



Μέθοδοι Στατιστικής Επεξεργασίας των βαθμολογιών των υποψηφίων που έλαβαν μέρος στις Γραπτές Εξετάσεις 2019 για Εγγραφή και Κατάταξη στους Πίνακες Διορισίμων

1. Εισαγωγή

Το 2019 πραγματοποιήθηκαν Εξετάσεις για Εγγραφή και Κατάταξη στους Πίνακες Διορισίμων για 41 ειδικότητες. Οι 20 από τις ειδικότητες που εξετάστηκαν το 2019 είχαν, επίσης, εξεταστεί το 2017, ενώ για τις υπόλοιπες 21 ειδικότητες το 2019 ήταν η πρώτη περίοδος εξετάσεων.

Σύμφωνα με τις πρόνοιες της σχετικής νομοθεσίας, οι βαθμολογίες των υποψηφίων έτυχαν στατιστικής επεξεργασίας, η οποία περιγράφεται πιο κάτω:

2. Περιγραφή της μεθοδολογίας για τη συγκρισιμότητα των βαθμολογιών των εξετάσεων των περιόδων 2017 και 2019

Προκειμένου να διασφαλιστεί η συγκρισιμότητα των βαθμολογιών των εξεταστικών περιόδων 2017 και 2019, για κάθε ειδικότητα εφαρμόστηκε στατιστική επεξεργασία των αντίστοιχων βαθμολογιών σε δύο στάδια. Στο πρώτο στάδιο, για κάθε ειδικότητα υπολογίστηκε ο βαθμός δυσκολίας της εξέτασης του 2019 σε σχέση με τον βαθμό δυσκολίας της αντίστοιχης εξέτασης του 2017 που αποτελεί και το σημείο αναφοράς. Αυτό το στάδιο ήταν απαραίτητο, προκειμένου να εξευρευθεί η βαθμολογία της εξέτασης του 2019 που αντιστοιχεί με το 50% της βαθμολογίας του 2017, που αποτελούσε και το βασικό κριτήριο επιτυχίας/αποτυχίας στην εξέταση. Στο δεύτερο στάδιο, οι βαθμολογίες μετασχηματίστηκαν σύμφωνα με τη μέθοδο που χρησιμοποιήθηκε το 2017, προκειμένου ο αριθμός μορίων των υποψηφίων του 2017 και του 2019 να είναι άμεσα συγκρίσιμος.

Στάδιο 1: Για κάθε ειδικότητα υπολογίστηκε η διαφορά του βαθμού δυσκολίας των εξετάσεων των εξεταστικών περιόδων 2017 και 2019. Για κάθε ειδικότητα υπολογίστηκε ένας δείκτης D όπου:

Αναπροσαρμοσμένος βαθμός του 2019 = Αρχική βαθμολογία εξέτασης του 2019 + D.

Νοείται ότι το D έχει θετική τιμή αν η εξέταση το 2019 ήταν πιο δύσκολη από αυτήν του 2017. Το D παίρνει αρνητική τιμή, αν η εξέταση το 2019 ήταν ευκολότερη από αυτήν του 2017. Στην περίπτωση, βέβαια, που οι εξετάσεις των δύο περιόδων ήταν της ίδιας δυσκολίας, η τιμή του D είναι 0. Η τιμή του D είναι, επίσης, 0 και στην περίπτωση της εξέτασης του γνωστικού αντικειμένου (όχι όμως και της εξέτασης γνώσης της Ελληνικής γλώσσας και της εξέτασης δεξιοτήτων) των 21 ειδικοτήτων που εξετάστηκαν για πρώτη φορά το 2019. Στην περίπτωση αυτή ισχύει:

Αναπροσαρμοσμένος βαθμός του 2019 = Αρχική βαθμολογία εξέτασης του 2019.

Στάδιο 2: Για κάθε ειδικότητα οι αναπροσαρμοσμένες βαθμολογίες που προέκυψαν από το Στάδιο 1 μετασχηματίστηκαν με τη μέθοδο που είχε χρησιμοποιηθεί το 2017. Η μέθοδος περιγράφεται αναλυτικά στο σχετικό κείμενο που είχε δημοσιευτεί στις 30.3.2018 και το οποίο είναι διαθέσιμο στον ακόλουθο ιστότοπο της Υπηρεσίας Εξετάσεων:

http://archeia.moec.gov.cy/ed/117/2018_03_30_statistiki_epexergasia_vathmologion.pdf



Πιο κάτω περιγράφονται αδρομερώς οι μέθοδοι που χρησιμοποιήθηκαν για τον υπολογισμό του δείκτη D για κάθε εξέταση και παρατίθεται σχετική βιβλιογραφία για περισσότερες τεχνικές λεπτομέρειες.

3. Μέθοδοι υπολογισμού του δείκτη D

Ο υπολογισμός του βαθμού δυσκολίας μιας εξέτασης σε σχέση με το βαθμό δυσκολίας μιας άλλης εξέτασης που προηγήθηκε χρονικά είναι διαδικασία πολύπλοκη. Συνήθως, η διεθνής βιβλιογραφία αλλά και η διεθνής πρακτική απαιτούν την αξιοποίηση «κοινών ερωτήσεων». Αυτό σημαίνει ότι κάποιες από τις ερωτήσεις των εξετάσεων του 2017 θα έπρεπε να επαναχρησιμοποιηθούν και στις εξετάσεις του 2019. Αν και σε κάποιες χώρες αυτό είναι υπόθεση ρουτίνας, η αξιοποίηση «κοινών ερωτήσεων» σε δημόσιες εξετάσεις είναι εκτός συζήτησης στην Κυπριακή πραγματικότητα. Για περισσότερες τεχνικές λεπτομέρειες για τη μέθοδο των «κοινών ερωτήσεων» παραπέμπουμε στους Kolen and Brennan (2014).

Μια άλλη συνηθισμένη πρακτική στο εξωτερικό, προκειμένου να υπολογιστεί ο βαθμός δυσκολίας μιας εξέτασης σε σχέση με το βαθμό δυσκολίας μιας άλλης εξέτασης που προηγήθηκε είναι η αξιοποίηση των αποτελεσμάτων των «κοινών υποψηφίων». Αυτό πρακτικά σημαίνει να συγκριθούν τα αποτελέσματα των υποψηφίων που παρακάθισαν στην εξέταση του 2017 με την εξέταση του 2019. Για παράδειγμα, στην περίπτωση των Εξετάσεων Διορισίμων για την εξέταση γνώσης Ελληνικής Γλώσσας, θα μπορούσε κανείς να συγκρίνει τις βαθμολογίες των 1710 υποψηφίων που παρακάθισαν στην εξέταση και το 2017 και το 2019. Επειδή ο αριθμός των υποψηφίων που εξετάστηκαν το 2017 και το 2019 ήταν πολύ μεγάλος, η μέθοδος των «κοινών υποψηφίων» προκρίθηκε ως μια από τις μεθόδους που αξιοποιήθηκαν για τον υπολογισμό του δείκτη D.

Αν και η μέθοδος των «κοινών υποψηφίων» είναι πολύ χρήσιμη, αποφασίστηκε να διασταυρωθούν τα αποτελέσματά της με αυτά άλλων τεχνικών προκειμένου να επιτευχθεί ελαχιστοποίηση του στατιστικού σφάλματος. Μελετώντας τη διεθνή πρακτική και τη διεθνή βιβλιογραφία προκρίθηκε η επιπρόσθετη χρήση των ακόλουθων τεχνικών: (α) propensity score matching, (β) pseudo anchor method, (γ) 3DC standard setting και (δ) Angoff standard setting. Αξίζει να αναφερθεί ότι οι μέθοδοι αυτές αξιοποιούνται διεθνώς σε εξετάσεις υψηλού διακυβεύματος. Για παράδειγμα, χρησιμοποιούνται για να διατηρηθούν σταθερά από χρόνο σε χρόνο τα επίπεδα δυσκολίας των γνωστών εξετάσεων GCE A' Level. Επίσης χρησιμοποιούνται σε διάφορες χώρες (π.χ. Αυστραλία, Αγγλία) για να διατηρείται από χρόνο σε χρόνο η συγκρισιμότητα των επιπέδων των εξετάσεων για πιστοποίηση άσκησης του ιατρικού επαγγέλματος (οι γνωστές «medical licensing exams»). Για πιο πολλές πληροφορίες παραπέμπουμε στους Ward, Chiavaroli, Fraser et al. (2018).

Για κάθε ειδικότητα αξιοποιήθηκε αριθμός από αυτές τις τεχνικές για τον υπολογισμό του D. Για κάθε μέθοδο υπολογίστηκε το D, αλλά και το τυπικό σφάλμα εκτίμησης του D. Για

κάθε ειδικότητα οι διάφορες εκτιμήσεις του D συνυπολογίστηκαν προκειμένου να υπολογιστεί ένα συνολικό D με το ελάχιστο δυνατό σφάλμα.

Πιο κάτω ακολουθούν σύντομες περιγραφές της κάθε μεθόδου:

3.1 Μέθοδος «κοινών υποψηφίων» (common persons method)

Για τις περισσότερες ειδικότητες υπήρχε μεγάλος αριθμός (χιλιάδες για κάποιες ειδικότητες) υποψηφίων που παρακάθισαν και στις δύο εξεταστικές περιόδους. Αν υποθέσουμε ότι η ομάδα των υποψηφίων δεν έχει αλλάξει σημαντικά στον ενδιάμεσο



χρόνο, προκύπτει ότι η διαφορά στη μέση επίδοσή τους το 2017 και το 2019 είναι μια καλή ένδειξη του βαθμού της σχετικής δυσκολίας των δύο εξετάσεων.

Ο αριθμός των κοινών υποψηφίων υποδηλώνεται με N . Η διαφορά στην επίδοση υπολογίζεται για κάθε υποψήφιο: $D_i = S_{17i} - S_{19i}$. Η μέση διαφορά επίδοσης στις δύο εξεταστικές περιόδους είναι μια εκτίμηση του δείκτη D ,

$$D_{resits} = \sum_i D_i / N$$

Το τυπικό σφάλμα εκτίμησης είναι συνάρτηση της τυπικής απόκλισης και του αντίστροφου της τετραγωνικής ρίζας του αριθμού των κοινών υποψηφίων,

$$SE_{resits} = sd(D_i) / \sqrt{N}$$

Για περισσότερες τεχνικές λεπτομέρειες για τη μέθοδο των «κοινών υποψηφίων» και για διάφορες παραλλαγές και υποπεριπτώσεις της μεθόδου παραπέμπουμε στους Kolen and Brennan (2014).

3.2 Μέθοδος αντιστοίχισης τάσης βαθμολογίας (Propensity score matching)

Για τους υποψηφίους χρησιμοποιήθηκαν πρόσθετες πληροφορίες όπως για παράδειγμα το φύλο, ο βαθμός του πτυχίου (Καλώς, Λίαν Καλώς και Άριστα), επιπλέον ακαδημαϊκά προσόντα (π.χ. MA, PhD), διδακτική εμπειρία κτλ. Στη διεθνή βιβλιογραφία και πρακτική, συχνά υιοθετείται η παραδοχή ότι υποψήφιοι με το ίδιο προφίλ θα έχουν παρόμοια επίδοση στην εξέταση. Συγκρίνοντας την επίδοση υποψηφίων με το ίδιο προφίλ στις εξετάσεις των δύο περιόδων, μπορούμε να αποδώσουμε τυχόν διαφορές στον διαφορετικό βαθμό δυσκολίας των εξετάσεων.

Η σχετική διαφορά στον βαθμό δυσκολίας των εξετάσεων των δύο περιόδων μοντελοποιήθηκε με τη μέθοδο της «ανάλυσης συνδιακύμανσης» (analysis of covariance-ANCOVA). Για σκοπούς ανάλυσης ορίστηκε ψευδομεταβλητή (dummy variable) dy . Στο μοντέλο η βαθμολογία των υποψηφίων «εξηγείται» από τη σταθερά c , την ψευδομεταβλητή dy και το σύνολο των μεταβλητών X που αντιπροσωπεύουν το προφίλ των υποψηφίων. Άρα,

$$S_i = c + \beta y * dy_i + \sum_j (\beta_j * X_{ji}) + e_i$$

Οι συντελεστές β του μοντέλου αξιοποιήθηκαν ώστε να εκτιμηθεί το D , με

$$DPSM = - \beta y$$

Το τυπικό σφάλμα εκτίμησης για τον συντελεστή βy είναι και το τυπικό σφάλμα του δείκτη D ,

$$SEPSM = SE_{\beta y}$$

Για περισσότερες πληροφορίες για τη μέθοδο αυτή παραπέμπουμε στους Livingston, Dorans & Wright (1990).

3.3 Μέθοδος οιονεί κοινών ερωτήσεων (Pseudo anchor method)

Για την εξεταστική περίοδο 2019 δόθηκαν οδηγίες προς τους θεματοθέτες ώστε να τηρηθεί πιστά ο Πίνακας Προδιαγραφών που είχε ανακοινωθεί, αλλά και να διατηρηθεί το στυλ και



το είδος των ερωτήσεων που είχαν χρησιμοποιηθεί για το εξεταστικό δοκίμιο του 2017. Όταν οι προδιαγραφές των εξεταστικών δοκιμών είναι σκοπίμως οι ίδιες, είναι δυνατόν να εντοπιστούν ζεύγη ερωτήσεων από τις δύο εξεταστικές περιόδους που να έχουν παρόμοια χαρακτηριστικά, π.χ. να εξετάζουν το ίδιο περιεχόμενο, να παίρνουν τον ίδιο αριθμό μονάδων, να έχουν περίπου τον ίδιο βαθμό δυσκολίας, να έχουν τον ίδιο τρόπο παρουσίασης κτλ. Σε τέτοιες περιπτώσεις, είναι δυνατόν να χρησιμοποιηθεί η μέθοδος των «οιονεί κοινών ερωτήσεων» (“pseudo anchor items”). Θεωρούμε ότι οι διαφορές μεταξύ των «οιονεί κοινών ερωτήσεων» πηγάζουν από στοιχεία που δεν έχουν κρίσιμη σημασία για τον σκοπό της αξιολόγησης. Επίσης, θεωρούμε ότι τυχόν μικρές αστοχίες στον εντοπισμό «οιονεί κοινών ερωτήσεων» (π.χ. σε ένα ζεύγος κοινών ερωτήσεων η ερώτηση της μιας εξεταστικής περιόδου, ίσως, είναι λίγο πιο δύσκολη από την ερώτηση της άλλης εξεταστικής περιόδου) θα εξουδετερώνουν η μια την άλλη, λόγω του αριθμού των ζευγών που θα χρησιμοποιούνται.

Έτσι, η βαθμολογία σε μια εξέταση (S) διαχωρίζεται στη βαθμολογία του μέρους του δοκιμίου που θα χρησιμοποιηθεί ως οιονεί κοινό (A) και στη βαθμολογία των υπολοίπων ερωτήσεων του δοκιμίου (R),

$$S_i = A_i + R_i$$

Επειδή αναμένουμε ότι τα ζεύγη των «οιονεί κοινών ερωτήσεων» έχουν συνολικά (κατά μέσον όρο) παρόμοιο βαθμό δυσκολίας στις δύο εξεταστικές περιόδους (εξ' ορισμού), οποιαδήποτε διαφορά στο βαθμό δυσκολίας των δοκιμών των δύο εξεταστικών περιόδων μπορεί να αποδοθεί, κυρίως, στη διαφορά του συνολικού βαθμού δυσκολίας των υπόλοιπων ερωτήσεων.

Η διαφορά στον βαθμό δυσκολίας περιγράφεται με συγκεκριμένο συντελεστή σε μοντέλο παλινδρόμησης. Αξιοποιείται ψευδομεταβλητή dy (δες περιγραφή προηγούμενης μεθόδου). Η συνολική βαθμολογία ενός υποψηφίου «εξηγείται» από τη σταθερά c , την ψευδομεταβλητή dy και τη βαθμολογία στο οιονεί κοινό μέρος των δοκιμών. Έτσι,

$$R_i = c + \beta y * dy_i + \beta a * A_i + e_i$$

Οι συντελεστές β χρησιμοποιούνται για να εκτιμήσουμε τον δείκτη D, με

$$DPA = - \beta y$$

Το τυπικό σφάλμα εκτίμησης του συντελεστή βy είναι το τυπικό σφάλμα του δείκτη D,

$$SEPA = SE\beta y.$$

Για περισσότερες πληροφορίες για τη μέθοδο αυτή και για παραλλαγές της παραπέμπουμε στους Bramley and Rodeiro (2014) καθώς και στον Bramley (2018).

3.4 Μέθοδος στάθμισης Angoff (Angoff standard setting)

Κατά τη στάθμιση Angoff, ομάδα ειδικών εκτιμά την αναμενόμενη βαθμολογία «άπειρων» υποψηφίων/εκπαιδευτικών σε κάθε μια από τις ερωτήσεις ενός εξεταστικού δοκιμίου. Το σύνολο των εκτιμήσεων υποδηλοί την αναμενόμενη επίδοση ενός «άπειρου» υποψηφίου στο συγκεκριμένο εξεταστικό δοκίμιο. Θεωρούμε ότι οι διαφορές στην αναμενόμενη



επίδοση ενός «άπειρου» υποψηφίου στις δύο εξεταστικές περιόδους μπορεί να αποδοθεί στον διαφορετικό βαθμό δυσκολίας της κάθε εξέτασης,

$$De = S17e - S19e$$

Έστω K ο αριθμός των ειδικών που χρησιμοποιούνται στη δραστηριότητα στάθμισης Angoff. Η εκτίμηση της τιμής του δείκτη D είναι ο μέσος όρος των διαφορών μεταξύ των αναμενόμενων επιδόσεων στις δύο εξεταστικές περιόδους από όλους τους ειδικούς,

$$D_{Angoff} = \sum e De / K$$

Το τυπικό σφάλμα της εκτίμησης είναι συνάρτηση της τυπικής απόκλισης και του αντίστροφου της τετραγωνικής ρίζας του αριθμού των ειδικών:

$$SE_{Angoff} = sd(De) / \sqrt{K}$$

Για περισσότερες πληροφορίες για τη μέθοδο αυτή και για παραλλαγές της παραπέμπουμε στους Hambleton and Plake (1995) και στον Brandon (2004).

3.5 Μέθοδος στάθμισης 3DC (3DC standard setting)

Στη μέθοδο στάθμισης 3DC, οι ειδικοί εκτιμούν την αναμενόμενη επίδοση «άπειρων» υποψηφίων σε ομάδες ερωτήσεων. Οι ειδικοί ενημερώνονται για τη σχετική δυσκολία της κάθε ομάδας ερωτήσεων, όπως αυτή προκύπτει από τα πραγματικά αποτελέσματα των εξετάσεων. Για αυτό τον λόγο, η συγκεκριμένη μέθοδος συχνά αναφέρεται και ως «στατιστικά ενημερωμένη» (statistically informed). Από μέρος της βιβλιογραφίας θεωρείται ότι η μέθοδος αυτή οδηγεί σε πιο αξιόπιστα αποτελέσματα σε σχέση με την μέθοδο Angoff.

Το σύνολο των αναμενόμενων βαθμολογιών σε κάθε ομάδα ερωτήσεων θεωρείται ως η αναμενόμενη επίδοση των υποψηφίων. Ο τρόπος υπολογισμού του δείκτη D είναι τεχνικά ο ίδιος με αυτόν της μεθόδου Angoff που έχει περιγραφεί πιο πάνω.

Για περισσότερες πληροφορίες για τη μέθοδο αυτή παραπέμπουμε στους Keuning, Straat and Feskens (2017).

4. Σύνθεση των δεικτών D από διαφορετικές μεθόδους

Δεν ήταν πρακτικά δυνατόν να αξιοποιηθούν όλες οι μέθοδοι στάθμισης για όλες τις ειδικότητες λόγω πρακτικών δυσκολιών. Για παράδειγμα, κάποιες ειδικότητες είχαν πολύ μικρό αριθμό υποψηφίων, άρα κάποιες μέθοδοι δεν μπορούσαν να χρησιμοποιηθούν. Ωστόσο, έγινε κάθε προσπάθεια να συνδυαστούν όσο το δυνατόν περισσότερες μέθοδοι για κάθε ειδικότητα, ώστε να έχουμε πολλαπλές πηγές πληροφόρησης και να πετύχουμε ελαχιστοποίηση του σφάλματος μέσω της τριγωνοποίησης των αποτελεσμάτων.

Κάθε μέθοδος αποδίδει μια εκτίμηση για το D και ένα τυπικό σφάλμα για αυτή την εκτίμηση. Προκειμένου να αποδώσουμε σε κάθε δείκτη D μια «δίκαιη» βαρύτητα, αξιοποιήσαμε το μέγεθος του τυπικού σφάλματος. Σε πιο επισφαλείς εκτιμήσεις (με μεγάλο σφάλμα) δόθηκε μικρότερη βαρύτητα. Ο μέσος όρος (weighted average), για κάθε ειδικότητα, υπολογίστηκε ως:

$$\text{Συνθετικός μέσος όρος } D = \sum w_m * D_m / \sum w_m$$

όπου

$$w_m = 1 / (SE_m * SE_m)$$



και w_m υποδηλώνει τη βαρύτητα του δείκτη για τη μέθοδο m (για όσες μεθόδους αξιοποιήθηκαν για τη στάθμιση των αποτελεσμάτων κάθε ειδικότητας).

5. Συμπερασματικά

Η διεθνής βιβλιογραφία και η διεθνής πρακτική προτείνουν πολλές μεθόδους για τη διαχρονική στάθμιση εξεταστικών δοκιμών. Προκειμένου να μεγιστοποιηθεί η ακρίβεια στους υπολογισμούς και προκειμένου να πετύχουμε την πιο δίκαιη διαχρονική στάθμιση, αξιοποιήθηκε μια μεγάλη συλλογή από μεθόδους, κατά τα πρότυπα άλλων Ευρωπαϊκών χωρών.

Κάθε μία από αυτές τις μεθόδους που αξιοποιήθηκαν έχει πλεονεκτήματα και μειονεκτήματα. Σε κάθε περίπτωση, η ομάδα των ειδικών που έφερε σε πέρας τη διαδικασία της στάθμισης διερεύνησε διεξοδικά τα μειονεκτήματα και τα πλεονεκτήματα της κάθε μεθόδου και αξιοποίησε μόνο τις πιο κατάλληλες για κάθε ειδικότητα.

Για περισσότερες πληροφορίες, ενδιαφερόμενοι με εξειδικευμένες γνώσεις παραπέμπονται σε σχετική βιβλιογραφία που παρατίθεται στο τέλος αυτού του κειμένου.

Αναφορές

Bramley, T. & Vidal Rodeiro, C.L. (2014). Using statistical equating for standard maintaining in GCSEs and A levels. Cambridge Assessment Research Report. Cambridge, UK: Cambridge Assessment.

Bramley, T. (2018, November). Evaluating the 'similar items method' for standard maintaining. Paper presented at the 19th annual conference of the Association for Educational Assessment in Europe, Arnhem-Nijmegen, The Netherlands.

Brandon, P. R. (2004). Conclusions about frequently studied modified Angoff standard-setting topics. *Applied Measurement in Education*, 17(1), 59-88.

Hambleton, R. K., & Plake, B. S. (1995). Using an extended Angoff procedure to set standards on complex performance assessments. *Applied Measurement in Education*, 8(1), 41-55.

Keuning, J., Straat, J. H., & Feskens, R. C. (2017). The Data-Driven Direct Consensus (3DC) Procedure: New Approach to Standard Setting. In *Standard Setting in Education* (pp. 263-278). Springer, Cham.

Kolen, M. J., & Brennan, R. L. (2014). *Test equating, scaling, and linking: Methods and practices*. Springer Science & Business Media.

Livingston, S. A., Dorans, N. J., & Wright, N. K. (1990). What combination of sampling and equating methods works best? *Applied Measurement in Education*, 3(1), 73-95.

Ward, H., Chiavaroli, N., Fraser, J. et al. (2018). Standard setting in Australian medical schools. *BMC Medical Education*, 18, 80. <https://doi.org/10.1186/s12909-018-1190-6>