ΚΕΦΑΛΑΙΟ Θ΄

ΑΝΑΛΥΣΗ ΕΡΩΤΗΣΕΩΝ ΚΑΙ ΠΡΟΣΔΙΟΡΙΣΜΟΣ ΤΩΝ ΒΑΣΙΚΩΝ ΧΑΡΑΚΤΗΡΙΣΤΙΚΩΝ ΤΩΝ ΤΕΣΤ ΣΤΟ ΠΛΑΙΣΙΟ ΤΗΣ ΚΛΑΣΙΚΗΣ ΘΕΩΡΙΑΣ ΓΙΑ ΤΙΣ ΕΞΕΤΑΣΤΙΚΕΣ ΔΟΚΙΜΑΣΙΕΣ

9.1. Εισαγωγή

Η προσπάθεια κατασκευής εξεταστικών δοκιμασιών που να εκφράζουν αξιόπιστα και έγκυρα αυτό το οποίο επιδιώκουν να μετρήσουν είναι παλαιά, όπως προκύπτει από όσα έχουν αναφερθεί σε προηγούμενα κεφάλαια. Για να επιτευχθεί το αποτέλεσμα αυτό, είναι, μεταξύ άλλων, απαραίτητο να γίνεται έλεγχος των ερωτήσεων. Ο έλεγχος αυτός αποσκοπεί στο να προσδιορίσει αν οι εν λόγω ερωτήσεις πληρούν ορισμένα στατιστικά και άλλα ποιοτικά χαρακτηριστικά, τα οποία έχουμε ήδη ευκαιριακά μνημονεύσει και θα τα εξετάσουμε συστηματικότερα στη συνέχεια του παρόντος κεφαλαίου.

Ανάλογος έλεγχος επιβάλλεται να γίνεται και επί των απαντήσεων στις κλειστού τύπου ερωτήσεις, προκειμένου να αποφευχθεί η αναγραφή αμφισβητούμενων επιλογών, οι οποίες μειώνουν την αξία του τεστ, προκαλώντας διαφωνίες μεταξύ των διορθωτών (Ebel, 1972: 359-406). Ο έλεγχος αυτός, ο οποίος είναι γνωστός ως **ανάλυση ερωτήσεων**,[1] γίνεται, κυρίως, κατά την προκαταρκτική φάση της εκπόνησης ενός τεστ. Είναι, όμως, δυνατόν να γίνει και μετά τη χορήγησή του, αλλά πριν αυτό βαθμολογηθεί (Livingston, 2006). Στην τελευταία περίπτωση, δεν υπολογίζονται στη βαθμολογία τα ερωτήματα που κρίθηκαν ακατάλληλα, διαδικασία η οποία, κατά τη γνώμη μας, δεν είναι απαλλαγμένη προβλημάτων. Για την ανάλυση των ερωτήσεων μπορούν να αξιοποιούνται δεδομένα από παλαιότερες εφαρμογές των τεστ, όταν πρόκειται για ερωτήσεις που διατηρούνται σε σχετικές τράπεζες, όπως ήδη αναφέραμε, ή συλλέγονται νέα.

---

1. Στην πραγματικότητα πρόκειται για ανάλυση των απαντήσεων στις ερωτήσεις. Εκτός από τη δυσκολία και τη διακριτικότητα των ερωτήσεων κατά τη σχετική ανάλυση ελέγχεται ακόμη αν τα ερωτήματα λειτουργούν διαφορετικά (differential item functioning) σε ομάδες με συγκεκριμένα χαρακτηριστικά π.χ. σε αγόρια και κορίτσια, σε άτομα διαφορετικής πολιτισμικής προέλευσης κ.ά. (Για το θέμα αυτό βλ. Κυριακίδης κ.ά. 2001).

Για την ανάλυση των ερωτήσεων και τον προσδιορισμό των ψυχομετρικών χαρακτηριστικών των τεστ έχουν αναπτυχθεί σχετικές θεωρίες και έχουν δημιουργηθεί αντίστοιχες πρακτικές εφαρμογές, για τις οποίες έχουν κατασκευαστεί και ανάλογα λογισμικά. Από τις θεωρίες αυτές εξετάζεται εκτενώς στο παρόν κεφάλαιο η κλασική θεωρία για τις μετρήσεις. Στο τέλος του κεφαλαίου γίνεται, επίσης, συνοπτική αναφορά στη θεωρία της γενικευσιμότητας, γνωστή και ως G-θεωρία από το αρχικό γράμμα τους αγγλοσαξονικού όρου generalizability (γενικευσιμότητα).[2] Στο επόμενο κεφάλαιο θα αναφερθούμε στη νεότερη θεωρία μέτρησης της ικανότητας ανταπόκρισης (ή απάντησης) σε ερωτήσεις, η οποία είναι ευρύτερα γνωστή ως «**Item Response Theory**» και αναφέρεται, συχνά, με τα αρχικά I.R.T.

9.2. Η κλασική θεωρία για την ανάλυση ερωτήσεων και τα χαρακτηριστικά των τεστ

Η κλασική θεωρία για τα τεστ αποτελεί την παλαιότερη και την πιο διαδεδομένη από τις υπάρχουσες προσεγγίσεις της ανάλυσης των ερωτήσεων και της αξιολόγησης των τεστ. Οι απαρχές της ανάγονται στις εργασίες του Speerman (1913, 1917), αλλά αναπτύχθηκε καθ' όλη τη διάρκεια του 1ου μισού του 20ου αιώνα (Gulliksen, 1950· Baker, 2001). Σύμφωνα με τους Lord και Novick (1968) αλλά και τους Allen και Yen (1979), η θεωρία αυτή βασίζεται στην άποψη ότι το αποτέλεσμα των μετρήσεων που διενεργούνται σχετικά με τις γνώσεις, τις δεξιότητες ή τις ικανότητες των ατόμων εμπεριέχει κάποιο σφάλμα. Κάθε μέτρηση είναι μια εμπειρική δειγματική εκτίμηση της πραγματικότητας, η οποία δεν συμπίπτει, κατ' ανάγκη, με την αληθινή κατάσταση των γνώσεων και των ικανοτήτων των ατόμων. Αυτό σημαίνει ότι η παρατηρούμενη επίδοση (μέτρηση) (X) αποτελεί προσεγγιστική εκτίμηση της **πραγματικής επίδοσης** (T), συνεκτιμουμένου του θετικού ή αρνητικού **σφάλματος** (E) που εμπεριέχει. Το σφάλμα αυτό η κλασική θεωρία δέχεται ότι είναι το ίδιο για όλους τους εξεταζομένους.

Η σχέση αυτή εκφράζεται, συνήθως, ως εξής:

$$X = T + E$$

Όσο επιτυχέστερα είναι τα χρησιμοποιούμενα ερωτήματα μιας δοκιμασίας, τόσο μικρότερο είναι το σφάλμα μέτρησης και τόσο ακριβέστερη θεωρείται η προσέγγιση της πραγματικότητας.

2. Άλλη ταξινόμηση των παραπάνω θεωριών βλ. στο Αλεξόπουλος (2004: 31-34 και 39-58).

Για να επιτευχθεί ο στόχος αυτός, επιδιώκεται να χρησιμοποιηθούν ερωτήματα, τα οποία να ανταποκρίνονται στα επιθυμητά επίπεδα των χαρακτηριστικών που αναφέρθηκαν προηγουμένως (δυσκολία και διακριτικότητα) και να εκπονηθούν εξεταστικά μέσα, τα οποία να πληρούν τα κριτήρια της εγκυρότητας, της αξιοπιστίας, και της αντικειμενικότητας.

Για τον υπολογισμό δεικτών σχετικών με τα παραπάνω χαρακτηριστικά έχουν προταθεί, κατά καιρούς, διάφοροι τρόποι στους οποίους θα αναφερθούμε παρακάτω.

9.3. Τρόποι δοκιμαστικού ελέγχου ερωτημάτων: μειονεκτήματα και πλεονεκτήματα

Σύμφωνα με όσα αναφέραμε σε προηγούμενο κεφάλαιο, ο έλεγχος των ερωτημάτων δεν είναι απαραίτητος μόνο για τα αρχικά ερωτήματα ενός τεστ. Είναι αναγκαίος και γι' αυτά που καταχωρίζονται σε τράπεζες, οι οποίες υποστηρίζουν ένα τεστ, καθώς και για την ανανέωση παλαιών τεστ ή για τη δημιουργία παράλληλων μορφών σταθμισμένων εξεταστικών δοκιμασιών.

Για τον έλεγχο των ερωτημάτων εφαρμόζονται τρεις, κυρίως, τρόποι, καθένας από τους οποίους έχει πλεονεκτήματα και μειονεκτήματα (Wendler & Walker, 2006). Αυτοί είναι οι εξής: α) η ένταξη σε ένα τεστ, που χορηγείται σε ορισμένο πληθυσμό, μιας χωριστής ενότητας, η οποία περιέχει νέα προς έλεγχο ερωτήματα, β) η διασπορά των προς έλεγχο ερωτημάτων ανάμεσα στα κύρια ερωτήματα ενός χορηγούμενου τεστ και γ) η χωριστή δοκιμαστική εξέταση νέων ερωτημάτων. Εξυπακούεται ότι στις περιπτώσεις α και β τα νέα ερωτήματα δεν λαμβάνονται υπόψη στη βαθμολόγηση του αντίστοιχου τεστ.

Τα πλεονεκτήματα του πρώτου τρόπου είναι: α) η εξασφάλιση της αντιπροσωπευτικότητας των δειγμάτων επί των οποίων προσδιορίζονται τα στατιστικά χαρακτηριστικά των ερωτημάτων, β) η μείωση του κόστους της δοκιμαστικής εφαρμογής και γ) η ύπαρξη κινήτρου εκ μέρους των εξεταζομένων να απαντήσουν στα ερωτήματα του τεστ, το αποτέλεσμα του οποίου θεωρείται σημαντικό απ'αυτούς, μη γνωρίζοντας ότι ένα τμήμα του δεν θα ληφθεί υπόψη. Μειονεκτήματά του είναι τα ακόλουθα: α) η αδυναμία δοκιμαστικής χορήγησης ερωτημάτων νέου τύπου, επειδή οι εξεταζόμενοι μπορεί να αντιληφθούν το σκοπό της ύπαρξής τους σε ένα τεστ, λόγω της διαφοροποίησής τους από τα υπόλοιπα, και να μην καταβάλουν τη δέουσα προσπάθεια να δώσουν ορθές απαντήσεις σ' αυτά, β) η επιμήκυνση των τεστ, τα οποία χορηγούνται σε επίσημες εξεταστικές δοκιμα-

σίες ή η μείωση του αριθμού των ερωτημάτων του κανονικού τεστ και γ) η αύξηση του χρόνου που απαιτείται για ορισμένη εξέταση, αν το τεστ που χρησιμοποιείται έχει πολύ επιμηκυνθεί.

Ο δεύτερος τρόπος έχει τα ίδια πλεονεκτήματα και μειονεκτήματα με τον πρώτο. Επιπρόσθετα, όμως, περιπλέκει το πρόβλημα της βαθμολόγησης των τεστ, όταν αυτό γίνεται με μηχανικά μέσα (π.χ. βαθμολόγηση απαντήσεων σε ερωτήσεις πολλαπλής επιλογής που σημειώνονται σε ειδικά φύλλα). Καθιστά, όμως, πιο δυσδιάκριτη την ύπαρξη παλαιών και νέων ερωτημάτων. Εξουδετερώνεται, ακόμη, η επίδραση που ασκεί στις απαντήσεις των εξεταζομένων η σειρά που κατέχουν τα ερωτήματα σε ένα τεστ (η πιθανότητα π.χ. ορθής απάντησης στα τελευταία ερωτήματα μειώνεται λόγω κόπωσης των εξεταζομένων).

Ο τελευταίος τρόπος έχει τα εξής πλεονεκτήματα: α) καθιστά δυνατή τη δοκιμαστική εφαρμογή εντελώς νέων μορφών ερωτημάτων και β) μειώνει το χρόνο που απαιτείται για τη δοκιμαστική εφαρμογή. Μειονεκτήματά του, όμως, είναι: α) η αύξηση του κόστους της δοκιμαστικής εφαρμογής, β) η δυσκολία εξασφάλισης της αντιπροσωπευτικότητας των δειγμάτων επί των οποίων γίνεται ο δοκιμαστικός έλεγχος και γ) η άμβλυνση του κινήτρου των εξεταζομένων να απαντήσουν στα συγκεκριμένα ερωτήματα, αν γνωρίζουν ότι η επίδοσή τους δεν θα έχει καμία επίπτωση γι' αυτούς.

Εξυπακούεται ότι ερωτήματα που περιέχονται σε τεστ, τα οποία έχουν ήδη χορηγηθεί και αξιολογηθεί ως προς τα ψυχομετρικά τους χαρακτηριστικά, μπορούν να επαναχρησιμοποιηθούν σε νέα τεστ, μετά την παρέλευση ορισμένου χρονικού διαστήματος, όπως έχουμε ήδη αναφέρει.

9.4. Ο δείκτης δυσκολίας (ή ευκολίας) των ερωτήσεων

Ο δείκτης δυσκολίας των ερωτήσεων έχει ιδιαίτερη σπουδαιότητα για τα σταθμισμένα τεστ που στηρίζονται σε νόρμες, κυρίως, στόχος των οποίων είναι η κατάταξη των εξεταζομένων σε αύξουσα ή φθίνουσα αξιολογική σειρά. Η γνώση, όμως, του στοιχείου αυτού μπορεί να φανεί χρήσιμη και στην κατασκευή των τεστ-κριτηρίων, καθώς και στα μη σταθμισμένα τεστ που εκπονούνται από τους εκπαιδευτικούς (Linn & Gronlund, 2000: 362). Για τους λόγους αυτούς θα αναφερθούμε στις επόμενες υποενότητες στη μεθοδολογία υπολογισμού του παραπάνω δείκτη, αρχίζοντας από τις αντικειμενικού ή κλειστού τύπου ερωτήσεις.

9.4.1. *Προσδιορισμός του δείκτη δυσκολίας των ερωτήσεων κλειστού τύπου*

Ο δείκτης δυσκολίας των ερωτήσεων κλειστού τύπου,[3] είναι ο λόγος των ατόμων που απάντησαν σωστά σε μια συγκεκριμένη ερώτηση προς το σύνολο των ατόμων, τα οποία έλαβαν μέρος στην αντίστοιχη εξέταση. Ο παραπάνω λόγος υπολογίζεται ταχύτατα, με βάση τον ακόλουθο τύπο:

$$\Delta = \frac{Nc}{N}$$

όπου Nc ο συνολικός αριθμός των ορθών απαντήσεων σε ορισμένη ερώτηση και N το πλήθος των ατόμων που πήραν μέρος στην εξέταση. Οι αναλογίες αυτές αναφέρονται στη διεθνή βιβλιογραφία ως p-values. Αν π.χ. σε μια εξέταση πήραν μέρος 150 μαθητές και σε μια ερώτηση πολλαπλής επιλογής που περιλαμβανόταν στη σχετική δοκιμασία απάντησαν σωστά 98 άτομα, ο βαθμός δυσκολίας της ερώτησης αυτής είναι:

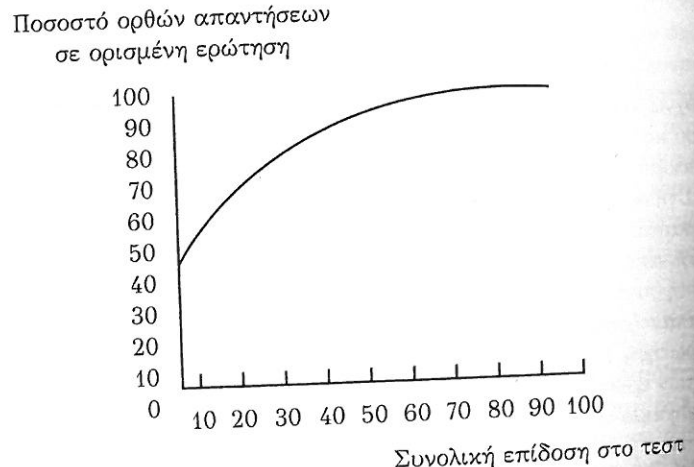
$$\Delta = \frac{98}{150} = 0.65 \text{ (ή } .65)$$

Συχνά, αντί αναλογιών αναγράφονται ποσοστά, τα οποία είναι περισσότερο κατανοητά από το ευρύ κοινό. Αυτά προκύπτουν από τον πολλαπλασιασμό του αποτελέσματος που εξάγεται από την παραπάνω διαίρεση επί 100. Στη συγκεκριμένη περίπτωση π.χ. το 0.65 γίνεται 65% (0.65 x 100) (Βαλσαμάκη-Ράλλη 1979: 149· Παπαϊωάννου, 1978: 42). Τούτο σημαίνει ότι το 35% από τους μαθητές του υποθετικού δείγματος δεν απάντησαν σωστά στη συγκεκριμένη ερώτηση. Όσο μεγαλύτερη είναι η αναλογία των ορθών απαντήσεων στο σύνολο των εξετασθέντων τόσο πιο εύκολη θεωρείται μια ερώτηση. Γι' αυτό, ορισμένοι συγγραφείς κάνουν λόγο για **δείκτη ευκολίας των ερωτήσεων** (Ebel, 1972: 395· Δημητρόπουλος, 2003: 290). Εμείς διατηρήσαμε τον όρο «δείκτης δυσκολίας, επειδή αυτός αναφέρεται συχνότερα στη σχετική βιβλιογραφία,[4] αν και αναγνωρίζουμε ότι από πλευράς νοήματος ο όρος «δείκτης ευκολίας των ερωτήσεων» ανταποκρίνεται περισσότερο στα πράγματα. Άλλοι, τέλος, συγγραφείς, για να αποφύγουν ακριβώς τη σύγχυση που ενδέχεται να προκληθεί από την ευρύτατη χρήση του όρου δυσκολία των ερωτήσεων, προτείνουν να προκύπτει αυτός από τον υπολογι-

3. Αφορά όλες τις περιπτώσεις στις οποίες δίνονται απαντήσεις που μπορούν, διχοτομικά, να χαρακτηρισθούν ως σωστές ή εσφαλμένες.
4. Βλ. ενδεικτικά Παπαϊωάννου (1978), Βαλσαμάκη-Ράλλη (1978), Ebel (1972), Αλεξόπουλος (1998).

σμό της αναλογίας (ή του ποσοστού) όχι εκείνων που απάντησαν σωστά, αλλά αυτών που απάντησαν λανθασμένα, περιλαμβανομένων και όσων δεν απάντησαν καθόλου στη συγκεκριμένη ερώτηση (Ebel, 1972: 395).[5]

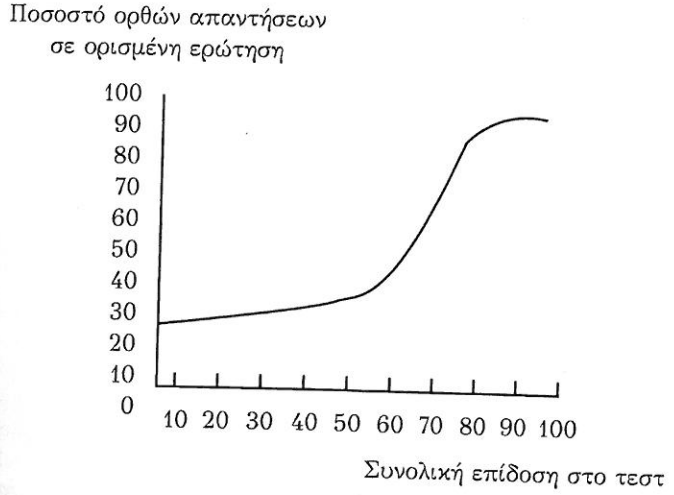
Εκτός από τον παραπάνω μαθηματικό τρόπο, ο δείκτης δυσκολίας μιας ερώτησης μπορεί να εκφρασθεί και με τη βοήθεια γραφικών παραστάσεων. Οι παραστάσεις αυτές κατασκευάζονται με τη βοήθεια δύο αξόνων: α) του οριζόντιου, που αντιστοιχεί στη συνολική επίδοση των εξετασθέντων με ένα τεστ, η οποία αναφέρεται από ορισμένους συγγραφείς ως **κριτήριο** (Livingston, 2006: 422-431) και β) του κάθετου άξονα, ο οποίος αντιπροσωπεύει το ποσοστό των αντίστοιχων ορθών απαντήσεων στην υπό εξέταση ερώτηση. Σχετικά παραδείγματα παρέχονται στη συνέχεια με βάση υποθετικές επιδόσεις μεγάλου αριθμού μαθητών που εξετάστηκαν με ένα τεστ αντικειμενικού τύπου, το οποίο περιείχε 100 ερωτήσεις πολλαπλής επιλογής που βαθμολογούνταν διχοτομικά (π.χ. 1/0).



Σχήμα 13. Καμπύλη κατανομής των απαντήσεων σε εύκολη ερώτηση κλειστού τύπου

5. Ο ίδιος συγγραφέας (όπ.π.: 385) προτείνει, ακόμη, έναν άλλο τρόπο υπολογισμού της δυσκολίας ενός ερωτήματος, ο οποίος συνίσταται στο εξής. Επιλέγεται από το σύνολο των εξετασθέντων το 27% εκείνων που έχουν την καλύτερη επίδοση και το 27% αυτών που έχουν τη χειρότερη. Υπολογίζεται το άθροισμα των ορθών απαντήσεων που δόθηκαν σε ορισμένο ερώτημα και από τις δύο ομάδες και αφαιρείται αυτό από το συνολικό τους πλήθος. Η διαφορά ανάγεται σε ποσοστό επί του συνολικού πλήθους των παραπάνω ομάδων, το οποίο

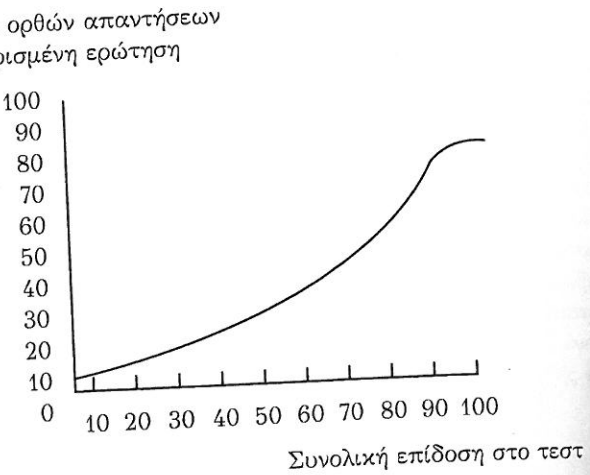
Η καμπύλη του σχήματος 13 απεικονίζει μια εύκολη ερώτηση, αφού αρκετοί από αυτούς που είχαν χαμηλή συνολική επίδοση απάντησαν σωστά σ' αυτήν και η συντριπτική πλειονότητα όσων είχαν μεσαίες και υψηλές επιδόσεις έδωσε σωστές απαντήσεις.



Σχήμα 14. Καμπύλη κατανομής των απαντήσεων σε ερώτηση κλειστού τύπου μέσης δυσκολίας

Η καμπύλη του σχήματος 14 απεικονίζει μια μέσης δυσκολίας ερώτηση, αφού λίγοι από αυτούς που είχαν χαμηλή συνολική επίδοση απάντησαν σωστά σ' αυτή, αρκετοί από όσους είχαν επίδοση γύρω στο μέσο όρο απάντησαν σωστά, ενώ η πλειονότητα των υψηλόβαθμων έδωσε σωστές απαντήσεις.

εκφράζει το βαθμό δυσκολίας του αντίστοιχου ερωτήματος. Σε μικρά δείγματα (π.χ. στους μαθητές μιας τάξης, είναι δυνατόν, κατά τους Linn και Gronlund (2000: 363-367), να λαμβάνονται οι δέκα πρώτοι και οι δέκα τελευταίοι ως προς την επίδοση μαθητές και να υπολογίζεται επ' αυτών ο δείκτης δυσκολίας των ερωτήσεων.



Σχήμα 15. Καμπύλη κατανομής των απαντήσεων σε δύσκολη ερώτηση κλειστού τύπου

Αντίθετα, το σχήμα 15 απεικονίζει την κατανομή των απαντήσεων σε μια δύσκολη ερώτηση. Αυτοί που είχαν χαμηλές επιδόσεις στο σύνολο του τεστ εμφανίζουν χαμηλά ποσοστά ορθών απαντήσεων στο συγκεκριμένο ερώτημα. Όσοι είχαν μεσαίες επιδόσεις παρουσιάζουν ελαφρώς υψηλότερα ποσοστά ορθών απαντήσεων, ενώ αυτοί που πέτυχαν υψηλές επιδόσεις εμφανίζουν ακόμη υψηλότερα ποσοστά ορθών απαντήσεων, αν και κάποιοι μεταξύ αυτών έδωσαν λανθασμένες απαντήσεις.

Οι γραφικές παραστάσεις αποτελούν ένα απλό τρόπο για την επισήμανση πολύ εύκολων ή πολύ δύσκολων ερωτήσεων, που οφείλονται σε ποικίλους λόγους και χρήζουν αναμόρφωσης ή πρέπει να εξαιρεθούν από το υπό ανάπτυξη τεστ.

Συχνά για τον υπολογισμό του δείκτη δυσκολίας των ερωτήσεων που πρόκειται να περιληφθούν σε ένα τεστ, χρησιμοποιούνται, όταν δεν υπάρχουν εμπειρικά δεδομένα ή για άλλους λόγους (π.χ. εξαιτίας του κόστους που προκαλείται), η μέθοδος των κριτών. Η μέθοδος αυτή εφαρμόζεται και για τον προσδιορισμό και άλλων ψυχομετρικών χαρακτηριστικών των ερωτημάτων, όπως θα δούμε παρακάτω. Για την ακρίβεια και την αξιοπιστία των εκτιμήσεων των κριτών έχουν γίνει πολλές έρευνες, οι παλαιότερες των οποίων ανάγονται ήδη στις πρώτες δεκαετίες του 20ου αιώνα (Hambleton

& Jirka, 2006).[6] Τα συμπεράσματα που προκύπτουν από αυτές ποικίλλουν. Η λεπτομερής παρουσίασή τους είναι εκτός των σκοπών της παρούσας εργασίας. Σημειώνουμε μόνο μερικές γενικές διαπιστώσεις.

Η εκτίμηση των δεικτών δυσκολίας και διακριτικότητας των ερωτημάτων είναι πιο έγκυρη, όταν προκύπτει από ομάδα κριτών παρά από μεμονωμένους κριτές. Στην αύξηση της ακρίβειας και της εγκυρότητας των εκτιμήσεων των κριτών συμβάλλουν τα εξής στοιχεία: α) η καλή γνώση του αντικειμένου στο οποίο αναφέρονται οι αντίστοιχες ερωτήσεις, β) η γνώση του πληθυσμού, στον οποίο απευθύνονται, γ) η προκαταρκτική ενημέρωσή των κριτών σχετικά με τους παράγοντες που επηρεάζουν τη δυσκολία των θεμάτων των εξετάσεων (π.χ. γλωσσική διατύπωση, πολυπλοκότητα, πρωτοτυπία, μορφή ερωτήσεων, σειρά των ερωτήσεων σε ένα τεστ ορισμένης χρονικής διάρκειας κτλ.), δ) η δοκιμαστική αξιολόγηση των ερωτημάτων, η οποία ακολουθείται από συζήτηση και ανατροφοδοτική ενημέρωση των κριτών, ε) η ύπαρξη κλιμάκων αξιολόγησης, η διαβάθμιση των οποίων είναι με σαφήνεια καθορισμένη, στ) η αντιστοίχιση δειγματικών απαντήσεων σχετικών με ορισμένο ερώτημα προς κάθε βαθμίδα ή επίπεδο της χρησιμοποιούμενης κλίμακας αξιολόγησης. Η διαδικασία, η οποία στηρίζεται σε κριτές, ακολουθείται, επίσης, στον καθορισμό των διαχωριστικών βαθμολογικών ορίων μεταξύ επιτυχίας/αποτυχίας σε μια εξεταστική δοκιμασία ή μεταξύ των διαβαθμίσεων της επιτυχίας.

9.4.2. *Προσδιορισμός του δείκτη δυσκολίας ερωτήσεων ανοικτού τύπου*

Ο δείκτης δυσκολίας σε περίπτωση ερωτήσεων ανοικτού τύπου (ανάπτυξης ή περιορισμένης σε έκταση απάντησης) που βαθμολογούνται με ισοδιαστημική κλίμακα μη διχοτομικού τύπου μπορεί να υπολογισθεί με βάση τον εξής τύπο:

$$\Delta = \frac{\Sigma fX - nX_{min}}{n(X_{max} - X_{min})}$$

όπου ΣfX το άθροισμα των επιδόσεων (X) των εξετασθέντων (n) στη συγκεκριμένη ερώτηση, X_{max} ο μέγιστος δυνατός βαθμός και X_{min} ο ελάχιστος δυνατός βαθμός.[7] Έστω π.χ. ότι 45 φοιτητές εξετάστηκαν με ερωτήσεις

6. Στο κεφάλαιο αυτό γίνεται σύνοψη των ευρημάτων των σχετικών ερευνών.

7. Βλ. The University of Iowa/EES/ACE Bulletin, Evaluation & Examination Service (www.uiowa.edu/~examserv/resources_fees/Technical_Bulletins/Preparing_and_Evaluating_essay_tests_Question.html – Ανακτήθηκε στις 1-8-2013. Στο περιεχόμενο της ιστοσελίδας αυτής στηρίζεται το παράδειγμα που ακολουθεί).

ανάπτυξης καθεμιά από τις οποίες έπαιρνε από το 0 έως 5 βαθμούς. Οι επιδόσεις τους φαίνονται στον πίνακα 26.

Πίνακας 26

Επιδόσεις 45 φοιτητών σε μια ερώτηση ανάπτυξης

Βαθμοί (X)	Συχνότητα (f)	Συνολικά μόρια (fX)
5	7	35
4	15	60
3	10	30
2	8	16
1	4	4
0	1	0
Σύνολο	n = 45	ΣfX = 145

Ο δείκτης δυσκολίας της συγκεκριμένης ερώτησης θα είναι:

$$\Delta = \frac{145-(45\cdot 0)}{45\cdot(5-0)} = \frac{145}{225} = 0.64$$

Από άλλους συγγραφείς (Δημητρόπουλος, 2003) προτείνεται, ως εναλλακτικός τρόπος υπολογισμού του δείκτη δυσκολίας σε ερωτήσεις, που βαθμολογούνται με βάση ορισμένη κλίμακα, η διαίρεση του μέσου όρου των βαθμών, τους οποίους πήραν οι εξετασθέντες στη συγκεκριμένη ερώτηση ($M_{ερ.}$) δια του μέσου όρου των μέγιστων δυνατών βαθμών ($M_{μεγ.}$) που θα μπορούσαν να δοθούν στην ερώτηση αυτή. Η εφαρμογή του τρόπου αυτού στο προηγούμενο παράδειγμα μας δίνει $M_{ερ}$= 3,2 (ήτοι 145:45=3,2) και $M_{μεγ}$=5, οπότε ο δείκτης δυσκολίας[7α] θα είναι:

$$\Delta = \frac{M_{ερ.}}{M_{μεγ.}} = \frac{3.2}{5} = 0.64$$

7α. Άλλο τρόπο υπολογισμού του δείκτη δυσκολία ανοικτού τύπου ερωτήσεων βλ. στα Reynolds et al. (2010: 154).

9.4.3. *Δείκτης δυσκολίας ολόκληρου του τεστ*

Για τον υπολογισμό του δείκτη δυσκολίας ολόκληρου του τεστ αρκεί η πρόσθεση των δεικτών δυσκολίας των επιμέρους ερωτήσεων και η διαίρεση του αθροίσματός τους δια του πλήθους των ερωτημάτων του τεστ. Εξυπακούεται ότι οι δείκτες δυσκολίας στους οποίους αναφερόμαστε ισχύουν για τον πληθυσμό αναφοράς από τον οποίο προέρχεται το δείγμα των εξετασθέντων, με βάση τις επιδόσεις των οποίων υπολογίστηκαν οι σχετικοί δείκτες. Ο δείκτης δυσκολίας των τεστ είναι ιδιαίτερα χρήσιμος στην κατασκευή ισοδύναμων παραλλαγών ενός τεστ, το οποίο χρησιμοποιείται σε επαναλαμβανόμενες εξεταστικές διαδικασίες.

Για την επίτευξη της συγκρισιμότητας των παραλλαγών αυτών επιδιώκεται η επίτευξη του ίδιου, περίπου, γενικού βαθμού δυσκολίας και παρόμοια κατανομή των δεικτών δυσκολίας των επιμέρους ερωτημάτων. Επειδή, όμως, ο δείκτης δυσκολίας (p-values) των ερωτημάτων, ο οποίος προκύπτει με τον τρόπο που εκθέσαμε προηγουμένως επηρεάζεται από το επίπεδο των εξεταζομένων, ο παραπάνω δείκτης μετατρέπεται, συνήθως, σε τιμές της τυποποιημένης κλίμακας Δέλτα (delta index) που έχει μέσο όρο 13 και τυπική απόκλιση 4 και ακολουθεί την κανονική κατανομή, ήτοι:

$$\Delta = 13 \pm 4z$$

όπου z η τιμή που προκύπτει από τον πίνακα της κανονικής κατανομής για αναλογία ίση με το (1-p) (Sireci & Allalouf, 2003· Wendler & Walker, 2006: 450).

Η πιο σημαντική αδυναμία του υπολογισμού του δείκτη δυσκολίας των ερωτημάτων, κατά την κλασική θεωρία, είναι ότι αυτός εξαρτάται από το επίπεδο του δείγματος, με βάση το οποίο προσδιορίζεται. Το ίδιο ερώτημα μπορεί να χαρακτηρισθεί ως εύκολο, αν το επίπεδο του δείγματος των εξετασθέντων είναι υψηλό, ή δύσκολο, αν το επίπεδό του είναι χαμηλό. Η αδυναμία αυτή επηρεάζει και τη διακριτικότητα των ερωτημάτων, η οποία εξετάζεται σε επόμενες ενότητες.

9.4.4. *Δείκτης δυσκολίας ερωτημάτων ενός τεστ που αποτελείται από υποτέστ*

Αν ένα τεστ περιλαμβάνει υποτέστ που αναφέρονται σε διαφορετικούς τομείς γνώσεων ή δεξιοτήτων (π.χ. ένα τεστ στις Θετικές Επιστήμες μπορεί να περιέχει χωριστά υποτέστ στη Φυσική, στη Χημεία και στη Βιολο-

γία), τότε είναι ορθότερο ο δείκτης δυσκολίας των ερωτημάτων να προσδιορίζεται με βάση τις επιδόσεις των μαθητών σε κάθε υποτέστ, υπό τον όρο ότι ο αριθμός των ερωτημάτων σε κάθε υποτέστ είναι ικανοποιητικός (μεγαλύτερος του 10) και η συνολική επίδοση σ'αυτά θεωρείται ότι εκφράζει επαρκώς τη μετρούμενη γνώση ή δεξιότητα.

Ο διαχωρισμός αυτός οφείλεται στην πιθανότητα κάποιος μαθητής να έχει διαφορετικές ικανότητες στους αντίστοιχους τομείς ή να έχει διαφορετικά προετοιμασθεί σ'αυτούς. Αν είναι ανάγκη να υπάρξει ενιαίος δείκτης δυσκολίας για ολόκληρο το τεστ, τότε αυτός μπορεί να προκύψει από το μέσο όρο των αντίστοιχων δεικτών κάθε υποτέστ.

9.4.5. *Διάταξη των ερωτήσεων με βάση το βαθμό δυσκολίας τους*

Οι ερωτήσεις που περιλαμβάνονται στις εξεταστικές δοκιμασίες δεν πρέπει να είναι του ίδιου βαθμού δυσκολίας, ιδιαίτερα όταν το σύστημα αξιολόγησης στηρίζεται στο συγκριτικό παράδειγμα. Πρέπει να υπάρχουν πολλές ερωτήσεις μέσης δυσκολίας (γύρω στο 0.50) και ορισμένες μικρής και μεγαλύτερης δυσκολίας. Επιβάλλεται, όμως, να αποφεύγονται οι πάρα πολύ εύκολες ερωτήσεις –εκτός από ειδικές περιπτώσεις ενθάρρυνσης των αδυνάτων μαθητών μέσω ενδοσχολικών δοκιμασιών διαμορφωτικού χαρακτήρα– καθώς και οι πάρα πολύ δύσκολες. Όσον αφορά στη διάταξη των ερωτήσεων, καλό είναι να προτάσσονται οι εύκολες ερωτήσεις και να έπονται οι δύσκολες και, μάλιστα, κατά αύξοντα βαθμό δυσκολίας, όπως έχουμε και αλλού αναφέρει.

Σκοπός αυτής της διάταξης είναι να μη αποθαρρυνθούν οι μαθητές στην προσπάθειά τους να απαντήσουν στα ερωτήματα που τους τίθενται. Είναι γνωστό ότι ορισμένοι μαθητές, όταν δεν μπορούν να απαντήσουν στις πρώτες ερωτήσεις, απογοητεύονται και εγκαταλείπουν κάθε προσπάθεια να απαντήσουν στις υπόλοιπες ερωτήσεις, ή δαπανούν το διαθέσιμο χρόνο στα πρώτα δύσκολα ερωτήματα, αφήνοντας αναπάντητα άλλα εύκολα, τα οποία έπονται.

9.4.6. *Παράγοντες που μπορεί να αλλοιώνουν το δείκτη δυσκολίας των ερωτήσεων*

Ο υπολογισμός του δείκτη δυσκολίας των ερωτήσεων έχει εγκυρότητα, αν εκφράζει την πραγματική δυνατότητα των εξεταζομένων να απαντήσουν σωστά σ' ένα ερώτημα ή να επιτύχουν την καλύτερη δυνατή επίδοση με βάση τις γνώσεις και τις ικανότητές τους. Αν, αντίθετα, η επιτυχία/απο-

τυχία σε κάποιο ερώτημα οφείλεται σε άλλους παράγοντες, τότε ο δείκτης δυσκολίας δεν είναι έγκυρος, κάτι που πρέπει να συνεκτιμάται από όσους ασχολούνται με την ανάλυση ερωτήσεων. Τέτοιου είδους παράγοντες μπορεί να είναι π.χ. η ανεπάρκεια του χρόνου που προβλέπεται για την απάντηση στα ερωτήματα ενός τεστ ή ενός υποτέστ, αν υπάρχει και καθορίζεται χωριστός χρόνος για το καθένα. Σε μια τέτοια περίπτωση, ενδέχεται οι εξεταζόμενοι να βιάζονται να απαντήσουν στα ερωτήματα του τεστ και να διαπράττουν λάθη όχι πάντοτε εξ αιτίας άγνοιας αλλά για άλλους λόγους (απροσεξίες, ανεπαρκή κατανόηση λόγω χρονικής πίεσης κτλ.). Για το λόγο αυτό απαιτείται να δίνεται ιδιαίτερη προσοχή στο χρόνο που απαιτεί ορισμένη εξέταση. Τα τεστ ταχύτητας (speed tests) χρησιμοποιούνται, άλλωστε, όλο και λιγότερο στο χώρο της εκπαίδευσης, ενώ επικρατούν τα επονομαζόμενα «τεστ ισχύος» (power tests), η χρονική διάρκεια των οποίων πρέπει να προσδιορίζεται με σχετική ευελιξία.

Η κούραση, επίσης, των εξεταζομένων μπορεί να έχει επιπτώσεις στην εκτίμηση του δείκτη δυσκολίας των ερωτήσεων, οι οποίες βρίσκονται στο τέλος μιας εξεταστικής δοκιμασίας, όπως και αλλού έχουμε αναφέρει.

Άλλος παράγοντας που, πιθανόν, να προκαλεί στρεβλώσεις στο ζήτημα αυτό είναι η ανομοιογένεια μεταξύ των θεματικών ενοτήτων ενός τεστ, όταν αυτό καλύπτει διαφορετικά αντικείμενα (π.χ. περισσότερα του ενός μαθήματα) ή διαφορετικές δεξιότητες. Σε ένα τεστ, που καλύπτει π.χ., γενικώς, τις Θετικές Επιστήμες, μπορεί οι εξεταζόμενοι να είναι καλύτεροι στη Φυσική απ' ό,τι στη Χημεία ή στη Βιολογία. Κατά συνέπεια, ο βαθμός δυσκολίας ενός ερωτήματος που αφορά στη Φυσική θα είναι διαφορετικός, αν ως κριτήριο ληφθεί το σύνολο των ερωτημάτων του τεστ, απ' ό,τι, αν ως κριτήριο ληφθούν μόνο οι ερωτήσεις στη Φυσική. Για τον παραπάνω λόγο συνιστάται στις περιπτώσεις αυτές ο δείκτης δυσκολίας των τεστ να υπολογίζεται κατά θεματικές ενότητες, εφόσον, όπως τονίσαμε προηγουμένως, ο αριθμός των σχετικών ερωτημάτων είναι επαρκής.

9.5. Προσδιορισμός του βαθμού διακριτικότητας των ερωτήσεων

9.5.1. *Γραφικές παραστάσεις της διακριτικότητας*

Μια γενική εικόνα για τη διακριτικότητα (ή τη διακριτική ισχύ) των ερωτήσεων μπορεί να προκύψει από τις γραφικές παραστάσεις που διαμορφώνονται με βάση το ποσοστό των εξεταζομένων, οι οποίοι απαντούν σωστά σε μια ερώτηση σε σχέση με τη συνολική τους επίδοση στο τεστ. Στις παραστάσεις αυτές αναφερθήκαμε προηγουμένως. Συμπληρώνοντας όσα έχουν

ήδη αναφερθεί, σημειώνουμε τα ακόλουθα. Όσο αυξάνει η ανοδική τάση της καμπύλης γραμμής, καθώς προχωρούμε από τα αριστερά προς τα δεξιά, τόσο πιο μεγάλη διακριτική ισχύ έχει το αντίστοιχο ερώτημα. Π.χ. η καμπύλη της γραφικής παράστασης 13 έχει μικρή διακριτικότητα, ενώ αυτή της παράστασης 15 έχει υψηλή διακριτικότητα. Αυτό σημαίνει ότι το ερώτημα στο οποίο αναφέρεται η γραφική παράσταση 13 δεν διαφοροποιεί σε σημαντικό βαθμό τους εξετασθέντες, ενώ συμβαίνει το αντίστροφο με το ερώτημα στο οποίο αναφέρεται η γραφική παράσταση 15.

Στην πράξη, όμως, και στο πλαίσιο της κλασικής θεωρίας για την ανάλυση των ερωτημάτων χρησιμοποιείται συχνότερα ο αριθμητικός δείκτης διακριτικότητας, στον υπολογισμό του οποίου αναφερόμαστε παρακάτω.

9.5.2. Δείκτης διακριτικότητας κλειστού τύπου ερωτήσεων

Για να προσδιοριστεί ο βαθμός διακριτικότητας μιας ερώτησης κλειστού τύπου (π.χ. πολλαπλής επιλογής), η ικανότητά της, δηλαδή, να διακρίνει τους μαθητές που πήραν μέρος σε μια εξέταση σε διαφορετικές κατηγορίες, ακολουθείται η εξής διαδικασία:[8]

Λαμβάνουμε το 27%[9] των γραπτών των μαθητών που πήραν μέρος στην εξέταση, τα οποία έχουμε βάλει σε φθίνουσα αξιολογική σειρά, αρχίζοντας από εκείνο που έχει το μεγαλύτερο βαθμό και σταματώντας σε εκείνο, με το οποίο συμπληρώνεται ο αριθμός των ατόμων που αντιπροσωπεύει το παραπάνω ποσοστό επί του συνολικού δείγματος. Έτσι σχηματίζεται η ομάδα Α, η οποία λέγεται ανώτερη ομάδα. Στη συνέχεια, λαμβάνουμε ίσο αριθμό γραπτών, αρχίζοντας από το τέλος, από εκείνο, δηλαδή, που έχει το μικρότερο βαθμό. Έτσι σχηματίζεται η ομάδα Β (κατώτερη ομάδα). Υπολογίζουμε τον αριθμό των ατόμων από κάθε ομάδα που απάντησαν σωστά στην ερώτηση, η οποία μας ενδιαφέρει. Η διαφορά μεταξύ των δύο αυτών αριθμών διαιρείται δια του πλήθους των ατόμων που αντιστοιχούν στο 27% του δείγματος. Το πηλίκο της διαίρεσης αντιπροσωπεύει το βαθμό

8. Περιγραφή της διαδικασίας υπολογισμού τόσο του δείκτη δυσκολίας όσο και του βαθμού διακριτικότητας των ερωτήσεων υπάρχει στα περισσότερα εγχειρίδια που ασχολούνται με την αξιολόγηση της επίδοσης (βλ. βιβλιογραφία).

9. Άλλοι προτείνουν τη λήψη ποσοστού δοκιμίων ίσο προς το 1/3 (33%) ή προς το 1/4 (25%) ή προς το μισό (50%) του συνόλου των εξετασθέντων. Όπως, όμως, απέδειξε ο Kelley (1939) η λήψη του 27% των ατόμων του συνόλου από την αρχή και από το τέλος της ομάδας εξασφαλίζει τη μεγαλύτερη πιθανότητα να υπάρχουν στην ομάδα Α άτομα κατά πολύ ανώτερα ως προς την ικανότητα που μετριέται με το χρησιμοποιούμενο τεστ από τα άτομα που βρίσκονται στην ομάδα Β.

διακριτικότητας (D) της αντίστοιχης ερώτησης. Η διαδικασία αυτή συνοψίζεται στον ακόλουθο τύπο:

$$D = \frac{O_A - O_B}{N}$$

όπου O_A ο αριθμός των ορθών απαντήσεων στην εξεταζόμενη ερώτηση των N_A καλύτερων μαθητών (ανώτερη ομάδα), O_B ο αριθμός αυτός μεταξύ των N_B χειρότερων μαθητών και N ίσο προς το 27% του συνόλου των μαθητών που εξετάστηκαν.

Έστω ότι σε μια εξέταση στην Ιστορία πήραν μέρος 178 μαθητές. Μετά την κατάταξη των φύλλων απαντήσεων σε σειρά επιτυχίας πάρθηκαν τα 48 (ποσοστό 27% του συνόλου) καλύτερα και τα 48 χειρότερα από αυτά και καταγράφηκαν οι απαντήσεις στην εξής ερώτηση: «*Ποιοι ίδρυσαν τη Φιλική Εταιρεία; (Βάλτε ένα σταυρό σε μια από τις ακόλουθες απαντήσεις την οποία θεωρείτε ορθή)*».

Η κατανομή των απαντήσεων που έδωσαν οι μαθητές των δύο ομάδων σημειώνεται παρακάτω. Η ορθή απάντηση φέρει ένα αστερίσκο.

Για να υπολογίσουμε το δείκτη διακριτικότητας της συγκεκριμένης ερώτησης, εφαρμόζουμε τον προηγούμενο τύπο και διαπιστώνουμε ότι η τιμή του δείκτη αυτού είναι 0,40.

	Απαντήσεις	
	Α΄ομάδα	Β΄ομάδα
Καποδίστριας, Κοραής και Σκουφάς	1	4
Καποδίστριας, Ρήγας και Σκουφάς	2	7
Σκουφάς, Τσακάλωφ και Ξάνθος*	39	20
Ρήγας, Τσακάλωφ και Ξάνθος	5	10
Καποδίστριας, Ρήγας, Κοραής	1	7
Σύνολο	48	48

$$D = \frac{39-20}{48} = \frac{19}{48} = 0.40$$

Σε περίπτωση πολύ μικρών δειγμάτων, στα οποία το 27% είναι λιγότερο από δέκα γραπτά, προτείνεται να περιλαμβάνονται στην πρώτη ομάδα οι 10 καλύτεροι μαθητές και στην δεύτερη ομάδα οι 10 χειρότεροι ως προς την επίδοσή τους στο συγκεκριμένο τεστ. Εάν υπάρχουν περιπτώσεις με τον ίδιο βαθμό που μπορούν να καταταχθούν και στις δύο κατηγορίες, τότε γίνεται τυχαία η κατανομή τους στις δύο ομάδες (Reynolds et al. 2010).

9.5.3. Δείκτης διακριτικότητας ανοικτού τύπου ερωτήσεων

Ο δείκτης διακριτικότητας για τις ερωτήσεις ανάπτυξης και περιορισμένης σε έκταση απάντησης ή άλλες, που δεν βαθμολογούνται διχοτομικά αλλά με βάση ισοδιαστημική κλίμακα περισσότερων διαβαθμίσεων, υπολογίζεται κατά παρόμοιο τρόπο, όπως και στις περιπτώσεις των ερωτήσεων κλειστού τύπου. Υπολογίζουμε το δείκτη δυσκολίας στην ανώτερη ομάδα Δ_A (βλ. προηγούμενη ενότητα) και τον αντίστοιχο δείκτη της κατώτερης ομάδα (Δ_K) ως προς ορισμένη ερώτηση. Η διαφορά τους μας δίνει το δείκτη διακριτικότητας για την υπό εξέταση ερώτηση.[10]

Έστω ότι η ανώτερη και η κατώτερη ομάδα 45 φοιτητών έχουν την ακόλουθη βαθμολογία σε μια ερώτηση που βαθμολογείται από 0 έως 5.

Πίνακας 27

Επίδοση της ανώτερης και κατώτερης ομάδας ενός δείγματος 45 φοιτητών σε μια ανοικτή ερώτηση που βαθμολογείται από 0-5 (0 = απουσία απάντησης)

Βαθμοί	Ανώτερη Ομάδα (27% του δείγματος)		Κατώτερη Ομάδα (27% του δείγματος)	
X	f_A	f_AX	f_K	f_KX
5	5	25	–	–
4	6	24	–	–
3	1	3	4	12
2	–	–	3	6
1	–	–	4	4
0	–	–	1	0
Σύνολο	n_A= 12	Σf_AX = 52	n_K = 12	Σf_KX = 22

Ο δείκτης διακριτικότητας (D) της συγκεκριμένης ερώτησης θα είναι:

$$D=[(\frac{\Sigma f_A X - nX_{min}}{n_A(X_{max}-X_{min})})-(\frac{\Sigma f_K X - nX_{min}}{n_B(X_{max}-X_{min})})]$$

όπου X_{max} και X_{min} έχουν την ίδια σημασία με αυτή που αναφέρθηκε στην αντίστοιχη περίπτωση του υπολογισμού του δείκτη δυσκολίας.

10. Βλ. The Iowa University, (όπου παρ.: 8-9). Στην ιστοσελίδα αυτή στηρίζεται το αναφερόμενο δικό μας παράδειγμα.

$$D=[(\frac{52-12\cdot(0)}{12\cdot(5-0)})-(\frac{22-12\cdot(0)}{12\cdot(5-0)})]=0.67-0.17=0.50$$

Από άλλους (Δημητρόπουλος, 2003, Reynolds et al. 2010: 163-64) προτείνεται ο υπολογισμός του δείκτη διακριτικότητας μιας ερώτησης να γίνεται με τη διαίρεση της διαφοράς των μέσων όρων των δύο ομάδων (M_A) και (M_B) δια του μέσου όρου των μέγιστων δυνατών βαθμών στην υπό εξέταση ερώτηση ($M_{μεγ.}$). Η εφαρμογή του τρόπου αυτού στο δικό μας παράδειγμα έχει ως εξής:

$$D = \frac{M_A - M_B}{M_{μεγ}} = \frac{4,33-1,83}{5} = \frac{2,5}{5} = 0.50$$

9.5.4. Δείκτης διακριτικότητας ολόκληρου του τεστ

Για να υπολογίσουμε το δείκτη διακριτικότητας ενός ολόκληρου τεστ, εφαρμόζουμε την ίδια μέθοδο, με βάση την οποία υπολογίζουμε το δείκτη δυσκολίας του. Προσθέτουμε, δηλαδή, τους δείκτες διακριτικότητας των επιμέρους ερωτήσεων και διαιρούμε το άθροισμά τους δια του πλήθους των.

Όσα αναφέρθηκαν προηγουμένως ως προς τους όρους ισχύος του βαθμού δυσκολίας των ερωτήσεων και των παραγόντων που τον επηρεάζουν έχουν εφαρμογή και στην περίπτωση του δείκτη διακριτικότητας.

9.5.5. Δείκτης διακριτικότητας σε περίπτωση ύπαρξης υποτέστ

Στην περίπτωση κατά την οποία ένα τεστ αποτελείται από υποτέστ που αναφέρονται σε χωριστούς τομείς, τότε πρέπει ο δείκτης διακριτικότητάς του να υπολογίζεται ανά υποτέστ για τους ίδιους λόγους και υπό τους ίδιους όρους με αυτούς που αναφέρθηκαν προηγουμένως κατά την εξέταση της δυσκολίας των ερωτήσεων.

9.5.6. Άλλος τρόπος υπολογισμού του δείκτη της διακριτικής ισχύος των ερωτημάτων

Άλλος τρόπος υπολογισμού του δείκτη της διακριτικής ισχύος ενός ερωτήματος είναι η εύρεση του δείκτη συνάφειας μεταξύ των βαθμών των εξετασθέντων σε μια ερώτηση και της συνολικής επίδοσής τους στο τεστ. Στην περίπτωση που μια ερώτηση αξιολογείται με 1 ή 0 (ορθή-εσφαλμένη, όπως συμβαίνει στις περισσότερες αντικειμενικού τύπου ερωτήσεις) υπάρχουν

δύο τρόποι υπολογισμού του παραπάνω δείκτη: α) η απλή συνάφεια μεταξύ του 1/0 και της συνολικής επίδοσης στο τεστ και β) η σειριακή συνάφεια (biserial correlation) που βασίζεται στη θεώρηση των απαντήσεων 1/0 ως ποιοτικών δεικτών της υποκείμενης ικανότητας, με βάση την οποία ποικίλλει η επίδοση των εξεταζομένων (Livingston, 2006). Στις υπόλοιπες περιπτώσεις, εφαρμόζονται οι τεχνικές υπολογισμού της συνάφειας που έχουμε ήδη εκθέσει. Από πρακτική άποψη και οι δύο τρόποι υπολογισμού του δείκτη διακριτικότητας έχουν παρόμοια χρησιμότητα. Απλώς ο δείκτης που προκύπτει από την υπό εξέταση μέθοδο είναι αποτέλεσμα της συνεκτίμησης των απαντήσεων όλων των υποψηφίων και όχι εκείνων της ανώτερης και κατώτερης ομάδας μόνον, όπως συμβαίνει στην προηγούμενη περίπτωση. Ερωτήματα με δείκτη διακριτικής ισχύος μικρότερο του 0.30 χρήζουν, συνήθως, επανεξέτασης.

9.6. Σχέση δυσκολίας και διακριτικότητας των ερωτήσεων

Από όσα αναφέρθηκαν προηγουμένως προκύπτει ότι υπάρχει σχέση μεταξύ των δύο βασικών χαρακτηριστικών των ερωτημάτων, ήτοι του βαθμού δυσκολίας τους και του δείκτη της διακριτικής τους ισχύος. Η σχέση, όμως, αυτή δεν είναι ευθύγραμμη. Ένα πολύ εύκολο ερώτημα που απαντάται από όλους σχεδόν τους εξεταζομένους δεν έχει διαφοροποιητική ισχύ. Το ίδιο συμβαίνει και για ένα πολύ δύσκολο ερώτημα, στο οποίο δεν απαντάει ορθά κανείς από τους εξεταζομένους. Δεν λειτουργεί ως κριτήριο διαφοροποίησης. Τα ερωτήματα που έχουν βαθμό δυσκολίας (γύρω στο 0.50 ή 50%) έχουν τους υψηλότερους δείκτες διακριτικότητας, όπως προκύπτει από τα στοιχεία του πίνακα 28.

Πίνακας 28

Μέγιστοι βαθμοί διακριτικότητας ανά διαφορετικά επίπεδα δυσκολίας των ερωτήσεων

Δείκτης δυσκολίας	Μέγιστος δείκτης διακριτικότητας
1.00	0.00
0.90	0.20
0.80	0.40
0.70	0.60
0.60	0.70
0.50	1.00

(συνέχεια)

0.40	0.70
0.30	0.60
0.20	0.40
0.10	0.20
0.00	0.00

Πηγή: Reynolds et al. (2010: 153)

9.7. Επιλογή των ερωτήσεων με βάση το βαθμό δυσκολίας και το δείκτη διακριτικότητάς τους

Ύστερα απ' όσα ελέχθησαν παραπάνω, γεννάται το ερώτημα: Τίνος βαθμού δυσκολίας ή τίνος δείκτη διακριτικότητας ερωτήσεις πρέπει να χρησιμοποιούνται στις διάφορες εξεταστικές δοκιμασίες;

Η απάντηση στο ερώτημα αυτό εξαρτάται από πολλούς παράγοντες, όπως από το είδος των ερωτήσεων που χρησιμοποιούνται, από την ύλη που εξετάζεται, από το σκοπό που επιδιώκεται κατά την εξέταση και άλλα παρόμοια.

Ο Lord (1952) προσδιόρισε για τους διάφορους τύπους των αντικειμενικών ερωτήσεων τους εξής ενδεικνυόμενους βαθμούς δυσκολίας:[11]

Τύποι ερωτήσεων	*Βαθμός δυσκολίας (% ορθών απαντήσεων)*
Συμπλήρωση κενών	50
Σωστό – Λάθος	85
Ερωτήσεις πολλαπλής εκλογής 5 εναλλακτικών απαντήσεων	70
Ερωτήσεις πολλαπλής εκλογής 4 εναλλακτικών απαντήσεων	74
Ερωτήσεις πολλαπλής εκλογής 3 εναλλακτικών απαντήσεων	77

Κατά γενικό, πάντως, κανόνα, επιλέγονται ερωτήσεις, των οποίων ο βαθμός δυσκολίας κυμαίνεται μεταξύ 40% και 70%.[12]

Ως προς τους δείκτες διακριτικότητας, ο Ebel (1972:399) δίνει την εξής ταξινόμηση:

11. Βλ. και Βαλσαμάκη-Ράλλη (1979: 149).
12. Συμφωνα με άλλους επιλέγονται ερωτήσεις που έχουν δείκτη δυσκολίας γύρω στο 50% (0.50) (Τσιμπούκης, 1979β: 62).

0.40 και άνω: Πολύ καλή ερώτηση.
0.30 - 0.39: Λογικά καλή ερώτηση, αλλά μπορεί να βελτιωθεί.
0.20 - 0.29: Περιθωριακές ερωτήσεις που έχουν ανάγκη βελτίωσης.
≤ 0.19: Πτωχές ερωτήσεις που συνήθως απορρίπτονται ή επανεξετάζονται σε μελλοντική αναθεώρηση του τεστ.[13]

Κατά τον ίδιο συγγραφέα, οι ερωτήσεις ενός αντικειμενικού τεστ εξέτασης πρέπει να έχουν δείκτη διακριτικότητας μεγαλύτερο του 0,40 σε ποσοστό 50% και άνω, μεταξύ 0.40 και 0.20 σε ποσοστό λιγότερο του 40% και μικρότερο του 0.20 σε ποσοστό λιγότερο του 10%.

Έστω ότι μια δοκιμαστική εξέταση 20 ερωτήσεων έδωσε τα αποτελέσματα που αναφέρονται στον πίνακα 29. Εάν ζητηθεί να κατασκευαστεί ένα τεστ κανονικής δυσκολίας και διακριτικής ικανότητας, τότε θα επιλεγούν από τις ερωτήσεις αυτές όσες έχουν δείκτη διακριτικότητας άνω του 0.40, τουλάχιστον, και βαθμό δυσκολίας από 40% έως 70%.

Με τα κριτήρια αυτά θα περιληφθούν στο υπό κατασκευή τεστ οι ερωτήσεις που περιλαμβάνονται μέσα σε κύκλο. Οι άλλες απορρίπτονται.

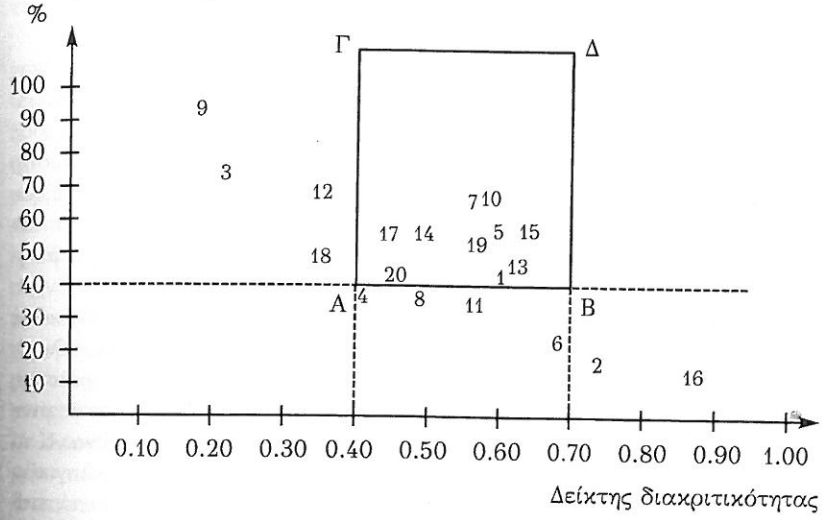
Πίνακας 29

Δείκτης δυσκολίας και διακριτικότητας 20 ερωτήσεων

Αύξ. αριθμός ερωτήσεως	Δείκτης δυσκολίας %	Δείκτης διακριτικότητας	Αύξ. αριθμός ερωτήσεως	Δείκτης δυσκολίας %	Δείκτης διακριτικότητας
(1)	40	0.60	11	35	0.53
2	24	0.70	12	70	0.35
3	80	0.22	(13)	45	0.61
4	36	0.40	(14)	52	0.50
(5)	60	0.57	(15)	58	0.65
6	20	0.68	16	15	0.89
(7)	68	0.54	(17)	53	0.45
8	36	0.50	18	48	0.36
9	90	0.18	(19)	59	0.58
(10)	62	0.56	(20)	40	0.46

13. Βλ και Αλεξόπουλος, (1998: 94).

Ένας άλλος τρόπος εύρεσης των ερωτήσεων που ενδείκνυνται να περιληφθούν σε ένα τεστ με προσδιορισμένα τα χαρακτηριστικά των ερωτήσεων του είναι να κατασκευαστεί η γραφική παράσταση που απεικονίζεται στο σχήμα 16.



Σχήμα 16. *Τρόπος επιλογής ερωτήσεων με βάση ορισμένο βαθμό δυσκολίας και διακριτικότητας*

Εκτός, όμως, από αυτούς τους προσδιορισμούς, οι τιμές των δεικτών δυσκολίας και διακριτικότητας των ερωτήσεων εξαρτώνται και από το σκοπό που επιδιώκεται κατά την εξέταση. Όπως σημειώνει η Βαλσαμάκη-Ράλλη (1979: 149), πρέπει να εξετάζεται αν «*πρόκειται για εξέταση που αποβλέπει στην επιλογή του 10% μόνο των εξεταζομένων ή για εξέταση που επιδιώκει να διαπιστώσει κατά πόσον οι μαθητές έχουν αφομοιώσει ορισμένες θεμελιώδεις αρχές και έννοιες; Μήπως η εξέταση αποβλέπει στην κατάταξη των μαθητών ανάλογα με τις γνώσεις τους; Ή μήπως επιδιώκει τη διάγνωση βασικών αδυναμιών; Φανερό είναι ότι διάφοροι θα είναι σε κάθε περίπτωση και οι επιθυμητοί δείκτες δυσκολίας και διακριτικότητας των ερωτήσεων*».

Η τοποθέτηση, βέβαια, του προβλήματος σε μια τέτοια βάση επαναφέρει το όλο θέμα στη διάκριση μεταξύ διαμορφωτικής και τελικής ή αθροιστικής αξιολόγησης των εξεταζόμενων, καθώς και στη διαφοροποίηση με-

ταξύ δοκιμασιών που στηρίζονται σε νόρμες και εκείνων που αξιολογούν την επίτευξη στόχων (τεστ-κριτήρια), θέματα τα οποία έχουν αναπτυχθεί σε προηγούμενα κεφάλαια.

9.8. Ποιοτικός έλεγχος των ερωτημάτων

Εκτός από τον υπολογισμό των αριθμητικών δεικτών για τη δυσκολία και τη διακριτική ισχύ των ερωτημάτων από ορισμένους συγγραφείς (Popham, 2000, Reynolds et al. 2010 κ.ά.) είναι αναγκαίος και ο ποιοτικός τους έλεγχος. Αυτός συνίσταται στη μελέτη της γλωσσικής τους διατύπωσης, στην εξάλειψη πιθανών σφαλμάτων ή παραλείψεων, στην αύξηση της σαφήνειάς τους και στη μεγαλύτερη σύνδεσή τους με το περιεχόμενο του εξεταζόμενου αντικειμένου. Για το σκοπό αυτό ορισμένοι (όπ. παρ.) εισηγούνται τα ακόλουθα:

1) Μετά την αρχική σύνταξη ενός τεστ, ο συντάκτης του το αφήνει κατά μέρος για μερικές ημέρες, για να απομακρυνθεί νοητικά και ψυχολογικά από αυτό. Επανέρχεται στο αρχικό του σχέδιο, μετά την παρέλευση ορισμένου χρονικού διαστήματος, και το διαβάζει δεύτερη ή και τρίτη φορά. Συμβαίνει, συχνά, κάποιος που έχει συντάξει ένα κείμενο να το διαβάζει επανειλημμένα και, λόγω της κόπωσης ή άλλων ψυχολογικών παραγόντων, να μην αντιλαμβάνεται λάθη που υπάρχουν σ' αυτό, ενώ η νοητική και ψυχοσυναισθηματική του απομάκρυνση από το κείμενο τον βοηθά να τα εντοπίσει.
2) Βοηθητική στην ποιοτική βελτίωση των ερωτημάτων ενός τεστ αποδεικνύεται, επίσης, η μελέτη τους από έναν ειδικό (π.χ. από ένα έμπειρο εκπαιδευτικό σχετικό με το αντικείμενο του τεστ). Αυτός μπορεί να είναι κάποιος συνάδελφος, ο οποίος υπηρετεί στο ίδιο σχολείο με το συντάκτη του τεστ, όταν πρόκειται για δοκιμασίες που εκπονούνται από τους διδάσκοντες ή ένας εμπειρογνώμονας στον οποίο ο φορέας ανάπτυξης του τεστ τού αναθέτει το έργο αυτό. Εξάλλου, οι επαγγελματίες που ασχολούνται με την παραγωγή τεστ, καθώς και οι ειδικές στον τομέα αυτόν υπηρεσίες διαθέτουν, συνήθως, ομάδες επιστημόνων που είναι αρμόδιοι για την ποιοτική ανάλυση των ερωτημάτων.

9.9. Ανάλυση των παρεμβολών σε ερωτήσεις πολλαπλής επιλογής

Μέχρι τώρα αναφερθήκαμε στην ανάλυση των ερωτημάτων ενός τεστ, με βάση το αν η απάντηση που δίνεται σ' αυτά είναι ορθή ή εσφαλμένη. Ορισμένοι, όμως, συγγραφείς (Linn & Gronlund, 2000: 367, Reynolds et al.

2010:157-158 κ.ά.) θεωρούν ότι τα ερωτήματα που συνθέτουν ένα τεστ, το οποίο αποτελείται από ερωτήσεις πολλαπλής επιλογής ή παραλλαγές τους, μπορεί να βελτιωθούν και με βάση την ανάλυση της συχνότητας με την οποία οι εξεταζόμενοι σημειώνουν τις παρεμβολές (distracters - μη ορθές εναλλακτικές απαντήσεις), που υπάρχουν σε τέτοιου είδους ερωτήματα.

Οι παρεμβολές θεωρούνται επιτυχείς: α) αν κάποιοι από τους εξεταζομένους τις σημειώνουν και β) αν ο αριθμός αυτών που τις σημειώνουν είναι σημαντικά μεγαλύτερος στην ομάδα με χαμηλή επίδοση απ' ό,τι μεταξύ εκείνων που ανήκουν στην ομάδα με υψηλή επίδοση. Αν οι συνθήκες αυτές δεν πληρούνται, τότε υπάρχει πρόβλημα. Αυτό σημαίνει ότι οι παρεμβολές πρέπει να επανεξεταστούν και, ενδεχομένως, να τροποποιηθούν. Αν π.χ. κάποια παρεμβολή δεν σημειώνεται από κανένα εξεταζόμενο, τότε αυτή δεν έχει λόγο ύπαρξης και πρέπει να αντικατασταθεί. Αν, επίσης, η συχνότητα εμφάνισης μιας παρεμβολής είναι μεγαλύτερη μεταξύ αυτών που έχουν υψηλές επιδόσεις στο τεστ σε σύγκριση με αυτούς που έχουν χαμηλές, τότε είναι πιθανόν να υπάρχει λάθος στη διατύπωση του ερωτήματος ή της παρεμβολής ή να συμβαίνει κάτι άλλο, το οποίο επιβάλλει την επανεξέταση του ερωτήματος.

Η ανάλυση των παρεμβολών μπορεί να γίνει συστηματικά στα οργανωμένα κέντρα παραγωγής τεστ. Είναι λιγότερο συχνή μεταξύ των εκπαιδευτικών που ετοιμάζουν τεστ για την τάξη τους, λόγω του χρόνου που απαιτείται, δεν πρέπει, όμως, να αγνοείται. Το συγκεκριμένο πρόβλημα, καθώς και εκείνο που σχετίζεται με τη γενικότερη ανάλυση των ερωτημάτων ενός τεστ αντιμετωπίζεται σήμερα με τη χρήση ειδικών προγραμμάτων ηλεκτρονικών υπολογιστών, τα οποία κάνουν ταχύτατα όλους τους αναγκαίους υπολογισμούς.[14]

9.10. Εκτίμηση της εγκυρότητας ενός εξεταστικού μέσου

Σε προηγούμενο κεφάλαιο ορίσαμε, σε γενικές γραμμές, την έννοια της εγκυρότητας των αποτελεσμάτων ενός τεστ και αναφέραμε τις σημαντικότερες από τις μορφές της. Στην παρούσα ενότητα θα ασχοληθούμε πιο συστηματικά με το ζήτημα αυτό και, κυρίως, με τη μεθοδολογία προσδιορισμού της, θέμα που δεν έχουμε προηγουμένως αναλύσει. Θα αναφερθούμε, ειδικότερα στις κυριότερες μορφές εγκυρότητας, όπως τις έχουμε ήδη προσδιορίσει, και συγκεκριμένα: α) στην **εγκυρότητα περιεχομένου**, β)

14. Βλ. στο διαδίκτυο τις εξής ιστοσελίδες: www.assess.com/Software/sItemTest.htm και Remark Office OMR: www.principiaproducts.com/office/index.html

στην **εγκυρότητα της εννοιολογικής κατασκευής** και γ) στην **εγκυρότητα με βάση κριτήρια** (περιλαμβάνεται και η προγνωστική εγκυρότητα).[15]

Παρά το γεγονός ότι δεν υπάρχει ομοφωνία μεταξύ των ειδικών ως προς το ποια μορφή εγκυρότητας είναι για κάθε περίπτωση αναγκαία, μπορεί να λεχθεί ότι η εγκυρότητα περιεχομένου είναι ιδιαιτέρως, απαραίτητη για τα τεστ σχολικής επίδοσης, η προγνωστική εγκυρότητα για τις διαδικασίες επιλογής και η εγκυρότητα εννοιολογικής κατασκευής για τη μέτρηση ψυχολογικών ή άλλων χαρακτηριστικών που προϋποθέτουν συγκεκριμένο θεωρητικό υπόβαθρο. Στις περισσότερες, όμως, περιπτώσεις διενέργειας αξιολογικών δοκιμασιών κρίσιμης σημασίας ή σε επιστημονικές έρευνες που σχετίζονται με θέματα τόσο της Ψυχολογίας όσο και της Παιδαγωγικής είναι αναγκαία η ύπαρξη περισσότερων της μιας από τις μορφές εγκυρότητας που σημειώθηκαν παραπάνω. Εξάλλου, σε προηγούμενο κεφάλαιο αναφέραμε ότι τα τελευταία χρόνια η εγκυρότητα προσεγγίζεται ως ενιαία έννοια, και τα διάφορα είδη της αποτελούν απλώς διαφορετικές όψεις της τεκμηρίωσής της. Η γνώση, επομένως, των μεθόδων, με βάση τις οποίες προσδιορίζονται όλα τα επιμέρους είδη εγκυρότητας είναι αναγκαία σε όλους όσοι ασχολούνται με τα θέματα αυτά. Την παραπάνω ανάγκη επιδιώκουν να ικανοποιήσουν όσα εκτίθενται στη συνέχεια.[16]

Πριν, όμως, εξετάσουμε τις επιμέρους μορφές εγκυρότητας, θεωρούμε απαραίτητο να υπογραμμίσουμε τα ακόλουθα:

Η εγκυρότητα καθορίζεται για τη συγκεκριμένη χρήση ή ερμηνεία των αποτελεσμάτων ενός τεστ και δεν έχει γενικό χαρακτήρα. Αφορά στην κατηγορία των ατόμων στα οποία δόθηκε το τεστ και **μπορεί να έχει διαβαθμίσεις** (χαμηλή, μεσαία, υψηλή εγκυρότητα). Η κρίση της εγκυρότητας γίνεται με σαφή, προκαθορισμένα σε κάθε περίπτωση κριτήρια και προκύπτει από τη συνολική αξιολόγηση ενός τεστ.

9.10.1. *Εγκυρότητα περιεχομένου*

Υπενθυμίζουμε στον αναγνώστη ότι η εγκυρότητα αυτή σχετίζεται με το βαθμό στον οποίο η επίδοση σε μια εξεταστική δοκιμασία εκφράζει το

15. Αναλυτικότερη κατηγοριοποίηση των διαφόρων μορφών εγκυρότητας βλ. στο Ebel (1972: 437-438). Μερικές από τις κατηγορίες αυτές αλληλοσχετίζονται. Άλλοι συγγραφείς ταξινομούν τις κατηγορίες εγκυρότητας σε δύο ευρύτερες ομάδες: α) τις άμεσες και β) τις παράγωγες (Thorndike & Hangen, 1955: 109-110· Ebel, 1972: 438).

16. Η ανάλυση που ακολουθεί διατηρεί την παραδοσιακή διάκριση των ειδών εγκυρότητας, υπογραμμίζουμε, όμως, το συμπληρωματικό τους χαρακτήρα.

πραγματικό επίπεδο κατοχής εκ μέρους των εξεταζομένων ορισμένου τομέα π.χ. ενός γνωστικού αντικειμένου ή τμήματός του. Υπό την έννοια αυτή, η εγκυρότητα περιεχομένου μιας εξέτασης είναι συνάρτηση της σχέσης των ερωτημάτων που περιέχει με την ύλη στην οποία αναφέρεται και τους στόχους που επιδιώκει.[17] Οι επιδιωκόμενοι στόχοι είναι δυνατόν να θεωρούνται ισότιμοι ή να διαφοροποιούνται από πλευράς βαρύτητας. Στην τελευταία περίπτωση διευκρινίζεται η βαρύτητα που έχει ο κάθε στόχος. Το ίδιο μπορεί να ισχύσει και για την εξεταστέα ύλη.

Οι διαδικασίες ελέγχου της εγκυρότητας περιεχομένου, οι οποίες εφαρμόζονται κατά την προετοιμασία ενός τεστ, έχουν ποιοτικό, κυρίως, χαρακτήρα. Σε περίπτωση που πρόκειται για τεστ-δασκάλου, η παραπάνω εγκυρότητα κρίνεται από τον κάθε διδάσκοντα. Όπου είναι δυνατόν μπορεί να ζητείται, συμβουλευτικά, και η άποψη εμπειροτέρων συναδέλφων. Χρήσιμα, από πρακτική άποψη, για την επίτευξη της εγκυρότητας περιεχομένου των εξεταστικών δοκιμασιών, που εκπονούνται από μεμονωμένους διδάσκοντες, είναι τα ακόλουθα:

Για να επιτευχθεί η ζητούμενη εγκυρότητα περιεχομένου, μπορεί να εφαρμοσθεί η εξής διαδικασία:[18] α) Καταγράφονται αρχικά οι στόχοι που επιδιώχτηκαν κατά τη διδασκαλία ορισμένης ύλης. Η καταγραφή των στόχων αυτών ενδείκνυται να γίνεται κατά διδακτική ενότητα. β) Δίπλα από κάθε κατηγορία στόχων προσδιορίζεται επί τοις % η βαρύτητα που θα δοθεί στον έλεγχο της πραγματοποίησής τους. Η βαρύτητα αυτή δεν θα πρέπει να απέχει πολύ από κείνη που δόθηκε σε κάθε κατηγορία στόχων κατά τη διδασκαλία της αντίστοιχης ύλης. γ) Προσδιορίζεται, επίσης, ο χρόνος που έχει στη διάθεσή του ο εξεταστής για την πραγματοποίηση της εξέτασης. Με βάση το διαθέσιμο χρόνο, κυρίως, αλλά συνυπολογίζοντας και τον παράγοντα όρια αντοχής ενός μέσου μαθητή, ο εξεταστής προχωρεί σε κατά προσέγγιση προσδιορισμό του αριθμού των ερωτήσεων που θα περιλαμβάνει η εξεταστική δοκιμασία.

17. Η κάλυψη στόχων που αντιστοιχούν σε γνωστικές δεξιότητες (π.χ. κατανόηση, ανάλυση, κριτική σκέψη, μαθηματική λογική κτλ.) μπορεί να θεωρηθεί ότι εμπίπτει και στην εγκυρότητα της εννοιολογικής κατασκευής, γεγονός που δείχνει τη σχέση μεταξύ των διαφόρων ειδών εγκυρότητας και ενισχύει τη θέση ότι η εγκυρότητα πρέπει να αντιμετωπίζεται κατά τρόπο ενιαίο. Τέλος, η εγκυρότητα περιεχομένου δεν πρέπει να συγχέεται με τη «φαινομενική εγκυρότητα» (face validity), η οποία αφορά στην εικόνα που σχηματίζει κάποιος από μια πρώτη επιφανειακή ανάγνωση ενός τεστ. Επειδή δεν θεωρούμε τη «φαινομενική εγκυρότητα» σημαντική, δεν ασχολούμαστε περαιτέρω μ' αυτήν.

18. Η διαδικασία αυτή ενδείκνυται για αθροιστικού χαρακτήρα εξετάσεις που καλύπτουν μεγάλα χρονικά διαστήματα.

Αν πρόκειται για την κατασκευή τεστ με ερωτήσεις αντικειμενικού τύπου, καλό είναι να γνωρίζει ο συντάκτης του ότι ένας μέσος μαθητής έχει τη δυνατότητα να συμπληρώσει, μέσα σε μια ώρα, περισσότερες από 100 ερωτήσεις του τύπου: «σωστό – λάθος» ή γύρω στις 60 ερωτήσεις πολλαπλής επιλογής μεσαίας δυσκολίας (Ebel, 1972· Ζαβλανός, 1978: 58).

Στην περίπτωση ερωτήσεων ανάπτυξης, ο εξεταστής πρέπει να υπολογίσει το χρόνο που απαιτεί η ανάπτυξη της καθεμιάς, αν και αυτό είναι εξαιρετικά αβέβαιο, γιατί κατά την απάντηση σε ερωτήσεις ανάπτυξης υπεισέρχονται πολλοί παράγοντες σχετικοί με τις ικανότητες, τις γνώσεις, την ταχύτητα γραφής του εξεταζομένου κ.ά.

Ο τελικός αριθμός των ερωτήσεων κατανέμεται ποσοτικά ανάλογα με τη βαρύτητα που δίδεται σε κάθε κατηγορία και υποκατηγορία στόχων, όπως φαίνεται στον πίνακα 30, ο οποίος παρουσιάζει την κατανομή 120 ερωτήσεων αντικειμενικού τύπου κατά στόχους, ανάλογα με τη σπουδαιότητα που δόθηκε σ' αυτούς σε μια υποθετική διδασκαλία ορισμένου μαθήματος.

Πίνακας 30

Ποσοτική κατανομή κατά κατηγορίες διδακτικών στόχων 120 αντικειμενικού τύπου ερωτήσεων

Κατηγορίες / Υποκατηγορίες στόχων	Ποσοστό της συνολικής προσπάθειας που αφιερώθηκε σε κάθε κατηγορία και υποκατηγορία στόχων κατά τη διδασκαλία. Κατηγ.	Υποκατηγ.	Ποσοστό ερωτήσεων κατά κατηγορία και υποκατηγορία στόχων. Κατηγ.	Υποκατηγ.	Αριθμός ερωτήσεων κατά κατηγορία και υποκατηγορία διδακτικών στόχων. Κατηγ.	Υποκατηγ.
1. Απόκτηση γνώσεων	30%		30%		36	
α. Γνώση επιμέρους δεδομένων		15%		15%		18
β. Γνώση μέσων που επιτρέπουν χρήση επί μέρους δεδομένων		10%		10%		12
γ. Γνώση αφηρημένων εννοιών και γενικών αρχών		5%		5%		6
2. Κατανόηση	15%		15%		18	
α. Μεταφορά		5%		5%		6
β. Ερμηνεία		5%		5%		6
γ. Προέκταση		5%		5%		6

(συνέχεια)

3. Εφαρμογή	20%	20%	24
4. Ανάλυση	10%	10%	12
α. Ανάλυση στοιχείων	3%	3%	3
β. Ανάλυση σχέσεων	3%	3%	4
γ. Ανάλυση οργανωτικών αρχών	4%	4%	5
5. Σύνθεση*	10%	10%	12
α. Παραγωγή προσωπικής δημιουργίας	4%	4%	5
β. Εκπόνηση ενός σχεδίου δράσεως	3%	3%	4
γ. Παραγωγή σειράς αφαιρετικών σχέσεων	3%	3%	3
6. Αξιολόγηση	15%	15%	18
α. Με βάση εσωτερικά κριτήρια	7%	7%	8
β. Με βάση εξωτερικά κριτήρια	8%	8%	10

*Για τη σύνθεση χρησιμοποιούνται ανοικτές ερωτήσεις περιορισμένης σε έκταση απάντησης.

Το αμέσως επόμενο κριτήριο που πρέπει να ληφθεί υπόψη κατά την εκπόνηση των ερωτήσεων, οι οποίες πρόκειται να ενταχθούν σε μια εξεταστική δοκιμασία είναι ο βαθμός κάλυψης της ύλης. Όσο περισσότερο καλύπτεται η εξεταστέα ύλη τόσο πιο έγκυρο, από άποψη περιεχομένου, είναι το εξεταστικό αποτέλεσμα. Βεβαίως, όλη η εξεταστέα ύλη δεν είναι δυνατόν να καλυφθεί. Για το λόγο αυτό επιδιώκεται η εκπόνηση ερωτήσεων που να είναι κατά το δυνατόν αντιπροσωπευτικές της εξεταστέας ύλης. Η απάντηση των εξεταζομένων σε πολύ λίγες ερωτήσεις, έστω και αν αυτές έχουν ληφθεί από διαφορετικά σημεία της αντίστοιχης ύλης, δεν μπορεί να δείξει, κατά τρόπο έγκυρο, σε ποιο βαθμό ο εξεταζόμενος την κατέχει.

Για την εκτίμηση της εγκυρότητας περιεχομένου μιας δοκιμασίας είναι, ακόμη, δυνατόν να εφαρμόζονται και άλλες διαδικασίες (Crocker & Algina, 1986: 218-224), οι οποίες οδηγούν σε ποσοτικούς δείκτες, όπως είναι π.χ. ο υπολογισμός από κριτές του ποσοστού των ερωτημάτων που κρίνονται σχετικές με την εξεταζόμενη ύλη ή με τους αντίστοιχους στόχους, ο υπολογισμός της συσχέτισης μεταξύ της βαρύτητας των στόχων και της βαρύτητας που θεωρούν οι κριτές ότι έχουν τα αντίστοιχα ερωτήματα, καθώς και άλλοι ποσοτικοί δείκτες που υπολογίζονται με βάση ειδικούς μαθηματικούς τύπους (για περισσότερες πληροφορίες βλ. τις εργασίες: Rovinelli & Hambleton, 1977· Hambleton, 1980). Για τα σταθμισμένα

τεστ, τα οποία πρόκειται να χρησιμοποιηθούν σε ευρεία κλίμακα, υπάρχουν συγκεκριμένες προδιαγραφές με βάση τις οποίες ελέγχεται η εγκυρότητά τους.

Θεωρούμε αυτονόητο ότι για την εξασφάλιση της επιθυμητής εγκυρότητας είναι απαραίτητη η τήρηση των αρχών σύνταξης των ερωτήσεων που εκτέθηκαν σε άλλο κεφάλαιο, τις οποίες δεν θεωρούμε απαραίτητο να επαναλάβουμε. Επιπρόσθετα, πρέπει να διασφαλίζονται όλες οι προϋποθέσεις που είναι απαραίτητες για τη διενέργεια μιας καλής εξέτασης στις οποίες αναφερόμαστε σε άλλο κεφάλαιο (βλ. Κεφάλαιο ΙΑ΄).

9.10.2. *Εγκυρότητα εννοιολογικής κατασκευής*

Η εγκυρότητα της εννοιολογικής κατασκευής, τον ορισμό της οποίας δώσαμε σε προηγούμενο κεφάλαιο, αφορά κυρίως ψυχολογικές δοκιμασίες. Έχει, όμως, εφαρμογές και σε εκπαιδευτικά τεστ, μέσω των οποίων επιδιώκεται να προσδιοριστεί έγκυρα ο βαθμός στον οποίο μια δοκιμασία μετρά ορισμένη νοητική δεξιότητα, την οποία η διδακτική διαδικασία επιδίωξε να καλλιεργήσει (π.χ. κατανόηση κειμένου, επικοινωνιακή δεξιότητα, κριτική σκέψη, αναλυτική ικανότητα, μαθηματική λογική, επίλυση προβλημάτων κτλ.). Τονίστηκε, ακόμη, σε προηγούμενο κεφάλαιο ότι τα μετρούμενα κατασκευαστικά χαρακτηριστικά ενός τεστ, τα οποία μπορεί να είναι περισσότερα του ενός, αποτελούν προϊόντα θεωρητικών συλλήψεων και στοχεύουν σε ερμηνευτικές γενικεύσεις, οι οποίες προχωρούν πέραν της διαπίστωσης της απλής αριθμητικής επίδοσης σε ορισμένο τεστ. Ενίοτε, η εγκυρότητα αυτού του είδους είναι αναγκαίος όρος και για τα διάφορα ερωτηματολόγια που χρησιμοποιούνται για έρευνες στον τομέα των Κοινωνικών Επιστημών, αφού το πόσο έγκυρα είναι τα στοιχεία που συγκεντρώνονται με τον τρόπο αυτό συνδέεται με την εννοιολογική κατασκευή στην οποία βασίζονται τα μέσα που χρησιμοποιούνται για τη συλλογή τους.

Ιδιαιτέρως εξετάζεται, κατά τον έλεγχο της υπό εξέταση εγκυρότητας, αν στα αποτελέσματα ενός τεστ υπεισέρχονται άλλοι παράγοντες, εκτός αυτών στους οποίους αναφέρεται το τεστ, οι οποίοι είναι δυνατόν να τα αλλοιώνουν. Είναι προφανές ότι ο έλεγχος αυτός είναι κρίσιμης σημασίας για τη σωστή ερμηνεία και την ενδεδειγμένη χρήση των αποτελεσμάτων μιας εξέτασης, λόγος για τον οποίο δίδεται σ' αυτό ιδιαίτερη σημασία τόσο κατά τη φάση της προετοιμασίας ενός εξεταστικού μέσου, όσο και κατά φάση της χορήγησής του. Βασική προϋπόθεση για τον έλεγχο αυτό είναι ο σαφής ορισμός της μεταβλητής ή των μεταβλητών που μετρά ένα τεστ, ώστε να είναι εύκολη η διάκρισή της/τους από άλλους παρεμφερείς παράγοντες.

Η εγκυρότητα της εννοιολογικής κατασκευής απαιτεί περισσότερα αποδεικτικά στοιχεία απ' ό,τι οι λοιπές μορφές εγκυρότητας. Οι κυριότερες διαδικασίες για τον έλεγχό της είναι:

α) Ο **υπολογισμός της συνάφειας** μεταξύ των αποτελεσμάτων της δοκιμασίας, που ελέγχεται ως προς την εγκυρότητά της, και άλλων δεδομένων που, θεωρητικά, μπορούμε να ισχυριστούμε ότι έχουν στενή συγγένεια μεταξύ τους, όπως είναι π.χ. η νοημοσύνη και οι σχολικές επιδόσεις, αν και δεν υπάρχει απόλυτη μεταξύ τους συσχέτιση.

Όπου είναι δυνατόν τα αποτελέσματα, που θεωρητικά μετρούν ορισμένη διάσταση, συγκρίνονται με τα δεδομένα, τα οποία προκύπτουν από άλλα τεστ που αναφέρονται στα ίδια ή σε παρόμοια πεδία (π.χ. παλαιότερες δοκιμασίες σχετικές με τον τομέα στον οποίο εμπίπτει η ελεγχόμενη δοκιμασία).

Συχνά χρησιμοποιείται ως στατιστική τεχνική στις περιπτώσεις που αναφέρθηκαν προηγουμένως η ανάλυση της παλινδρόμησης, η οποία επιτρέπει τη μέτρηση της επίδρασης της ελεγχόμενης μεταβλητής σε σχέση με αυτή άλλων παραγόντων (Linn & Gronlund, 2000: 82-85). Άλλες στατιστικές τεχνικές έχουν, επίσης, προταθεί από διάφορους συγγραφείς (Darlington, 1970).

β) Η **διαφοροποίηση μεταξύ ομάδων**. Η τεχνική αυτή εφαρμόζεται, όταν ζητείται ο προσδιορισμός της διαφοράς μεταξύ ομάδων στις οποίες έχει γίνει παρέμβαση, με στόχο τη διαφοροποίησή τους, ή όταν πρόκειται για αξιολόγηση χαρακτηριστικών που είναι από τη φύση τους διαφορετικά μεταξύ των ομάδων (π.χ. ομάδες που έχουν και δεν έχουν δεχθεί ενισχυτική διδασκαλία ή άλλης μορφής αντισταθμιστική εκπαίδευση). Η μη εύρεση διαφοράς στις περιπτώσεις αυτές μπορεί να οφείλεται όχι μόνο σε αποτυχία των σχετικών παρεμβάσεων αλλά και σε θεωρητικές αδυναμίες, με βάση τις οποίες διαμορφώθηκε το μέσο αξιολόγησης.

γ) Η **ανάλυση παραγόντων**. Με τη στατιστική αυτή τεχνική ελέγχεται αν τα ερωτήματα ενός αξιολογικού μέσου ομαδοποιούνται σε παράγοντες, κατά τρόπο, σύμφωνο με τη θεωρία στην οποία βασίζεται η κατασκευή του. Είναι, ακόμη, δυνατόν να διαμορφώνεται, με τη βοήθεια της ανάλυσης παραγόντων, μια μήτρα συσχετίσεων για ένα σύνολο Ν διαφορετικών τεστ ή μετρήσεων, με στόχο τον καθορισμό του βαθμού στον οποίο οι συσχετίσεις των δεδομένων μπορούν να αποδοθούν στη διακύμανση ενός ή περισσότερων κοινών παραγόντων (Crocker & Algina (1986: 232). Για τον τρόπο διενέργειας της ανάλυσης παραγόντων παραπέμπουμε στα βιβλία Στατιστικής Εφαρμοσμένης στις Κοινωνικές Επιστήμες, αρκετά από τα οποία έχουν κυκλοφορήσει τα τελευταία χρόνια και στη χώρα μας (βλ. βιβλιο-

γραφία). Ορισμένα, μάλιστα, έχουν και παραδείγματα τρόπων εφαρμογής στατιστικών προγραμμάτων με τη βοήθεια ηλεκτρονικών υπολογιστών.

δ) Η **πολυχαρακτηριστική-πολυμεθοδική μήτρα** (multitrait-multi-method matrix). Η μέθοδος αυτή προτάθηκε από τους Campbell & Fiske (1959) ως διαδικασία κατασκευής των τεστ, αλλά χρησιμοποιήθηκε και για τον έλεγχο της κατασκευαστικής εγκυρότητάς τους. Σύμφωνα με την τεχνική αυτή, καθορίζονται δύο τουλάχιστον διαφορετικοί τρόποι για τη μέτρηση του χαρακτηριστικού που μας ενδιαφέρει. Επιπρόσθετα, απαιτείται ο προσδιορισμός άλλων διακριτών χαρακτηριστικών που μπορούν να αξιολογηθούν με τις ίδιες μεθόδους και επιλέγεται ένα δείγμα υποκειμένων στα οποία εφαρμόζονται οι συγκεκριμένες μέθοδοι για την αξιολόγηση των χαρακτηριστικών αυτών. Στη συνέχεια υπολογίζονται οι συντελεστές συσχέτισης ανά ζεύγη μετρήσεων (ίδιο χαρακτηριστικό μετρούμενο με δύο διαφορετικές μεθόδους, δυο διαφορετικά χαρακτηριστικά μετρούμενα με την ίδια μέθοδο). Οι παραπάνω συντελεστές συσχέτισης αναγράφονται σε ένα πίνακα διπλής εισόδου (μέθοδοι-χαρακτηριστικά) και με βάση τη μήτρα που σχηματίζεται αποφαινόμαστε για την εγκυρότητα της εννοιολογικής κατασκευής των χρησιμοποιούμενων μέσων, εφόσον δεν παρατηρούνται σημαντικές αποκλίσεις. Οι Campbell και Fiske (1959) προτείνουν την οπτική θεώρηση της παραπάνω μήτρας, τρόπος που, όπως παρατηρούν οι Crocker και Algina (1986), μπορεί να αποδειχθεί προβληματικός, διότι δεν συνεκτιμάται το σφάλμα της δειγματοληψίας. Για το λόγο αυτό έχουν αναπτυχθεί άλλες μέθοδοι που επιτρέπουν ακριβέστερες και ασφαλέστερες εκτιμήσεις (Smitt, 1978· Lomax & Algina, 1979· Marsh & Hocevar, 1983, αναφέρονται στο Crocker & Algina, 1986: 234).

Πρέπει, ακόμη, να σημειώσουμε ότι και η εγκυρότητα της εννοιολογικής κατασκευής, όπως και οι λοιπές μορφές εγκυρότητας, μπορεί να έχει διαβαθμίσεις (χαμηλή, μεσαία, υψηλή εγκυρότητα), ανάλογα με τη βεβαιότητα που υπάρχει ως προς την εκτίμησή της.

9.10.3. *Εγκυρότητα βάσει συγκριτικού κριτηρίου*

Η εγκυρότητα στην περίπτωση αυτή μπορεί να είναι: α) **προγνωστική** ή β) **συγχρονική**. Όπως αναφέραμε σε προηγούμενο κεφάλαιο, μια δοκιμασία έχει προγνωστική εγκυρότητα, όταν οι μεταγενέστερες επιδόσεις των ίδιων ατόμων στον ίδιο τομέα επιβεβαιώνουν τις αρχικές. Π.χ. οι επιδόσεις στην Α΄ τάξη του Λυκείου σχετίζονται σε μεγάλο βαθμό με αυτές της Γ΄ τάξης ή με εκείνες των πανελλαδικών εξετάσεων. Συγχρονική είναι η

διαδικασία εγκυρότητας, όταν η επίδοση σε ένα τεστ, επιβεβαιώνεται από την ταυτόχρονη επίδοση σε αντίστοιχη πρακτική εφαρμογή ή σε σύγκριση με άλλες επιδόσεις που θεωρούνται έγκυρες π.χ. με την προφορική βαθμολογία των εκπαιδευτικών ή με τα αποτελέσματα ενός άλλου τεστ, που χορηγείται ταυτόχρονα, για το οποίο υπάρχουν επαρκείς αποδείξεις εγκυρότητας.

Η διαδικασία ελέγχου της εγκυρότητας αυτής σε μεγάλης σημασίας τεστ ακολουθεί, κατά τους Crocker & Algina (1986: 224), τα παρακάτω βήματα:

1) Καθορισμός της σχετικής με το συγκεκριμένο κριτήριο συμπεριφοράς, καθώς και της μεθόδου μέτρησής της.
2) Επιλογή ενός δείγματος αντιπροσωπευτικού του πληθυσμού στον οποίο θα χρησιμοποιηθεί το τεστ.
3) Πιλοτική χορήγησή του στο δείγμα αυτό.
4) Εκτίμηση της επίδοσης που αφορά στο κριτήριο, το οποίο έχει επιλεγεί, όταν θα έχουν συγκεντρωθεί επαρκή δεδομένα.
5) Καθορισμός του βαθμού της σχέσης μεταξύ των επιδόσεων στο τεστ και της εξεταζόμενης μεταβλητής-κριτηρίου.

Η διαδικασία, πάντως, αυτή δεν είναι απαλλαγμένη προβλημάτων, τα οποία μπορεί να σχετίζονται είτε με την εύρεση του κατάλληλου κριτηρίου, είτε με την αντιπροσωπευτικότητα του δείγματος είτε με την αξιοπιστία των μετρήσεων που αφορούν στο κριτήριο.

Για την εύρεση του βαθμού προγνωστικής ή συγχρονικής εγκυρότητας είναι δυνατόν να χρησιμοποιηθούν διάφοροι μέθοδοι υπολογισμού του συντελεστή συσχέτισης μεταξύ των αποτελεσμάτων ενός τεστ και των επιδόσεων των ίδιων υποκειμένων σ' ό,τι έχει καθοριστεί ως κριτήριο. Η επιλογή του είδους του συντελεστή συσχέτισης εξαρτάται από τη μορφή των διαθέσιμων δεδομένων (ποσοτικά συνεχή δεδομένα, δεδομένα που εκφράζουν σειρά κατάταξης, ποιοτικά δεδομένα κτλ.).

Αν ο συντελεστής συσχέτισης υπολογίζεται μεταξύ των επιδόσεων των εξεταζομένων που προκύπτουν από δύο τεστ, τότε γίνεται λόγος για συντελεστή εγκυρότητας, το τετράγωνο του οποίου αποτελεί το **συντελεστή επιβεβαίωσης**. Έστω ότι θεωρούμε πώς η επίδοση στα Μαθηματικά μπορεί να προδιαγνώσει την επίδοση στη Φυσική. Ο βαθμός συνάφειας των αποτελεσμάτων σε δύο αντίστοιχα τεστ βρέθηκε ότι είναι 0.85, τότε ο συντελεστής επιβεβαίωσης θα είναι 0.72, που σημαίνει ότι το 72% της διακύμανσης της επίδοσης στη Φυσική είναι δυνατόν να ερμηνευθεί με βάση τη διακύμανση της επίδοσης στα Μαθηματικά. Αυτό σημαίνει ότι η επίδοση στα Μαθηματικά μπορεί να θεωρηθεί ότι έχει υψηλή προγνωστική εγκυρότητα για την επίδοση στη Φυσική και αντίστροφα.

Χρήσιμο είναι να αναφέρεται στις περιπτώσεις αυτές και το τυπικό σφάλμα μέτρησης ($\sigma_{\chi\cdot\gamma}$), το οποίο δίνεται από τον τύπο:

$$\sigma_{\chi\gamma} = \sigma_{\gamma}\sqrt{1 - r_{\chi\gamma}^{2}}$$

όπου σ_{γ} το τυπικό σφάλμα μέτρησης του συγκριτικού κριτηρίου και $r_{\chi\cdot\gamma}$ ο συντελεστής συσχέτισης μεταξύ των δύο επιδόσεων.

Η εγκυρότητα βάσει συγκριτικού κριτηρίου δεν χρησιμοποιείται μόνο για λόγους πρόγνωσης. Εφαρμόζεται και σε άλλες περιπτώσεις, όπως είναι π.χ. η υποκατάσταση ενός κοπιώδους ή δαπανηρού τρόπου εξέτασης από κάποιο άλλο, που απαιτεί λιγότερο κόπο ή λιγότερες δαπάνες και εξασφαλίζει εξίσου έγκυρο αποτέλεσμα.

Για την ερμηνεία των συντελεστών συσχέτισης μεταξύ των αποτελεσμάτων μιας δοκιμασίας και εκείνων που προέρχονται από τη δοκιμασία που χρησιμεύει ως κριτήριο πρέπει να συνεκτιμώνται ποικίλοι παράγοντες, όπως είναι: α) ο βαθμός ομοιότητας των δύο δοκιμασιών, β) το εύρος διασποράς των δεδομένων, γ) η σταθερότητα των μετρήσεων και δ) ο χρόνος που μεσολαβεί μεταξύ των δύο δοκιμασιών (Linn & Gronlund, 2000: 94).

9.10.4. *Θεώρηση της εγκυρότητας μιας εξεταστικής διαδικασίας από την άποψη των συνεπειών της*

Σε προηγούμενο κεφάλαιο κάναμε νύξη σχετική με το γεγονός ότι ορισμένοι ειδικοί σε θέματα εκπαιδευτικής αξιολόγησης (π.χ. Messick, 1989, 1994· Linn & Gronlund, 2000: 97-99) προσεγγίζουν τα ζητήματα της εγκυρότητας μιας εξεταστικής διαδικασίας και από την άποψη των συνεπειών που η ερμηνεία και η χρήση των αποτελεσμάτων της μπορεί να έχει στους εμπλεκόμενους (consequential validity). Συμπληρώνοντας τη συνοπτική αυτή αναφορά, σημειώνουμε τα εξής:

Στόχος της αξιολόγησης, ιδιαίτερα της σύγχρονης μορφής της, είναι να λειτουργήσει θετικά τόσο για τους διδάσκοντες όσο και για τους διδασκόμενους, βοηθώντας τους πρώτους να γίνουν πιο αποτελεσματικοί στο έργο τους και τους δεύτερους να αποκτήσουν συνείδηση των ικανοτήτων τους, να εντείνουν την προσπάθειά τους, να αποκτήσουν μαθησιακά κίνητρα και να αναπτυχθούν σφαιρικά ως προσωπικότητες. Αυτό σημαίνει ότι η αξιολόγηση πρέπει να λειτουργεί, όπως επανειλημμένα έχουμε τονίσει, σε αρμονία με τη διδακτική πράξη, την οποία όχι μόνον πρέπει να υποβοηθεί αλλά οφείλει και να συμπληρώνει. Αν μια αξιολογική διαδικασία λειτουργεί έτσι, τότε έχει θετική «**εφαρμοστική**» **εγκυρότητα**. Αν, όμως, δημιουργεί στρεβλώσεις με το να προκαλεί υπέρμετρη αγωνία στους μαθητές, να οδηγεί σε αποθάρρυνση μεγάλο ποσοστό διδασκομένων ή να περιορίζει τους στόχους της διδασκαλίας και το περιεχόμενό της μόνο σε ό,τι γίνεται αντικείμενο εξέτασης, με αποτέλεσμα να μην κατακτάται επαρκώς ευρύτερο φάσμα γνώσεων και να μην αναπτύσσεται πολυμορφία δεξιοτήτων, τότε έχει αρνητική εγκυρότητα εφαρμογής. Το στοιχείο αυτό οφείλουν να έχουν υπόψη τους τόσο οι εκπαιδευτικοί όσο και οι συντάκτες των διαφόρων τεστ και να λαμβάνουν μέτρα για την αποφυγή ενδεχόμενων αρνητικών συνεπειών που προκύπτουν από τη χρήση τους.

Τέτοια μέτρα μπορούν να είναι η παροχή κατάλληλων οδηγιών για την έγκυρη ερμηνεία και τη σωστή χρήση των αποτελεσμάτων ενός τεστ, η αποφυγή τεχνητής ενίσχυσης της αγωνίας των μαθητών για την επίδοσή τους σε επικείμενες εξεταστικές δοκιμασίες, η έμφαση στους γενικούς σκοπούς της διδασκαλίας, ανεξάρτητα από τα πιθανά θέματα των εξετάσεων, η αποφυγή εξεζητημένων λεπτομερειών κατά τις εξετάσεις, ο συνδυασμός μεθόδων αξιολόγησης στο πλαίσιο της συνεχούς ενδοσχολικής αξιολόγησης και άλλα παρόμοια. Τα ζητήματα αυτά βρίσκονται, συνεχώς, στην επικαιρότητα, τα τελευταία χρόνια, υπό την επίδραση του κινήματος της αυθεντικής αξιολόγησης, κυρίως.

9.10.5. *Αρνητικές επιδράσεις στην εγκυρότητα ενός εξεταστικού μέσου*

Ποικίλοι παράγοντες είναι δυνατόν να επηρεάζουν αρνητικά την εγκυρότητα ενός εξεταστικού μέσου. Αυτοί μπορεί να σχετίζονται: α) με το περιεχόμενο και τη δομή της δοκιμασίας (ακατάλληλα ερωτήματα, μονόπλευρη έμφαση σε ορισμένα μόνο ζητήματα, ακατάλληλη διευθέτηση των απαντήσεων σε ερωτήσεις κλειστού τύπου, ανεπαρκής αποσαφήνιση της «κατασκευαστικής έννοιας» στην οποία στηρίζεται, μεγάλη συντομία του τεστ και άλλα παρόμοια), β) με τον τρόπο χορήγησης ενός τεστ (ανεπαρκείς οδηγίες, περιορισμένη χρονική διάρκεια εξέτασης, ακατάλληλες κλιματολογικές συνθήκες στο χώρο εξέτασης κ.ά.), γ) με τη διόρθωση και τη βαθμολογία της δοκιμασίας (ασαφείς ή ανεπαρκείς οδηγίες διόρθωσης, παραλείψεις επισήμανσης σφαλμάτων κτλ.) και δ) με προσωπικούς παράγοντες των εξεταζομένων (υπέρμετρη αγωνία και άγχος, προβλήματα υγείας, ειδικές ανάγκες των ατόμων). Έχοντας όλα αυτά υπόψη, οι διδάσκοντες, οι συντάκτες τεστ, καθώς και όσοι χορηγούν ή βαθμολογούν τεστ οφείλουν να λαμβάνουν όλα τα αναγκαία μέτρα για να αποφεύγονται μειωμένης εγκυρότητας αποτελέσματα σχετικά με την αξιολόγηση των μαθητών.

9.11. Αξιοπιστία

Σε προηγούμενο κεφάλαιο τονίσαμε ότι η αξιοπιστία αφορά στη συνέπεια και στη σταθερότητα των αποτελεσμάτων ενός μέσου αξιολόγησης που επιτυγχάνονται σε διαφορετικές χρονικές φάσεις στο ίδιο δείγμα εξεταζομένων, εφόσον οι όροι εξέτασης δεν έχουν σημαντικά μεταβληθεί. Η αξιοπιστία των αποτελεσμάτων μιας δοκιμασίας αποτελεί, όπως και η εγκυρότητα, προϋπόθεση για τη σωστή ερμηνεία τους και κυρίως για την εξαγωγή γενικεύσιμων συμπερασμάτων. Εξάλλου, η εγκυρότητα των αποτελεσμάτων μιας εξέτασης συναρτάται, μαζί με άλλους παράγοντες, με το βαθμό αξιοπιστίας τους.

Στην πράξη, όμως, είναι αδύνατη η πλήρης ταύτιση των αποτελεσμάτων που επιτυγχάνονται σε διαφορετικές φάσεις εφαρμογής του ίδιου μέσου, έστω και αν το δείγμα των εξεταζομένων είναι το ίδιο. Τούτο οφείλεται στο γεγονός ότι οι συνθήκες και οι όροι υπό τους οποίους διεξάγονται οι αντίστοιχες εξετάσεις δεν ταυτίζονται απόλυτα, ακόμη κι αν έχει ληφθεί μέριμνα για την επίτευξη της αναγκαίας ομοιομορφίας. Ποικίλοι παράγοντες, που σχετίζονται τόσο με τα εξεταζόμενα άτομα όσο και με τις συνθήκες του περιβάλλοντος και τη φύση της εξέτασης, μπορούν να επηρεάζουν και, κατά συνέπεια, να διαφοροποιούν τα αντίστοιχα αποτελέσματα.

Με άλλα λόγια, μεταξύ διαδοχικών μετρήσεων με το ίδιο μέσο υπάρχει πάντα κάποιο σφάλμα ως προς το βαθμό στον οποίο προσεγγίζουν τη θεωρητική πραγματική επίδοση των μαθητών. Το ζητούμενο είναι να προσδιορισθεί το μέγεθος του σφάλματος αυτού. Αν το σφάλμα μεταξύ διαδοχικών μετρήσεων είναι πολύ μεγάλο, είναι επόμενο να διαφέρουν μεταξύ τους οι κατανομές των αντίστοιχων μετρήσεων, οπότε μπορεί κανείς να υποστηρίξει την άποψη ότι η εξεταστική δοκιμασία που εφαρμόζεται δεν παρέχει αξιόπιστα αποτελέσματα.[19]

Επιδίωξη, λοιπόν, είναι να κατασκευαστεί ένα εξεταστικό μέσο, στο οποίο το παραπάνω σφάλμα να είναι όσο γίνεται μικρότερο. Αυτό απαιτείται ιδιαίτερα σε εξεταστικές δοκιμασίες, με βάση τις οποίες λαμβάνο-

19. Ορισμένοι συγγραφείς (π.χ. Linn & Gronlund, 2000: 108) θεωρούν ότι είναι αρκετό να γίνεται λόγος για την αξιοπιστία των «σκορ» ενός τεστ παρά για το ίδιο το τεστ. Αυτή μας είναι ότι η αξιοπιστία των «σκορ» εξαρτάται σε σημαντικό βαθμό και από την καταλληλότητα του οργάνου μέτρησης και τη διασφάλιση των αναγκαίων όρων (ανπρόσπαστη κότητα, σαφήνεια, ακρίβεια ερωτημάτων, καταλληλότητα εναλλακτικών απαντήσεων σε κλειστού τύπου ερωτήσεις και άλλα παρόμοια). Επομένως, η αξιοπιστία αφορά και το όργανο μέτρησης.

νται σημαντικές εκπαιδευτικές αποφάσεις. Για να επιτευχθεί κάτι τέτοιο, πρέπει να ρυθμιστούν πολλά στοιχεία, όπως είναι το πλήθος των ερωτήσεων του τεστ, ο βαθμός δυσκολίας τους, ο χρόνος που διατίθεται για τη χορήγησή του, η σαφήνεια των οδηγιών εξέτασης, η ομοιομορφία στη διόρθωση και πολλά άλλα, από τα οποία εξαρτάται η αξιοπιστία του εξεταστικού μέσου. Ο δείκτης αξιοπιστίας ενός τεστ αποτελεί βασικό χαρακτηριστικό του και στα σταθμισμένα τεστ πρέπει να αναγράφεται, μαζί με τη μεθοδολογία υπολογισμού του.

Η αξιοπιστία είναι στατιστικό χαρακτηριστικό, το οποίο προκύπτει από τη συνάφεια των σχετικών δεδομένων. Δηλώνεται, συνήθως, με ένα δείκτη, οι τιμές του οποίου κυμαίνονται μεταξύ 0 και 1.0. Δείκτης που πλησιάζει προς το 0, σημαίνει ότι η αξιοπιστία του τεστ είναι ελάχιστη. Το αντίστροφο συμβαίνει, όταν η τιμή πλησιάζει προς 1.0. Ο δείκτης αυτός μπορεί να θεωρηθεί ότι εκφράζει την αναλογία της διακύμανσης των αποτελεσμάτων του τεστ που μπορεί να αποδοθεί στην πραγματική διακύμανση. Οι τρόποι υπολογισμού του δείκτη αυτού παρουσιάζονται παρακάτω:

9.11.1. *Επανάληψη της εξέτασης*

Κατά τη διαδικασία αυτή, επιδίδουμε το ίδιο τεστ στα ίδια άτομα σε διαφορετικά χρονικά διαστήματα και υπό παρόμοιες συνθήκες και υπολογίζουμε στη συνέχεια το βαθμό συνάφειας μεταξύ των αποτελεσμάτων που επιτυγχάνονται στις δύο διαδοχικές εξετάσεις. Όσο υψηλότερος είναι ο δείκτης συνάφειας, τόσο πιο αξιόπιστο είναι το χρησιμοποιούμενο τεστ.

Συνιστάται η ύπαρξη διαφορετικών αξιολογητών των αποτελεσμάτων κατά τις δύο εξετάσεις, για να μην υπάρχει επίδραση των πρώτων αποτελεσμάτων στη δεύτερη εξέταση, υπό τον όρο ότι παρέχονται οι ίδιες οδηγίες διόρθωσης και βαθμολόγησης και, γενικώς, καταβάλλεται προσπάθεια να μη διαφοροποιούνται σημαντικά οι συνθήκες εξέτασης μεταξύ πρώτης και δεύτερης φοράς.

Η μέθοδος, όμως, αυτή παρουσιάζει πολλές αδυναμίες και γι' αυτό αποφεύγεται. Οι σπουδαιότερες από αυτές οφείλονται στην επίδραση που μπορεί να έχει ο χρόνος στις γνώσεις των μαθητών, οπότε οι συνθήκες εξέτασης δεν είναι οι ίδιες στις δύο περιπτώσεις. Αν η χρονική απόσταση μεταξύ των δύο εξετάσεων είναι μεγάλη, είναι πιθανόν οι γνώσεις και οι δεξιότητες των εξεταζομένων να έχουν τροποποιηθεί. Αν πάλι οι δύο εξετάσεις γίνουν σε κοντινά χρονικά διαστήματα, είναι δυνατόν να θυμούνται οι μαθητές τις ερωτήσεις από την προηγούμενη εξέταση, οπότε το αποτέλεσμα της δεύτερης δοκιμασίας δεν είναι έγκυρο και συγκρίσιμο με το αποτέλεσμα της πρώτης.

9.11.2. *Κατασκευή ισοδύναμων τεστ*

Μια άλλη μέθοδος, η οποία προτείνεται για τον υπολογισμό της αξιοπιστίας ενός τεστ, που μπορεί να γίνει χωρίς την επίδοσή του δύο φορές, είναι η σύνταξη δύο διαφορετικών αλλά ισοδύναμων μορφών του. Η ισοδυναμία τους επιδιώκεται να επιτευχθεί με την κατά το δυνατόν εξίσωση των χαρακτηριστικών των δύο τεστ και του τρόπου σύνταξής τους (ίδιες οδηγίες σύνταξης, ίδιοι στόχοι, ίδια ύλη αναφοράς, ερωτήσεις ίδιου βαθμού δυσκολίας, ίδιοι συντάκτες κτλ.). Τα δύο παρόμοια τεστ επιδίδονται στην ίδια ομάδα και, κατόπιν, υπολογίζεται ο βαθμός συσχέτισης των επιδόσεών τους (δείκτης αξιοπιστίας), όπως και στην προηγούμενη περίπτωση.

Η μέθοδος αυτή, αν και χρησιμοποιείται συχνά στα σταθμισμένα τεστ, δεν είναι απαλλαγμένη αδυναμιών, όπως είναι π.χ. ο χρόνος, ο κόπος και το κόστος κατασκευής διπλών τεστ. Επιπρόσθετα, ούτε αυτή εγγυάται ότι η μετρούμενη επίδοση των εξεταζομένων θα παραμείνει σταθερή το επόμενο διάστημα. Άρα, αποτελεί και αυτή, όπως και οι λοιπές μέθοδοι, μια εκτίμηση της αξιοπιστίας σε συγκεκριμένη χρονική φάση, η οποία δεν μπορεί να γενικευθεί χωρίς επιφυλάξεις.

Εξυπακούεται ότι κατά τη χορήγηση παράλληλων μορφών του ίδιου τεστ λαμβάνονται όλα τα ενδεδειγμένα μέτρα για την πληρότητα και την εγκυρότητά της σχετικής διαδικασίας και διασφαλίζονται οι αναγκαίες συνθήκες για την ομοιόμορφη βαθμολόγησή τους.

9.11.3. *Διαίρεση του τεστ σε δύο ισοδύναμα μέρη*

Προκειμένου να αποφευχθεί η διπλή εξέταση των μαθητών με ένα ή δύο τεστ και να παρακαμφθούν οι δυσκολίες που παρουσιάζει, υιοθετείται, συνήθως, άλλη τακτική. Ολόκληρο το τεστ επιδίδεται μια μόνο φορά. Στη συνέχεια υπολογίζεται χωριστά το σύνολο των επιτυχών απαντήσεων στις ερωτήσεις με μονό αριθμό και χωριστά στις ερωτήσεις με ζυγό αριθμό και βρίσκεται ο δείκτης συνάφειας μεταξύ των δύο αυτών κατηγοριών για το σύνολο των μαθητών (Ebel, 1972: 412-413).

Ο τελικός δείκτης αξιοπιστίας του τεστ δίδεται από τον ακόλουθο τύπο των Spearman–Brown:

$$R = \frac{2r}{r+1}$$

όπου r ο συντελεστής συνάφειας μεταξύ των δύο μερών.

Έστω το ακόλουθο παράδειγμα. Τεστ 8 ερωτήσεων (I-VIII) δόθηκε σε 10 μαθητές (A-I). Τα αποτελέσματα σε κάθε ερώτηση σημειώνονται στον

πίνακα 31. Με 1 σημειώνεται η ορθή απάντηση και με 0 η εσφαλμένη. Να υπολογιστεί ο δείκτης αξιοπιστίας του τεστ με τη μέθοδο της διαίρεσης του σε δύο ισοδύναμα μέρη.

Πίνακας 31

Αποτελέσματα 10 μαθητών (A-I) σε 8 αντικειμενικού τύπου ερωτήσεις (I-VIII)

	I	II	III	IV	V	VI	VII	VIII	X	Υ	X^2	$Υ^2$	X Υ	X+Υ	$(X+Υ)^2$
A	1	1	1	1	0	1	0	1	2	4	4	16	8	6	36
B	0	0	1	0	0	0	1	0	2	0	4	0	0	2	4
Γ	0	0	0	1	0	1	0	0	0	2	0	4	0	2	4
Δ	1	0	0	1	1	0	0	1	2	2	4	4	4	4	16
E	1	1	1	1	0	0	0	0	2	2	4	4	4	4	16
ΣΤ	0	1	1	1	1	1	1	1	3	4	9	16	12	7	49
Z	0	1	0	0	0	0	0	1	0	2	0	4	0	2	4
H	1	0	0	1	1	1	0	1	2	3	4	9	6	5	25
Θ	0	1	1	1	1	1	1	1	3	4	9	16	12	7	49
I	1	0	1	0	0	1	1	1	3	2	9	4	6	5	25
p	0,5	0,5	0,6	0,7	0,4	0,6	0,4	0,7	Σχ 19	Σy 25	$Σχ^2$ 47	$Σy^2$ 77	Σχy 52	Σ(χ+y) 44	$Σ(χ+y)^2$ 228
q	0,5	0,5	0,4	0,3	0,6	0,4	0,6	0,3							
pq	0,25	0,25	0,24	0,21	0,24	0,24	0,24	0,21							

(Όπου X οι βαθμοί στις ερωτήσεις[20] με μονό αριθμό (1,3,5,7), y οι βαθμοί στις ερωτήσεις με ζυγό αριθμό (2,4,6,8), p η αναλογία των ορθών απαντήσεων σε κάθε ερώτηση και q= 1- p)

$$M = \frac{44}{10} = 4,4$$

$σ^2 = 3,44$

$Σpq = 1,88$

$rxy = 0,36$

Ο βαθμός συνάφειας rxy = 0,36 υπολογίστηκε με βάση το γνωστό τρόπο και δείχνει τη σχέση που υπάρχει μεταξύ των δύο τμημάτων του τεστ.

20. Ο μικρός αριθμός ερωτήσεων και ατόμων στο συγκεκριμένο παράδειγμα διευκολύνει τους αναγκαίους υπολογισμούς. Στην πράξη ο αριθμός αυτός είναι, συνήθως, πολύ μεγαλύτερος.

Στην πραγματικότητα πρόκειται για το δείκτη αξιοπιστίας του μισού τεστ. Ολόκληρο το τεστ θα έχει δείκτη αξιοπιστίας:

$$r = \frac{2 \cdot 0{,}36}{1+0{,}36} = \frac{0{,}72}{1{,}36} = 0{,}53$$

Αν το τεστ αυτό αυξηθεί σε μήκος σε σχέση προς το αρχικό, με την προσθήκη ισοδύναμων ερωτήσεων στις παλιές, τότε θα αυξηθεί και ο δείκτης αξιοπιστίας του. Στην περίπτωση αυτή, για να ευρεθεί ποιος θα είναι ο δείκτης του νέου αυτού τεστ, χρησιμοποιείται ο αρχικός τύπος των Spearman-Brown, ο οποίος έχει ως εξής:

$$r_n = \frac{nr_s}{(n-1)r_s + 1},$$

όπου n το μέγεθος του νέου τεστ σε σύγκριση με το αρχικό και r_s ο δείκτης αξιοπιστίας του (Ebel, 1972: 413).

Εάν π.χ. ένα συγκεκριμένο τεστ έχει δείκτη αξιοπιστίας 0.50 και αυξηθεί τρεις φορές σε μέγεθος με την προσθήκη νέων ερωτήσεων ισοδύναμων προς τις προηγούμενες, η αξιοπιστία του θα γίνει:

$$r_3 = \frac{3 \cdot 0.50}{(3-1) \cdot 0.50 + 1} = \frac{1.5}{2} = 0.75$$

Ο τύπος αυτός χρησιμεύει, συνήθως, για να προσδιοριστεί το μέγεθος ενός υπό εκπόνηση τεστ, προκειμένου να επιτευχθεί ο επιθυμητός συντελεστής αξιοπιστίας. Έστω π.χ. ότι τεστ 20 ερωτήσεων δίδει στο στάδιο της δοκιμαστικής εφαρμογής του συντελεστή αξιοπιστίας 0.30. Πόσο πρέπει να αυξηθεί για να έχει αξιοπιστία 0.60. Αντικαθιστούμε τις γνωστές τιμές στον προηγούμενο τύπο και έχουμε:

$$0.60 = \frac{n \cdot 0.30}{(n-1) \cdot 0.30 + 1}$$

Λύνοντας την εξίσωση ως προς n, βρίσκομε ότι το ζητούμενο τεστ πρέπει να είναι 3,5 φορές μεγαλύτερο του αρχικού (20 X 3,5 = 70 ερωτήσεις).

Άλλος συνήθης τρόπος υπολογισμού του δείκτη αξιοπιστίας με τη μέθοδο του χωρισμού του σε δύο ισοδύναμα τμήματα είναι εκείνος που προτάθηκε από το Rulon στα 1939 (Rulon, 1939). Ο τύπος του Roulon έχει ως εξής

$$r = 1 - \frac{\sigma_d^2}{\sigma_s^2}$$

όπου σ_d^2 η διακύμανση των διαφορών μεταξύ των βαθμολογιών των μονών

και των ζυγών ερωτήσεων[21] και σ_s^2 η συνολική διακύμανση των βαθμών όλων των ερωτήσεων του τεστ.

Εκτός από τους τρόπους που αναφέρθηκαν χρησιμοποιούνται και άλλοι τύποι, σπανιότερα όμως, για τον υπολογισμό της αξιοπιστίας ενός τεστ.[22]

9.11.4. *Η μέθοδος των Kuder-Richardson*

Οι Kuder και Richardson σε άρθρο τους, που δημοσίευσαν, το 1937, στο Psychometrika, περιέλαβαν διάφορους τύπους υπολογισμού της αξιοπιστίας ενός τεστ, από τους οποίους δυο έγιναν ευρύτερα αποδεκτοί και χρησιμοποιούνται πολύ συχνά από τους ειδικούς. Ο ένας από τους τύπους αυτούς, γνωστός ως **τύπος 20** (KR formula 20), έχει ως εξής (Ebel, 1972: 414):

$$r = \frac{K}{K-1}\left(1 - \frac{\Sigma pq}{\sigma^2}\right)$$

όπου K ο αριθμός των ερωτήσεων του τεστ, p η αναλογία των ορθών απαντήσεων σε κάθε ερώτηση, q = 1-p και σ^2 η ολική διακύμανση των βαθμών στο τεστ.

Ο τύπος είναι εφαρμόσιμος μόνο στα τεστ, στα οποία βαθμολογείται με ένα βαθμό η κάθε ορθή ερώτηση και με 0 η εσφαλμένη ή εκείνη που έχει παραλειφθεί.

Επαναλαμβάνουμε τον υπολογισμό του δείκτη αξιοπιστίας του τεστ του πίνακα 31. Κάτω από κάθε ερώτηση αναγράφεται η αναλογία (p) των ορθών απαντήσεων μεταξύ των 10 μαθητών, καθώς και η αντίστοιχη αναλογία των εσφαλμένων (q). Για κάθε ερώτηση ανευρίσκεται το γινόμενο pq και τελικά αθροίζονται τα γινόμενα αυτά. Το άθροισμα αυτών Σpq είναι 1.88 Η διακύμανση των τελικών βαθμών των 10 μαθητών είναι $\sigma^2 = 3.44$, ο αριθμός δε των ερωτήσεων (K) του τεστ είναι 8. Αντικαθιστούμε τα δεδομένα αυτά στον τύπο και βρίσκουμε:

$$r = \frac{8}{8-1} \cdot \left(1 - \frac{1.88}{3.44}\right) = 0.52$$

Συνάγεται, λοιπόν, ότι ο δείκτης αξιοπιστίας του τεστ αυτού είναι 0.52.

Εάν δεχθούμε ότι ο βαθμός δυσκολίας δεν διαφέρει πάρα πολύ μεταξύ των ερωτήσεων του τεστ, τότε μπορούμε να θεωρήσουμε ότι το πηλίκο:

21. Παραδείγματα εφαρμογής του τύπου αυτού βλέπε στο βιβλίο: Παπαϊωάννου (1977: 125-126).

22. Τους τύπους αυτούς βλέπε στο άρθρο του Thorndike (1966: 560-621). Βλ. ακόμη Μάνος (1980).

$$\frac{M(K-m)}{K\sigma^2}$$

(όπου Μ ο μέσος όρος των αποτελεσμάτων του τεστ και Κ ο αριθμός των ερωτήσεών του) προσεγγίζει το μέγεθος:

$$\frac{\Sigma pq}{\sigma^2}$$

τότε ο τύπος που αναφέραμε προηγουμένως μπορεί να γραφεί ως εξής (Ebel, 1972: 415):

$$r = \frac{K}{K-1}\left(1-\frac{M(K-M)}{K\sigma^2}\right)$$

Αν αντικαταστήσουμε στον τύπο αυτό τα δεδομένα του παραδείγματος του πίνακα 31, έχουμε:

$$r = \frac{8}{8-1}\cdot\left(1-\frac{4{,}4\cdot(8-4.4)}{8\cdot 3.44}\right)=0.48$$

9.11.5. *Ο συντελεστής αξιοπιστίας Cronbach' a*

Όταν οι υπάρχουσες ερωτήσεις δεν έχουν διχοτομικό χαρακτήρα, για το υπολογισμό της αξιοπιστίας ενός τεστ χρησιμοποιείται ο συντελεστής Cronbach'a. Οφείλει το όνομά του στον εισηγητή του, τον Cronbach, ο οποίος τον επινόησε το 1951, και δείχνει την εσωτερική συνοχή των ερωτημάτων ενός τεστ (internal consistency coefficient). Ο συντελεστής αυτός δίνεται από τον τύπο:

$$\text{Cronbach' a} = \frac{K}{K-1}\left(1-\frac{\Sigma^{K}_{i=1}\ \sigma^2_{yi}}{\sigma^2_{x}}\right)$$

όπου Κ ο αριθμός των ερωτήσεων (*K-items* or *testlets*), σ^2_x η ολική διακύμανση του συνόλου των ερωτήσεων ενός τεστ και σ^2_{yi} η διακύμανση του στοιχείου i για συγκεκριμένο δείγμα ατόμων.[23] Ο τύπος αυτός αποτελεί ουσιαστικά επέκταση του τύπου Κ-20 του Kuder-Richarson για διχοτομούμενες μετρήσεις (1/0).

Εναλλακτική μορφή του παραπάνω τύπου είναι η ακόλουθη:

$$\text{Cronbach'a} = \frac{Kc}{(\upsilon+(K-1)c)}$$

23. Βλ. http://en.wikipedia.org/wiki/Cronbach's_alpha

όπου Κ ο αριθμός των ερωτήσεων, υ η μέση διακύμανση (average variance) και c ο μέσος όρος των συνδιακυμάνσεων (covariances) μεταξύ των ερωτήσεων, ενώ ο σταθμισμένος συντελεστής Cronbach'a δίδεται από τον τύπο:

$$\text{σταθμισμένο Cronbach'a} = \frac{K\bar{r}}{(1+(K-1)\bar{r})}$$

όπου Κ ο αριθμός των ερωτήσεων και r̄ ο μέσος όρος των συντελεστών συνάφειας μεταξύ των ερωτημάτων ενός τεστ (inter-item correlation among the items).[24]

Σημειώνουμε ότι τόσο ο συντελεστής των Kuder-Richarson όσο και αυτός του Cronbach δεν ενδείκνυνται για τον υπολογισμό της αξιοπιστίας των τεστ ταχύτητας.

Τόσο ο συντελεστής Cronbach' a όσο και οι προηγούμενοι μπορούν σήμερα να υπολογίζονται πολύ εύκολα με τη χρήση στατιστικών προγραμμάτων που χρησιμοποιούνται με τη βοήθεια ηλεκτρονικών υπολογιστών.

Υψηλός συντελεστής α σημαίνει ότι τα ερωτήματα ενός τεστ σχετίζονται μεταξύ τους. Τιμές του α από 0.90 και άνω θεωρούνται εξαιρετικές, μεταξύ 0.80 και 0.89 καλές, μεταξύ 0.70 και 0.79 αποδεκτές, μεταξύ 0.60 έως 0.69 αμφισβητήσιμες, μεταξύ 0.50 και 0.59 χαμηλές και κάτω του 0.5 μη αποδεκτές.

Τέλος, πρέπει να αναφέρουμε ότι στους δείκτες, οι οποίοι προκύπτουν από τις μεθόδους που παρουσιάστηκαν προηγουμένως, δίνονται, συχνά, διαφορετικά ονόματα, που υποδηλώνουν τη σημασία τους, όπως: α) δείκτης (ή συντελεστής) **εμπιστοσύνης**, όταν εφαρμόζεται η μέθοδος επανάληψης της χορήγησης ενός τεστ, μέσα σε διάστημα λιγότερο των δύο μηνών, β) δείκτης **σταθερότητας**, όταν εφαρμόζεται ή ίδια μέθοδος αλλά το παραπάνω διάστημα είναι μεγαλύτερο των δύο μηνών, γ) **εσωτερικής συνέπειας ή συνοχής**, όταν χρησιμοποιείται η μέθοδος της διαίρεσης του τεστ σε δύο ίσα μέρη ή η μέθοδος Kuder-Richarson ή η μέθοδος Cronbach και γ) δείκτης **ισοδυναμίας**, όταν εφαρμόζεται η μέθοδος των παράλληλων ομάδων (κατασκευή ισοδύναμων τεστ) (Berrut, n.d).

24. Στη σχετική βιβλιογραφία αναφέρονται και άλλοι δείκτες αξιοπιστίας, όπως είναι π.χ. ο συντελεστής αξιοπιστίας του Hoyt, ο συντελεστής Tau του Woodbury κ.ά. Συγγενής είναι και ο συντελεστής ομοιογένειας του Lovenger. Περισσότερες λεπτομέρειες για τους συντελεστές αυτούς και τον τρόπο υπολογισμού τους βλ. στο Hashway (1998: 71-72).

9.11.6. Υπολογισμός της αξιοπιστίας μιας εξέτασης με ερωτήσεις ανάπτυξης

Με τη μέθοδο που εισηγήθηκαν οι Kuder-Richardson μπορεί, ακόμη, να υπολογισθεί η αξιοπιστία εξεταστικών μέσων παραδοσιακού τύπου, όπως είναι οι εξετάσεις στις οποίες οι μαθητές υποχρεούνται να δώσουν ελεύθερες απαντήσεις σε ορισμένο αριθμό ερωτήσεων. Αυτό ισχύει υπό την προϋπόθεση ότι κάθε ερώτηση βαθμολογείται χωριστά.

Ο τύπος που χρησιμοποιείται στην περίπτωση αυτή είναι ο ακόλουθος (Ebel, 1972: 419).

$$r = \frac{K}{K-1}\left(1-\frac{\Sigma\sigma_i^2}{\sigma_t^2}\right)$$

όπου Κ ο αριθμός των ερωτήσεων που περιλαμβάνει η εξέταση, $\Sigma\sigma_i^2$ το άθροισμα της διακύμανσης των βαθμών (χ) των ερωτήσεων και σ_τ^2 η διακύμανση του συνόλου των τελικών βαθμών.

Έστω ότι τρεις ερωτήσεις ανάπτυξης δόθηκαν σε 10 μαθητές[25] για εξέταση. Η καθεμιά από αυτές βαθμολογήθηκε με βάση την κλίμακα 1-5 ως εξής:

Πίνακας 32

Αποτελέσματα 10 μαθητών σε 3 ερωτήσεις ανάπτυξης που βαθμολογούνται από 1 έως 5

Μαθητές → / Ερωτήσεις ↓	Α	Β	Γ	Δ	Ε	ΣΤ	Ζ	Η	Θ	Ι	ΣX_i	$(\Sigma X)_i^2$
1	2	3	1	4	3	2	1	3	4	3	26	676
2	5	4	3	3	2	4	3	4	3	4	35	1225
3	1	2	3	3	2	3	2	3	2	2	23	529
Βαθμοί (X_t)	8	9	7	10	7	9	6	10	9	9	84	
$(X_t)^2$	64	81	49	100	49	81	36	100	81	81	722	

Σύμφωνα με τα παραπάνω δεδομένα θα έχουμε:

$\Sigma(\chi 2) = 264$ (το άθροισμα των τετραγώνων όλων των βαθμών)

$\Sigma(Xt) = 84$

$\Sigma(Xt)^2 = 722$

$\Sigma(Xi)^2 = (676 + 1225 + 529) = 2430$

Οπότε:

$$\sigma_t^2 = \frac{\Sigma X_t^2}{N} - \frac{(\Sigma X_t)^2}{N^2} = \frac{722}{10} - \frac{84^2}{10^2} = 1.64$$

$$\Sigma\sigma_t^2 = \frac{\Sigma(x^2)}{N} - \frac{\Sigma X_t^2}{N^2} = \frac{264}{10} - \frac{2430}{10^2} = 2.1$$

και ο δείκτης αξιοπιστίας θα είναι:

$$r = \frac{3}{3-1}\cdot\left(1-\frac{1.64}{2.1}\right) = 0.33$$

Η αξιοπιστία, όμως, στην περίπτωση ερωτήσεων ανάπτυξης δεν είναι σταθερή και επηρεάζεται πάρα πολύ από την υποκειμενικότητα της αξιολόγησης των απαντήσεων.

9.11.7. Η αξιοπιστία με βάση κριτές

Σε ορισμένες μορφές τεστ, όπως είναι π.χ. οι δοκιμασίες που αποτελούνται από ερωτήσεις ανάπτυξης, ο δείκτης αξιοπιστίας μπορεί να υπολογισθεί με βάση την αξιολόγηση δύο ή περισσότερων βαθμολογητών. Εκφράζεται δε από το δείκτη συνάφειας μεταξύ των βαθμών που δίδονται από τους παραπάνω κριτές στο σύνολο ενός τεστ ή στις επιμέρους ερωτήσεις του, σύμφωνα με όσα αναφέραμε στο περί Στατιστικής κεφάλαιο. Η διαδικασία αυτή δεν χρησιμοποιείται για τα τεστ που αποτελούνται από αντικειμενικού τύπου ερωτήσεις, επειδή θεωρείται ότι οι αποκλίσεις της βαθμολογίας μεταξύ διαφορετικών βαθμολογητών περιορίζονται στο ελάχιστον.

Μια πολύ απλή μέθοδος υπολογισμού της αξιοπιστίας της βαθμολογίας μεταξύ δύο βαθμολογητών είναι ο υπολογισμός του ποσοστού συμφωνίας τους στο σύνολο των αξιολογουμένων. Αν θεωρήσουμε ως αποδεκτή ορισμένη μεταξύ τους διαφορά (π.χ. 1 μονάδα στην κλίμακα 0-10), τότε στον υπολογισμό του ποσοστού αυτού εντάσσονται όλες οι περιπτώσεις για τις οποίες ισχύει η συνθήκη αυτή (Linn & Gronlund, 2000: 116-117).

9.11.8. Επιλογή του τρόπου υπολογισμού του δείκτη αξιοπιστίας ενός τεστ

Για την επιλογή μιας μεθόδου υπολογισμού του δείκτη αξιοπιστίας ενός τεστ πρέπει να λαμβάνονται υπόψη το είδος του τεστ, η μορφή των σχετικών ερωτημάτων, οι προσφερόμενες σε κάθε περίπτωση δυνατότητες, αλλά και το είδος του σφάλματος με το οποίο σχετίζεται ο κάθε τρόπος υπολογισμού του. Ο πίνακας 33 που ακολουθεί δείχνει τις πηγές σφάλματος που αντανακλά καθεμιά από τις μεθόδους, οι οποίες αναφέρθηκαν προηγουμένως.

25. Βλ. υποσημείωση 20.

Πίνακας 33

Πηγές του σφάλματος της διακύμανσης που συνδέονται με τον αντίστοιχο τρόπο υπολογισμού του δείκτη αξιοπιστίας

Τρόπος υπολογισμού της αξιοπιστίας	Σφάλμα διακύμανσης
1. Επανάληψη του τεστ.	1. Σφάλμα χρονικής δειγματοληψίας.
2. Εναλλακτικές μορφές του ίδιου τεστ: ταυτόχρονη χορήγηση.	2. Σφάλμα δειγματοληψίας περιεχομένου.
3. Εναλλακτικές μορφές του ίδιου τεστ: χορήγηση σε διαφορετικές φάσεις.	3. Σφάλμα και χρονικής δειγματοληψίας και περιεχομένου.
4. Διαίρεση του τεστ σε δύο ίσα μέρη.	4. Σφάλμα δειγματοληψίας περιεχομένου.
5. Τύπος KR-20 και Cronbach' a.	5. Σφάλμα δειγματοληψίας περιεχομένου και ετερογένειας ερωτημάτων.
6. Αξιοπιστία μεταξύ διαφορετικών αξιολογητών	6. Διαφορές που οφείλονται στους κριτές.

Πηγή: Reynolds et al. (2010: 105)

9.11.9. *Τυπικό σφάλμα μέτρησης*

Στην αρχή του υποκεφαλαίου περί αξιοπιστίας αναφέραμε ότι οι μετρήσεις που πραγματοποιούνται σε μια συγκεκριμένη περίπτωση εμπεριέχουν ορισμένο σφάλμα, σε σύγκριση προς τη θεωρητικά πραγματική μέτρηση που έπρεπε να επιτυγχάνεται με ένα συγκεκριμένο τεστ.

Το σφάλμα αυτό είναι αντιστρόφως ανάλογο προς την αξιοπιστία ενός τεστ. Αν, δηλαδή, η αξιοπιστία είναι μεγάλη, τότε το σφάλμα μέτρησης είναι μικρό, Αν αντίθετα η αξιοπιστία είναι μικρή, τότε το σφάλμα μέτρησης είναι μεγάλο.

Το σφάλμα αυτό, το οποίο ονομάζεται τυπικό σφάλμα μέτρησης (E), υπολογίζεται ως εξής:

$$E = \sigma \sqrt{1-r}$$

όπου σ η τυπική απόκλιση της κατανομής των βαθμών ενός τεστ και r ο δείκτης αξιοπιστίας του.

Έστω ότι η τυπική απόκλιση των βαθμών των μαθητών ενός σχολείου που εξετάστηκαν με ένα τεστ δείκτη αξιοπιστίας 0.80, είναι 2. Ποιο είναι το σφάλμα μέτρησης;

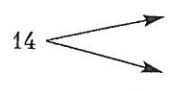
$$E = 2 \sqrt{1-0,80} \quad E = 0.89$$

Αν το ίδιο τεστ είχε αξιοπιστία 0,25, τότε το σφάλμα θα ήταν

$$E = 2 \sqrt{1-0,25} \quad E = 1.73$$

Παρατηρείται ότι η ελάττωση του δείκτη αξιοπιστίας από 0,80 σε 0,25 διπλασίασε σχεδόν το τυπικό σφάλμα μέτρησης.

Το τυπικό σφάλμα μέτρησης βοηθά σημαντικά στην ερμηνεία των βαθμών που επιτυγχάνονται σ' ένα τεστ ή σε μια συνηθισμένη εξέταση ερωτήσεων ανάπτυξης, προσδιορίζοντας τα όρια, μεταξύ των οποίων κείται ο πραγματικός βαθμός του κάθε μαθητή. Αν π.χ. στην πρώτη περίπτωση ένας μαθητής πήρε βαθμό 14, ο οποίος υποθέτομε ότι κείται σε απόσταση μιας τυπικής απόκλισης από το μέσο όρο, εμπεριέχεται δηλαδή στο 68% των δεδομένων, τα όρια εμπιστοσύνης του βαθμού αυτού θα είναι:



Ένας άλλος τρόπος για να εκφραστεί η ακρίβεια των βαθμών ενός τεστ είναι η χρησιμοποίηση του *πιθανού σφάλματος* μέτρησης. Το πιθανό σφάλμα μέτρησης είναι κάπως μικρότερο από το τυπικό σφάλμα μέτρησης. Στην πραγματικότητα είναι: το τυπικό σφάλμα μέτρησης πολλαπλασιαζόμενο επί 0.6745 (Ebel, 1972: 424).

9.11.10. *Παράγοντες που επηρεάζουν την αξιοπιστία*

Η αξιοπιστία των μετρήσεων μπορεί να επηρεάζεται, όπως συμβαίνει και με την εγκυρότητα, από ποικίλους παράγοντες. Ανάμεσα σ' αυτούς περιλαμβάνονται: α) Το πλήθος των ερωτημάτων στα οποία καλούνται να απαντήσουν οι εξεταζόμενοι. Όσο μεγαλύτερος είναι ο αριθμός τους, τόσο πιο αξιόπιστα είναι τα αποτελέσματα, κατά την κλασική θεωρία των μετρήσεων, επειδή με τον τρόπο αυτό το δείγμα των γνώσεων και δεξιοτήτων των μαθητών που ελέγχονται είναι πιο αντιπροσωπευτικό του αντίστοιχου τομέα. Το πλήθος, βέβαια, των ερωτημάτων μιας δοκιμασίας εξαρτάται και από άλλους παράγοντες, όπως είναι ο διαθέσιμος χρόνος, το κόστος της εξέτασης, η ηλικία των μαθητών, η βαρύτητα της εξέτασης και άλλα παρόμοια, στοιχεία που πρέπει να συνεκτιμούνται. β) Το εύρος της διασποράς των αποτελεσμάτων της εξέτασης. Εξετάσεις με μεγάλο εύρος αποτελεσμάτων έχουν μεγαλύτερο βαθμό αξιοπιστίας, παρά αυτές με μικρό εύ-

ρος επιδόσεων. γ) Η αντικειμενικότητα της αξιολόγησης των απαντήσεων. Όσο πιο αντικειμενικά αξιολογούνται αυτές, τόσο αυξάνεται η αξιοπιστία μιας δοκιμασίας.

9.11.11. Άλλος τρόπος υπολογισμού της αξιοπιστίας μιας εξέτασης με αποτελέσματα διχοτομικής μορφής

Όταν τα αποτελέσματα μιας εξέτασης εκφράζονται με διχοτομικό τρόπο, με βάση προκαθορισμένο επίπεδο επίδοσης (performance standard), όπως είναι π.χ. η περίπτωση της κατάταξης των εξετασθέντων σε αποτυχόντες/επιτυχόντες, τότε η αξιοπιστία τους μπορεί να εκτιμηθεί κατά τον ακόλουθο τρόπο (Linn & Gronlund, 2000: 129-131). Επαναλαμβάνεται η ίδια δοκιμασία μια δεύτερη φορά στο ίδιο δείγμα και υπολογίζεται το ποσοστό που αντιπροσωπεύουν τα άτομα εκείνα που και τις δύο φορές κατατάχθηκαν στην ίδια κατηγορία (πραγματικό ποσοστό). Το ποσοστό αυτό συγκρίνεται προς το πιθανό τυχαίο ποσοστό συμφωνίας, το οποίο υπολογίζεται ως ακολούθως: α) Πολλαπλασιάζεται η αναλογία αυτών που κατατάχθηκαν στην ανώτερη κατηγορία (επιτυχόντες) την πρώτη φορά επί την αναλογία αυτών που κατατάχθηκαν στην ίδια κατηγορία τη δεύτερη φορά. β) Γίνεται το ίδιο και γι' αυτούς που εντάχθηκαν στη χαμηλότερη κατηγορία (αποτυχόντες). γ) Προστίθενται τα δύο γινόμενα (α και β) και το αποτέλεσμα πολλαπλασιάζεται επί 100. Από το μέγεθος της διαφοράς πραγματικού και τυχαίου ποσοστού εκτιμάται η αξιοπιστία των αποτελεσμάτων της συγκεκριμένης δοκιμασίας. Αν π.χ. το πραγματικό ποσοστό υπερβαίνει κατά πολύ το τυχαίο, τότε η δοκιμασία θεωρείται πολύ αξιόπιστη.[26]

Η παραπάνω διαδικασία είναι δυνατόν να εφαρμόζεται στα τεστ-κριτήρια, με βάση τα οποία ελέγχεται αν οι μαθητές πέτυχαν τους επιδιωκόμενους διδακτικούς στόχους σε βαθμό που ανταποκρίνεται σε προκαθορισμένο επίπεδο επίδοσης. Εξυπακούεται ότι, αν τα αποτελέσματα των τεστ αυτών εκφράζονται και με αριθμητικές επιδόσεις ανά εξεταζόμενο, τότε μπορούν να εφαρμόζονται και οι τεχνικές ελέγχου της αξιοπιστίας, τις οποίες αναφέραμε προηγουμένως.

26. Υπάρχουν και τρόποι υπολογισμού της αξιοπιστίας στις περιπτώσεις αυτές που δεν απαιτούν επανάληψη της εξέτασης. Η παρουσίασή τους δεν περιλαμβάνεται στα περιεχόμενα του παρόντος βιβλίου. Θα αναλυθούν στην επόμενη έκδοσή του. Σ' αυτήν θα εξετασθούν, επίσης, οι κύριες διαφορές που υπάρχουν ως προς τις διαδικασίες μέτρησης της αξιοπιστίας των τεστ που αναφέρονται σε νόρμες και αυτών τα οποία βασίζονται σε κριτήρια (criterion referenced tests), ζήτημα που δεν καλύπτεται επαρκώς από το παρόν σύγγραμμα.

9.11.12. Σχέση διακριτικότητας και αξιοπιστίας

Όσο μεγαλύτερη είναι η τιμή του μέσου όρου των δεικτών διακριτικότητας τόσο μεγαλύτερη είναι η διασπορά της βαθμολογίας και, κατά συνέπεια, και η αξιοπιστία του τεστ, όπως φαίνεται στον ακόλουθο πίνακα που παρουσιάζει τη σχέση αυτή για ένα τεστ 100 ερωτήσεων.

Πίνακας 34

Σχέση διακριτικότητας και αξιοπιστίας[27]

Μέσος δείκτης διακριτικότητας	Τυπική απόκλιση βαθμών	Αξιοπιστία βαθμών
0.1225	5.0	0,00
0.16	6.53	0.42
0.20	8.16	0.63
0.30	12.25	0.84
0.40	16.32	0.915
0.50	20.40	0.949

Πηγή: Ebel (1972: 400)

Η σχέση, βέβαια, αυτή μεταξύ δείκτη διακριτικότητας και αξιοπιστίας των αποτελεσμάτων ισχύει μόνο αν ο αριθμός των ερωτήσεων που περιλαμβάνονται στα τεστ είναι μεγάλος.

Αν μια εξέταση περιλαμβάνει μικρό αριθμό ερωτήσεων ($N < 10$), η ύπαρξη υψηλού μέσου όρου διακριτικότητας δε σημαίνει, κατ' ανάγκη, και υψηλή αξιοπιστία. Ο πίνακας 35 δείχνει πόσο μεγάλος μπορεί να είναι ο μέσος όρος διακριτικότητας σε τεστ διαφόρων μεγεθών, των οποίων η αξιοπιστία είναι 0.

Η σχετικότητα αυτή του μέσου δείκτη διακριτικότητας πρέπει να λαμβάνεται υπόψη κατά την ερμηνεία της διακριτικής ισχύος μιας εξεταστικής δοκιμασίας. Σε αντίθετη περίπτωση υπάρχει ο κίνδυνος πλάνης.

27. Η πρόβλεψη της αξιοπιστίας στηρίζεται στην υπόθεση ότι όλα τα ερωτήματα έχουν δείκτη δυσκολίας 50%.

Πίνακας 35
Τιμή μέσου δείκτη διακριτικότητας σε τεστ μηδενικής αξιοπιστίας σε συνάρτηση με το μέγεθός τους

Αριθμός ερωτήσεων	Μέσος δείκτης διακριτικότητας
6	0.50
9	0.40
17	0.30
37	0.20
67	0.15
150	0.10
600	0.05

Πηγή: Ebel (1972: 401)

9.12. Αντικειμενικότητα

Μεταξύ των στοιχείων που πρέπει να χαρακτηρίζουν μια καλή εξεταστική διαδικασία περιλαμβάνεται και η αντικειμενικότητας της αξιολόγησης των εξεταζομένων. Η ιδιότητα αυτή σχετίζεται με την εγκυρότητα και την αξιοπιστία των αποτελεσμάτων που προκύπτουν από τις εξεταστικές δοκιμασίες. Αποτελεί, επίσης, προϋπόθεση για τη συγκρισιμότητα των επιδόσεων διαφορετικών μαθητών και, κατ' επέκταση, και για την ερμηνεία τους, με βάση τους κανόνες που έχουν καθορισθεί από την κλασική θεωρία των μετρήσεων.

Αντικειμενικά θεωρούνται τα αποτελέσματα μιας εξέτασης, όταν δεν επηρεάζονται από παράγοντες που σχετίζονται με τις διαθέσεις, τις υποκειμενικές αντιλήψεις και άλλα προσωπικά χαρακτηριστικά των αξιολογητών, όπως έχουμε αλλού σημειώσει.

Η εξασφάλιση της συνθήκης αυτής συνδέεται, ασφαλώς, με τη μέθοδο εξέτασης (π.χ. με τη μορφή των ερωτήσεων που χρησιμοποιούνται), αλλά σχετίζεται και με την ισότιμη αντιμετώπιση όλων των εξεταζομένων. Η τελευταία επιτυγχάνεται, όταν υπάρχει ομοιομορφία τόσο ως προς τον τρό-

πο εξέτασης των ατόμων όσο και ως προς τη βαθμολόγηση των απαντήσεών τους. Η ομοιομορφία της εξέτασης αναφέρεται όχι μόνο στη χρήση του ίδιου εξεταστικού μέσου για όλους, αλλά και στην ομοιότητα των διευκρινίσεων και λοιπών οδηγιών που τους δίνονται, στην ισότητα της χρονικής διάρκειας της εξέτασης, καθώς και στον ενιαίο τρόπο γραφής των απαντήσεων στις αντίστοιχες ερωτήσεις, σε περίπτωση γραπτών δοκιμασιών.

Η ομοιομορφία ως προς τη βαθμολόγηση αφορά: α) στην εκτίμηση της ορθότητας και στη βαθμολογική αποτίμηση των απαντήσεων των εξεταζομένων με βάση κοινά, ενιαία και σαφή κριτήρια, και β) στο υπολογισμό της βαρύτητας που έχει η κάθε ερώτηση στον τελικό βαθμό, κατά το ίδιο τρόπο από όλους τους βαθμολογητές.

Για τους τρόπους με τους οποίους μπορούν να επιτευχθούν οι παραπάνω όροι, στο βαθμό του δυνατού –αφού απόλυτη αντικειμενικότητα δεν είναι εφικτή με βάση όσα ισχύουν μέχρι σήμερα– έγινε εκτενής λόγος στα κεφάλαια που προηγήθηκαν. Η επανάληψή τους εδώ θα συνιστούσε παλιλλογία. Συμπληρωματικές διευκρινίσεις σχετικές με τις συνθήκες διενέργειας των εξεταστικών δοκιμασιών, οι οποίες πρέπει, επίσης, να διασφαλίζουν όχι μόνο το αδιάβλητό τους, αλλά και την ισότιμη αντιμετώπιση των εξεταζομένων παρέχονται σε επόμενο κεφάλαιο (βλ. κεφάλαιο ΙΑ΄), μαζί με ορισμένα πρόσθετα στοιχεία που αφορούν στη δίκαιη και αντικειμενική βαθμολόγηση των γραπτών.

9.13. Πρακτικότητα και οικονομία

Στα κριτήρια, με βάση τα οποία αξιολογείται μια εξεταστική δοκιμασία ή μια σχετική διαδικασία μπορούν, ακόμη, να προστεθούν η πρακτικότητά της και το κόστος που συνεπάγεται η εφαρμογή της. Η πρακτικότητα ή χρηστικότητα ορισμένης εξεταστικής διαδικασίας προσδιορίζεται από την ευκολία προετοιμασίας και εφαρμογής των αναγκαίων μέσων, καθώς και από την ευχέρεια ως προς την εξαγωγή των σχετικών αποτελεσμάτων και το βαθμό κατανόησής τους από τους άμεσα ενδιαφερομένους. Η οικονομική της διάσταση σχετίζεται με την απασχόληση του αναγκαίου εκπαιδευτικού και λοιπού προσωπικού, καθώς και με την ποσότητα του χρόνου, που διατίθεται για το σκοπό αυτό, και τον όγκο της εργασίας και των δαπανών που απαιτούνται για τη διεξαγωγή της και, κυρίως, για τη διόρθωση και τη βαθμολόγηση των γραπτών.

Τα στοιχεία αυτά δεν έχουν, βέβαια, άμεση σχέση με την κλασική θεωρία, την οποία εξετάσαμε στο κεφάλαιο αυτό. Διαδραματίζουν, όμως, συχνά σημαντικό ρόλο ως προς το ποιες διαδικασίες υιοθετούνται για την

ανάπτυξη των διαφόρων δοκιμασιών και το δοκιμαστικό έλεγχό τους (π.χ. ως προς τον προσδιορισμό του μεγέθους των δειγμάτων στα οποία σταθμίζονται, την επιλογή της μεθοδολογίας που υιοθετείται για τον έλεγχο της αξιοπιστίας τους ή της εγκυρότητάς τους και άλλα παρόμοια).

9.14. Αδυναμίες της κλασικής θεωρίας

Αναφέραμε ήδη ότι η κλασική θεωρία βασίζεται στην παραδοχή ότι κάθε εξεταστική διαδικασία συνιστά ένα είδος μέτρησης, η οποία αποτελεί προσεγγιστική εκτίμηση της πραγματικότητας. Άρα, εμπεριέχει κάποιο σφάλμα, το οποίο, κατά την εν λόγω θεωρία, είναι το ίδιο για όλα τα άτομα που συμμετέχουν σε ορισμένη εξέταση. Η παραδοχή αυτή δεν φαίνεται ορθή ως προς το δεύτερο σκέλος της, ως προς την ταυτότητα, δηλαδή, του σφάλματος για όλους τους εξεταζομένους, ανεξάρτητα από τα ατομικά τους χαρακτηριστικά και, ιδιαίτερα, από την ικανότητα του καθενός. Η αθαίρετη αυτή υπόθεση έχει γίνει, συχνά, αντικείμενο κριτικής.

Τα ψυχομετρικά, επίσης, χαρακτηριστικά των ερωτημάτων, που περιλαμβάνονται στις διάφορες εξεταστικές δοκιμασίες, δεν είναι ανεξάρτητα ούτε από τον τρόπο διατύπωσής τους ούτε από τη σχέση τους με το γνωστικό τομέα στον οποίο αναφέρονται. Πολύ περισσότερο δεν είναι ανεξάρτητα από το επίπεδο του δείγματος των ατόμων, με βάση το οποίο καθορίζονται. Η δυσκολία π.χ. ενός ερωτήματος εξαρτάται, όπως έχουμε τονίσει, σε σημαντικό βαθμό από το επίπεδο της ομάδας, με βάση την οποία γίνεται η εκτίμησή της. Το ίδιο ερώτημα θα θεωρηθεί εύκολο, αν δοθεί σε μια ομάδα ατόμων υψηλής ικανότητας, ενώ θα κριθεί δύσκολο, αν δοθεί σε μια ομάδα χαμηλής ικανότητα. Κάτι ανάλογο ισχύει και για τον υπολογισμό των δεικτών διακριτικότητας των ερωτημάτων. Η σχετικότητα αυτή καθιστά δύσκολη την έγκυρη γενίκευση του βαθμού δυσκολίας και διακριτικότητας που εκτιμάται σύμφωνα με τις μεθόδους, οι οποίες στηρίζονται στην κλασική θεωρία. Εύλογο, είναι ότι ανάλογα προβλήματα απαντώνται και κατά την εκτίμηση των συνολικών χαρακτηριστικών ενός τεστ.

Τα δείγματα, βέβαια, στα οποία σταθμίζεται μια τυποποιημένη δοκιμασία, καταβάλλεται προσπάθεια να είναι αντιπροσωπευτικά του αντίστοιχου πληθυσμού, για να αποφεύγονται, όσο είναι δυνατόν, τέτοιου είδους προβλήματα. Παρά την προσπάθεια, όμως, αυτή, δεν είναι πάντα εφικτή η εξάλειψη αποκλίσεων μεταξύ ομάδων, οι οποίες ανήκουν στον ίδιο πληθυσμό. Πιθανές, όμως, αποκλίσεις μιας ομάδας αξιολογουμένων από τα χαρακτηριστικά του δείγματος, στο οποίο σταθμίστηκε ορισμένη δοκιμασία, εγείρουν ερωτήματα ως προς την εγκυρότητα της ερμηνείας των

αποτελεσμάτων της. Αυτό ισχύει ιδιαίτερα, όταν το τεστ βασίζεται σε νόρμες, οι οποίες έχουν καθορισθεί με βάση συγκεκριμένο δείγμα. Όλα αυτά συνιστούν αδυναμίες της υπό εξέταση θεωρίας.

Επιπρόσθετα, η κλασική θεωρία δέχεται την ύπαρξη παράλληλων μορφών του ίδιου τεστ και τις εκλαμβάνει ως ισοδύναμες κατά την ερμηνεία των σχετικών αποτελεσμάτων, εφόσον αυτές έχουν γίνει με τις ίδιες βασικές προδιαγραφές. Η ισοδυναμία, όμως, αυτή, αξιολογούμενη με αυστηρά κριτήρια, τίθεται, αρκετές φορές, υπό αμφισβήτηση. Είναι πολύ δύσκολο να υπάρξουν απόλυτα ισοδύναμα αποτελέσματα, όταν χρησιμοποιούνται δοκιμασίες με διαφορετικά ερωτήματα, ακόμη και την περίπτωση που έχουν παρόμοια ψυχομετρικά χαρακτηριστικά.

Για όλους αυτούς τους λόγους έχει ασκηθεί κριτική κατά της κλασικής θεωρίας. Ο Wright (1967) έφτασε, μάλιστα, μέχρι του σημείου να αποκαλεί τα αποτελέσματα που προκύπτουν από σταθμισμένα τεστ, τα οποία βασίζονται σε νόρμες, ή άλλες ανάλογες δοκιμασίες ως «εύκαμπτους δείκτες μέτρησης» (rubber yardsticks).

Για την υπέρβαση των αδυναμιών αυτών, οι ειδικοί στα θέματα των ψυχολογικών και εκπαιδευτικών μετρήσεων εισηγούνται την εφαρμογή μιας νέας θεωρίας, η οποία μετρά, πιο έγκυρα, την ικανότητα των ατόμων να απαντούν σε ερωτήματα. Τη θεωρία αυτή εξετάζουμε αναλυτικά στο επόμενο κεφάλαιο. Ορισμένες, όμως, αδυναμίες της κλασικής θεωρίας επιχειρεί να καλύψει και η θεωρία της γενικευσιμότητας, στην οποία γίνεται συνοπτική αναφορά παρακάτω.

9.15. Η θεωρία της γενικευσιμότητας[28]

Η δυνατότητα γενίκευσης των διαπιστώσεων που προκύπτουν από διαδικασίες μέτρησης, όπως είναι η εφαρμογή των διαφόρων ψυχολογικών και εκπαιδευτικών τεστ, αποτελεί σημαντικό στοιχείο για τη λήψη ποικίλων αποφάσεων, στην υποβοήθηση των οποίων στοχεύει, μεταξύ άλλων, η αξιολόγηση. Όπως έχουμε ήδη αναφέρει, κάθε μέτρηση που διενεργείται θεωρείται, κατά την κλασική θεωρία, ως προσεγγιστική εκτίμηση της πραγμα-

28. Η παρούσα ενότητα δεν εντάσσεται, ουσιαστικά, στο περί της κλασικής θεωρίας κεφάλαιο. Κανονικά θα έπρεπε να αποτελέσει αντικείμενο χωριστού κεφαλαίου. Μη θέλοντας, όμως, να αυξήσουμε περαιτέρω τον όγκο του παρόντος συγγράμματος, με την προσθήκη ενός ακόμη κεφαλαίου, από το ένα μέρος, και λαμβάνοντας, από το άλλο, υπόψη το γεγονός ότι πρόκειται περί συνοπτικής παρουσίασης του θέματος, αποφασίσαμε να επισυνάψουμε, σ' αυτή τη φάση, τη σχετική ανάλυση στο τέλος του παρόντος κεφαλαίου.

τικότητας. Η πραγματική επίδοση των εξεταζομένων, κατά τη θεωρία αυτή, θα προέκυπτε από τις απαντήσεις που θα έδιναν, αν ήταν δυνατόν να απαντήσουν σε όλα τα πιθανά ερωτήματα, τα οποία θα μπορούσαν να τους τεθούν σε σχέση με ορισμένο γνωστικό αντικείμενο. Κάτι τέτοιο, όμως, δεν είναι ρεαλιστικό. Αναγκαστικά, λοιπόν, η επίδοση των εξεταζομένων προκύπτει από τις απαντήσεις τους σε περιορισμένο αριθμό ερωτήσεων. Αυτό έχει ως συνέπεια την απόκλιση της επιτυγχανόμενης επίδοσης από την «πραγματική». Ο συνυπολογισμός του εκτιμώμενου σφάλματος θεωρείται ότι καλύπτει την απόκλιση αυτή.

Στην αύξηση της δυνατότητας γενίκευσης των αποτελεσμάτων, που προκύπτουν από τις εξεταστικές δοκιμασίες, και στην προσέγγιση της «αληθούς επίδοσης» συμβάλλουν η αξιοποίηση στοιχείων που προέρχονται από πολλαπλές πηγές και καλύπτουν ευρύ χρονικό διάστημα, καθώς και η χρήση δοκιμασιών με υψηλούς δείκτες εγκυρότητας και αξιοπιστίας (Oosterfof, 2010: 95-99). Παρά τη λήψη, όμως, των μέτρων αυτών, το σφάλμα της μέτρησης παραμένει.

Επιπρόσθετα, η κλασική θεωρία δεν καθορίζει την πηγή των διαφόρων σφαλμάτων. Τα ενοποιεί όλα σε ένα γενικό σφάλμα. Το γεγονός αυτό δεν διευκολύνει τη λήψη αποφάσεων, όταν απαιτείται π.χ. να γίνουν βελτιώσεις ή αλλαγές σε συγκεκριμένους τομείς. Αντίθετα, η θεωρία της γενικευσιμότητας βασίζεται στην άποψη ότι ποικίλες πηγές σφάλματος μπορούν ταυτόχρονα, να επηρεάζουν, κατά διαφορετικό τρόπο, τα δεδομένα των μετρήσεων και παρέχει, επιπλέον, τη δυνατότητα εκτίμησής τους. Η προσέγγιση αυτή εξυπηρετεί τις μελέτες λήψης αποφάσεων, γνωστές ως D-studies (το D αποτελεί το αρχικό γράμμα της αγγλικής λέξης Decision = απόφαση), οι οποίες επιδιώκουν να απαντήσουν σε ερωτήματα του τύπου: τι θα συνέβαινε, εάν άλλαζαν κάποια πράγματα σε μια δεδομένη κατάσταση ή στη λειτουργία ενός συστήματος; Η απάντηση σ' αυτά είναι ιδιαίτερα χρήσιμη σε όσους ασχολούνται με την έρευνα στον τομέα της εκπαιδευτικής αποτελεσματικότητας (Creemers et al. 2010).

Στην αντιμετώπιση τέτοιου είδους ζητημάτων στοχεύει η ανάπτυξη της θεωρίας για τη γενικευσιμότητα (generalizability theory), γνωστής στη διεθνή βιβλιογραφία ως G-θεωρία (G-theory), καθώς και η συσχέτισή της με τη θεωρία λήψης αποφάσεων. Πρόκειται για στατιστική θεωρία, την οποία εισήγαγαν στις εκπαιδευτικές και ψυχολογικές μετρήσεις οι Cronbach, Nageswari και Glerer (1963).[29] Αλλά και άλλοι ειδικοί συνέβαλαν στην απο-

29. Βλ., επίσης, Gronbach et al., (1972).

σαφήνισή της, όπως οι Crocker & Algina (1986), Shavelson & Webb (1991), Brennan, (2001), Chiu (2001), Marculides & Kyriakides (2010).

Ουσιαστικά, η θεωρία αυτή επιδιώκει να αξιολογήσει την εξαρτησιμότητα (dependability) της αξιοπιστίας των μετρήσεων από ελεγχόμενους παράγοντες και να προσδιορίσει κατά πόσον αυτό, το οποίο διαπιστώθηκε σε συγκεκριμένες συνθήκες ή καταστάσεις μέτρησης, ισχύει για άπειρο πλήθος (universe) ανάλογων συνθηκών ή καταστάσεων.

Με άλλα λόγια, η G-θεωρία στοχεύει στο να απομονώσει και να αξιολογήσει τις επιμέρους όψεις (facets) των σφαλμάτων των μετρήσεων. Οι πτυχές αυτές αντιστοιχούν, κατά κάποιο τρόπο, σ' αυτό που ονομάζουμε στην ανάλυση της διακύμανσης «*παράγοντες*». Τέτοιοι παράγοντες μπορούν να είναι, στην περίπτωση των εκπαιδευτικών μετρήσεων π.χ., οι εξεταζόμενοι, οι βαθμολογητές, τα ερωτήματα που τίθενται στους εξεταζόμενους, οι φορές συμμετοχής τους σε μια εξέταση και άλλα παρόμοια.

Η θεωρία της γενικευσιμότητας ανήκει, όπως και η κλασική, στην κατηγορία των θεωρητικών προσεγγίσεων των μετρήσεων που στηρίζονται στην τυχαία δειγματοληψία (random sampling theories). Ο υπολογισμός της αξιοπιστίας των αποτελεσμάτων των τεστ, σύμφωνα με την κλασική θεωρία, την οποία παρουσιάσαμε προηγουμένως, αντικαθίσταται στην προκειμένη περίπτωση από τη θεωρία της γενικευσιμότητας (Shavelson, Webb & Rowley, 1989).

Ως προς το ποιες πτυχές των σφαλμάτων εξετάζονται κάθε φορά, αυτό προσδιορίζεται από το σχέδιο της μελέτης που αφορά στη λήψη απόφασης (D-study). Με βάση αυτήν, εφαρμόζεται ο κατάλληλος τρόπος υπολογισμού των συντελεστών γενίκευσης (ρ), οι οποίοι είναι ανάλογοι προς τους δείκτες αξιοπιστίας, τους οποίους εξετάσαμε λεπτομερώς προηγουμένως.

Η τιμή των συντελεστών αυτών υπολογίζεται με βάση τα στατιστικά μεγέθη που προκύπτουν από την ανάλυση της διακύμανσης. Στον πίνακα 36 σημειώνουμε τους τρόπους υπολογισμού για τέσσερα ενδεικτικά μονοπαραγοντικά σχέδια λήψης απόφασης (single facet D-study designs), όπως αυτοί δίδονται από τους Crocker & Algina (1986). Η σημασία των συμβόλων του παραπάνω πίνακα είναι η εξής: p = εξεταζόμενοι, ι = συνθήκες εξέτασης (π.χ. εξεταστές), σ_p^2 η διακύμανση των παρατηρούμενων μετρήσεων μεταξύ των εξεταζομένων, σ_i^2 η διακύμανση μεταξύ των συνθηκών εξέτασης και σ_ε^2 η διακύμανση του καταλοίπου (residual), όπως προκύπτουν από τους αντίστοιχους πίνακες ανάλυσης των πηγών της διακύμανσης. Το σημείο * έχει την έννοια ότι ο συντελεστής είναι κατάλληλος για τη μελέτη λήψης απόφασης (D-study).

Πίνακας 36

Τρόποι υπολογισμού συντελεστών γενικευσιμότητας για τέσσερα μονοπαραγοντικά (sigle-facet) μοντέλα λήψης απόφασης

Σχέδια	Περιγραφή	Αριθμός συνθηκών μέτρησης	Διακύμανση παρατηρούμενων «σκορ»	Συντελεστής γενικευσιμότητας
1	pxi	1	$\sigma_p^2 + \sigma_e^2$	$\rho_i{}^{*2} = \frac{\sigma_p^2}{\sigma_p^2 + \sigma_e^2}$
2	pxi	n_i'	$\sigma_p^2 + \sigma_e^2 / n_i'$	$\rho_i{}^{*2} = \frac{\sigma_p^2}{\sigma_p^2 + \sigma_e^2 / n_i'}$
3	i : p	1	$\sigma_p^2 + \sigma_i^2 + \sigma_e^2$	$\rho_i{}^{*2} = \frac{\sigma_p^2}{\sigma_p^2 + \sigma_i^2 + \sigma_e^2}$
4	i : p	n_i'	$\sigma_p^2 + (\sigma_i^2 + \sigma_e^2) / n_i'$	$\rho_i{}^{*2} = \frac{\sigma_p^2}{\sigma_p^2 + (\sigma_i^2 + \sigma_e^2) / n_i'}$

Πηγή: Crocker & Algina (1986: 170)

Για τη μέθοδο της ανάλυσης της διακύμανσης παραπέμπουμε στα ειδικά κεφάλαια των εγχειριδίων Στατιστικής Εφαρμοσμένης στις Κοινωνικές Επιστήμες, αρκετά από τα οποία αναφέρονται στη βιβλιογραφία.

Ανάλογοι τύποι, κατάλληλα διαμορφωμένοι, χρησιμοποιούνται σε άλλες μορφές σχεδίων υπολογισμού των δεικτών γενικευσιμότητας. (Για περισσότερες πληροφορίες ως προς το ζήτημα αυτό παραπέμπουμε τον αναγνώστη στο βιβλίο των Crocker & Algina, 1986: 170-212).

Στην πράξη, οι δείκτες αυτοί υπολογίζονται σήμερα με τη βοήθεια ειδικών προγραμμάτων ηλεκτρονικών υπολογιστών, όπως είναι π.χ. το "WINSTEPS", το MULTILOG", το SPSS, το SAS, το GENOVA και άλλα.

Η σύγχρονη θεωρία της γενικευσιμότητας διαφέρει, ακόμη από την κλασική ως προς το ότι συνεκτιμά το πώς επηρεάζεται η συνέπεια των αποτελεσμάτων μιας εξεταστικής δοκιμασίας, εάν αυτή χρησιμοποιείται για απόλυτες ή για σχετικές συγκρίσεις. Απόλυτες θεωρούνται οι συγκρίσεις που προκύπτουν από τεστ-κριτήρια, με βάση τα οποία ένα άτομο κρίνεται

αν υπερβαίνει ή όχι το προκαθορισμένο βαθμολογικό όριο που διαχωρίζει τους εξετασθέντες σε επιτυχόντες και αποτυχόντες. ή τους κατατάσσει σε άλλες βαθμολογικές κατηγορίες. Σχετικές είναι οι συγκρίσεις με τους συνομήλικους ορισμένου ατόμου, οι οποίες γίνονται με βάση τα τεστ, τα οποία στηρίζονται σε νόρμες, καθώς και οι ενδοατομικές συγκρίσεις.