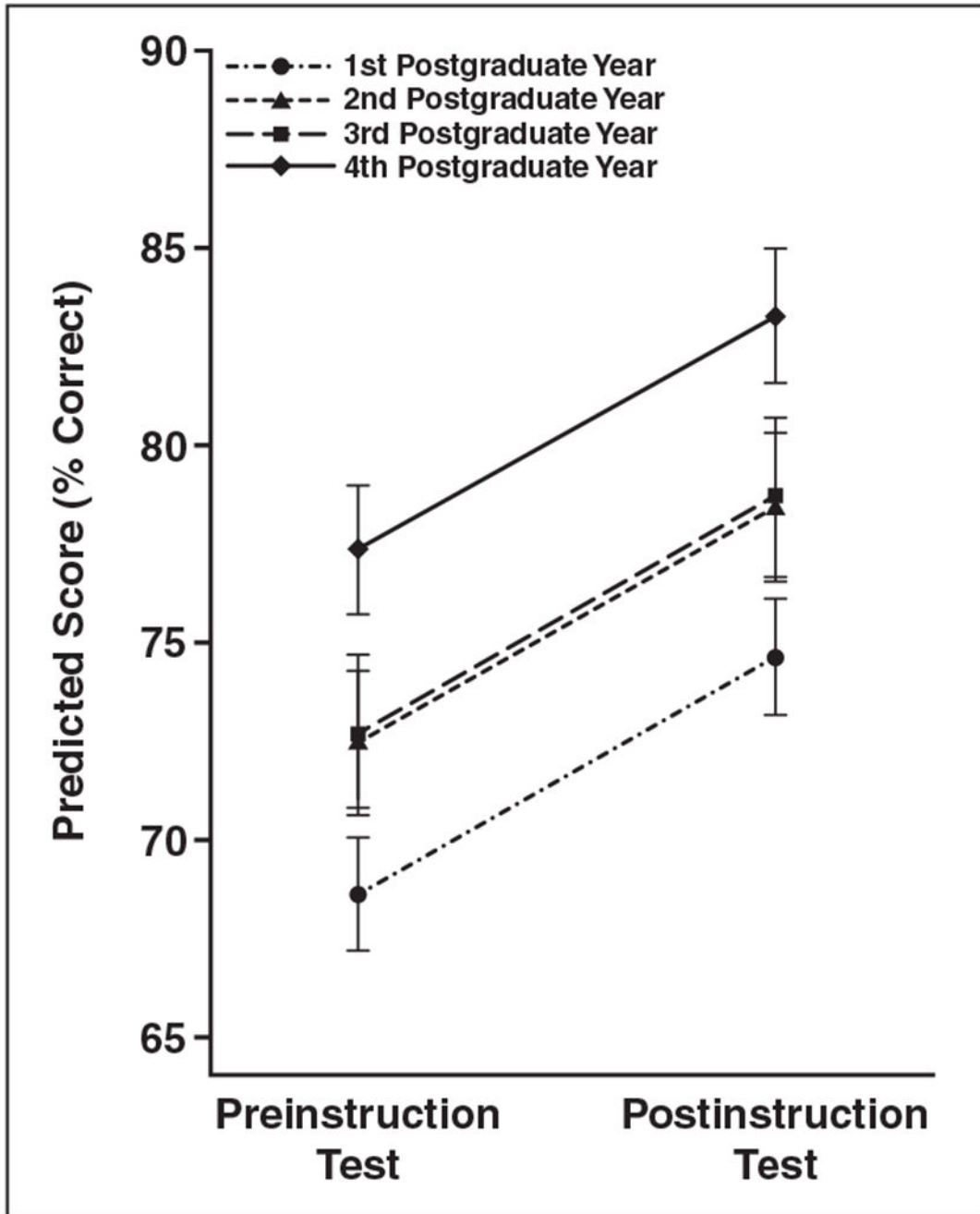
Fig. 9—.
Predicted score (percentage correct) and standard error from linear mixed model for pre- and postinstruction tests, by postgraduate year of participants.

TABLE 1:

Representation of Ultrasound Descriptors in Pre- and Postinstruction Questions

Category (Total No. of Questions), Descriptor	No. of Questions for Each Descriptor
Echo pattern (*n* = 15 questions)	
Anechoic	3
Hypoechoic	3
Isoechoic	3
Hyperechoic	3
Complex	3
Lesion boundary (*n* = 6 questions)	
Abrupt interface	3
Echogenic halo	3
Margin (*n* = 15 questions)	
Circumscribed	3
Angular	3
Spiculated	3
Indistinct	3
Microlobulated	3
Orientation (*n* = 6 questions)	
Parallel	3
Nonparallel	3
Posterior acoustic features (*n* = 6 questions)	
Enhancement	3
Shadowing	3
Shape (*n* = 9 questions)	
Round	3
Oval	3
Irregular	3
BI-RADS assessment (*n* = 15 questions)	
1	3
2	3

Category (Total No. of Questions), Descriptor	No. of Questions for Each Descriptor
3	3
4	3
5	3
BI-RADS 4 subcategories (*n* = 18 questions)	
4a	6
4b	6
4c	6
Total	90

Note—See also Figure S1, a Microsoft PowerPoint presentation of the 90 questions, in supplemental data online available at www.ajronline.org.

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

TABLE 2:

Breast Imaging Experience per Training Level Examined

Training Level	Average No. of Weeks in Breast Imaging Rotation	No. (%) of Participants ($n = 34$ Total)
Postgraduate year 2	4	12 (35.29)
Postgraduate year 3	4	7 (20.59)
Postgraduate year 4	8	6 (17.65)
Postgraduate year 5	12	9 (26.47)

TABLE 3:

Numbers and Percentages of Correct Answers on Pre- and Postinstruction Tests by Question Category

Question Group, Test	Median No. (%) Correct	IQR (No. of Questions)	Minimum No. (%) Correct	Maximum No. (%) Correct	p^a
All 90 questions					< 0.0001
Preinstruction	65.5 (72.8)	7	52 (57.8)	78 (86.7)	
Postinstruction	72 (80)	8	56 (62.2)	81 (90)	
Questions 1–9, shape					0.98
Preinstruction	9 (100)	1	7 (77.8)	9 (100)	
Postinstruction	9 (100)	0	7 (77.8)	9 (100)	
Questions 10–24, echo pattern					0.07
Preinstruction	12 (80)	2	9 (60)	15 (100)	
Postinstruction	13 (86.7)	2	9 (60)	15 (100)	
Questions 25–30, lesion boundary					0.0050
Preinstruction	5.5 (91.7)	1	0 (0)	6 (100)	
Postinstruction	6 (100)	0	3 (50)	6 (100)	
Questions 31–45, margin					0.0003
Preinstruction	10 (66.7)	2	7 (46.7)	14 (93.3)	
Postinstruction	12 (80)	3	7 (46.7)	15 (100)	
Questions 46–51, orientation					0.0104
Preinstruction	5 (83.3)	1	3 (50)	6 (100)	
Postinstruction	6 (100)	0	0 (0)	6 (100)	
Questions 52–57, posterior acoustic features					0.50
Preinstruction	6 (100)	0	5 (83.3)	6 (100)	
Postinstruction	6 (100)	0	5 (83.3)	6 (100)	
Questions 58–72, final BI-RADS assessment					0.0005
Preinstruction	7 (46.7)	3	2 (13.3)	14 (93.3)	
Postinstruction	9 (60)	3	3 (20)	14 (93.3)	
Questions 73–90, final BI-RADS 4 subcategories assessment					0.24
Preinstruction	11 (61.1)	2	6 (33.3)	16 (88.9)	
Postinstruction	12 (66.7)	3	5 (27.8)	15 (83.3)	

Note—IQR = interquartile range.

a
p values were calculated from a linear mixed model for test score and Wilcoxon signed rank test for all the others.