ABSTRACT
CONTEXT: Differential item function (DIF) refers to the different probability of giving certain responses in a test when test takers from different groups have the same ability. Items showing DIF indicate potential bias, which should be further reviewed during the test development stage. Because it is a high-stakes test to assess candidates’ ability to practice osteopathic medicine, it is essential to ensure test fairness for the COMLEX-USA examination series.

METHODS: In this study, three groups are examined (gender, race, and school) for the COMLEX-USA Level 1, Level 2-Cognitive Evaluation (CE), and Level 3 examinations. The study also aims to check whether there is a difference between Osteopathic Principles and Practice (OPP) items and non-OPP items in terms of showing DIF for the above three groups. OPP items intend to specifically test candidates’ understanding and integration of the osteopathic philosophy of health care practices. When developing the OPP items, their content might be either distinct from non-OPP items. This DIF detection research was conducted using the Mantel-Haenszel approach, which is one of the most popular methods used in testing programs because of its computational simplicity and ease of implementation.

RESULTS: Overall, less than 5% of the items for all three groups in all three COMLEX-USA levels are identified by DIF statistics. Consistent with best practices in assessments used by other physician testing and regulatory organizations worldwide (AERA, APA, NCME, 2014), the NBOME reviews the DIF statistics along with other psychometric information, such as key validation and candidate comments, and concludes the fairness across groups (gender, race, school) in COMLEX-USA Level 1, Level 2-CE and Level 3 examinations.

REFERENCES: A total of 1,800 items from COMLEX-USA Level 1, Level 2-CE and Level 3 examinations are examined. The sample sizes are \( N_1 = 6,362 \), \( N_2 = 6,451 \), and \( N_3 = 5,712 \) respectively.

METHODS/MEASURES
DIFFERENTIAL ITEM FUNCTION

DEFINITION
Differential item function (DIF) occurs when test takers from different groups (e.g., gender / ethnicity) have the same ability but show a different probability of giving certain responses in a test. For instance, if an item uses the scenario of football season, candidates who have the same ability to answer the question correctly may not have an equal probability of answering correctly due to gender (e.g., female candidates might be less familiar with football season).

DETECTION METHOD
2. Non-parametric: Mantel-Haenszel test (Holland and Thayer, 1988; Mantel & Haenszel, 1959), generalized Mantel-Haenszel test (Mantel & Haenszel, 1959), z-score (Conover, 1999), z-score (Zwick, Donoghue, & Grima, 1992)

METHODS/MEASURES
DIFFERENTIAL ITEM FUNCTION

REFERENCE VERSUS FOCAL GROUP
Reference Group
Chosen to provide standard performance on the item of interest

Focal Group
Its differential performance, if any, is to be detected and measured

MANTEL-HAENSZEL TEST
• Non-iterative method based on contingency tables
• One of the most popular detection methods for dichotomous items
• Computationally simple, inexpensive, and easy to implement

MANTEL-HAENSZEL CHI-SQUARE TEST STATISTIC
where
\[ \chi^2 = \frac{(|\text{C}_{11} - |\text{C}_{21}|)^2}{\text{Var}(|\text{C}_{11} | |\text{C}_{21}|)} \]

and
\[ \text{Var}(|\text{C}_{11} | |\text{C}_{21}|) = \frac{\text{Var} (\text{C}_{11}) \text{Var} (\text{C}_{21})}{\text{Var} (\text{C}_{11} + |\text{C}_{21}|)} \]

REFERENCE VERSUS FOCAL GROUP

RESULTS

<table>
<thead>
<tr>
<th>Gender</th>
<th>Race</th>
<th>School</th>
<th>Level 1</th>
<th>Level 2-CE</th>
<th>Level 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>3.05%</td>
<td>3.00%</td>
<td>3.00%</td>
<td>2.89%</td>
<td>2.85%</td>
<td>2.81%</td>
</tr>
</tbody>
</table>

NBOME STANDARDS FOR QUALITY ASSURANCE ON TEST FAIRNESS

• Standard 3.2: “The examination products should “Address fairness in the design, development, and administration.”
• Standard 3.5: “When a construct is measured in different ways that are reasonably equally valid, reliable, practical and affordable, consider available evidence of group differences in assessment results in determining how to measure the construct.”

FUTURE DIRECTIONS
Use other methods that can better accommodate non-uniform DIF (e.g., logistic regression).

INVESTIGATION OF DIFFERENTIAL ITEM FUNCTIONING ON COMPLEX-USA EXAMINATION SERIES
Can Shao, Ph.D., Yi Wang, M.S., Silu Liu, M.P.H., Tsung-Hsun Tsai, Ph.D.
National Board of Osteopathic Medical Examiners

AACOM ANNUAL 2018 CONFERENCE
National Board of Osteopathic Medical Examiners