Rejoinder to 'Applicability of Entry to Practice Examinations for Optometry in Canada and the United States – Optometry Examining Board of Canada and National Board of Examiners in Optometry' by Woo, Hrynchak, & Hutchings (2022) [published in the Canadian Journal of Optometry]

March 1, 2022

Brooke Houck, Ph.D.
Director of Psychometrics and Research
National Board of Examiners in Optometry

NATIONAL BOARD OF EXAMINERS IN OPTOMETRY®
200 South College Street
Suite 2010
Charlotte, North Carolina 28202
www.optometry.org

In January 2020, Woo, Hrynchak, and Hutchings released a white paper advocating for the use of the Optometry Examining Board of Canada's (OEBC) licensure exams over the National Board of Examiners in Optometry (NBEO) series of licensure exams. The authors did not contact NBEO with requests for information. Their white paper was turned into an article published by the Canadian Journal of Optometry (CJO).[1] The purpose of this paper is to correct factual inaccuracies and to provide information about the NBEO licensure exam series that was not included in the Woo et al. paper, and is in response to a pre-publication version of the article appearing in CJO.

Woo et al. (2022) provides a high-level overview of both the OEBC and NBEO exams and attempts to compare them in an apples-to-apples fashion within a framework for validating test score interpretations and uses. We contend that, while the authors' intentions were to ascertain a clear comparison, their paper reflects certain errors and a lack of information about NBEO processes and examinations, which this paper will address. This article provides a supplement to the work of Woo et al. (2022) by both providing missing information and by correcting misinformation. The format of this paper will mostly follow that of Woo et al. (2022) for greater ease of understanding about which facts and further information the NBEO is supplying, as they relate to specific sections of the paper.

Introduction

NBEO agrees with the authors that the schools and colleges of optometry throughout North America strive to provide sound training to optometrists such that they are prepared to begin safe, effective practice upon entering the profession. NBEO further agrees that validity and reliability are paramount in assessments. Specifically, in high-stakes assessments, the measurement properties of an assessment are especially consequential. The licensure examinations offered by both OEBC and NBEO are used by governmental bodies to make decisions to grant or not grant licenses to practice. In the context of examinations, the stakes cannot be higher – the results of these exams determine if one is allowed to enter the profession of optometry, thus utilizing their extensive and costly training.

Because licensure examinations are high stakes, the validity and reliability of these exams should be of ongoing interest to all parties charged with developing, implementing, and maintaining the examinations. Woo et al. (2022) are correct in their assertion that reliability and validity are important so that correct decisions about licensure can be made by regulatory

[1] In-text citations listing page numbers for Woo et al. are approximate and are based on the pre-publication draft reviewed by NBEO.

boards. However, the psychometric concepts of validity and reliability are complex and merit deeper discussion than is provided by the authors.

We contend that the overall argument within Woo et al.'s (2022) paper is that the NBEO series of licensure exams suffer from a validity issue within the Canadian context, and that this argument is unsubstantiated. The following sections will provide information and evidence of the validity of NBEO exams.

How are high-stakes summative assessments constructed?

We agree with the authors that all high-stakes assessments should be founded upon the knowledge, skills, and abilities the assessment is intended to measure. Given that NBEO examinations are used to determine if a candidate is competent to enter into the safe, effective independent practice of optometry, the competencies that make up the foundation of the exam series are derived from the competencies necessary for unsupervised practice.

NBEO follows the test development process set forth in the Standards for Educational and Psychological Testing (American Educational Research Association et al., 2014). Exam development refers to a process whereby a measurement of an individual's knowledge, skills, and abilities (KSAs) is collected through the use of a test formed "according to a specified plan" (American Educational Research Association et al., 2014, p. 75). The development process steps are compiled in a test design plan, and test design begins with an evaluation of the intended uses of the test scores and expectations for how scores will be interpreted. That is, before beginning to define competencies to be measured, an initial step of determining the purpose of the test and what inferences from the test score are needed. Once this is known, the four phases of test development can begin. The *Standards* state,

> "Test design and development procedures must support the validity of the interpretations of test scores for their intended uses" (p. 75).

The validity of the intended inferences drawn from test scores must be upheld in all stages of test development. That is, the path for the creation of an examination should run as shown in Figure 1, and all steps should support validity.

Figure 1

The steps between determining the purpose of the exam and producing an operational exam are shown in Figure 2, and represent the guidance set forth in the *Standards* (American Educational Research Association et al., 2014).

Figure 2

For a sufficiently deep understanding of the exam development process that is easily accessible to someone who is not a testing professional, please see Balogh (2016). One remaining point that merits attention is that test design is and should be an iterative process wherein empirical data from pilot testing and operational use is continuously incorporated. For example, NBEO provides all candidates the opportunity to point out any issues found with any exam question and reviews that feedback for future exam form creation. Additionally, item statistics are reviewed, and changes are made to reflect best practices based on item statistics.

Woo et al. (2022) assert, "In optometry, entry to practice assessments have used the traditional knowledge test and technical skills assessment approach, and have been slower to adopt a competency-based approach" (p. 4). It is difficult to understand the authors' evidence of this statement as NBEO examinations are built around competency. The Association of Regulatory Boards of Optometry (ARBO) state,

> "Assembling a quality optometrist population to meet the needs of the public begins with licensure…the state ensures all practicing optometrists have appropriate education and training, and they abide by recognized standards (emphasis added) of professional conduct while serving their patients….Candidates for licensure must also complete a rigorous examination, designed to assess an optometrist's ability to apply knowledge, concepts and principles that are important in health and disease and that constitute the basis of safe and effective patient care" (ARBO FAQ, 2021).

Obtaining licensure to practice optometry, as indicated by ARBO, requires candidates to demonstrate competency through examination. The NBEO exam series rests upon determining if a candidate is minimally competent to enter into independent practice.

Exam Development for Entry to Practice Assessments for Optometry in North America

The information provided by Woo et al. (2022) about the components, structure, and processes of NBEO licensure examinations contains inaccuracies and a lack of information.

Woo et al. (2022) did not request information from NBEO to ensure the accuracy or sufficiency of the information they present in their article.

The NBEO series of licensure exams consists of three separate parts. Part I Applied Basic Science (ABS) is a multiple-choice, computer-based exam that assesses candidates' mastery of the underlying basic science concepts necessary for entry into optometric practice. The exam consists of 370 questions, 20 of which are unscored, pre-test items, and is administered in two sessions of 4 hours each. Part II Patient Assessment and Management (PAM) examination assesses clinical thinking and decision-making, along with knowledge of diagnosis and treatment. The Part II PAM exam is also a computer-based, multiple-choice exam. It contains 350 items and is administered over two sessions of 3.5 hours each. Part II PAM questions frequently are shown as part of an overall case wherein candidates are given clinical information, sometimes including diagnostic images. The questions for the case follow a sequence that mimics clinical thinking and decision-making; however, examinees are able to select from a list of possible answers while thinking through the case and appropriate treatment steps. The Treatment and Management of Ocular Disease (TMOD) examination can be completed as part of Part II PAM (embedded within the exam) or may be taken as a standalone examination.

Lastly, Part III Clinical Skills Exam (CSE) is a performance-based exam wherein examinees are required to perform optometric clinical skills that reflect practice. These skills are performed at four different stations; all stations rely on standardized patients on whom the examinee performs the skills for each station. Candidates have 30 minutes at three stations, and 15 minutes at one station, making the total testing time 1 hour and 45 minutes, not including time for check-in, orientation, time between stations, and checkout. Each station is housed within an examination room that is designed to simulate real-life optometric exam rooms. The equipment, placement of materials, and room dimensions are standardized, and the NBEO follows a multilayered protocol for quality assurance throughout the examination process.

Given that every knowledge, skill, and ability necessary for entry into the independent practice of optometry cannot be tested in the same format, the examination series provides a scaffolded path for the assessment of overall competency. Figure 3 provides a graphical representation of this holistic assessment. Each exam within the series covers an aspect of optometric competency, but it is the combination of the series of exams that represents overall competency.

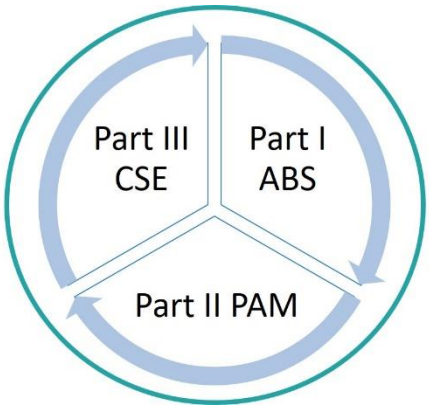

Figure 3. Three-part series of optometric licensure exams, when combined, measure overall optometric competency.

Each examination in the series contributes to the holistic assessment of competency. Following Miller's Pyramid of assessment (Miller, 1990), the NBEO exam series can be mapped along the pyramid by the area in which candidates demonstrate competency in each exam (see Figure 4). However, in optometry, the top of the pyramid, "Does," is truncated. The final assessment of competency in optometry is at the "Shows How" level because examinees are able to, after completing this level of assessment, apply for and receive a license to practice independently. In some other healthcare professions, examinees similarly progress through a series of licensure examinations to determine competency, but then must also undergo a period of supervised practice outside of their graduate medical education. This period of supervised practice, or, residency, falls within the category "Does" on the pyramid. Supervised practice operates as an additional layer to the overall assessment of clinical competency; the optometry profession does not require a residency.

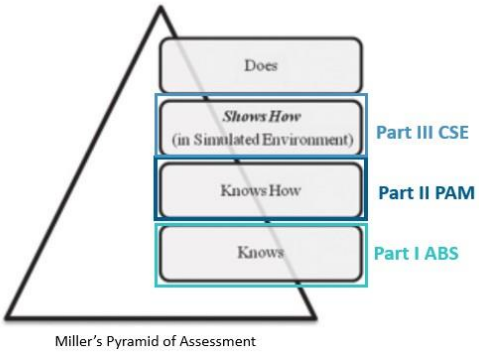

Figure 4. NBEO exam series as mapped onto Miller's pyramid.

The content matrices for each examination can be easily accessed at www.optometry.org, listed under the first tab, "NBEO Exams."

Blueprint Development and Task Analyses

Woo et al. (2022) misrepresent the processes by which they believe NBEO undertakes job analyses. They state, "The condition areas, disciplines and skills were updated in 2016. The process used the current framework used by NBEO® in lieu of a clean slate for blueprint

development…Since the NBEO relied on a prior framework, the examination missed an opportunity to incorporate a more contemporary, competency profile of abilities of an entry-level practitioner" (p. 13). This statement belies a misunderstanding of the various reasons for job analyses (also often called "practice analyses" or "job task analyses"). In the initial development of an examination, a job analysis should be undertaken to ensure that the examination is being created in such a way that it reflects the necessary KSAs for score interpretation to support intended inferences (such as decisions about licensure). However, job analyses are conducted on regular cycles to determine if the weighting of exam domains still accurately reflect practice.

There are several approaches to job analyses to consider. These include the critical incident technique (Flanagan, 1954), the task inventory approach (Newman et al., 1999), the professional practice model (LaDuca, 1994), functional job analysis (Fine & Wiley, 1971), as well as variations on the approaches listed including incorporating a cognitive task analysis. Whichever approach is taken, the goal of the analysis should be to link the KSAs needed for safe, effective independent practice of optometry to the KSAs measured on the examination (Clauser & Raymond, 2017). Historically, the NBEO has conducted job analyses with a focus on a matrix of task frequency and task criticality.

For examinations that are in regular, ongoing maintenance, such as those covered under the job analysis conducted in 2016 (Foley, 2016) and referred to by Woo et al. (2022), the job analysis is used to ascertain the accuracy of domain weights – if they still reflect practice. An entirely new examination blueprint, consisting of possibly new or different domains, constitutes either the generation of a new examination or a major restructure of an existing examination. Because NBEO is currently in the process of restructuring Part III CSE, the 2019 job analysis was conducted with a different purpose.

In 2019, NBEO conducted extensive research among a wide variety of stakeholders regarding the possible content of the restructured examination. A survey for stakeholders was deployed utilizing pairwise preference modeling (Bradley, 1984; Brown & Maydeu-Olivares, 2013; Hatzinger & Dittrich, 2012; Kendall & Smith, 1940; Zerman et al., 2018). The response rate was 36.9%. The sample contained respondents from all governmental jurisdictions who utilize the NBEO licensure exam series, including Canadian jurisdictions. The full analysis of the survey data was provided to a task force convened to specifically work on the examination restructure. This task force was comprised of a balance of practitioners ranging from a great deal of experience to only several years of experience, and included representation from Canada. After extensive work by this task force, NBEO conducted a job analysis targeted

specifically for Part III of the examination series for the construction and validation of the exam blueprint.

The job analysis conducted in 2016 and the one conducted in 2019 had different purposes. The appropriate time for a "clean slate for blueprint development" is when an exam is being created or undergoing major restructuring. The current restructure of Part III has in fact capitalized on the opportunity for the creation and validation of a new examination blueprint. While a job analysis is desirable at certain intervals, undertaking the creation of a new exam blueprint should only be done when there has been either a significant change in practice or it becomes desirable to completely revamp an exam.

Standard Setting

The current cut scores used to determine if a candidate passes or fails an examination were set at standard settings conducted for each specific part of the examination series (i.e. different standard settings held for Part I ABS, Part II PAM, and Part III CSE). The standard setting is held upon the initial administration of an exam form after it has been revised after a job analysis study – whether revised heavily to the extent the examination is being restructured and has a new blueprint or content outline, or revised in domain weighting or other areas of the existing blueprint. The exam form that constitutes the initial administration is thereafter known as the reference form. Subsequent exam forms are psychometrically equated to the reference form to ensure equality of difficulty. The standard setting method deployed should always directly relate to the type of examination for which it is being conducted. Overall, NBEO utilizes the Angoff (1971) method for all standard settings; during the most recent standard setting, NBEO used Impara & Plake's (1997) modified-Angoff method. Because all multiple-choice questions are scored dichotomously (correct or incorrect), this method allows subject-matter experts (SMEs) serving as panelists for the standard setting to review each item and indicate, based on their expertise in the field, the likelihood a minimally qualified candidate[2] will correctly answer the question. For the purposes of the standard setting procedure, this likelihood does not refer to a specific statistical, item response probability. For a more thorough explanation of the standard setting procedure, see Impara & Plake (1997), Plake et al. (2012), and Cizek & Bunch (2007).

Reliability

As previously stated, reliability in testing and measurement is an overarching term that generally means consistency – consistency with regards to many aspects of testing and measurement. Woo et al. (2022) discuss reliability within Kane's (2013) framework under the umbrella of generalization. As Kane (2013) discusses at length, exam reliability is a necessary

[2] For NBEO a minimally qualified candidate represents a candidate who is minimally competent to enter into the safe, effective independent practice of optometry.

but insufficient criterion for exam validity. Kane (2013) writes, "Evidence of generalizability (or reliability) therefore is rarely sufficient for validity" (p.3). That is, validity evidence must contain evidence of reliability; however, evidence of reliability alone does not constitute sufficient evidence of validity.

Woo et al. (2022) correctly establish that generalization refers to the relationship between observed test scores and true test scores; however, they do not accurately describe the role of sampling error within generalizability theory. Drawing directly from Kane's (2013) work, "Universes of generalization include observations that can vary in a number of ways, involving, for example, samples of tasks, testing contexts [location at which the test is given], occasions in which the test is administered, and possibly raters who score the responses" (p. 26). Sampling error typically derives from multiple possible sources such as the variance in the facets just described. The authors, though, describe sampling in terms of sampling test items over an array of exam domains, referring to "the likelihood of obtaining similar scores if new items are used" (p. 7). The use of different items on different exam forms while maintaining the same level of exam difficulty is more appropriately discussed in the process of statistically equating exam forms. In terms of exam reliability, sampling error refers to the "uncertainty in the generalization" (Kane, 2013, p. 26); that is, the imprecision of the relationship between observed scores and true scores. Standard errors and confidence intervals around them provide information about the strength of that relationship.

Setting aside the treatment of error within generalization theory, Woo et al.'s (2022) interpretation of Kane's (2013) framework describes reliability mostly in terms of internal exam reliability. Additional aspects of reliability include but are not limited to the following: test-retest reliability, parallel forms reliability, and intra- and inter-rater reliability. Woo et al. (2022) indicate that "The overall reliability is determined using Livingston's criterion-referenced coefficient alpha (Livingston, 1972)" for the OEBC examination. This statistic represents a measure of internal reliability. NBEO reviews several measures with every test form administration including Cronbach's alpha (1951) and Livingston-Lewis decision consistency and decision accuracy (1995). Further, reliability is assessed by reviewing the standard error of measurement (SEM) for each test form as well as the conditional SEM at the cut score. Of note, regardless of which alpha reliability coefficient used, the alpha estimate is conditional to the sample of test scores from which it was derived. That is, internal reliability should be thought of less as "a property of the test itself, but of the scores from the test that are obtained from a particular sample" (Bandalos, 2018, p. 186). It is for this reason that NBEO measures internal reliability along different dimensions for each administration of a new test form.

Method

Norcini et al.'s (2018) consensus framework of assessment in healthcare is used to review OEBC and NBEO licensure exams. Under this framework, Woo et al. (2022) draw on Kane's (2013) discussion of validation and interpretations of *uses* of test scores. Woo et al. (2022) write, "The understanding of validity has changed from considering separate types of validity to a single concept of construct validity" (p. 6). This is an oversimplification of Kane's (2013) article which details at length a framework for Interpretation and Use Arguments (IUAs) to keep the interpretation and uses of test scores in alignment with evidence that supports those uses and interpretations. The *Standards* (2014) state, "Validity is a unitary concept" (p.14); Kane's framework (2013) is in alignment with this. Specifically, in the 2014 edition, the *Standards* continue in the usage of nomenclature around types of validity evidence rather than types of validity, "… (i.e., the use of the terms content validity or predictive validity)" (p.14). Kane (2013) does not dismiss altogether various aspects of validity (or types of validity evidence); however, his framework does emphasize four critical inferences needed for valid interpretations and uses of test scores. There is continued discussion within the field regarding Kane's framework (Cook et al., 2015). For example, the *Standards* (2014) continue to emphasize the five sources of validity evidence introduced by Messick (1984) rather than the key inferences from Kane's (2013) framework.

Kane's (2013) Inferences

In his description of an IUA for an observable attribute of an exam, Kane (2013) states that three main inferences are typically necessary: scoring, generalization, and "some kind of extrapolation to nontest performances or competencies" (p. 25).

Scoring

Kane (2013) describes the scoring inference as resting upon assumptions about the process of scoring items or tasks. First, the scoring inference assumes that scoring procedures are appropriate (for example, multiple choice items should be scored using a procedure that is different from that of scoring an essay). Next, the scoring inference assumes that the scoring procedures are applied correctly. Lastly, the inference assumes that the scoring procedure are free of bias. The evidence to support these assumptions exists primarily within the initial exam development process. For example, documentation of the construction of a scoring procedure when a test is first developed can provide evidence that the assumption of appropriateness is met. Steps to ensure raters are consistently applying scoring rubrics would constitute evidence that scoring procedures are being applied correctly.

Woo et al. (2022) describe item analyses and internal reliability statistics as the evidence for validity in terms of the scoring inference for the OEBC. While these aspects of exam development and maintenance are crucial, they do not fall within Kane's (2013) taxonomy of

validity evidence under scoring. Woo et al. (2022) do not provide information about the appropriateness of scoring procedures, evidence of correct application of scoring procedures, or steps to ensure that scoring procedures are not biased.

Generalization

As previously stated in this paper, generalization refers to the relationship between observed scores and true scores. Kane (2013) describes the generalization inference stating, "…[it] treats the observed score as an estimate of the universe score (the mean over the universe of generalization for the test taker)" (p.25). More simply stated, the generalization inference refers the assumption that the observed score (Sally's score on the exam on a particular day) is representative of the true score (the average of Sally's scores on the exam if she took it at the same time an infinite number of times). Evidence for the strength of the relationship between the observed and true score can be estimated by sampling error. Statistics such as the SEM, Cronbach's alpha (1951), and Livingston-Lewis decision consistency and decision accuracy (1995) measure.

The authors describe the overall reliability of the OEBC in terms of Livingston's criterion-referenced coefficient alpha (Livingston, 1972) under the scoring inference rather than the generalization inference. Additionally, according to Kane (2013), "To simply estimate coefficient alpha and some other measure of internal consistency and assume that the question of reliability / generalizability has been addressed is to beg the question of generalizability" (p. 20). That is, providing basic information about internal reliability falls into one of the validity fallacies Kane (2013) discusses (see pgs. 18-19). The authors' discussion of the development of a set of competencies for the OEBC, provided as validity evidence in terms of generalization, is important though misplaced within Kane's (2013) framework.

Extrapolation

Within Kane's framework, being able to draw inferences from test scores about how a candidate will perform as an independent optometrist in the real-world, is categorized as extrapolation. Kane (2013) describes the extrapolation inference as extending,

> "…the interpretation of the universe of generalization to the target domain. The score does not necessarily change, but its interpretation is broadened to include "real-world" performances. The extrapolation inference does not involve a simple statistical generalization, but rather a more-ambitious leap from claims about test performances to claims about the full range of performance in the target domain, including nontest performances in nontest contexts" (p. 28)

A more simplistic description of extrapolation is, assuming the observed score represents the true score, how well does the observed score represent possible content knowledge or performance of skills in practice. For licensure testing programs, the target domain that Kane

(2013) describes is independent practice. Support for the extrapolation evidence, as described by Kane (2013) can be both analytic and empirical. Analytic evidence, broadly, is developed during initial exam development and is descriptive of how well the content and design of the exam matches the target domain.

We contend that Woo et al.'s (2022) paper largely focuses on validity evidence in terms of extrapolation, based on analytic evidence, asserting that the OEBC has a significantly stronger extrapolation inference above that of the licensure series administered by the NBEO. Their evidence rests largely on the contention that the OEBC is developed by Canadian optometrists for the Canadian context whereas the NBEO exam is not. This is factually inaccurate. In each cycle of stakeholder input (as described previously), NBEO conscientiously strives to obtain feedback from the Canadian provinces in which NBEO examinations are accepted. Woo et al.'s (2022) statement that, "the sample of practitioners used to provide the context were limited to their own countries. In other words, the OEBC worked with Canadian optometrists while the NBEO worked with American optometrists" (p. 15) is simply incorrect with regards to the procedures NBEO follows.

Further, Woo et al. (2022) point out, "There are differences between the US and Canada in the practice of optometry including the legislated scope of practice, availability and naming of pharmaceuticals, standards of practice, regulations and the system of healthcare - among others" (p. 15). NBEO does not find this to be a strong argument against the use of our examinations. In the United States, each state and federal jurisdiction (i.e. Washington, D.C. and Puerto Rico) has its own regulations. There is no *one* U.S. context; there is a plurality around which NBEO examinations have been designed and continuously updated. To that end, candidates are provided with an updated list of medications that they may access while completing Parts I and II.[3] The differences in medication availability, bottle sizes, concentrations of active ingredients, and different names for the same formulation of medications the authors cite are not differences that make NBEO examinations inapplicable in the Canadian context. In fact, given that candidates have access to an annually updated reference list of medications for NBEO Parts I and II, NBEO examinations offer quite a lot of flexibility to accommodate a variety of candidate contexts. Additionally, NBEO offers advanced exams when the scope of optometric expands based on the needs of jurisdictions.[4]

Lastly, NBEO disagrees that a review of the impact of the National Council Licensure Examination – Registered Nurses (NCLEX-RN®) constitutes indirect evidence of how the use

[3] Part III CSE does not require candidates to give specific medication names or dosages thus eliminating the need for a medication reference list.
[4] For example, the NBEO Laser & Surgical Procedures Exam (LSPE) and Injection Skills Exam (ISE) are offered as stand-alone exams.

of NBEO examinations has impacted or may in the future impact Canadian candidates or the Canadian healthcare system. Nursing represents a qualitatively different profession from optometry. Generally, the care provided by a nurse is not fully independent – it is typically supported by a physician who is ultimately responsible for patient safety and well-being, supervising directly or indirectly nursing practice. Optometrists, however, are physicians who work independently, are not required to study for a period of supervised practice through residency, and are not required to work under a supervising optometrist upon receiving a license to practice. These differences inherently make comparisons between the two professions troublesome.

However, given that Woo et al. (2022) compare the two, a review of their argument is revealing. They indicate that pass rates among Canadian candidates on the NCLEX-RN® were lower than those of U.S. candidates. This data alone is insufficient to substantiate an argument that cultural differences are at cause for the differences in pass rates between the two groups. A number of factors could play a role in generating different pass rates, not the least of which is that any exam administered among a new population can be expected to have lower pass rates in the new population for a time period. The only way to properly compare the performance of Canadian candidates on the NCLEX-RN® to the Canadian Registered Nurse Exam (CRNE) would be through a statistical analysis of the difficulty of the two exams when examinee ability, along with other influencing factors such as demographics, are held constant. The evidence provided by Woo et al. (2022) is qualitative in nature and is oriented to demonstrate that the NCLEX-RN® was not designed for the Canadian context. However, the issue at hand is not whether it was designed for the context, but rather, whether or not it works equally well within the Canadian context as the CRNE.

Implication

The author's assertions regarding the implications of accepting the NCLEX-RN®, regardless of the appropriateness of comparison, can be correctly classified under Kane's (2013) treatment of the implications of test score interpretations and uses. Kane (2013) provides an in-depth discussion of the role of the consequences of test score interpretations and uses in evaluating those interpretations and uses. The three main negative outcomes from evaluations of score-based decision rules include the following:

> "…(1) the extent to which the intended outcomes are achieved, (2) differential impact on groups (particularly adverse impacts on legally protected groups, and (3) positive and negative systemic effects (particularly in education)" (Kane, 2013, p. 48).

The author's assertion that the acceptance of the NCLEX-RN® has had "an adverse impact on public perception of the profession" represents the third category of negative outcomes that Kane (2013) provides for evaluating the consequences of score-based decision procedures.

The author's discussion of the standard setting procedure that is provided in their discussion of validity evidence for implications would better reflect support for the scoring inference. Specifically, standard setting procedures offer evidence of the appropriateness of the scoring procedures used for an exam.

Consistency & Equivalency

The section of the article under the headings "OEBC - Consistency & Equivalency" and "NBEO - Consistency & Equivalency" are not aligned with Kane's main inferences for evaluating score interpretations and uses. The authors describe the use of simulations within the OEBC and the lack of simulations within the NBEO series in these sections. Simulations are described as, generally, increasing the overall reliability of the exam. While the NBEO does not necessarily disagree with this assertion, it is worthwhile to note Kane's (2013) discussion of the reliability / validity paradox.

Kane (2013) writes, "…standardization can increase the expected correlation of the observed score with the universe score, but it tends to decrease the expected correlation of the universe score with the target score" (p. 31). Using different terminology, standardization improves generalization (how well the observed core matches the true score), at the expense of extrapolation (how well the observed score matches the target domain, i.e., independent practice). In other words, as an exam becomes increasingly standardized, our confidence in the generalization inference goes up but our confidence in the extrapolation inference goes down. This paradox occurs because high standardization reduces variance in the observed score. However, a strictly standardized exam is less similar to the real-world, the target domain about which we wish to make extrapolation inferences. For example, an exam that measures a certain content domain is given in the format of multiple-choice questions. This format is highly standardized and thus has very strong evidence for generalization inferences. We know this because multiple-choice formats tend to have very high internal reliability statistics and low SEMS. Imagine now a different exam that measures the same content domain, but instead of multiple-choice questions the test-taker has to perform specific tasks that are necessary for safe, effective practice in the real world. This format will definitionally have less evidence for generalization. For example, this format requires not only internal reliability but also inter-rater and intra-rater reliability. However, this format is much more representative of the real-world than the multiple-choice format. Thus, as generalization increases, extrapolation decreases – or, as the name of the paradox suggests, as reliability increases, validity decreases.

Conclusion

In sum, we remind the reader that the that the absence of a formal validity study to scientifically measure the appropriateness of NBEO examinations in the Canadian context does not mean NBEO examinations are necessarily inappropriate. It simply means there has yet to be a scientific study of either analytic or empirical evidence for the extrapolation inference for either OEBC or NBEO. Until comparable validity studies are available for both NBEO and OEBC exams, no conclusions, at least no conclusions grounded in empirical evidence, can be drawn about stronger or weaker validity within the Canadian context.

Below is a reproduction of Table 4 from Woo et al. (2022), modified to provide accurate information about the NBEO exam series and reclassifying the author's assertions correctly within Kane's (2013) framework. Of note, the author's item, "No correlation with other measures" listed in the row labeled "Extrapolation" was removed due to insufficient information accompanying the statement. It is possible that this statement is referring to evidence of discriminant validity, and if so, the statement would most appropriately be classified under the extrapolation inference. Evidence of discriminant validity speaks to the degree to which the exam does not measure constructs for which it was not designed to measure (for example, a mathematics examination may incidentally measure reading ability due to the inclusion of word problems). If an exam assesses constructs other than the intended construct (i.e., has poor evidence of discriminant validity), the evidence backing extrapolation inferences is weakened due to the incongruence between what the observed score measures and what the observed score is intended to measure.

Table 4: Evidence for validity and of OEBC and NBEO® Assessment using Kane's validity framework. *Adapted from Woo et al. (2022).*

	OEBC	**NBEO®**
Scoring Inference *Assumes that scoring procedures are,* *1) appropriate to the test* *2) applied correctly* 3) *minimize bias in scoring*	• MCQ and OSCE (skills and higher reasoning) • Assessors trained for OSCE stations • Canadian practitioners with support of psychometrician • Cut-Score standard setting • Angoff method • Criterion-referenced • Question developers trained to write questions and do assessments for entry-level competency in Canada	• Angoff & Modified-Angoff standard setting methods for criterion-referenced examinations (W. H. Angoff, 1971; Plake et al., 2012); standard setting based on the minimally qualified candidate (MQC) and competencies needed for safe, effective independent practice • Scoring validated both by internal psychometrician and independent, external psychometrician. • Scoring procedures based on multiple-choice questions, multiple response questions, case scenarios, and performance-based skills assessment (currently being restructured to include OSCE format) • Exam blueprinting process includes stakeholder survey, task force of subject matter experts, JTA, task force review of JTA, review by Board of Directors • Exam maintenance includes review of items for content and statistical properties upon every test administration by layers of content experts, continuous training and calibration of raters
Generalization Inference *the relationship between observed scores and true scores*	• Internal review process using Livingstone coefficient • Simulations for greater reliability	• Cronbach's alpha • SEM • Conditional SEM at the cut score • Livingston-Lewis decision consistency • Livingston-Lewis accuracy
Extrapolation Inference *how well does the observed score represent possible content knowledge or performance of skills in practice; supported by analytic and empirical evidence*	• Bilingual (French and English) • Canadian competency statements	• Optometrists representing various levels of experience, various modes of practice, various academic institutional affiliations, and diversity of demographic categories participate in exam development and maintenance. • Exams designed for plurality of target domain contexts • Exam given in English only
Implications	• potential loss of a viable, bilingual OEBC assessment	• Insufficient evidence of adverse effects of the use of the NBEO licensure series

In conclusion, Woo et al.'s (2022) statement that, "Although intrinsically satisfactory for their respective jurisdictions, the NBEO® does not appear to satisfy the critical criteria of validity, equivalency and acceptability for Ontario or, more broadly, Canada" (p.24) lacks merit. Insufficient evidence is provided to substantiate this claim, and the evidence the authors do provide is often misclassified within the frameworks they reference. We encourage the reader to review the updated information about NBEO Parts I, II, & III provided in this paper and on the NBEO website (www.optometry.org) in order to more accurately inform their understanding of the content of and validity evidence for the NBEO examination series.

References

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education (Eds.). (2014). *Standards for Educational and Psychological Testing*. American Educational Research Association.

Angoff, W. A. (1971). Scales, norms and equivalent scores. In R. L. Thorndike (Ed.), *Educational Measurement* (pp. 508–600). American Council on Education.

Angoff, W. H. (1971). *Educational measurement*. American Council on Education.

Balogh, J. E. (2016). *A practical guide to creating quality exams*.

Bandalos, D. L. (2018). *Measurement theory and applications for the social sciences*. Guilford Press.

Bradley, R. (1984). Paired comparisons: Some basic procedures and examples. In P. Krishnaiah & P. Sen (Eds.), *Handbook of Statistics* (Vol. 4, pp. 299–326). https://doi.org/10.1016/S0169-7161(84)04016-5

Brown, A., & Maydeu-Olivares, A. (2013). How IRT can solve problems of ipsative data in forced-choice questionnaires. *Psychological Methods*, *18*(1), 36.

Cizek, G. J., & Bunch, M. B. (2007). *Standard setting: A guide to establishing and evaluating performance standards on tests*. Thousand Oaks, Calif. ; London : SAGE, c2007. https://catalog.lib.unc.edu/catalog/UNCb7068686

Clauser, A. L., & Raymond, M. R. (2017). Specifying the Content of Credentialing Examinations. In S. Davis-Becker & C. W. Buckendahl (Eds.), *Testing in the professions: Credentialing polices and practice* (pp. 64–84). Routledge, Taylor & Francis Group.

Cook, D. A., Brydges, R., Ginsburg, S., & Hatala, R. (2015). A contemporary approach to validity arguments: A practical guide to Kane's framework. *Medical Education*, *49*(6), 560–575. https://doi.org/10.1111/medu.12678

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, *16*(3), 297–334. https://doi.org/10.1007/BF02310555

Fine, S. A., & Wiley, W. W. (1971). *An Introduction to Functional Job Analysis: A Scaling of Selected Tasks from the Social Welfare Field. Methods for Manpower Analysis No. 4.*

Flanagan, J. C. (1954). The critical incident technique. *Psychological Bulletin*, *51*(4), 327.

Foley, B. JP. (2016). *Job Analysis Survey Report for the National Board of Examiners in Optometry* (pp. 1–25) [Job Analysis]. Alpine Testing Solutions.

Hatzinger, R., & Dittrich, R. (2012). Prefmod: An R package for modeling preferences based on paired comparisons, rankings, or ratings. *Journal of Statistical Software*, *48*(10), 1–31.

Impara, J. C., & Plake, B. S. (1997). Standard setting: An alternative approach. *Journal of Educational Measurement*, *34*(4), 353–366.

Kane, M. T. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement*, *50*(1), 1–73.

Kendall, M. G., & Smith, B. B. (1940). On the method of paired comparisons. *Biometrika*, *31*(3/4), 324–345.

LaDuca, A. (1994). Validation of professional licensure examinations: Professions theory, test design, and construct validity. *Evaluation & the Health Professions*, *17*(2), 178–197.

Livingston, S. A. (1972). Criterion-referenced applications of classical test theory 1, 2. *Journal of Educational Measurement*, *9*(1), 13–26.

Livingston, S. A., & Lewis, C. (1995). Estimating the consistency and accuracy of classifications based on test scores. *Journal of Educational Measurement*, *32*(2), 179–197.

Messick, S. (1984). The Psychology of Educational Measurement. *Journal of Educational Measurement*, *21*(3), 215–237. JSTOR.

Miller, G. E. (1990). The assessment of clinical skills / competency / performance. *Academic Medicine*, *65*(9), S63-67.

Newman, L., Slaughter, R., & Taranath, S. (1999). *The selection and use of rating scales in task surveys: A review of current job analysis practice*. Annual meeting of the National Council of Measurement in Education, Montreal, QC.

Norcini, J., Anderson, M. B., Bollela, V., Burch, V., Costa, M. J., Duvivier, R., Hays, R., Palacios Mackay, M. F., Roberts, T., & Swanson, D. (2018). 2018 Consensus framework for good assessment. *Medical Teacher*, *40*(11), 1102–1109.

Plake, B. S., Cizek, G. J., & Cizek, G. (2012). The modified Angoff, extended Angoff, and Yes/No standard setting methods. *Setting Performance Standards. Foundations, Methods, and Innovations*, 181–253.

Woo, S., Hrynchak, P., & Hutchings, N. (2020). *Applicability of Entry to Practice Examinations for Optometry in Canada and the United States – Optometry Examining Board of Canada and National Board of Examiners in Optometry* (pp. 1–18). University of Waterloo School of Optometry & Vision Science.

Woo, S., Hrynchak, P., & Hutchings, N. (2022). Applicability of Entry to Practice Examinations for Optometry in Canada and the United States – Optometry Examining Board of Canada and National Board of Examiners in Optometry. *Canadian Journal of Optometry*, *84*(1), 1–18.

Zerman, E., Hulusic, V., Valenzise, G., Mantiuk, R. K., & Dufaux, F. (2018). The relation between MOS and pairwise comparisons and the importance of cross-content comparisons. *Electronic Imaging*, *2018*(14), 1–6.