# Flaws in the design of the *Examen Nacional para Aspirantes a Residencias Médicas* produce inequity

Aldo Barajas-Ochoa, MD,[1] César Ramos-Remus, MD, MSc,[2] José Dionisio Castillo-Ortiz, MD,[3] José Yáñez, MD, MHD,[4] Zalathiel Barajas-Ochoa, MD,[1] Jorge Manuel Sánchez-González, MD, PhD,[5] Mauricio Hernández-Ávila, MD, MSc, DSc,[6] José Ángel Córdova-Villalobos, MD,[7] Lilia Patricia Bustamante-Montes, MD, MSc, DrPH.[8]

**Barajas-Ochoa A, Ramos-Remus C, Castillo-Ortiz JD, Yáñez J, Barajas-Ochoa Z, Sánchez-González JM, Hernández-Ávila M, Córdova-Villalobos JA, Bustamante-Montes LP.**
**Flaws in the design of the *Examen Nacional para Aspirantes a Residencias Médicas* produce inequity.**
**Salud Publica Mex. 2019;61:125-135.**
*https://doi.org/10.21149/9790*

## Abstract

**Objective.** To assess the assumption of 'equity' of Mexico's resident-selection assessment tool, the *Examen Nacional para Aspirantes a Residencias Médicas* (ENARM). **Materials and methods.** Official ENARM-2016 and -2017 databases were analyzed. Differences in the absolute number of correct answers (multivariable linear regression) and the number of applicants reaching their specialty minimum score (SMS) per test day (odds ratio [OR]) were calculated. Applicants affected by test-day inequity were estimated. **Results.** There were 36 114 applicants in 2016, and 38 380 in 2017. In 2016, day-2 applicants had significantly higher scores and more reached the SMS than on days 1-3-4 (OR 1.55), and 5 (OR 3.8); 3 565 non-passing applicants were affected by inequity (equivalent to 44.64% of those selected). In 2017, day-1 and -2 applicants had significantly higher scores and more reached the SMS than

**Barajas-Ochoa A, Ramos-Remus C, Castillo-Ortiz JD, Yáñez J, Barajas-Ochoa Z, Sánchez-González JM, Hernández-Ávila M, Córdova-Villalobos JA, Bustamante-Montes LP.**
**Fallas en el diseño del Examen Nacional para Aspirantes a Residencias Médicas generan inequidad.**
**Salud Publica Mex. 2019;61:125-135.**
*https://doi.org/10.21149/9790*

## Resumen

**Objetivo.** Evaluar el atributo de "equidad" asignado al Examen Nacional para Aspirantes a Residencias Médicas (ENARM). **Material y métodos.** Se analizaron las bases de datos oficiales del ENARM 2016 y 2017. Se compararon las diferencias inter-día de respuestas correctas (regresión linear multivariable) y de sustentantes que alcanzaron el puntaje mínimo de su especialidad (PME) (razón de momios [RM]). Se estimó a los afectados por la inequidad. **Resultados.** Hubo 36 114 sustentantes en 2016 y 38 380 en 2017. Los días 2 (ENARM-2016) y 1-2 (ENARM-2017) registraron puntajes significativamente más altos, y más sustentantes alcanzaron el PME que en los días 1-3-4 (RM .55) y 5 (RM 3.8) en 2016, y los días 3-4 (RM 1.85) y 5 (RM 4.04) en 2017. Se estimó que cuatro de cada diez sustentantes que aprobaron el ENARM no lo hubieran hecho si el examen

(1) Unidad de Investigación en Enfermedades Crónico-Degenerativas. Guadalajara, Jalisco, México.
(2) Universidad Autónoma de Guadalajara. Zapopan, Jalisco, México.
(3) Unidad de Investigación en Enfermedades Crónico-Degenerativas. Guadalajara, Jalisco, México.
(4) Universidad Iberoamericana. Mexico City, México.
(5) Instituto Nacional del Aprendizaje, Habilidades e Investigación de las Ciencias, S.C. Zapopan, Jalisco, México.
(6) Centro Universitario de los Altos, Universidad de Guadalajara. Tepatitlán de Morelos, Jalisco, México.
(7) Universidad de Guanajuato. Mexico City, México.
(8) Universidad Autónoma de Guadalajara. Zapopan, Jalisco, México.

**Received on:** June 7, 2018 • **Accepted on:** August 31, 2018
Corresponding autor: Aldo Barajas Ochoa. Unidad de Investigación en Enfermedades Crónico-Degenerativas. Colomos 2292, Providencia. 44620 Guadalajara, Jalisco, Mexico.
E-mail: aldouch@gmail.com

on days 3-4 (OR 1.85), and 5 (OR 4.04); 3,155 non-passing applicants were affected by inequity (37.2% of those selected). **Conclusion.** Analysis of official ENARM databases does not support the official attribution of equity, suggesting the test should be redesigned.

Keywords: education; graduate medical; specialty; residency and internship; personnel selection; academic test score; psychometrics

fuera equitativo. **Conclusión.** Los resultados sugieren que el atributo de equidad del ENARM está en duda.

Palabras clave: educación de postgrado en medicina; especialización; internado y residencia; selección de personal; rendimiento académico; psicometría

The availability and quality of facilities and human resources in healthcare is of utmost importance,[1] with physicians comprising one of the main human resources. Most medical doctors (MDs) seek to continue their training within a specialty. However, not all MDs pursuing specialty training can enter an official program.

Most Western countries have regulatory agencies that are responsible for evaluating, selecting, and placing MDs to continue training. These agencies use multiple high-stakes processes to select the best candidates.[2-4] For example, the selection process in the United States involves approving different components of the United States Medical Licensing Examination, interviewing, and participating in a national matching program. In France, students present a national ranking examination and choose the specialty they wish to pursue according to their rank.[4] In Mexico, MDs must first register on the registry of the *Examen Nacional para Aspirantes a Residencias Médicas* (National Exam for Applicants to Medical Residencies, ENARM).[5] This requires registration of an individual's ID, the specialty for which they are competing (one of 27 options), and the preferred date to present the exam (one of five consecutive days on which the test is available yearly, for 2016 and 2017). An online registration platform is enabled at a predetermined day and time, and closes around 10 days later. As registration and date-selection occurs on a first-come, first-served basis, 36 000+ applicants typically register during the first 24 hours. The ENARM is a norm-referenced test,[6-8] and is the assessment tool used in Mexico to select those who will enter specialty training. There are approximately 8 000 official training positions available nationwide. These positions vary by specialty, and are determined by the certified training programs of the healthcare institutions that provide them. Applicants for each specialty are ranked from highest to lowest according to their total ENARM score. Ranked applicants receive a 'pass' certificate until the quota is met according to that specialty's available positions, and present it to the institution of their choice to apply for an appointment.

The quality and performance of such selection and placement processes cannot be overemphasized, as they have a beneficial or harmful impact on the candidate, patients, healthcare systems, and ultimately, the public health of a given country. The *Comité de Posgrado y Educación Continua* (Committee of Postgraduate and Continuous Education [CPEC]), the agency in charge of the ENARM, confers the attributes of "equitable, transparent, objective, and valid" to the ENARM.[9,10] The CPEC declares that the five different test forms (one for each test day, each comprising an exclusive set of items) used for the ENARM in a particular year are equivalent because they are created with "the same objectivity, quality, balance, and academic level".[9,10] However, our group recently reported that the test development processes and the theoretical basis under which the ENARM (2016 and 2017) was built do not support the stated attributes of validity, reliability, and equity.[11] This contrasts with important advances that have resulted in a change in established standards for educational testing,[12-14] and supports the statement that, for some systems of educational assessment in medicine, tradition weighs more than evidence-based methods and best practices.[15]

The CPEC does not precisely define 'equity', but the sense of this attribute is close to that of 'fairness' (preferred in the English language and in the literature related to education), as defined by Lane and colleagues,[16] "fairness in testing is achieved if a given test score has the same meaning for all examinees and is not substantially influenced by factors not relevant to the examinee's performance."

This study aimed to assess whether the ENARM met the attribute of 'equity' (i.e. fairness) as ascribed by the CPEC, by analyzing the databases of the results for all applicants in 2016 and 2017.

## Materials and methods

This study analyzed the official databases of ENARM 2016 and 2017. The study took place in Guadalajara, Mexico from February to May of 2018.

## ENARM characteristics

The ENARM test is available yearly and measures knowledge in general medicine. Five test forms are created each year, each comprising 450 multiple-choice single-best answer items; no item is used in more than one test form. Each form is used nationwide for one of the five consecutive days on which applicants can present the test. All forms comprise the same number of items per area of knowledge (specialty/subspecialty), with an approximate item distribution of 37.5% internal medicine, 25% pediatrics, 22% gynecology-obstetrics, and 15% surgery. Furthermore, all test forms include 25% low-difficulty (n=113), 50% medium-difficulty (n=225), and 25% high-difficulty items (n=112).[17] The processes of item development and test assembly have been the same since at least 2010.[18] These processes are performed by "at least eight expert professors, whom are selected according to specific profiles to participate in different steps of item-development such as creation, analysis and calibration, quality-control, decision of difficulty category, and final approval".[9,10] These experts assign items to the three difficulty categories, based on "assessing the number of correct answers of five expert clinicians and their experience," without testing the items on the target population[17] or using psychometric theories.

Applicants receive their ENARM score report immediately after finishing the test. The score report includes the number of correct answers (NoCA) per item difficulty-category, NoCA per area of knowledge, absolute number of correct answers (ANoCA: 0–450), and the total score (0–100). The total score is calculated by dividing the ANoCA by the total number of items (i.e., ANoCA/450). After all tests have been completed, applicants' registers are clustered by nationality (Mexican or foreign medical graduate) and chosen specialty (the same 27 direct-entry specialties in 2016 and 2017), and ranked from best to worst according to total scores. Applicants are then selected until the quota for each specialty is met. When a tie in the total score occurs, the tiebreak process considers (in successive order) applicants' scores in internal medicine, pediatrics, gynecology-obstetrics, surgery, and the level of difficulty of correctly-answered items.[7,8]

## Characteristics of the ENARM 2016 and 2017 databases

The anonymized complete official ENARM databases for 2016 and 2017 were obtained from the *Dirección General de Calidad y Educación en Salud* (General Direction of Quality and Education in Health; DGCES), a federal institution that participates in ENARM development, through the *Plataforma Nacional de Transparencia* (National Transparency Platform). This platform is supported by the *Instituto Nacional de Transparencia, Acceso a la Información Pública y Protección de Datos Personales* (National Institute of Transparency, Access to Information and Protection of Personal Data). In Mexico, the federal law on transparency and access to public information (*Ley General de Transparencia y Acceso a la Información Pública*) allows citizens to obtain data from publicly funded institutions, provided the information is not considered confidential and does not affect the privacy of third persons.

The databases included all applicants of the 2016 and 2017 ENARMs. Available data per applicant included age, sex, medical school, nationality (Mexican, foreign), chosen specialty (the same 27 specialties for 2016 and 2017), test date (five days for each year), overall rank in the test, NoCA in low- (0–113), medium- (0–225), and high-difficulty (0–112) items, and total score (0–100).

## Statistical analysis

This ecological study of the ENARM databases performed the same analysis separately for 2016 and 2017. The primary end-point was 'equity,' which was assumed if no significant inter-day differences in ANoCA existed. If equity was present, test forms were assumed to have the same difficulty. If lack of equity was suggested by the primary end-point, the secondary end-point was to estimate the number of applicants dubiously classified as not passing.

For the primary end-point, an exploratory analysis was performed to assess inter-day differences on the average ANoCA. To assess this, the variable 'ANoCA' (0–450) was created by adding applicants' NoCA for low-, medium-, and high-difficulty items. The average ANoCA per day was calculated and compared using one-way ANOVA. If inter-day differences in the average ANoCA suggested a lack of equity, further analysis with estimation of effect size was performed.

The effect size for lack of equity was assessed by two methods. For the first method, the categorical variable 'applicant reached the specialty minimum score' (ARSMS) was created, as data for pass/not pass status was not available. Specialty minimum scores (SMS) were obtained from reports published by the *Comisión Interinstitucional para la Formación de Recursos Humanos para la Salud* (CIFRHS),[19,20] the federal agency that reports ENARM outcomes. Each SMS was converted into its correspondent ANoCA, which represented the minimum ANoCA required to obtain a pass certificate for that specialty in that year (mANoCA). Next, whether an applicant's ANoCA was equal or higher to

their selected specialty mANoCA was coded as yes/no. The odds ratios (OR) for the proportion of ARSMS by specialty per day were calculated. The days with similar proportions of ARSMS were grouped, and the OR between the day/group with the higher proportion of ARSMS and the other days/groups were calculated.

The second effect size method involved constructing a multivariable linear regression model that calculated the inter-day differences in ANoCA and in NoCA per item's ascribed difficulty. By themselves, these values represent an appropriate measure of the magnitude of the effect. The model included available variables that could explain the differences: sex, age (potential surrogate for years since graduation and the number of test attempts), nationality, and chosen specialty (because of auto-selection). A sensitivity analysis that also included the medical school 'public/private' status was performed.

A third effect size definition was used to assess the secondary end-point. The number of applicants dubiously classified as not passing (i.e., number of affected applicants) was estimated using the coefficients obtained in the regression model. Days on which the regression model showed similar results (differences of ≤2 items) or overlap of the confidence intervals (CI) for the ANoCA were grouped. A reference group (*R*) was defined when ≥3 days met the former condition, or when two days met the condition and their average ANoCA was similar to that year's average ANoCA. The differences between the rounded averages of *R* and days *i* and *j* ('*i*' for day[s] with higher scores than *R*, and '*j*' for day[s] with lower scores), and between days *i* and *j* were used to calculate the ANoCA range where affected applicants lay. For each specialty, the SMS (CIFRHS 2016b, 2017b) was expressed as the mANoCA. To calculate the lower limits of each range, differences between *i* and *R* ($\Delta iR$), *i* and *j* ($\Delta ij$), and *R* and *j* ($\Delta Rj$) were subtracted from each specialty's mANoCA. For the upper limits, 1 was subtracted from the mANoCA. Therefore, the limits of the ranges for each specialty were ([mANoCA – $\Delta iR$] to [mANoCA – 1]), ([mANoCA – $\Delta ij$] to [mANoCA – 1]), and ([mANoCA – $\Delta Rj$] to [mANoCA – 1]). The sum of the number of applicants per specialty that lay within the ranges ([mANoCA – $\Delta iR$] to [mANoCA – 1]) and ([mANoCA – $\Delta ij$] to [mANoCA – 1]) reflected the number of applicants affected by the lack of equity in different ENARM test forms; that is, those that could have obtained a pass score if they had completed the test on the easiest day.

Confidence intervals (95%) and effect size were used to define significance, as statistically significant *p*-values were assumed likely to appear because of the population size.[21,22]

This study used an audit approach with publicly-available, anonymized, official databases. Informed consent and institutional review board approval was not required.

## Results

In the 2016 ENARM, 36 114 applicants from 112 medical schools applied for 27 specialties, mainly general surgery, internal medicine, gynecology-obstetrics, pediatrics, anesthesiology, and family medicine. Table I shows applicants' demographics, NoCA by item difficulty, and ANoCA. Applicants' age and sex distribution was similar across all ENARM days, except for day 5, in which applicants were on average 2 years older. Significant differences were found (according to the 95%CI) in scores for the low-, medium-, and high-difficulty questions; day 2 had higher scores (easier test form) and day 5 had lower scores (more difficult test form). These differences were also reflected in the ANoCA. Univariate linear regression showed that the ANoCA decreased significantly for each year of increase in age (−3.26 [95%CI −3.37 to −3.15], being female (−5.45 [95%CI −6.26 to −4.64]), being non-Mexican (−5.12 [95%CI −7.57 to −2.66]), and having studied in a private medical school (−3.36 [95%CI −4.3 to −2.42]).

In the 2017 ENARM, there were 6% more applicants and three more medical schools, but the same number of specialties. For most specialties, the number of applicants varied slightly. Table II shows demographics, NoCA by item difficulty, and the ANoCA. Applicants' age and sex distribution was similar across all days. Significant differences were found (according to the 95%CI) in the scores for low-, medium-, and high-difficulty items. The NoCA and ANoCA showed patterns similar to 2016. Days 1 and 2 showed higher scores (easier test forms) and day 5 showed lower scores (more difficult test form). The univariate linear regression showed that ANoCA decreased significantly in similar way to the previous year: for each year of increase in age (−2.99 [95%CI −3.1 to −2.88]), being female (−5.29 [95%CI −4.47 to −6.1]), being non-Mexican (−4.97 [−7.29 to −2.65]), and having studied in a private medical school (−3.29 [95%CI −4.23 to −2.34]).

The identified inter-day differences influenced the number of applicants who achieved the SMS for the specialty for which they were competing. Table III highlights the number and proportion of ARSMS (overall and per day) for the 12 specialties with the greatest number of applicants (data on all specialties available upon request). Grouping days with similar proportions of ARSMS showed that applicants on day

**Table I**

**ENARM 2016 applicants' characteristics nationwide, with overall and per test-day scores. Guadalajara, Mexico 2018**

| | *Overall* | *Day 1* | *Day 2* | *Day 3* | *Day 4* | *Day 5* |
|---|---|---|---|---|---|---|
| Applicants, n (%) | 36 114 | 7 161 (19.8) | 7 571 (21) | 7 415 (20.5) | 7 067 (19.6) | 6 900 (19.1) |
| Female, n (%) | 18 529 (51) | 3 649 (50) | 3 929 (52) | 3 816 (51.5) | 3 659 (52) | 3 476 (50) |
| Mexican physicians*, n (%) | 35 104 (97) | 7 056 (98.5) | 7 494 (99) | 7 349 (99) | 6 975 (99) | 6 230 (90) |
| Age, mean ± SD (min; max.) [95%CI] | 27 ± 3.5 (21; 58) [26.9 to 27] | 27 ± 3 (21; 50) [26.9 to 27] | 26 ± 3 (22; 58) [25.9 to 26] | 27 ± 3 (22; 55) [26.9 to 27] | 27 ± 3 (21; 53) [26.9 to 27] | 29 ± 4 (22; 58) [28.9 to 29] |
| NoCA, mean ± SD (min; max) [95%CI] | | | | | | |
| Low difficulty (0 to 113) | 78.6 ± 11.3 (0; 107) [78.5 to 78.7] | 78.8 ± 10.1 (0; 104) [78.6 to 79] | 82.5 ± 10.3 (31; 107) [82.3 to 82.7] | 81.4 ± 10.5 (0; 106) [81.2 to 81.6] | 81.4 ± 9.4 (26; 107) [81.2 to 81.6] | 68 ± 9.8 (7; 99) [67.8 to 68.2] |
| Medium difficulty (0 to 225) | 136.3 ± 20 (0; 206) [136 to 136.5] | 138.9 ± 21 (0; 196) [138.4 to 139.3] | 140.5 ± 19.6 (31; 206) [140 to 140.9] | 135 ± 21.3 (0; 197) [134.5 to 135.5] | 135.6 ± 19.5 (34; 194) [135.1 to 136] | 131.3 ± 17 (4; 190) [130.9 to 131.7] |
| High difficulty (0 to 112) | 60.6 ± 11 (0; 100) [60.5 to 60.7] | 58.9 ± 10.2 (0; 91) [58.6 to 59.1] | 67.5 ± 10.7 (12; 100) [67.2 to 67.7] | 61.8 ± 9.4 (0; 94) [61.6 to 62] | 58.4 ± 10.6 (16; 94) [58.1 to 58.6] | 55.5 ± 9.7 (1; 93) [55.3 to 55.7] |
| ANoCA (0 to 450), mean ± SD (min; max.) [95%CI] | 275.5 ± 39.3 (0; 410) [275 to 275.9] | 276.6 ± 39 (0; 390) [276.6 to 277.5] | 290.5 ± 38.4 (74; 410) [289.6 to 291.3] | 278.15 ± 38.9 (0; 391) [277.3 to 279] | 275.5 ± 37.1 (83; 381) [274.6 to 276.7] | 254.9 ± 33.9 (12; 371) [254.1 to 255.7] |

* The rest represents foreign medical graduates

SD: standard deviation
CI: confidence interval
NoCA: number of correct answers
ANoCA: absolute number of correct answers
ENARM: *Examen Nacional para Aspirantes a Residencias Médicas*

2 were more likely to achieve their SMS when compared with days 1, 3, 4 (OR 1.55 [95%CI: 1.46 to 1.64]), and 5 (OR 3.8 [95%CI: 3.47 to 4.16]). The inter-day differences also had an effect for the 2017 ENARM. After grouping days with similar proportions of ARSMS, applicants on days 1 and 2 were more likely to reach their SMS when compared with days 3, 4 (OR 1.85 [95%CI 1.75 to 1.94]), and 5 (OR 4.04 [95%CI 3.72 to 4.38]).

The inter-day differences among item difficulty categories and the ANoCA were adjusted for age, gender, selected-specialty, and nationality using multivariable linear regression. The 2016 ENARM showed that days 1, 3, and 4 had similar coefficients (differences in the ANoCA), whereas day 2 showed an advantageous coefficient (e.g., 11.66 higher ANoCA than day 1) and day 5 showed a disadvantageous coefficient (16.18 lower ANoCA than day 1) (table IV). The 2017 ENARM showed days 3 and 4 had similar inter-day coefficients, days 1 and 2 had similar but higher (advantaged) coefficients, and day 5 had a lower coefficient (disadvantaged). Furthermore, the inter-day coefficients within the item difficulty categories also significantly differed (table IV).

Since the significant inter-day variability described above precludes the assumption of equity for the ENARM in 2016 and 2017, we estimated its impact as the number of applicants that could have been dubiously classified as not passing (secondary end-point). Table V shows that for 2016, *i* (day 2) presupposed a disadvantage for 2 070 (9.5%) applicants of *R* (days 1, 3, and 4) and for 1 495 (21.7%) applicants of *j* (day 5). In consequence, 3 565 (12.49%) applicants were affected by the lack of equity in the test forms. In other words, this last figure represents 44.64% of the 7 986 applicants that received a pass certificate that year.[23-25] Table V shows the estimate of affected applicants in 2017, where *i* (days 1 and 2) presupposed a disadvantage for 1 644 (13.6%) applicants of *R* (days 3 and 4) and 1 511 (19.79%) applicants of *j* (day 5), indicating 3 155 (13.64%) applicants were affected. This represents 37.2% of the 8 480 applicants that received a pass certificate in 2017.[26-28]

**Table II**

**ENARM 2017 applicants' characteristics nationwide, with overall and per test-day scores. Guadalajara, Mexico 2018**

| | *Overall* | *Day 1* | *Day 2* | *Day 3* | *Day 4* | *Day 5* |
|---|---|---|---|---|---|---|
| Applicants, n (%) | 38 380 | 7 391 (19.3) | 7 862 (20.5) | 7 803 (20.3) | 7 687 (20) | 7,637 (19.9) |
| Female, n (%) | 19 687 (51) | 3 792 (51) | 4 071 (52) | 4 005 (51) | 3 944 (51) | 3,875 (51) |
| Mexican physicians*, n (%) | 37 147 (97) | 7 078 (96) | 7 639 (97) | 7 618 (98) | 7 425 (97) | 7,387 (97) |
| Age, mean ± SD (min; max.) [95%CI] | 27 ± 3.5<br>(19; 63)<br>[27 to 27] | 27 ± 3.5<br>(22; 63)<br>[27 to 27] | 26.5 ± 3<br>(21; 63)<br>[26 .4 to 26.5] | 27 ± 3<br>(21; 59)<br>[26.8 to 26.9] | 27 ± 4<br>(21; 62)<br>[27.1 to 27.3] | 28 ± 4<br>(19; 59)<br>[27.7 to 27.8] |
| NoCA, mean ± SD (min; max) [95%CI] | | | | | | |
| Low difficulty (0 to 113) | 80.2 ± 10.75<br>(0; 111)<br>[80 to 80.3] | 80.4 ± 10.35<br>(29; 105)<br>[80.1 to 80.6] | 83.1 ± 10.8<br>(2; 111)<br>[82.8 to 83.3] | 80.9 ± 10.8<br>(10; 109)<br>[80.6 to 81.1] | 81 ± 11<br>(0; 108)<br>[80.7 to 81.2] | 75.4 ± 9<br>(24; 100)<br>[75.1 to 76.6] |
| Medium difficulty (0 to 225) | 141.8 ± 22<br>(0; 205)<br>[141.5 to 142] | 144.8 ± 23<br>(59; 202)<br>[144.2 to 145.3] | 151.4 ± 21.3<br>(3; 205)<br>[150.9 to 151.8] | 139.7 ± 19.5<br>(9; 190)<br>[139.2 to 140.1] | 142.2 ± 20.6<br>(0; 204)<br>[141.7 to 142.6] | 130.8 ± 20<br>(27; 196)<br>[130.3 to 131.2] |
| High difficulty (0 to 112) | 59.9 ± 11<br>(0; 96)<br>[59.8 to 60] | 65.8 ± 10.7<br>(24; 96)<br>[65.5 to 66] | 60.9 ± 10.2<br>(0; 96)<br>[60.6 to 61.1] | 60.2 ± 9.4<br>(1; 89)<br>[59.9 to 60.4] | 55.9 ± 10<br>(0; 93)<br>[55.6 to 56.1] | 56.9 ± 10.7<br>(12; 94)<br>[56.6 to 57.4] |
| ANoCA (0 to 450), mean ± SD (min; max.) [95%CI] | 281.8 ± 41<br>(0; 400)<br>[281.3 to 282.2] | 291 ± 42.1<br>(120; 399)<br>[290 to 292] | 295.4 ± 40.3<br>(5; 400)<br>[294.5 to 296.2] | 280.8 ± 37.5<br>(20; 379)<br>[279.9 to 281.6] | 279 ± 39.4<br>(0; 395)<br>[278.1 to 279.8] | 263.1 ± 37.2<br>(69; 379)<br>[262.2 to 263.9] |

* The rest represents foreign medical graduates

SD: standard deviation
CI: confidence interval
NoCA: number of correct answers
ANoCA: absolute number of correct answers
ENARM: *Examen Nacional para Aspirantes a Residencias Médicas*

A sensitivity analysis for the multivariable linear regression model that also adjusted to the status of 'public/private' medical school in addition to the other variables was performed (data not shown, available upon request), but excluded approximately 1 200 and 1 400 applicants (2016 and 2017, respectively) from foreign medical schools for which the "public/private" status could not be identified. It yielded similar statistically significant results and identical grouping of the days, but a higher estimate of applicants affected by inequity. Considering these findings, the authors decided to opt for the more discrete estimate.

Additional data on the ANoCA for each specialty for both years is available upon request.

## Discussion

The effort required for an MD to meet the selection requirements is vast. The selection process is based on different constructs. A construct of particular importance is fairness. This ecological study evaluated the equity (i.e. 'fairness')[16] of the ENARM over two consecutive years.

Our first important finding was that there were significant inter-day differences in scores for items classified as low-, medium-, and high-difficulty, and thus the ANoCA. These differences, although significant, may seem small. However, a difference of just one correct item has an impact on an applicant's pass or no-pass status. Consequently, the variable 'day on which an applicant presented the test' in both years gave applicants an advantage or disadvantage. For 2016, day 2 applicants were 55% more likely to reach their SMS than applicants on days 1, 3, and 4, and almost four times more likely than on day 5. For 2017, day 1 and 2 applicants were 85% more likely to reach their SMS than applicants on days 3 and 4, and four times more likely than on day 5. Therefore, the stated premise of equity was not sustained.

The methods used to create the different test forms may explain these inter-day differences. The agency that creates the ENARM does not perform item analysis with established psychometric theories, such as classical test theory or item response theory. Instead, experts classify an item's difficulty into three broad categories defined

**Table III**

**Applicants that reached a passing score for the twelve specialties with the higher number of applicants in ENARMs 2016 and 2017. Guadalajara, Mexico 2018**

| | *ENARM 2016, Total number of applicants; ARSMS number (%)* | | | | | | *ENARM 2017, Total number of applicants; ARSMS number (%)* | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| *Specialty* | *Overall* | *Day 1* | *Day 2* | *Day 3* | *Day 4* | *Day 5* | *Overall* | *Day 1* | *Day 2* | *Day 3* | *Day 4* | *Day 5* |
| 2 | 3 683; 832 (22.6) | 735; 150 (20.4) | 708; 276 (39) | 746; 173 (23.2) | 719; 167 (23.2) | 775; 66 (8.5) | 3 981; 972 (24.4) | 802; 279 (34.8) | 773; 243 (31.4) | 760; 177 (23.3) | 837; 180 (21.5) | 809; 93 (11.5) |
| 5 | 5 394; 713 (13.2) | 1 061; 129 (12.2) | 1 313; 293 (22.3) | 1 151; 170 (14.8) | 993; 105 (10.6) | 876; 16 (1.8) | 5 794; 769 (13.3) | 1 075; 256 (23.8) | 1 347; 280 (20.8) | 1 219; 100 (8.2) | 1 077; 99 (9.2) | 1 076; 34 (3.2) |
| 9 | 4 263; 710 (16.7) | 861; 166 (19.3) | 946; 265 (28) | 956; 143 (15) | 826; 108 (13.1) | 674; 28 (4.2) | 4 746; 712 (15) | 881; 195 (22.1) | 1 022; 249 (24.4) | 985; 111 (11.3) | 965; 121 (12.5) | 893; 36 (4) |
| 10 | 1 508; 349 (23.1) | 270; 74 (27.4) | 294; 101 (34.4) | 282; 62 (22) | 298; 75 (25.2) | 364; 37 (10.2) | 1 503; 349 (23.2) | 292; 91 (31.2) | 290; 96 (33.1) | 290; 63 (21.7) | 322; 72 (22.4) | 309; 27 (8.7) |
| 13 | 1 603; 626 (39.1) | 316; 128 (40.5) | 231; 118 (51.1) | 308; 146 (47.4) | 360; 148 (41.1) | 388; 86 (22.2) | 2 151; 649 (30.2) | 417; 163 (39.1) | 351; 140 (39.9) | 461; 149 (32.3) | 450; 140 (31.1) | 472; 57 (12.1) |
| 15 | 3 647; 1 653 (45.3) | 661; 320 (48.4) | 449; 242 (53.9) | 627; 315 (50.2) | 767; 405 (52.8) | 1 143; 371 (32.5) | 3 295; 1 712 (52) | 634; 389 (61.4) | 463; 269 (58.1) | 637; 353 (55.4) | 741; 409 (55.2) | 820; 292 (35.6) |
| 16 | 4 530; 1 078 (23.8) | 905; 222 (24.5) | 1 072; 387 (36.1) | 963; 244 (25.3) | 892; 185 (20.7) | 698; 40 (5.7) | 4 835; 1 146 (23.7) | 941; 338 (35.9) | 1 077; 381 (35.4) | 967; 185 (19.1) | 930; 159 (17.1) | 920; 83 (9) |
| 21 | 902; 159 (17.6) | 181; 31 (17.1) | 255; 71 (27.8) | 207; 32 (15.5) | 148; 22 (14.9) | 111; 3 (2.7) | 835; 190 (22.8) | 173; 67 (38.7) | 224; 76 (33.9) | 175; 17 (9.7) | 138; 20 (14.5) | 125; 10 (8) |
| 22 | 703; 91 (12.9) | 144; 18 (12.5) | 182; 32 (17.6) | 141; 22 (15.6) | 134; 17 (12.7) | 102; 2 (2) | 688; 93 (13.5) | 130; 33 (25.4) | 196; 40 (20.4) | 146; 9 (6.2) | 125; 9 (7.2) | 91; 2 (2.2) |
| 24 | 4 115; 812 (19.7) | 841; 166 (19.7) | 962; 290 (30.1) | 885; 185 (20.9) | 789; 146 (18.5) | 638; 25 (3.9) | 3 979; 816 (20.5) | 751; 231 (30.8) | 870; 270 (31) | 846; 137 (16.2) | 773; 127 (16.4) | 739; 51 (6.9) |
| 25 | 727; 162 (22.3) | 143; 38 (26.6) | 134; 47 (35.1) | 157; 34 (21.7) | 140; 30 (21.4) | 153; 13 (8.5) | 813; 176 (21.6) | 180; 46 (25.6) | 156; 54 (34.6) | 159; 32 (20.1) | 150; 32 (21.3) | 168; 12 (7.1) |
| 27 | 2 202; 338 (15.3) | 436; 65 (14.9) | 508; 117 (23) | 465; 72 (15.5) | 453; 70 (15.5) | 340; 14 (4.1) | 2 573; 456 (17.7) | 489; 140 (28.60) | 517; 142 (27.5) | 554; 76 (13.7) | 503; 71 (14.1) | 510; 27 (5.3) |
| Total* | 36 114; 8 128 (22.51) | 7 161; 1 654 (23.10) | 7 571; 2 398 (31.67) | 7 415; 1 732 (23.36) | 7 067; 1 594 (22.56) | 6 900; 750 (10.87) | 38 377;‡ 8 679 (22.62) | 7 391; 2 393 (32.38) | 7 861; 2 414 (30.71) | 7 803; 1 523 (19.52) | 7 687; 1 568 (20.4) | 7 635; 781 (10.23) |

Specialty codes are as follows: 2= Anaesthesiology; 5= General surgery; 9= Gynaecology-obstetrics; 10= Diagnostic and therapeutic radiology; 13= Emergency medicine; 15= Family medicine; 16= Internal medicine; 21= Ophthalmology; 22= Otorhinolaryngology, and head and neck surgery; 24= Paediatrics; 25= Psychiatry; 27= Traumatology and orthopaedics

* Includes all specialties

‡ Three applicants that did not choose specialty were excluded

ARSMS: Applicants that reached the minimum score of their selected specialty
ENARM: *Examen Nacional para Aspirantes a Residencias Médicas*

**Table IV**

**Multivariable linear regression of ENARM 2016 and ENARM 2017 for inter-day differences on the number of correct answers per ascribed item difficulty categories, and on the overall number of correct answers. Guadalajara, Mexico 2018**

| | *ENARM 2016 (n=36 114)* | | | | *ENARM 2017 (n=38 377*)* | | | |
|---|---|---|---|---|---|---|---|---|
| | *Ascribed Difficulty of the Items* | | | *Overall correct items* | *Ascribed Difficulty of the Items* | | | *Overall correct items* |
| | *Low ($R^2$=0.2867)* | *Medium ($R^2$=0.132)* | *High ($R^2$=0.2284)* | *($R^2$=0.1913)* | *Low ($R^2$=0.1493)* | *Medium ($R^2$=0.2024)* | *High ($R^2$=0.1976)* | *($R^2$=0.1889)* |
| Day 2‡ | 3.22§ (2.91 to 3.53) | 0.37 (-0.22 to 0.98) | 8.06§ (7.74 to 8.37) | 11.66§ (10.51 to 12.8) | 2.02§ (1.71 to 2.34) | 5.14§ (4.51 to 5.76) | -5.54§ (-5.85 to -5.23) | 1.62# (0.45 to 2.8) |
| Day 3‡ | 2.5§ (2.19 to 2.82) | -4.14§ (-4.75 to -3.53) | 2.85§ (2.54 to 3.17) | 1.22& (0.07 to 2.37) | 0.32& (0.01 to 0.64) | -5.52§ (-6.14 to -4.89) | -5.77§ (-6.08 to -5.47) | -10.97§ (-12.14 to -9.8) |
| Day 4‡ | 2.94§ (2.62 to 3.25) | -2.77§ (-3.39 to -2.16) | -0.16 (-0.48 to 0.14) | 0 (-1.16 to 1.15) | 0.67§ (0.36 to 0.99) | -2.39§ (-3.02 to -1.77) | -9.82§ (-10.13 to -9.51) | -11.54§ (-12.72 to -10.36) |
| Day 5‡ | -9.39§ (-9.72 to -9.06) | -4.68§ (-5.33 to -4.04) | -2.09§ (-2.42 to -1.76) | -16.18§ (-17.39 to -14.96) | -4.54§ (-4.86 to -4.22) | -13§ (-13.63 to -12.37) | -8.39§ (-8.7 to -8.08) | -25.95§ (-27.13 to -24.77) |

The results shown are adjusted for age in years (continuous), sex (male, female), nationality (Mexican, non-Mexican), and selected specialty (categorical, 1 to 27). Values are reported as β coefficient (95% confidence interval)
$R^2$: Adjusted R-squared

* Three applicants that did not choose specialty were excluded
‡ Day 1 for each year is the reference group
§ $p< 0.001$
# $p< 0.01$
& $p< 0.05$

ENARM: *Examen Nacional para Aspirantes a Residencias Médicas*

*a priori*, without performing field testing on the target population, whereas item discrimination is not calculated (data on applicants' answers are erased after the test).[17,29] The only consideration in assembling different test forms is including an exclusive set of items for each test form that accounts for the same number of items per area of knowledge and includes 25% low-difficulty, 50% medium-difficulty, and 25% high-difficulty items. Consequently, the problem is in the definition of difficulty. For example, the low-difficulty category has in itself a range of difficulty values; one test form may have low-difficulty items that are, on average, located at the top of this range, whereas another test form may contain items that are located, on average, at the bottom of the range. Furthermore, for a test where the score is computed as the sum of *N* dichotomously scored items, such as the ENARM, the test's mean score (i.e., average ANoCA) directly relates to the average difficulty of item scores.[30]

If research-supported methods in educational assessment already in use had been used for the ENARM, the attribution of equity would not have been compromised. Best practice involves performing item tryout and item analysis to assess item difficulty and discrimination, with the purpose (among others) of calibrating items to design alternate test forms as if they were on the same test score scale.[31] Standard 5.12 of the Educational Standards for Educational and Psychological Testing[12] states that "a clear rationale and supporting evidence should be provided for any claim that scale score earned on alternate forms of a test may be used interchangeably." However, in the ENARM, although 'equivalence among different test forms' is stated, no supporting evidence is provided and no test-equating methods are used.[17]

Our second finding concerns estimating the impact on applicants due to differences according to the day they took the exam. Since the ENARM is a selection exam for specialties and each specialty has a predetermined number of positions, it is expected that applicants with high scores will not be impacted. The expected impact would be on those scoring closely below and above each SMS. Our estimation showed that around four out of ten applicants that received a pass certificate might have not received a pass if the test forms were equitable. The impact might go beyond these estimates; that is, on applicants' professional development, quality of healthcare provided by the training institutions, and ultimately, population health.

**Table V**
**Overall and per specialty estimation of the applicants affected by the inequity of ENARMs 2016 and 2017. Guadalajara, Mexico 2018**

| Specialty | ENARM 2016 | | | | ENARM 2017 | | | |
|---|---|---|---|---|---|---|---|---|
| | Referent group* affected by Day 2 | | Day 5 affected by Day 2 | | Referent group‡ affected by Days 1 & 2 | | Day 5 affected by Days 1 & 2 | |
| | Applicants | Affected, n (%) | Applicants | Affected, n (%) | Applicants | Affected, n (%) | Applicants | Affected, n (%) |
| 1 | 179 | 26 (14.5) | 66 | 13 (19.7) | 78 | 6 (7.7) | 52 | 14 (26.9) |
| 2 | 2 200 | 208 (9.5) | 775 | 194 (25) | 1 597 | 189 (11.8) | 809 | 163 (20.1) |
| 3 | 58 | 3 (5.2) | 27 | 3 (11.1) | 38 | 2 (5.3) | 17 | 7 (41.2) |
| 4 | 25 | 2 (8) | 15 | 1 (6.7) | 8 | 1 (12.5) | 3 | 0 (0) |
| 5 | 3 205 | 256 (8) | 876 | 76 (8.7) | 2 296 | 181 (7.9) | 1 076 | 126 (11.7) |
| 6 | 99 | 15 (15.2) | 47 | 7 (14.9) | 74 | 14 (18.9) | 45 | 6 (13.3) |
| 7 | 69 | 13 (18.8) | 28 | 4 (14.3) | 41 | 4 (9.8) | 26 | 9 (34.6) |
| 8 | 338 | 25 (7.4) | 98 | 17 (17.3) | 272 | 23 (8.5) | 136 | 19 (14) |
| 9 | 2 643 | 222 (8.4) | 674 | 108 (16) | 1 950 | 172 (8.8) | 893 | 138 (15.5) |
| 10 | 850 | 103 (12.1) | 364 | 100 (27.5) | 612 | 79 (12.9) | 309 | 73 (23.6) |
| 11 | 45 | 2 (4.4) | 16 | 1 (6.3) | 38 | 2 (5.3) | 18 | 0 (0) |
| 12 | 326 | 23 (7.1) | 107 | 13 (12.1) | 194 | 29 (14.9) | 103 | 18 (17.5) |
| 13 | 984 | 117 (11.9) | 388 | 145 (37.4) | 911 | 129 (14.2) | 472 | 155 (32.8) |
| 14 | 155 | 16 (10.3) | 65 | 18 (27.7) | 128 | 14 (10.9) | 62 | 12 (19.4) |
| 15 | 2 055 | 249 (12.1) | 1 143 | 407 (35.6) | 1 378 | 187 (13.6) | 820 | 299 (36.5) |
| 16 | 2 760 | 275 (10) | 698 | 122 (17.5) | 1 897 | 194 (10.2) | 920 | 180 (19.6) |
| 17 | 66 | 1 (1.5) | 32 | 1 (3.1) | 31 | 3 (9.7) | 28 | 6 (21.4) |
| 18 | 35 | 6 (17.1) | 18 | 5 (27.8) | 20 | 2 (10) | 15 | 3 (20) |
| 19 | 38 | 5 (13.2) | 24 | 3 (12.5) | 11 | 0 (0) | 18 | 4 (22.2) |
| 20 | 127 | 7 (5.5) | 49 | 11 (22.4) | 151 | 15 (9.9) | 64 | 12 (18.8) |
| 21 | 536 | 41 (7.6) | 111 | 13 (11.7) | 313 | 22 (7) | 125 | 18 (14.4) |
| 22 | 419 | 29 (6.9) | 102 | 5 (4.9) | 271 | 19 (7) | 91 | 11 (12.1) |
| 23 | 26 | 2 (7.7) | 12 | 3 (25) | 28 | 1 (3.6) | 25 | 7 (28) |
| 24 | 2 515 | 272 (10.8) | 638 | 139 (21.8) | 1 619 | 206 (12.7) | 739 | 125 (16.9) |
| 25 | 440 | 40 (9.1) | 153 | 26 (17) | 309 | 34 (11) | 168 | 29 (17.3) |
| 26 | 96 | 13 (13.5) | 34 | 14 (41.2) | 168 | 9 (5.4) | 91 | 12 (13.2) |
| 27 | 1 354 | 99 (7.3) | 340 | 46 (13.5) | 1 057 | 107 (10.1) | 510 | 65 (12.7) |
| Total | 21 643 | 2 070 (9.56) | 6 900 | 1 495 (21.67) | 15 490 | 1 644 (10.61) | 7 635 | 1 511 (19.79) |

* Referent group for 2016 = Grouping of days 1, 3 and 4
‡ Referent group for 2017 = Grouping of days 3 and 4

ENARM: *Examen Nacional para Aspirantes a Residencias Médicas*

Regarding the variables associated with obtaining a lower ANoCA, we can only speculate. We have no explanation as to why female applicants obtained lower scores, but older applicants may most likely represent MDs that graduated years ago and may not be as updated as their recently-graduated counterparts. Regarding the private status of a medical school, a possible explanation is that acceptance into some private schools may not be as competitive and selective as that of public schools.

The strengths of this study that lend weight to the conclusions are that the official ENARM 2016 and 2017 databases were analyzed and the attribution of equity was assessed using different approaches. Nonetheless, several limitations should be noted. First, this was an ecological study with inherent limitations of that design.

Second, differences in the average ANoCA among test forms do not necessarily prove a lack of equity, but do suggest that further review is needed.[32] The ENARM is a norm-referenced test by which applicants are selected based on a quota. The inter-day differences in the proportions of ARSMS strongly suggest that the different test forms jeopardized the possibility that a better (or at least equally) prepared number of applicants entered specialty training. Third, the official databases did not include data on the correct/incorrect status for each answered item, which precluded calculating the standard error of measurement of each test form to assess its reliability.[33] In addition, we could not use the method proposed by Raykov and Marcoulides[34] to assess whether the different ENARM test forms were consistent with the model of parallel tests in the classical test theory framework, as the way applicants' scores are compared fits this model (i.e., no estimation of random error of measurement and no use of test-equating methods). Fourth, other relevant variables that might have influenced test scores were not available for analysis, such as applicants' number of test attempts and the grade point average obtained during medical school. Fifth, it is assumed that applicants distributed in a more or less randomized manner among test days. This assumption is supported by the registration method. Test-registration occurs on a first-come, first-served basis; most applicants try to register immediately, which leads the web-server to saturate, delaying successful registration for applicants to a random fashion.

In many science-related areas, there is a concern that a gap between research and practice exists, even with the growing volume and quality of evidence along with the technological and organizational improvements in information management and synthesis.[35] The field of educational assessment seems to be no exception, as attested by the example of the ENARM. Furthermore, although this study aimed to assess equity among ENARM test forms for 2016 and 2017, it has to be considered that the exam methods have been the same since at least the 2010 ENARM.[18] Changes newly introduced for ENARM 2018 are likely to decrease the lack of equity by using the same test form for all applicants competing for the same specialty.[36] This may be the best immediate approach to improve the equity of the test. However, improvements in many other aspects of the ENARM are needed in the long term. Examples include: a) increasing transparency by creating a publicly-available test's development technical report to provide evidence of validity, reliability and fairness; b) improving item-writing, as applicants' opinions suggest that poorly-written items introduce construct-irrelevant variation (e.g., those written in English); c) improving item development by using item tryout and item analysis; d) assessing the effect on test scores of the nine different item arrangements used for each test form,[18] as the location of items on a test form can affect item statistics, particularly in testing programs that report immediately or shortly after test administration;[16] e) addressing sources of systematic errors (applicants' opinions suggest that a 2017's test form included an item that required applicants to identify a clinical sign in an image, but the image was veiled); and f) estimating the standard error of measurement, which requires the non-deletion of answered items after the test.

Evidence points toward the need to use a robust, evidence-based test-development process on the ENARM if the test's stated objectives and characteristics are to be achieved. Estimates suggest that average items for high-stakes tests may be valued around 300 USD, with some items valued above 1 000 USD (e.g., those that undergo extensive statistical analysis or that need complex development processes such as simulations); overall, an item bank with a few thousand items may be valued around 1 000 000 USD.[37] The revenues obtained from the ENARM registration fees are around 110 000 000 MXP per year (approximately 6 000 000 USD). With these funds, it would be possible to redesign the test to improve its quality.

In summary, the ENARM has a paramount and multidimensional impact on Mexico's healthcare system and on applicants wishing to undergo specialty training. However, the analysis of the official ENARM databases does not support the attribution of equity. For this reason, it is necessary to redesign the test using evidence-based test-development processes that support the fairness, validity, and reliability of the ENARM.

*Declaration of conflict of interests*. The authors declare that they have no conflict of interests.

## References

1. Makary MA, Daniel M. Medical error-the third leading cause of death in the US. BMJ. 2016;353:i2139. https://doi.org/10.1136/bmj.i2139
2. Roberts C, Khanna P, Rigby L, Bartle E, Llewellyn A, Gustavs J, *et al.* Utility of selection methods for specialist medical training: a BEME (best evidence medical education) systematic review: BEME guide no. 45. Med Teach. 2018;40(1):3-19. https://doi.org/10.1080/0142159X.2017.1367375
3. mrcpuk.org [internet] London: Membership of the Royal Colleges of Physicians of the United Kingdom, 2018 [cited 2018 Jun 4] [cited 2018 Jun 4]. Available from: https://www.mrcpuk.org/
4. Centre National de Gestion [internet]. Paris: CNG, 2018. Épreuves Classantes Nationales (ECN) [cited 2018 Jun 4]. Available from: https://www.cng.sante.fr/concours-examens/epreuves-classantes-nationales-ecn
5. Comisión Interinstitucional para la Formación de Recursos Humanos para la Salud [internet]. Mexico City: CIFRHS, 2018. Examen Nacional para Aspirantes a Residencias Médicas [cited 2018 Jun 4]. Available from: http://cifrhs.salud.gob.mx/

6. Norcini J, Friedman Ben-David N. Concepts in assessment. In: Dent JA, Harden RM (eds). A practical guide for medical teachers. 4th ed. Churchill Livingstone, 2013:285-91.
7. Comité de Posgrado y Educación Continua. XL ENARM: Metodología proceso de selección. Mexico City: CIFRHS, 2016 [cited 2018 Jun 4]. Available from: http://cifrhs.salud.gob.mx/site1/enarm/docs/2016/E40_met_proceso_seleccion_2016.pdf
8. Comité de Posgrado y Educación Continua. XLI ENARM: Metodología proceso de selección. Mexico City: CIFRHS, 2017 [cited 2018 Jun 4]. Available from: http://cifrhs.salud.gob.mx/site1/enarm/docs/2017/E41_met_proceso_seleccion_2017.pdf
9. Comisión Interinstitucional para la Formación de Recursos Humanos para la Salud. Convocatoria XL Examen Nacional para Aspirantes a Residencias Médicas. Mexico City: CIFRHS, 2016 [cited 2018 Jun 4]. Available from: http://cifrhs.salud.gob.mx/site1/enarm/docs/2016/E40_convo_2016.pdf
10. Comisión Interinstitucional para la Formación de Recursos Humanos para la Salud. Convocatoria XLI Examen Nacional para Aspirantes a Residencias Médicas. Mexico City: CIFRHS, 2017 [cited 2018 Jun 4]. Available from: http://cifrhs.salud.gob.mx/site1/enarm/docs/2017/E41_convo_2017.pdf
11. Barajas-Ochoa A, Ramos-Remus C. Equidad, validez y confiabilidad del Examen Nacional para Aspirantes a Residencias Medicas (ENARM): oportunidades para mejorar. Salud Publica Mex. 2017;36:501-2. https://doi.org/10.21149/8769
12. American Educational Research Association, American Psychological Association, National Council on Measurement in Education. Standards for educational and psychological testing. Washington, DC: American Educational Research Association, 2014.
13. Lane S, Raymond MR, Haladyna TM (eds). Handbook of test development. 2nd ed. New York: Routledge, 2016.
14. Bennett RE, von Davier M. Advancing human assessment: The methodological, psychological and policy contributions of ETS [eBook]. Springer, 2017. https://doi.org/10.1007/978-3-319-58689-2
15. Schuwirt LWT, van der Vleuten CPM. General overview of the theories used in assessment: AMEE Guide No. 57. Med Teach. 2011;33(10):783-97. https://doi.org/10.3109/0142159X.2011.611022
16. Lane S, Raymond MR, Haladyna TM, Downing SM. Test development process. In: Lane S, Raymond MR, Haladyna TM (eds). Handbook of test development. 2nd ed. New York: Routledge, 2016:3-18.
17. Dirección General de Calidad y Educación en Salud. Departamento de Transparencia de la Secretaría de Salud: Reply to request No. 0001200055817 on public information regarding the development of the ENARM. Mexico City: DGCES, 2017. Available from: https://drive.google.com/file/d/0B4Qe4F_sYmX0bHM0dHlpWFBObnM/view
18. Comisión Interinstitucional para la Formación de Recursos Humanos para la Salud. ENARM: Características y evolución a su formato electrónico. Mexico City: CIFRHS, 2012 [cited 2018 Jun 4]. Available from: http://cifrhs.salud.gob.mx/descargas/pdf/enarm_caracteristicas_evolucion.pdf
19. Comisión Interinstitucional para la Formación de Recursos Humanos para la Salud. XL Examen Nacional para Aspirantes a Residencias Médicas: Puntajes máximos y mínimos por especialidad. Mexico City: CIFRHS, 2016 [cited 2018 Jun 4]. Available from: http://cifrhs.salud.gob.mx/site1/enarm/docs/2016/E40_puntajes_max_min_2016.pdf
20. Comisión Interinstitucional para la Formación de Recursos Humanos para la Salud. XLI Examen Nacional para Aspirantes a Residencias Médicas: Puntajes máximos y mínimos por especialidad. Mexico City: CIFRHS, 2017 [cited 2018 Jun 4]. Available from: http://cifrhs.salud.gob.mx/site1/enarm/docs/2017/E41_puntajes_max_min_2017.pdf
21. Ferguson CJ. An effect size primer: A guide for clinicians and researchers. Prof Psychol Res Pract. 2009:40:532-8. https://doi.org/10.1037/a0015808
22. Sullivan GM, Feinn R. Using effect size—or why the P value is not enough. J Grad Med Educ. 2012;4:279-82. https://doi.org/10.4300/JGME-D-12-00156.1
23. Comisión Interinstitucional para la Formación de Recursos Humanos para la Salud. XL Examen Nacional para Aspirantes a Residencias Médicas: Plazas para médicos seleccionados. Mexico City: CIFRHS, 2016 [cited 2018 Jun 4]. Available from: http://cifrhs.salud.gob.mx/site1/enarm/docs/2016/E40_plazas_mex_lugares_ext_2016.pdf
24. Comisión Interinstitucional para la Formación de Recursos Humanos para la Salud. XL Examen Nacional para Aspirantes a Residencias Médicas: Folios de médicos seleccionados – Médicos con categoría mexicana. Mexico City: CIFRHS, 2016 [cited 2018 Jun 4]. Available from: http://cifrhs.salud.gob.mx/site1/enarm/docs/2016/E40_folio_sel_mex_2016.pdf
25. Comisión Interinstitucional para la Formación de Recursos Humanos para la Salud. XL Examen Nacional para Aspirantes a Residencias Médicas: Folios de médicos seleccionados – Médicos con categoría extranjera. Mexico City: CIFRHS, 2016 [cited 2018 Jun 4]. Available from: http://cifrhs.salud.gob.mx/site1/enarm/docs/2016/E40_folio_sel_ext_2016.pdf
26. Comisión Interinstitucional para la Formación de Recursos Humanos para la Salud. XLI Examen Nacional para Aspirantes a Residencias Médicas: Plazas para médicos seleccionados. Mexico City: CIFRHS, 2017 [cited 2018 Jun 4]. Available from: http://cifrhs.salud.gob.mx/site1/enarm/docs/2017/E41_plazas_mex_lugares_ext_2017.pdf
27. Comisión Interinstitucional para la Formación de Recursos Humanos para la Salud. XLI Examen Nacional para Aspirantes a Residencias Médicas: Folios de médicos seleccionados – Médicos con categoría mexicana. Mexico City: CIFRHS, 2017 [cited 2018 Jun 4]. Available from: http://cifrhs.salud.gob.mx/site1/enarm/docs/2017/E41_folio_sel_mex_2017.pdf
28. Comisión Interinstitucional para la Formación de Recursos Humanos para la Salud. XLI Examen Nacional para Aspirantes a Residencias Médicas: Folios de médicos seleccionados – Médicos con categoría extranjera. Mexico City: CIFRHS, 2017 [cited 2018 Jun 4]. Available from: http://cifrhs.salud.gob.mx/site1/enarm/docs/2017/E41_folio_sel_ext_2017.pdf
29. Dirección General de Calidad y Educación en Salud. Departamento de Transparencia de la Secretaría de Salud: Reply to request No. 0001200357817 on the petition of the official databases of the 2016 and 2017 ENARM. Mexico City: DGCES, 2017. Available from: https://drive.google.com/open?id=1SkzaK20MXXShKTVzK42AIc4jbLWfB8kg
30. Moses T. A review of developments and applications in item analysis. In: Bennett RE, von Davier M (eds). Advancing human assessment: The methodological, psychological and policy contributions of ETS [eBook]. Springer Open, 2017:19-46. https://doi.org/10.1007/978-3-319-58689-2
31. Haladyna TM. Item analysis for selected-response test items. In: Lane S, Raymond MR, Haladyna TM (eds). Handbook of test development. 2nd ed. New York: Routledge, 2016:392-409.
32. Zieky MJ. Developing fair tests. In: Lane S, Raymond MR, Haladyna TM (eds). Handbook of test development. 2nd ed. New York: Routledge, 2016:81-99.
33. Tighe J, McManus IC, Dewhurst NG, Chis L, Mucklow J. The standard error of measurement is a more appropriate measure of quality for postgraduate medical assessments that is reliability: an analysis of MRCP(UK) examinations. BMC Med Educ. 2010;10:40. https://doi.org/10.1186/1472-6920-10-40
34. Raykov T, Marcoulides GA. Introduction to psychometric theory. Classical test theory. New York, NY: Routledge, 2011:115-36.
35. Green LW, Ottoson JM, Garcia C, Hiatt RA. Diffusion theory and knowledge dissemination, utilization, and integration in public health. Annu Rev Public Health. 2009;30:151-74. https://doi.org/10.1146/annurev.publhealth.031308.100049
36. Comisión Interinstitucional para la Formación de Recursos Humanos para la Salud. Convocatoria XLII Examen Nacional para Aspirantes a Residencias Médicas. Mexico City: CIFRHS, 2018 [cited 2018 Jun 4]. Available from: http://cifrhs.salud.gob.mx/site1/enarm/docs/2018/E42_convo_2018.pdf
37. Timothy JM. Web-based item development and banking. In: Lane S, Raymond MR, Haladyna TM (eds). Handbook of test development. 2nd ed. New York: Routledge, 2016: 241-56.