ΚΕΦΑΛΑΙΟ 10

# Η ΣΥΓΧΡΟΝΗ ΘΕΩΡΙΑ ΜΕΤΡΗΣΗΣ ΤΗΣ ΙΚΑΝΟΤΗΤΑΣ ΑΠΑΝΤΗΣΗΣ ΣΕ ΕΡΩΤΗΜΑΤΑ

## 10.1. Εισαγωγή

Η Θεωρία Μέτρησης της Ικανότητας Απάντησης σε Ερωτήματα[1] (Θ.Μ.Ι.Α.Ε.) έχει εφαρμογή τόσο σε ψυχολογικά τεστ όσο και σε τεστ επίδοσης και αποτελεί, σήμερα, σημαντική πτυχή της ενασχόλησης με τις μετρήσεις σε διάφορες Επιστήμες. Παρά το γεγονός ότι η παραπάνω θεωρία χρησιμοποιείται τις τελευταίες, κυρίως, δεκαετίες (μετά το 1980), οι απαρχές της ανάγονται στα πρώτα χρόνια του δεύτερου μισού του 20ού αιώνα. Η εξέλιξή της και η επέκταση των εφαρμογών της συνδέονται με ονόματα σημαντικών επιστημόνων, όπως είναι ο Αμερικανός Lord (1980), ο Δανός μαθηματικός Rasch (1960), ο μαθητής του Andersen (1972), ο Αυστριακός κοινωνιολόγος Lazarfeld (Lazarfeld & Henry 1968), ο Αυστριακός, επίσης, Fisher (1973), ο Andrich (1988, 1989), ο Αμερικανός Wright (1992) και οι μαθητές του (Wright & Douglas, 1977, Masters, 1982, Wilson, 1989), ο Van der Linden, ο Hambleton (1997), ο Thissen, ο Wainer (2001) και πολλοί άλλοι.

Η συγκεκριμένη θεωρία είναι γνωστή και ως «**θεωρία των λανθανόντων χαρακτηριστικών**», καθώς και ως θεωρία της «**ισχυρής πραγματικής επίδοσης**». Εστιάζεται περισσότερο στην ανάλυση των απαντήσεων των εξεταζομένων στα επιμέρους ερωτήματα (items) που συνθέτουν ένα τεστ, παρά στα συνολικά αποτελέσματά του (Baker, 2001). Η μέτρηση ενός χαρακτηριστικού εξαρτάται, κατά τη θεωρία αυτή, συγχρόνως από τις απα-

---

1. Ο όρος αποτελεί μετάφραση του αγγλικού όρου: Item Response Theory (I.R.T.). Στην ελληνική βιβλιογραφία δεν έχει ευρέως καθιερωθεί. Ορισμένοι τον αποδίδουν ως «*θεωρία απολογισμού ικανοτήτων*» (Αλεξόπουλος, 1998), απόδοση την οποία δεν θεωρούμε δόκιμη, επειδή δεν πρόκειται για υπολογισμό ικανοτήτων γενικώς, αλλά της ικανότητας απάντησης σε συγκεκριμένα ερωτήματα (ζητήματα). Άλλοι αποδίδουν τον παραπάνω όρο ως «*θεωρία αντίδρασης σε ερωτήματα*» (Κουλάκογλου, 2002: 110) ή ως «*πιθανολογική θεωρία*» (Ingenkamp, 1993: 159) ή ως «*θεωρία λανθανόντων χαρακτηριστικών*» (Μυλωνάς, 2012: 311).

ντήσεις των εξεταζόμενων και από τις ιδιότητες των σχετικών ερωτημάτων. Στόχος της είναι η βελτίωση των διαδικασιών μέτρησης των διαφόρων χαρακτηριστικών των ατόμων, περιλαμβανομένων και αυτών που σχετίζονται με τις γνώσεις και τις δεξιότητές τους, τις οποίες επιδιώκει να αναπτύξει το σχολείο.

Η υπό εξέταση θεωρία είναι πιο σύνθετη σε σύγκριση με την κλασική. Η εφαρμογή της απαιτεί πολύπλοκους μαθηματικούς υπολογισμούς, οι οποίοι δεν είναι εύκολο να παρουσιασθούν λεπτομερώς στο πλαίσιο της περιορισμένης έκτασης του παρόντος κεφαλαίου και, σε αρκετές περιπτώσεις, υπερβαίνουν τις επιστημονικές μας αρμοδιότητες. Για τους λόγους αυτούς θα περιοριστούμε σε βασικά στοιχεία, παραπέμποντας όσους ενδιαφέρονται να εμβαθύνουν στα ζητήματα αυτά σε πιο εξειδικευμένα, ξενόγλωσσα, κυρίως, βιβλία.[2]

## 10.2. Βασικές παραδοχές της θεωρίας μέτρησης της ικανότητας απάντησης σε ερωτήματα. Χαρακτηριστική καμπύλη ερωτημάτων

Κεντρικός άξονας της Θ.Μ.Ι.Α.Ε. είναι ο προσδιορισμός του βαθμού της καταλληλότητας των ερωτημάτων που χρησιμοποιούνται σε μια δοκιμασία, για να επιτευχθεί όσο γίνεται καλύτερα το επιδιωκόμενο αποτέλεσμα (Lord, 1980, Hulin et al., 1983, Hambleton et al. 1991). Η αξιολόγηση της καταλληλότητάς τους είναι κρίσιμης σημασίας ζήτημα όχι μόνο για την κατασκευή και την αποτίμηση της αξίας ενός τεστ, αλλά και για την οργάνωση των ερωτημάτων στις σχετικές τράπεζες θεμάτων, καθώς και για τις διάφορες συγκριτικές διαχρονικές μελέτες αξιολόγησης ή για τις αντίστοιχες

2. Από την ελληνική βιβλιογραφία βλ. Ingenkamp, 1993, Αλεξόπουλος, 2004, Κουλάκογλου, 2002 και Κατσής κ.ά. 2010. Από την ξενόγλωσση βιβλιογραφία σημειώνουμε, ενδεικτικά, τις παρακάτω πηγές: Lord & Novick, 1968, Baker, 1992, Van Linden & Hambleton, 1997, Embretson & Reise, 2000. Η παρουσίαση της Θ.Μ.Ι.Α.Ε., που ακολουθεί, βασίζεται εν μέρει στην εργασία του Baker (The basic of item response theory, Eric, 2001), η οποία υπήρχε στο διαδίκτυο κατά τη συγγραφή της παρούσας ενότητας (Μάιος, 2012 - http://info.worldbank.org/etools/ docs/library/117765/Item%20 Response%20 Theory%20-%20F%20Baker.pdf). Στην εργασία αυτή ο Baker παραθέτει τα κύρια στοιχεία του προμνημονευθέντος βιβλίου του (1992). Παράλληλα, όμως, αξιοποιήθηκαν και άλλες πηγές (Der Linden & Hambleton, 1997· Hashway, 1998· Emertson & Reise, 2000· Livingston, 2006· Wu & Adams, 2007· Demars, 2010· Κατσής κ.ά. 2010· Μυλωνάς, 2012), οι οποίες σημειώνονται στα κατάλληλα σημεία του κειμένου. Για την απόδοση μέρους των μαθηματικών στοιχείων της εν λόγω θεωρίας και την παρουσίαση ορισμένων τεχνικών ζητημάτων (π.χ. λογικά διαγράμματα) συνεργαστήκαμε με το μαθηματικό Ιωάννη Οικονομίδη, η βοήθεια του οποίου υπήρξε πολύτιμη. Η τελική σύνθεση έγινε από το γράφοντα.





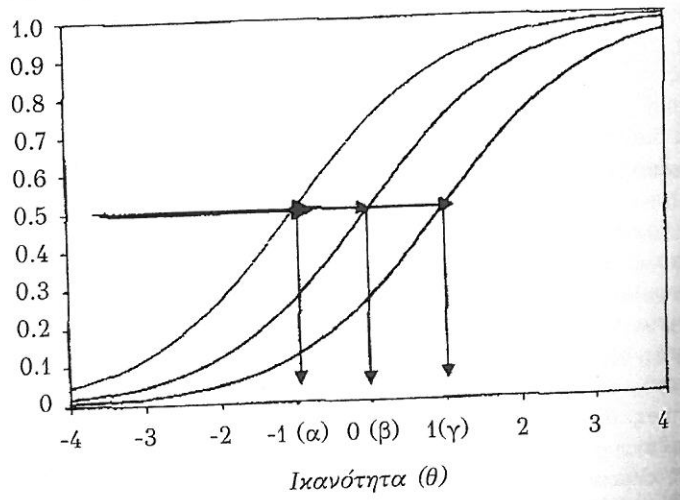
μετα-αναλύσεις. Ιδιαίτερα, η κατασκευή των λεγόμενων προσαρμοζόμενων τεστ με τη βοήθεια υπολογιστών (computerized adaptive tests) (Wainer et al., 2000), καθώς και ορισμένες μέθοδοι υπολογισμού του σημείου των αποτελεσμάτων των τεστ που διαχωρίζουν τους εξεταζομένους σε επίπεδα (cut off score) αξιοποιούν σημαντικά τη νέα θεωρία των μετρήσεων.

Η Θ.Μ.Ι.Α.Ε. στηρίζεται στις εξής, κυρίως, παραδοχές: α) Οι απαντήσεις των εξεταζομένων στα ερωτήματα ενός τεστ αποτελούν παρατηρήσιμες μορφές συμπεριφοράς που αντανακλούν μη εμφανείς (λανθάνουσες) ικανότητές τους, οι οποίες καθορίζουν την πιθανότητα που έχουν να απαντήσουν σωστά στα ερωτήματα, τα οποία τους τίθενται (π.χ. επίπεδο σχετικών γνώσεων και δεξιοτήτων, στάσεις έναντι ορισμένων ζητημάτων, νοητική ικανότητα κτλ.). Η δυνατότητα αυτή (χαρακτηριστικό) συμβολίζεται, συνήθως, με το ελληνικό γράμμα $\theta$, και μπορεί να μετρηθεί με βάση μια κλίμακα λογιστικών μονάδων (logits) που έχουν μέσο όρο μηδέν (0) και τυπική απόκλιση[3] ίση με ±1. Οι συνήθεις τιμές που λαμβάνει το λανθάνον χαρακτηριστικό κυμαίνονται από –3 έως +3, αλλά είναι δυνατόν να υπάρχουν και μεγαλύτερες τιμές, αφού, θεωρητικά, η κλίμακα αυτή εκφράζει διάστημα που εκτείνεται από το $-\infty$ έως το $+\infty$, β) Προϋποθέτει την ανεξαρτησία μεταξύ των διαφόρων ερωτημάτων (local independence), η οποία μπορεί να ελεγχθεί με διάφορους τρόπους στους οποίους αναφερόμαστε παρακάτω. γ) Οι παράμετροι των ερωτημάτων (δυσκολία, διακριτικότητα κτλ.) παραμένουν οι ίδιες σε διαφορετικά δείγματα εξεταζομένων (population invariance). Για το λόγο αυτό δεν απαιτείται τυχαία δειγματοληψία, όπως επιβάλλουν η κλασική θεωρία και η θεωρία της γενικευσιμότητας, θέματα στα οποία θα επανέλθουμε. Χρειάζονται, ωστόσο, δείγματα που να διασφαλίζουν ικανό εύρος της μετρούμενης λανθάνουσας ικανότητας. δ) Οι τιμές των παραμέτρων των ερωτημάτων και ο βαθμός της λανθάνουσας ικανότητας των εξεταζομένων εκφράζονται με την ίδια μετρική κλίμακα και τοποθετούνται στον ίδιο οριζόντιο άξονα. ε) Η απάντηση του εξεταζομένου σε ένα ερώτημα μπορεί να εκφρασθεί ως λογιστική συνάρτηση (item response function-I.R.F.), η οποία δηλώνει την πιθανότητα ($p_i$) ενός ατόμου, με ορισμένο $\theta$ χαρακτηριστικό, να απαντήσει σωστά σ' ένα ερώτημα, το οποίο έχει συγκεκριμένες ιδιότητες (παραμέτρους).[4] Η συνάρτηση αυτή εκφράζεται με τη μορφή μιας σιγμοειδούς καμπύλης, όπως

3. Η κλίμακα αυτή μοιάζει με εκείνη των z-τιμών (κανονική κατανομή), δεν πρέπει όμως να συγχέεται μ' αυτήν, επειδή υποδηλώνει τελείως διαφορετικά πράγματα.
4. Βλ. και http://www.Wikipedia.org?wiki/Item response theory - ανακτήθηκε στις 8/5/2012. Βλ., επίσης, Hashway (1998: 53) και Baker (2001: 7).

είναι αυτές που απεικονίζονται στο σχήμα 17. Η εν λόγω καμπύλη ονομάζεται **Χαρακτηριστική Καμπύλη ενός Ερωτήματος** (Χ.Κ.Ε. - Item Characteristic Curve - I.C.C.).

Παρατηρούμε ότι η θέση των καμπυλών, οι οποίες αντιστοιχούν στα τρία ερωτήματα (α, β, γ), ίδιας διακριτικότητας, διαφοροποιείται ανάλογα με τη δυσκολία τους και το επίπεδο ικανότητας των εξετασθέντων. Οι κλίσεις (slopes) των καμπυλών, οι οποίες συγκλίνουν, αλλά δεν διασταυρώνονται, είναι ίδιες. Τέλος, όλες οι καμπύλες εμφανίζουν προοδευτική ανιούσα τάση ως προς την πιθανότητα ορθής απάντησης, η οποία αρχίζει να κάμπτεται, μόλις η παραπάνω πιθανότητα υπερβεί το 0.5 (50%).




***Σχήμα 17.*** *Καμπύλες κατανομής απαντήσεων σε τρία ερωτήματα ίδιας διακριτικότητας αλλά διαφορετικής δυσκολίας σε σχέση με την ικανότητα των εξεταζόμενων να απαντήσουν σ' αυτά (α = εύκολο ερώτημα, β = μέσης δυσκολίας και γ = υψηλής δυσκολίας)*[5]

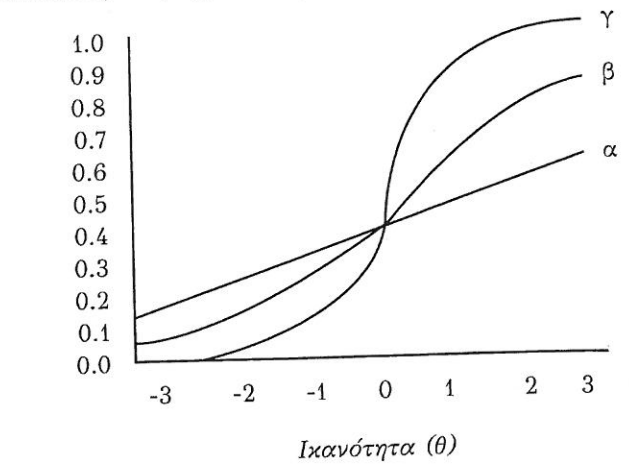
5. Οι Χ.Κ.Ε. βασίζονται σε μονοπαραμετρικό μοντέλο ανάλυσης (βλ. παρακάτω).





Η διαμόρφωση της χαρακτηριστικής καμπύλης για κάθε ερώτημα ενός τεστ, είναι, όπως σημειώνει ο Baker (2001: 7), «*βασικό δομικό στοιχείο της Θ.Μ.Ι.Α.Ε. Όλα τα υπόλοιπα στοιχεία της σχετίζονται μ' αυτήν*». Η πιθανότητα απάντησης σε ένα ερώτημα, με βάση την οποία διαμορφώνεται η εν λόγω καμπύλη, σχετίζεται: α) με το βαθμό δυσκολίας του και β) με τη διακριτική του ισχύ. Η δυσκολία ενός ερωτήματος προσδιορίζει σε ποιο εύρος της κλίμακας ικανότητας αυτό λειτουργεί. Π.χ. ένα εύκολο ερώτημα λειτουργεί μεταξύ ατόμων με χαμηλή ικανότητα (θ), ενώ ένα δύσκολο ερώτημα λειτουργεί για άτομα με υψηλή ικανότητα. Έτσι, η δυσκολία ενός ερωτήματος είναι, όπως αναφέραμε προηγουμένως, δείκτης θέσης στην ενιαία κλίμακα ικανότητας των εξεταζομένων και η σημασία της δεν καθορίζεται, όπως απαιτεί η κλασική θεωρία, από τις επιδόσεις των άλλων. Ο δείκτης δυσκολίας ενός ερωτήματος προσδιορίζεται από το σημείο της εν λόγω κλίμακας στο οποίο το 50% των εξεταζομένων μπορεί να απαντήσει ορθώς στο αντίστοιχο ερώτημα.

Από το σχήμα 17 προκύπτει ότι η πιθανότητα επιτυχίας των ατόμων αυξάνει όσο μεγαλώνει το επίπεδο της λανθάνουσας ικανότητάς τους να απαντήσουν στο σχετικό ερώτημα. Η πιθανότητα αυτή για άτομα πολύ υψηλής ικανότητας είναι, σύμφωνα με το εν λόγω σχήμα, ίση, σχεδόν, με το 1.0, ενώ για τα άτομα πολύ χαμηλής ικανότητας είναι σχεδόν ίση με 0. Για τους μέσης ικανότητας εξεταζομένους η πιθανότητα αυτή είναι και στις τρεις περιπτώσεις ίση με 0.5. Η δυσκολία, όμως, απάντησης στα τρία ερωτήματα είναι διαφορετική, όπως φαίνεται από τη θέση που κατέχουν. Το πρώτο (α) είναι το πιο εύκολο και προϋποθέτει μικρότερη ικανότητα απάντησης, ενώ το τρίτο (γ) είναι το πιο δύσκολο και απαιτεί μεγαλύτερη ικανότητα απάντησης. Σε όλες τις περιπτώσεις η ικανότητα αυτή προσδιορίζεται με βάση την ίδια κλίμακα, όπως σημειώθηκε ήδη. Αυτό σημαίνει ότι, αν μας είναι γνωστή η ικανότητα ενός εξεταζομένου, μπορούμε να προβλέψουμε την πιθανότητα απάντησής του χωρίς να τού δώσουμε το σχετικό ερώτημα (Wu & Adams, 2007: 14).




***Σχήμα 18.*** *Καμπύλες κατανομής απαντήσεων σε τρία ερωτήματα ίδιας δυσκολίας αλλά διαφορετικής διακριτικότητας(α = ερώτημα χαμηλής διακριτικότητας, β = μέσης διακριτικότητας και γ = υψηλής διακριτικότητας).*

Η διακριτικότητα σχετίζεται με το πόσο καλά ένα ερώτημα μπορεί να διαχωρίζει τους εξεταζομένους σ' αυτούς που έχουν ικανότητα κάτω και πάνω από τη θέση που αυτό κατέχει στην κλίμακα ικανότητας. Όσο πιο απότομη γίνεται η κάμψη της καμπύλης στο μέσο επίπεδο ικανότητας, τόσο καλύτερη είναι η διακριτικότητα του αντίστοιχου ερωτήματος (βλ. σχήμα 18). Για τον παραπάνω λόγο ο δείκτης αυτός ονομάζεται και **κύρτωση του ερωτήματος**.

Πρέπει να διευκρινισθεί ότι οι δύο αυτοί δείκτες (δυσκολία και διακριτικότητα) χρησιμεύουν απλώς για τον προσδιορισμό της καμπύλης κατανομής των πιθανοτήτων απάντησης σε ένα ερώτημα και δεν επιβεβαιώνουν ότι το συγκεκριμένο ερώτημα μετρά το προς εξέταση λανθάνον χαρακτηριστικό, ζήτημα το οποίο σχετίζεται με την εγκυρότητά του, για την οποία έχουμε κάμει λόγο στο περί κλασικής θεωρίας κεφάλαιο.

Με βάση τους δείκτες αυτούς, τα ερωτήματα μπορούν να χαρακτηρισθούν ως πολύ εύκολα, εύκολα, μέσης δυσκολίας, δύσκολα και πολύ δύ-





σκολα, ενώ από πλευράς διακριτικότητας δύνανται να καταταχθούν σε μηδενικής, χαμηλής, μέσης, υψηλής και άριστης διακριτικής ισχύος (Baker, 2001: 11).

Υπενθυμίζουμε στον αναγνώστη ότι οι γραφικές παραστάσεις χρησιμοποιούνται συχνά στην πράξη ως μια εύκολη αρχική μέθοδος για την απεικόνιση του βαθμού δυσκολίας και διακριτικότητας των ερωτήσεων που συνθέτουν ένα τεστ (Kingston & Dorans, 1985).

Η μαθηματική λογιστική συνάρτηση, η οποία εκφράζει τη σχέση μεταξύ των μεταβλητών, που αναφέρθηκαν προηγουμένως, εμφανίστηκε το 1844 και εφαρμόστηκε στο τομέα των Θετικών Επιστημών και, ειδικότερα, στη Βιολογία, με στόχο τη μοντελοποίηση της ανάπτυξης των φυτών και των ζώων. Στον τομέα των Ψυχοπαιδαγωγικών Επιστημών άρχισε να εφαρμόζεται μετά τα τέλη του 1950. Η χρήση της, όμως, διευρύνθηκε τις τελευταίες, κυρίως, δεκαετίες, όπως ήδη αναφέραμε, χάρη στην αύξηση των δυνατοτήτων των ηλεκτρονικών υπολογιστών, η οποία κατέστησε δυνατούς τους αναγκαίους υπολογισμούς.[6]

Η νέα θεωρία για τις μετρήσεις έχει διαμορφώσει νέους κανόνες σχετικούς με τα τεστ, οι οποίοι διαφέρουν σημαντικά από αυτούς στους οποίους στηρίζεται η κλασική θεωρία. Οι κύριες διαφορές τους συνίστανται, σύμφωνα με τους Embretson & Reise (2000: 15-39), στα εξής: 1) Το τυπικό σφάλμα μέτρησης δεν είναι το ίδιο για όλους τους εξεταζομένους που ανήκουν σε ορισμένο πληθυσμό, όπως δέχεται η κλασική θεωρία, αλλά διαφοροποιείται μεταξύ των επιμέρους «σκορ». Επιπρόσθετα, έχει γενικό χαρακτήρα και δεν αφορά ειδικά το συγκεκριμένο πληθυσμό στον οποίο ανήκουν οι εξεταζόμενοι. 2) Τα εκτενή τεστ δεν είναι πάντα πιο αξιόπιστα από τα σύντομα, όπως δέχεται η κλασική θεωρία. 3) Η σύγκριση των αποτελεσμάτων που επιτυγχάνονται με διαφορετικές μορφές τεστ είναι εφικτή, ακόμη και αν δεν υπάρχει ισοδύναμη παραλληλία τους, όπως επιβάλλει η κλασική θεωρία. 4) Η αρχή της κλασικής θεωρίας, κατά την οποία, για να γίνουν ακριβείς εκτιμήσεις, απαιτείται αντιπροσωπευτικότητα των δειγμάτων στα οποία σταθμίζεται ένα τεστ, ανατρέπεται από τη Θ.Μ.Ι.Α.Ε., αφού δέχεται ότι μπορούν να γίνουν έγκυρες (unbiased) εκτιμήσεις και χωρίς την προϋπόθεση αυτή. 5). Τα αποτελέσματα ενός τεστ μπορούν να ερμηνευθούν όχι με βάση νόρμες, αλλά με βάση τη σύγκριση των αποστάσεών τους στην ενιαία κλίμακα ικανότητας. 6) Ο όρος ότι οι ιδιότητες των ισοδιαστημικών κλιμάκων ισχύουν, κατά την κλασική θεωρία, όταν υπάρχουν κανο-

6. Σύμφωνα με τους Embretson & Reise (2000: 6-7) προηγήθηκε η εφαρμογή της στον τομέα της Εκπαίδευσης και ακολούθησε η εφαρμογή της στην Ψυχολογία.

νικές κατανομές των μετρήσεων μεταβάλλεται, αφού, κατά τη Θ.Μ.Ι.Α.Ε., αυτό ισχύει, όταν εφαρμόζονται τα κατάλληλα (justifiable) μοντέλα μέτρησης. 7) Μικτού τύπου ερωτήματα επηρεάζουν, κατά την κλασική θεωρία, ανομοιόμορφα τα τελικά αποτελέσματα, ενώ, κατά τη Θ.Μ.Ι.Α.Ε., αυτό δεν ισχύει. 8) Η αλλαγή επιδόσεων μπορεί να συγκριθεί ακόμη και στην περίπτωση ανομοιογενούς αρχικής αφετηρίας, άποψη που δεν γίνεται δεκτή στο πλαίσιο της κλασικής θεωρίας. 9) Η παραγοντική ανάλυση παρέχει πληρέστερη πληροφόρηση στο πλαίσιο της Θ.Μ.Ι.Α.Ε. απ' ό,τι στο πλαίσιο της κλασικής θεωρίας. 10) Τέλος, τα ιδιαίτερα χαρακτηριστικά των ερωτημάτων (item stimulus features) σχετίζονται, κατά τη Θ.Μ.Ι.Α.Ε., με τις ψυχομετρικές τους ιδιότητες, ενώ υποστηρίζεται το αντίθετο από την κλασική θεωρία.

Οι νέοι αυτοί κανόνες αναμορφώνουν εκ βάθρων την ψυχομετρική θεωρία για τα τεστ, καθιστούν πιο ευέλικτες τις διαδικασίες εφαρμογής τους, συντελούν στην εξασφάλιση πιο έγκυρων και αξιόπιστων δεδομένων και επιλύουν ορισμένα εγγενή προβλήματα της παραδοσιακής κλασικής θεωρίας (π.χ. το ζήτημα των παράλληλων μορφών ενός τεστ). Επιπρόσθετα, ο εντοπισμός στρεβλώσεων της λειτουργίας των ερωτημάτων, που ενδέχεται να προκαλούνται από πολιτισμικούς, φυλετικούς ή άλλους παράγοντες, καθίσταται πιο αποτελεσματικός με την εφαρμογή της νέας θεωρίας.

## 10.3. Τα μοντέλα ανάλυσης των ερωτημάτων, με βάση τη Θ.Μ.Ι.Α.Ε.

Η Θ.Μ.Ι.Α.Ε. είναι μια θεωρία που στηρίζεται σε **μοντέλα μέτρησης** με βάση τα οποία αποτιμάται κατά πόσο οι ιδιότητες των ερωτημάτων (δυσκολία, διακριτικότητα, βαθμός πιθανότητας να δοθεί ορθή απάντηση στην τύχη) και το επίπεδο κατοχής του μετρούμενου λανθάνοντος χαρακτηριστικού από ένα άτομο καθορίζουν την πιθανότητα που έχει να απαντήσει επιτυχώς στα ερωτήματα.[7] Για το λόγο αυτό αναφέρεται και ως **μέθοδος ισχυρής μοντελοποίησης** (Embretson & Reise, 2000: 43).

Αξίζει να σημειωθεί ότι η εκτίμηση του αξιολογούμενου χαρακτηριστικού βασίζεται, σύμφωνα με την κλασική θεωρία, στο άθροισμα του τελικού αριθμού των ορθών απαντήσεων σε ένα τεστ και στην αναγωγή του, όπου απαιτείται, σε πρότυπους βαθμούς. Αντίθετα, η εκτίμηση αυτή είναι, κατά την Θ.Μ.Ι.Α.Ε., περισσότερο μια διαδικασία αναζήτησης του μοντέ-

7. Και η κλασική θεωρία στηρίζεται σε μοντέλα, αλλά η λογική τους διαφέρει ουσιωδώς από εκείνη των μοντέλων της Θ.Μ.Ι.Α.Ε. (περισσότερες λεπτομέρειες βλ. στο Embretson & Reise (2000: 40 κ.έ.).





λου που εκφράζει καλύτερα τη μετρούμενη ικανότητα ενός εξεταζομένου, παρά ένα απλό άθροισμα επιτυχών απαντήσεων. Για το λόγο αυτό τα ερωτήματα ενός τεστ ενδέχεται να μην έχουν ίση βαρύτητα στην τελική εκτίμηση της ικανότητας αυτής, ανάλογα με το μοντέλο που εφαρμόζεται.

Στο πλαίσιο της υπό εξέταση θεωρίας, έχουν διαμορφωθεί διάφορα λογιστικά (logistic) μοντέλα, τα οποία κατηγοριοποιούνται σε: α) **μονοδιάστατα** και β) **πολυδιάστατα** (Embrertson & Reise, 2000· Baker, 2001· Wu & Adams, 2007· Demars, 2010). Τα πρώτα βασίζονται στην υπόθεση ότι η δυνατότητα των εξεταζομένων να απαντήσουν στα ερωτήματα ενός τεστ προσδιορίζεται από ένα, κυρίως, λανθάνον χαρακτηριστικό, σε σχέση με το οποίο έχουν διατυπωθεί τα ερωτήματα που περιλαμβάνονται σε ένα τεστ. Το χαρακτηριστικό αυτό μπορεί, σε ορισμένες περιπτώσεις, να έχει σύνθετο χαρακτήρα (π.χ. ένα τεστ μπορεί να μετρά την ικανότητα ανάγνωσης και την κατανόηση κειμένου).

Στη δεύτερη περίπτωση (πολυδιάστατα μοντέλα), ισχύει η υπόθεση ότι τα μετρούμενα λανθάνοντα χαρακτηριστικά των ατόμων είναι περισσότερα του ενός και διαφέρουν μεταξύ τους κατά τρόπο διακριτό. Ένα τεστ π.χ. μπορεί να μετρά την αναγνωστική και τη μαθηματική ικανότητα.

Για τον έλεγχο της ύπαρξης μιας ή περισσότερων διαστάσεων στα στοιχεία που μετρούν τα ερωτήματα ενός τεστ εφαρμόζεται η παραγοντική ανάλυση. Αν από την ανάλυση αυτή, η οποία γίνεται με τη βοήθεια κατάλληλων λογισμικών, προκύπτει ότι ο πρώτος παράγοντας που εξάγεται από τη σχετική διαδικασία ερμηνεύει περισσότερο από το 60% της συνολικής διακύμανσης και ο δεύτερος λιγότερο από το 5%, τότε μπορούμε να πούμε ότι ισχύει η συνθήκη της μοναδικής διάστασης ως προς το λανθάνον χαρακτηριστικό (Κατσής, κ.ά. 2010: 362).

Για τον ίδιο σκοπό είναι δυνατόν να χρησιμοποιηθεί το τεστ του Stout (Stout's test of essential unidimensionality),[8] καθώς και η ανάλυση των καταλοίπων της μήτρας διωνυμικής αναλογίας των ορθών απαντήσεων (analysis of residuals of the bivariate proportion-correct matrix).[9]

Λόγω της πολύ μεγάλης πολυπλοκότητας που εμφανίζουν τα πολυδιάστατα μοντέλα, δεν θα ασχοληθούμε περαιτέρω με αυτά στην παρούσα εργασία, παραπέμποντας τον αναγνώστη σε ειδικά βιβλία.[10]

8. Είναι γνωστό και ως DIMTEST από το ομώνυμο λογισμικό πρόγραμμα που χρησιμοποιείται για τον υπολογισμό του, όταν τα δεδομένα είναι διχοτομικού τύπου. Επί πολυτιμοτόμων δεδομένων χρησιμοποιείται το POLY-DIMET (Stout, 1999, 2005).

9. Για περισσότερες λεπτομέρειες βλ. (Demars, 2010: 43-47).

10. Βλ. Van der Linden & Hambleton (1997), όπου παρατίθεται και σχετική βιβλιογραφία.

Τα μονοδιάστατα μοντέλα κατηγοριοποιούνται σ'αυτά που εφαρμόζονται σε: α) **διχοτομούμενες απαντήσεις** του τύπου ναι/όχι ή σωστό/λάθος ή ισχύει/δεν ισχύει ή αποδεκτή/απορριπτέα απάντηση ή συμφωνώ/διαφωνώ,[11] και β) σε **πολυτεμνόμενα ή βαθμολογικά διατάξιμα ερωτήματα** (polytomous (ενίοτε polychotomous) items), στα οποία η κάθε απάντηση μπορεί να έχει διαφορετική βαθμολογική βαρύτητα (π.χ. στις κλίμακες μέτρησης στάσεων τύπου Likert, στη βαθμολογία ερωτήσεων ανοικτού τύπου κτλ.). Οι ερωτήσεις πολλαπλής επιλογής θεωρούνται, συνήθως, ως διχοτομούμενες, έστω και αν έχουν περισσότερες από δύο εναλλακτικές δυνατότητες απάντησης, εφόσον η απάντηση σ' αυτές μπορεί να χαρακτηρισθεί ως σωστή ή λανθασμένη και να βαθμολογηθεί, αντίστοιχα, με 1/0. Πρέπει να σημειωθεί ότι, αρχικά, δόθηκε έμφαση στα μοντέλα που εφαρμόζονται σε διχοτομούμενες απαντήσεις, λόγω της κυριαρχίας των τεστ πολλαπλής επιλογής. Αργότερα, όμως, αξιοποιήθηκαν και μοντέλα που έχουν εφαρμογή στην ανάλυση ερωτημάτων με βαθμολογική διαβάθμιση.

Τα μονοδιάστατα διχοτομούμενα μοντέλα μπορούν να υποδιαιρεθούν περαιτέρω: α) στα **τριπαραμετρικά**, τα οποία περιλαμβάνουν τρία χαρακτηριστικά των ερωτημάτων μιας εξεταστικής δοκιμασίας, ήτοι τη δυσκολία τους, τη διακριτική τους ισχύ και την πιθανότητα που έχει ένα άτομο να απαντήσει στην τύχη σωστά σε ένα ερώτημα, β) στα **διπαραμετρικά** που αναφέρονται σε δύο μόνο χαρακτηριστικά των ερωτημάτων και συγκεκριμένα στη θέση τους στον άξονα δυσκολίας και στη διαχωριστική τους ισχύ (διακριτικότητα) και γ) στα **μονοπαραμετρικά** μοντέλα που εξετάζουν μόνο τη θέση των ερωτημάτων στον άξονα δυσκολίας, σε συνάρτηση με τον οποίο βρίσκεται το επίπεδο ικανότητας που απαιτείται για να απαντήσει κάποιος σωστά σ' ένα ερώτημα.[12] Στα μονοπαραμετρικά[13] μοντέλα κατατάσσεται και αυτό του Δανού Rasch, το οποίο, κυρίως, χρησιμοποιείται στη σύγχρονη διαδικασία των μετρήσεων και στο οποίο θα δώσουμε περισσότερη έμφαση.

Η λογική που διέπει τα τριπαραμετρικά μοντέλα εκφράζεται μαθηματικά ως ακολούθως:

11. Μη διχοτομικού τύπου απαντήσεις σε ερωτήματα μπορούν να γίνουν αντικείμενο επεξεργασίας με αντίστοιχα μοντέλα, αν, λογικά, είναι δυνατόν να χωριστούν σε δύο κατηγορίες.

12. Στη σχετική βιβλιογραφία αναφέρονται ακόμη και τα τετραπαραμετρικά μοντέλα με την προσθήκη μιας ασύμπτωτης (asymptote) υπερκείμενης παραμέτρου, αλλά δεν χρησιμοποιούνται στην εκπαιδευτική πράξη και για το λόγο αυτό δεν εξετάζονται εδώ.

13. Ορισμένοι συγγραφείς (π.χ. Hashway, 1998: 58) υποστηρίζουν την άποψη ότι και το μοντέλο Rasch μπορεί να θεωρηθεί ως τριπαραμετρικό στο οποίο οι δύο παράμετροι b και c έχουν προκαθορισμένες τιμές (b = 1 και c = 0).





$$p(\theta) = c + (1-c)\frac{1}{1+e^{-a(\theta-b)}} = c+(1-c)\ \frac{1}{1+e^{-L}}$$

ή

$$p(\theta) = c + (1-c)\frac{1}{1+exp(-a(\theta-b))} = c+(1-c)\ \frac{1}{1+exp(-L)}$$

όπου: α) Το a είναι η παράμετρος της διακριτικότητας ενός ερωτήματος, το εύρος της οποίας ορίζεται από τη σχέση $-\infty \le a \le +\infty$. Στην πράξη, όμως, λαμβάνει, συνήθως, τιμές που εμπεριέχονται στο διάστημα: $-3.0 \le a \le +3.0$. Ερωτήματα με αρνητικό δείκτη διακριτικότητας, χρήζουν προσεκτικής εξέτασης, επειδή ενδέχεται να περιέχουν κάποιο σφάλμα, και είναι δυνατόν να αφαιρούνται από ένα τεστ. β) Το b εκφράζει την παράμετρο της δυσκολίας ενός ερωτήματος. Αντιστοιχεί στο σημείο της κλίμακας ικανότητας στο οποίο η πιθανότητα σωστής απάντησης είναι $p(\theta) = c + (1 - c)\ 0.5 = (1 + c) / 2$ (ενώ στα μοντέλα της μιας και των δύο παραμέτρων το $p(\theta)$ είναι ίσο με 0.5). Το θεωρητικό εύρος του b ορίζεται από την ανισότητα $-\infty \le b \le +\infty$, αλλά στην πράξη παίρνει και αυτό τιμές από –3.0 έως +3.0 (και κατ' άλλους από –2.0 έως +2.0). γ) Το c αντιστοιχεί στην πιθανότητα εύρεσης μιας σωστής απάντησης στην τύχη. Το θεωρητικό εύρος του c ορίζεται από τη σχέση $0.0 \le c \le 1.0$. Στην πράξη, όμως, δεν γίνονται αποδεκτές τιμές πάνω από το 0.35 (Baker, 2001: 38). δ) Το $\theta$ εκφράζει, όπως αναφέρθηκε ήδη, το επίπεδο ικανότητας, των εξεταζομένων και δ) το $L$ είναι η λογιστική απόκλιση $a(\theta - b)$ (Logit (L): logistic deviation).[14]

Έστω ότι ζητείται η πιθανότητα ενός εξεταζομένου με $\theta = -2$ να απαντήσει σωστά σε ένα ερώτημα πολλαπλής επιλογής με τέσσερις δυνατές επιλογές ($c = 0.25$), διακριτικής ισχύος $a = 1.5$ και βαθμού δυσκολίας $b = 1.0$.

Για να βρούμε το ζητούμενο αυτό, που αντιστοιχεί σε τριπαραμετρικό μοντέλο, εργαζόμαστε ως ακολούθως:

Υπολογίζουμε πρώτα το Logit: $L = 1.5\ (-2 - 1) = -4.5$.

Βρίσκουμε το $\exp(-L) = \exp(-(-4.5) = \exp(4.5) = 90.017$.

Αντικαθιστούμε τις τιμές αυτές στον τύπο που αναφέρθηκε προηγουμένως και έχουμε:

$$p(-2) = 0.25 + (1-0.25)\frac{1}{1+90.017} = 0.25 + (0.75 \cdot 0.011) = 0.25 + 0.008 = 0.258$$

14. Το e συμβολίζει τη βάση των νεπερίων λογαρίθμων.

Η πιθανότητα ορθής απάντησης από το συγκεκριμένο άτομο στο υπό εξέταση ερώτημα είναι 0.26 (ή 26%).

Ένας άλλος εναλλακτικός τρόπος για τον υπολογισμό της πιθανότητας ($p_i(\theta)$ ενός ατόμου, με ορισμένο $\theta$ χαρακτηριστικό, να απαντήσει σωστά σ' ένα ερώτημα, σύμφωνα με τη Θ.Ι.Μ.Α.Ε., είναι ο εξής:[15]

$$p(\theta) = \Phi\left(\frac{\theta - b_i}{\sigma_i}\right)$$

όπου Φ είναι η αθροιστική κατανομή της συνάρτησης της τυπικής κανονικής κατανομής και $\sigma_i$ η τυπική απόκλιση του σφάλματος μέτρησης για το ερώτημα i, ενώ τα θ και b έχουν την ίδια σημασία με αυτήν που αναφέρθηκε προηγουμένως. Ο τρόπος αυτός είναι γνωστός ως μοντέλο της κανονικής κατανομής (normal ogive model) και η εφαρμογή του, όπως και αυτή των προηγούμενων τρόπων, υποστηρίζεται από σχετικά στατιστικά προγράμματα ηλεκτρονικών υπολογιστών.

Στα διπαραμετρικά μοντέλα εξετάζονται, όπως ήδη αναφέρθηκε, δύο μόνο παράμετροι (η διακριτικότητα (a) και η θέση (b) του ερωτήματος στην κλίμακα δυσκολίας), θεωρώντας ότι το c, δηλαδή οι σωστές απαντήσεις στην τύχη, ισούται με 0, οπότε ο τύπος, που αναφέρθηκε προηγουμένως, μπορεί να γραφεί ως εξής:

$$p(\theta) = \frac{1}{1+e^{-L}} = \frac{1}{1+e^{-a(\theta-b)}} = 1/(1+\exp(-a(\theta-b)))$$

όπου a και b έχουν την ίδια σημασία μ' αυτήν που μνημονεύθηκε παραπάνω. Το b, όμως, στην περίπτωση αυτή αντιπροσωπεύει το σημείο της κλίμακας ικανότητας στο οποίο η πιθανότητα (p(θ)) να δοθεί σωστή απάντηση αντιστοιχεί σε 0.5 (50%).

Έστω το ακόλουθο παράδειγμα: Ποια είναι η πιθανότητα ενός εξεταζομένου να απαντήσει σωστά σε μια ερώτηση βαθμού δυσκολίας 3, όταν το θ = 1 και α = 0.5.

Υπολογίζουμε το Logit: L = 0.5 (1 – 3) = –1

Βρίσκουμε το exp (–L) = exp (–(–1) = exp(1) = 2.718 και έχουμε:

$$p(1) = \frac{1}{1+2.718} = \frac{1}{3.718} = 0.269.$$

Η ζητούμενη πιθανότητα είναι 0.27 (ή 27%).

15. Βλ. http://en.wikipedia.org/wiki/Item_response_theory – (ανακτήθηκε στις 19 Σεπτεμβρίου 2012).





## 10.4. Το μοντέλο του Rasch

Το συγκεκριμένο μοντέλο αποτελεί ένα από τα πιο σημαντικά μαθηματικά μοντέλα που χρησιμοποιούνται, σήμερα, για την ανάλυση των ερωτημάτων των διαφόρων τεστ.[16] Βασίζεται στο σκαλόγραμμα του Guttman και αναπτύχθηκε, παράλληλα, αλλά χωριστά από την ανάπτυξη της Θ.Μ.Ι.Α.Ε. Για το λόγο αυτό, ορισμένοι συγγραφείς δεν το θεωρούν ως υποπερίπτωση της εν λόγω θεωρίας (Demars, 2010: 15), αν και αντανακλά τη φιλοσοφία της και ακολουθεί τη λογική του μονοπαραμετρικού της μοντέλου, οπτική υπό την οποία το εξετάζουμε και εμείς εδώ. Ενίοτε, επίσης, χρησιμοποιούνται για το μοντέλο αυτό διαφορετικά ορισμένα από τα σύμβολα που αναφέρθηκαν προηγουμένως (π.χ. δ αντί b και β αντί θ).

Το μοντέλο του Rasch, που δημοσιεύθηκε το 1960, δεν λαμβάνει υπόψη τις σωστές απαντήσεις στην τύχη, θεωρώντας ότι αυτές δεν επηρεάζουν σημαντικά τα δεδομένα της κατάταξης των εξεταζομένων, εφόσον υπόκεινται στο νόμο της τυχαιότητας. Οι δύο παράμετροι που προσδιορίζουν τα χαρακτηριστικά ενός ερωτήματος, ήτοι η διακριτικότητα και η ευχέρεια ενός εξεταζομένου να απαντήσει, εκφράζονται με μια ενιαία τιμή a = 1.0. Κατα τον τρόπο αυτό, μόνο το χαρακτηριστικό της δυσκολίας μπορεί να λαμβάνει διαφορετικές τιμές, λόγος για τον οποίο το μοντέλο του Rasch κατατάσσεται, όπως ήδη αναφέρθηκε, στα μονοπαραμετρικά μοντέλα. Το πλεονέκτημά του συνίσταται στο ότι μπορεί να εφαρμοσθεί και σε μικρά δείγματα (100-200 άτομα), ενώ τα πολυπαραμετρικά μοντέλα απαιτούν μεγάλα δείγματα, προκειμένου να εξασφαλίζεται σχετικό εύρος ως προς το θ, γεγονός που μειώνει την πρακτική τους χρησιμότητα (Livingston, 2006).[17]

Το μοντέλο του Rasch για ερωτήματα με διχοτομούμενες απαντήσεις μπορεί, από μαθηματική άποψη, να γραφεί ως εξής:

$$p(\theta) = \frac{1}{1+e^{-1(\theta-b)}} = 1/(1+\exp(-1(\theta-b)))$$

Η σημασία των διαφόρων συμβόλων είναι η ίδια με αυτήν που αναφέρθηκε προηγουμένως. Επαναλαμβάνουμε ότι στο μοντέλο του Rasch η παράμετρος που αντιστοιχεί στη διακριτικότητα είναι ίση με 1.0 (ήτοι a = 1.0).

16. Για το μοντέλο αυτό βλ. από την ελληνική βιβλιογραφία το οικείο κεφάλαιο του βιβλίου: Κατσής κ.ά. (2010: 347 κ.έ.) Βλ., επίσης, Αλεξόπουλος (1998: 110 κ.έ.) και (Μυλωνάς, 2012: 315-318).

17. Το συνιστώμενο μέγεθος των σχετικών δειγμάτων ποικίλλει, ανάλογα με τη μορφή του εφαρμοζόμενου μοντέλου. Στις περισσότερες περιπτώσεις υπερβαίνει τα 500-1000 άτομα (περισσότερες λεπτομέρειες βλ. στο Demars, 2010: 34-37).

Έστω ότι ζητείται ο υπολογισμός του p(θ), όταν b = 1.0 και θ = 2.0.
Logit: L = 1 (2 – 1) = 1.
Exp (–L) = exp (–1) = 0,3679.

$$p(2)=\frac{1}{1+0.3679}=\frac{1}{1.3679}=0.731.$$

Άρα η πιθανότητα ενός εξεταζόμενου, με χαρακτηριστικό ικανότητας 2, να απαντήσει σε ένα ερώτημα δυσκολίας 1 είναι 0.73.

Η συχνότητα της χρήσης του συγκεκριμένου μοντέλου οφείλεται στις εξής, κυρίως, μαθηματικές του ιδιότητες: Ο αριθμός των παρατηρούμενων ορθών απαντήσεων θεωρείται επαρκές στοιχείο για το στατιστικό υπολογισμό της μετρούμενης λανθάνουσας ικανότητας (θ). Αυτό σημαίνει ότι όλα τα άτομα που απάντησαν σωστά σε ίδιο αριθμό ερωτημάτων έχουν τον ίδιο βαθμό ικανότητας, ακόμη κι αν τα ερωτήματα αυτά ήσαν διαφορετικά, αλλά μετρούσαν την ίδια λανθάνουσα ικανότητα. Τούτο δεν ισχύει, όταν χρησιμοποιούνται άλλα μοντέλα, γεγονός που δυσκολεύει την ερμηνεία και την κατανόηση των αποτελεσμάτων ενός τεστ από μη εξοικειωμένους με την υπό εξέταση θεωρία. Κατ' ανάλογο τρόπο, το ποσοστό των ορθών απαντήσεων αποτελεί επαρκή στατιστικό δείκτη της δυσκολίας των ερωτημάτων. Μια άλλη ιδιότητά του σχετίζεται με το γεγονός ότι οι χαρακτηριστικές καμπύλες των ερωτημάτων δεν διασταυρώνονται και δείχνουν, κατά τρόπο πιο κατανοητό το διαφορετικό βαθμό, δυσκολίας τους.

Παράγοντες που ενδέχεται να στρεβλώνουν τις υποθέσεις του παραπάνω μοντέλου είναι: α) η πιθανότητα ανεύρεσης σωστών απαντήσεων στην τύχη, β) η αλληλεπίδραση μεταξύ ερωτημάτων, γ) η διαφοροποιημένη λειτουργία τους (differential item functioning) μεταξύ διαφορετικών ομάδων και δ) άλλα χαρακτηριστικά τους (π.χ. ερωτήματα στα μαθηματικά, τα οποία είναι δυνατόν να μετρούν εννοιολογική κατανόηση και υπολογιστική ακρίβεια) (Wu & Adams, 2007: 74).[18]

## 10.5. Έλεγχος της ανεξαρτησίας των ερωτημάτων

Σημειώσαμε προηγουμένως ότι μια από τις βασικές παραδοχές της νέας θεωρίας για τις μετρήσεις είναι η τοπική ανεξαρτησία (local independence) των ερωτημάτων, τα οποία συνθέτουν ένα τεστ, υπό την έννοια ότι δεν

18. Άλλα μονοδιάστατα λογιστικά μοντέλα βλ. στο Embretson & Reise (2000: 76-81). Στο ίδιο έργο (σσ. 82- 92) παρουσιάζονται, επίσης, πολυδιάστατα μοντέλα που έχουν εφαρμογές σε διχοτομικά δεδομένα (binary data). Όπως έχουμε ήδη αναφέρει, η παρουσίασή τους εκφεύγει των στόχων της παρούσας εργασίας.





επικαλύπτονται μεταξύ τους ως προς τη μετρούμενη ικανότητα (θ). Αυτό σημαίνει ότι, έχοντας υπό τον έλεγχο τη μεταβλητή θ, δεν πρέπει να υπάρχει συσχέτιση μεταξύ δύο ερωτημάτων. Αν αυτό δεν ισχύει, τότε είναι πιθανόν να υπεισέρχεται στη διαδικασία μέτρησης και άλλη διάσταση, πλην εκείνης στην οποία αυτή αναφέρεται, γεγονός που επηρεάζει αρνητικά την εγκυρότητα των δεδομένων. Για τον έλεγχο της ανεξαρτησίας αυτής εφαρμόζεται το τεστ Q του Yen (βλ. περισσότερες πληροφορίες στο Yen, 1984).

## 10.6. Εκτίμηση των παραμέτρων των ερωτημάτων και έλεγχος της προσαρμογής τους στο επιλεγόμενο μοντέλο

Τα μοντέλα που βασίζονται στη Θ.Μ.Ι.Α.Ε. προσαρμόζουν τη διακριτική ισχύ των ερωτημάτων και παρέχουν ενδείξεις για το βαθμό εναρμόνισής τους με το ακολουθούμενο μαθηματικό πρότυπο έτσι, ώστε να επιτυγχάνεται ο έλεγχος της καταλληλότητάς τους. Η αξία της διάγνωσης αυτής για τους κατασκευαστές των τεστ και τους δημιουργούς τραπεζών ερωτήσεων είναι προφανής και δεν χρήζει περαιτέρω επισημάνσεων.

Ως μια από τις θετικές συμβολές των υπό εξέταση μοντέλων θεωρείται, επίσης, η διεύρυνση της έννοιας της αξιοπιστίας ενός εξεταστικού μέσου. Όπως αναφέραμε ήδη, η νέα θεωρία για τις μετρήσεις έδειξε ότι η αξιοπιστία ενός τεστ δεν έχει ενιαίο χαρακτήρα για το σύνολο του τεστ. Δεδομένα ακραία π.χ. ως προς τη διασπορά των αποτελεσμάτων του απηχούν λάθη σε μεγαλύτερο βαθμό απ'ό,τι δεδομένα που επικεντρώνονται στο μέσο της κλίμακας μέτρησης.

Μέχρι τώρα αναφέραμε ότι οι παράμετροι των ερωτημάτων (διακριτικότητα, δυσκολία, τυχαιότητα ορθής απάντησης) και η παράμετρος που αντιστοιχεί στη λανθάνουσα ικανότητα των εξεταζομένων εκφράζονται στην ίδια μετρική κλίμακα που έχει μέσο το 0 και μονάδα μέτρησης το 1 με απεριόριστα, θεωρητικώς, περιθώρια θετικής και αρνητικής κατεύθυνσης σε σχέση με το 0. Χρησιμοποιήσαμε, επίσης, παραδείγματα για να εξηγήσουμε βασικές πτυχές της Θ.Μ.Ι.Α.Ε. στα οποία δώσαμε υποθετικές τιμές στις εν λόγω παραμέτρους. Στην πράξη, όμως, οι τιμές αυτές υπολογίζονται σε συνάρτηση με τις απαντήσεις διχοτομικού τύπου (1/0), τις οποίες δίνουν οι εξεταζόμενοι σε συγκεκριμένα ερωτήματα. Ο υπολογισμός των παραμέτρων αυτών συνιστά τη διαδικασία «**βαθμονόμησης**» (calibration) του τεστ. Οι τιμές των παραπάνω παραμέτρων υπολογίζονται χωριστά για κάθε ερώτημα (item), αφού, όπως έχουμε ήδη τονίσει, καθένα μετρά διαφορετική πτυχή της λανθάνουσας ικανότητας, και χωριστά για κάθε εξεταζόμενο, επειδή κάθε άτομο κατέχει σε διαφορετικό βαθμό τη μετρούμενη ικανότητα.

Σημειώνουμε ότι ο προσδιορισμός των παραμέτρων αυτών δεν είναι δυνατός, όταν σε ένα ερώτημα απαντούν όλοι οι εξεταζόμενοι ή κανείς από αυτούς. Το ίδιο ισχύει και στην περίπτωση στην οποία κάποιος εξεταζόμενος απαντά σε όλα τα ερωτήματα που του τίθενται ή δεν απαντά σε κανένα απ' αυτά. Για τον παραπάνω λόγο, τέτοιου είδους περιπτώσεις δεν λαμβάνονται υπόψη κατά τη βαθμονόμηση ενός τεστ.

Ο υπολογισμός των τιμών των υπό εξέταση παραμέτρων γίνεται θεωρητικά σε δύο φάσεις (stages) σύμφωνα με την τεχνική που προτάθηκε από τον Birnbaum (αναφέρεται στο Baker, 2001: 134). Στην πράξη, όμως, οι δύο αυτές φάσεις πραγματοποιούνται ταυτόχρονα. Για το σκοπό αυτό εφαρμόζεται μια επαναληπτική διαδικασία (iterative procedure) που στηρίζεται στην εκτίμηση της μέγιστης πιθανοφάνειας (maximum likehood estimation), με τη βοήθεια ειδικών λογισμικών. Η διαδικασία αυτή ακολουθεί τα βήματα που περιγράφονται, αδρομερώς, παρακάτω.

1) Ορίζεται μια αρχική επιτρεπτή τιμή για κάθε παράμετρο (a, b, θ), καθώς για καθεμιά από τις μεταβλητές: «dstopa», «dstopb» και «dstopθ», οι οποίες χρησιμοποιούνται για τον έλεγχο τερματισμού της σχετικής διαδικασίας.[19]
2) Καταχωρίζεται σε μια βοηθητική μεταβλητή $\theta_0$, η τρέχουσα τιμή της θ, υπολογίζεται, στη συνέχεια, η πρώτη μερική παράγωγος ως προς το $\theta_0$ της συνάρτησης $\ln(L(\theta))$ και καταχωρίζεται στη μεταβλητή Lθ1. Υπολογίζεται, κατόπιν, η δεύτερη μερική παράγωγος ως προς $\theta_0$ της συνάρτησης $\ln(L(\theta))$ και καταχωρίζεται στη μεταβλητή Lθ2. Υπολογίζεται η διαφορά $\theta_0$ - Lθ1 / Lθ2 και καταχωρίζεται στη μεταβλητή $\theta_{new}$.
3) Αν η απόλυτη διαφορά $| \theta_0 - \theta_{new} |$ είναι μεγαλύτερη ή ίση της τιμής της μεταβλητής dstopθ, τότε στη θ προσδίδεται η τιμή της $\theta_{new}$.
4) Κατά τον ίδιο τρόπο ορίζεται σε μια βοηθητική μεταβλητή $a_0$, η τρέχουσα τιμή της a, υπολογίζεται η πρώτη μερική παράγωγος ως προς $a_0$ της συνάρτησης $\ln(L(\theta))$ και καταχωρίζεται στη μεταβλητή La1. Υπολογίζεται κατόπιν η δεύτερη μερική παράγωγος ως προς $a_0$ της συνάρτησης $\ln(L(\theta))$ και καταχωρίζεται στη μεταβλητή La2. Προσδιορίζεται η διαφορά $| a_0 - La1/La2 |$ και καταχωρίζεται στη μεταβλητή $a_{new}$. Αν η απόλυτη διαφορά $| a_0 - a_{new} |$ είναι μεγαλύτερη ή ίση της τιμής της dstopa, τότε στη μεταβλητή a προσδίδεται η τιμή της $a_{new}$.
5) Κατ' ανάλογο τρόπο, ορίζεται σε μια βοηθητική μεταβλητή $b_0$, η τρέχουσα τιμή της b, υπολογίζεται η πρώτη μερική παράγωγος ως προς $b_0$ της συνάρτησης $\ln(L(\theta))$ και καταχωρίζεται στη μεταβλητή Lb1. Υπολογίζεται κατόπιν η δεύτερη μερική παράγωγος ως προς $b_0$ της συνάρτησης $\ln(L(\theta))$ και καταχωρίζεται στη μεταβλητή Lb2. Υπολογίζεται η διαφορά $b_0$ – Lb1/ Lb2 και καταχωρίζεται στη μεταβλητή $b_{new}$. Αν η απόλυτη διαφορά $| b_0 - b_{new} |$ είναι μεγαλύτερη ή ίση της τιμής της dstopb, τότε στη μεταβλητή b προσδίδεται η τιμή της $b_{new}$.
6) Αν $| a - a_{new} | <$ dstopa, $| b - b_{new} | <$ dstopb και $| \theta - \theta_{new} | <$ dstopθ, τότε η διαδικασία τερματίζεται. Ειδάλλως επαναλαμβάνεται από το βήμα 3 και εξής, συμπεριλαμβανομένου και αυτού.

Στο θέμα της αληθούς μέτρησης της λανθάνουσας ικανότητας θα επανέλθουμε παρακάτω.

Η παραπάνω διαδικασία, η οποία με τη μορφή λογικού διαγράμματος παρατίθεται στο παράρτημα, πραγματοποιείται σήμερα με τη βοήθεια ειδικών λογισμικών, τα οποία μετά την εισαγωγή στον ηλεκτρονικό υπολογιστή των δεδομένων που αφορούν στις απαντήσεις των εξετασθέντων με ένα τεστ, δίνουν τις τιμές όλων των υπό εξέταση παραμέτρων. Τα πιο γνωστά από τα λογισμικά αυτά είναι τα εξής: το RUMM, το MULTILOG, το PARSCALE, το BILOG, το QUEST, το CONQUEST, το BILOG-GM, το TESTFACT, το PARAM-3PL, το ICL, το JMETRIK, το FLEXMIRT, το XCALIBRE και το WINSTEPS (Embretson & Reise, 2000, www.en.wikipedia.org/wiki/Psychometric_software).[20]

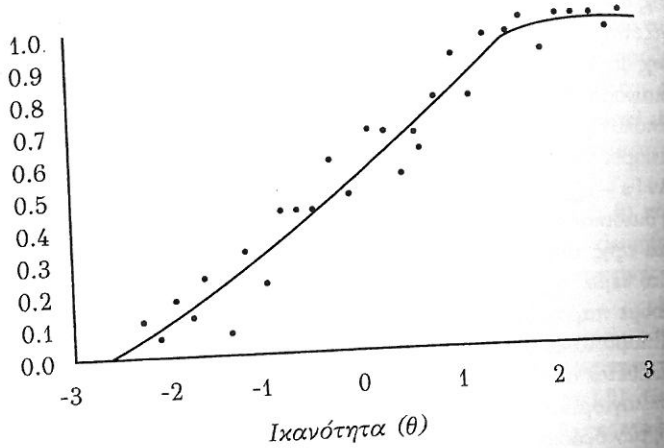
Με ανάλογο τρόπο υπολογίζεται και η τιμή της παραμέτρου c. Σχετικές λεπτομέρειες υπάρχουν στο παράρτημα.

Το ζητούμενο είναι να δούμε κατά πόσο η καμπύλη που προκύπτει από τις παρατηρηθείσες τιμές (O) της πιθανότητας $p_o(\theta)$ ταιριάζει με τα χαρακτηριστικά της καμπύλης P(θ), που υπολογίζονται με βάση τους τύπους, τους οποίους ήδη αναφέραμε, ανάλογα με το χρησιμοποιούμενο μοντέλο. Επαναλαμβάνουμε ότι οι τιμές των παραμέτρων a (διακριτικότητα), b (δυσκολία) και c (πιθανότητα να δοθεί στην τύχη ορθή απάντηση) ορίζονται

19. Στις αναλυτικές μεθόδους ο υπολογισμός των ζητούμενων τιμών γίνεται με απόλυτη ακρίβεια. Επειδή, όμως, δεν είναι πάντα εφικτή η εφαρμογή τους, χρησιμοποιούμε αριθμητικές μεθόδους, σύμφωνα με τις οποίες ο υπολογισμός των ζητούμενων τιμών γίνεται προσεγγιστικά. Η μέθοδος Newton Raphson, στην οποία στηρίζεται η αναφερόμενη διαδικασία είναι αριθμητική. Κατά συνέπεια, ο υπολογισμός των ζητούμενων τιμών γίνεται προσεγγιστικά. Η «dstop» είναι μια μεταβλητή στην οποία καταχωρίζουμε μια τιμή που εκφράζει τον επιθυμητό βαθμό προσέγγισης της ακριβούς τιμής της αντίστοιχης παραμέτρου. Ως τιμή, λοιπόν, μπορεί να τεθεί π.χ. dstop = 0.001 ή άλλη για μεγαλύτερη ή μικρότερη ακρίβεια προσέγγισης. Ανεξάρτητα από την «dstop», στις παραμέτρους a, b, θ δίνονται, όπου χρειάζεται, αρχικές τιμές από το πεδίο των επιτρεπτών γι' αυτές αριθμών.

20. Αναλυτικότερα στοιχεία για την εφαρμογή του Winsteps βλ. στο Κατσής κ.ά. (2010).

με βάση τη μέγιστη πιθανότητα εκτίμησης. Σύμφωνα με τη διαδικασία αυτή, γίνεται προσθήκη νέων ερωτημάτων στο εφαρμοζόμενο μοντέλο, έως ότου επιτευχθεί η μέγιστη δυνατή συμφωνία ανάμεσα στην καμπύλη των παρατηρούμενων και των θεωρητικών αναλογιών.




***Σχήμα 19.*** *Η καμπύλη η οποία βρίσκεται στη μικρότερη δυνατή απόσταση από τα σημεία τομής εκφράζει την καμπύλη κάθε ερωτήματος*

Η διαδικασία αυτή συνεχίζεται με τη βοήθεια των λογισμικών που αναφέρθηκαν, έως ότου επιτευχθεί η μέγιστη σύγκλιση μεταξύ παρατηρούμενων και θεωρητικών δεδομένων, οπότε σταματάει η προσαρμογή και δίνονται από το πρόγραμμα του υπολογιστή οι τιμές των ζητούμενων παραμέτρων a και b, σύμφωνα με όσα μνημονεύθηκαν προηγουμένως.

Για τον έλεγχο της συμφωνίας μεταξύ της κατανομής των παρατηρούμενων για ένα ερώτημα πραγματικών και θεωρητικών πιθανοτήτων εφαρμόζεται το κριτήριο προσαρμοστικότητας $x^2$, το οποίο δίνεται από τον τύπο:





$$x^2 = \sum_{j=1}^{k} \Sigma m_j \frac{[p(\theta_j) - P(\theta_j)]^2}{p(\theta_j)P(\theta_j)}$$

όπου k ο αριθμός των ομάδων διαφορετικής ικανότητας, ενός πληθυσμού εξεταζομένων,[21] θ το επίπεδο ικανότητας κάθε ομάδας j (j = 1 ... k), $m_j$ ο αριθμός των εξετασθέντων που έχουν ικανότητα $\theta_j$, ήτοι των ατόμων της κάθε ομάδας, p(θ) η παρατηρούμενη αναλογία των ορθών απαντήσεων σε κάθε ομάδα και P(θ) η εκτιμούμενη πιθανότητα ορθών απαντήσεων σε κάθε ομάδα, με βάση το χρησιμοποιούμενο μοντέλο της Θ.Μ.Ι.Α.Ε. (Baker, 2001: 50).

Εάν η τιμή ενός επιτυγχανόμενου δείκτη αποκλίνει σημαντικά απ' αυτήν που έχει προσδιορισθεί ως κριτήριο, η κατανομή των χαρακτηριστικών των ερωτημάτων, με βάση τη Θ.Μ.Ι.Α.Ε, εντοπίζει ποια από τα ερωτήματα ενός τεστ δεν εναρμονίζονται με το μοντέλο που έχει επιλεγεί. Αυτό μπορεί να συμβαίνει είτε λόγω λάθους στο συγκεκριμένο ερώτημα είτε λόγω ευρείας διασποράς των σχετικών απαντήσεων. Αν μεγάλος αριθμός ερωτημάτων δείχνει σημαντικές αποκλίσεις από τις προσδοκίες του μοντέλου, επιβάλλεται να εφαρμοσθεί άλλο μοντέλο ανάλυσης ή να γίνουν μεταβολές στα ερωτήματα του τεστ, ώστε να επιτευχθεί το άριστο αποτέλεσμα. Εάν τα πρωτογενή δεδομένα μετατρέπονται σε τιμές ικανότητας, με βάση ένα μοντέλο της Θ.Μ.Ι.Α.Ε., τα οποία αποκλίνουν από αυτό, τότε τα αποτελέσματα δεν είναι έγκυρα.

Η παραπάνω αξιολόγηση των ερωτημάτων γίνεται με τη χρήση των στατιστικών δεικτών: infit και outfit (Master & Wright, 1982· Κατσής κ.ά. 2010: 350). Ο πρώτος αφορά ερωτήματα που βρίσκονται στο μέσο ενός τεστ, ενώ ο δεύτερος αφορά ερωτήματα που βρίσκονται στα άκρα του. Οι τιμές των δεικτών αυτών για το μοντέλο Rasch δίνονται από τους ακόλουθους τύπους:

$$\text{InfitMS}_i = \frac{\sum_{n=1}^{N} y^2_{in}}{\sum_{n=1}^{N} W_{in}}$$

$$\text{OutfitMS}_i = \frac{\sum z^2_{ni}}{N},$$

21. Κάθε ομάδα περιλαμβάνει εξεταζόμενους ίδιας ικανότητας (θ).

όπου το $\sum_{n=1}^{N} y_{in}^2$ αναφέρεται στα παρατηρούμενα σφάλματα, το $\sum_{n=1}^{N} W_{in}$ σχετίζεται με τα προσδοκώμενα σφάλματα και το $\frac{\sum z_{ni}^2}{N}$ εκφράζει το μέσο όρο των σταθμισμένων σφαλμάτων τόσο για τους συμμετέχοντες όσο και για τα ερωτήματα του τεστ. Αν τα ερωτήματα ταιριάζουν με το μοντέλο, τότε οι παραπάνω δείκτες πρέπει να παίρνουν τιμές γύρω στο 1.0. Αν οι τιμές αυτές είναι πολύ μεγαλύτερες (πάνω από 1.5), τότε υπάρχει σημαντικό σφάλμα και το αντίστοιχο ερώτημα πρέπει ή να αφαιρεθεί από το τεστ ή να διορθωθεί. Αν οι τιμές τους βρίσκονται κοντά στο μηδέν, τότε τα αντίστοιχα ερωτήματα δεν δίδουν σημαντικές πληροφορίες γι' αυτό που μετρά το τεστ. Κατά τον Linacre (αναφέρεται στο Κατσής κ.ά. 2010: 352), οι παραπάνω δείκτες θα πρέπει να βρίσκονται μεταξύ 0.5 και 1.5, για να θεωρηθεί ότι τα αντίστοιχα ερωτήματα συμβάλλουν στην εκτίμηση του εξεταζόμενου λανθάνοντος χαρακτηριστικού. Οι σχετικοί δείκτες χρησιμεύουν, ακόμη, στον έλεγχο της σειράς με την οποία τίθενται τα ερωτήματα, στοιχείο το οποίο χρησιμεύει στον καθορισμό των βαθμολογικών στάνταρ (βλ. οικείο κεφάλαιο).

Σημειώνεται, όμως, ότι κατά τον έλεγχο της καταλληλότητας των ερωτημάτων και τις ενδεχόμενες αναμορφώσεις τους θα πρέπει να συνεκτιμώνται και άλλα στοιχεία που έχουν σχέση με την εγκυρότητα αξιολόγησης του μετρούμενου χαρακτηριστικού. Η αξιολόγηση αυτή διευκολύνεται για τα τεστ επίδοσης από τους σκοπούς, τους στόχους και τις προδιαγραφές της ύλης που καθορίζουν τα ισχύοντα αναλυτικά προγράμματα.

Οι ίδιοι δείκτες χρησιμοποιούνται και για τον έλεγχο του βαθμού στον οποίο οι εξεταζόμενοι, συμπεριφέρονται ή όχι κατά τρόπο που ταιριάζει με το μοντέλο.[22] Ο έλεγχος αυτός παρέχει τη δυνατότητα αφαίρεσης από το δείγμα, με βάση το οποίο αξιολογείται ένα τεστ, των ατόμων εκείνων η συμπεριφορά των οποίων δεν εναρμονίζεται με τη λογική του μοντέλου που εφαρμόζεται, αν και υπάρχουν απόψεις, σύμφωνα με τις οποίες η αφαίρεση αυτή πρέπει να γίνεται με πολύ προσοχή και ύστερα από τη συνεξέταση και άλλων στοιχείων (Κατσής, κ.ά. 2010). Πρέπει, όμως, να σημειωθεί ότι οι δείκτες αυτοί έχουν σχετικό και όχι απόλυτο νόημα και ισχύουν, όταν τα ερωτήματα εμφανίζουν χαρακτηριστικές καμπύλες με την ίδια κλίση. Επιπρόσθετα, απαιτείται ικανοποιητικό μέγεθος του δείγματος στο οποίο υπολογίζονται οι δείκτες αυτοί, όπως έχουμε ήδη σημειώσει.

22. Στην περίπτωση αυτή, η άθροιση γίνεται με βάση των αριθμό των ερωτημάτων (βλ. Wu & Adams, 2001: 75).





## 10.7. Ερμηνεία των τιμών των διαφόρων παραμέτρων

Οι ποιοτικές κατηγορίες της διακριτικότητας των ερωτημάτων, που αναφέραμε σε προηγούμενη υποενότητα, μπορούν να αντιστοιχηθούν σε αριθμητικές τιμές, το εύρος των οποίων έχει ως ακολούθως: α) μηδενική διακριτικότητα: 0, β) χαμηλή: 0.01 – 0.34, γ) μέση: 0.35 – 1.34, δ) υψηλή: 1.35 – 1.69 ε) πολύ υψηλή: > 1.70 και στ) άριστη = άπειρο (Baker, 2001: 35).

Αν επιθυμούμε να μετατρέψουμε τις παραπάνω αριθμητικές τιμές σε τιμές, οι οποίες ακολουθούν το μοντέλο της κανονικής κατανομής, τότε μπορούμε να τις διαιρέσουμε δια του 1.7.

Για την καλύτερη κατανόηση και την ερμηνεία των τιμών που λαμβάνουν οι εξεταζόμενες παράμετροι στο πλαίσιο της Θ.Μ.Ι.Α.Ε. θεωρούμε απαραίτητο να συμπληρώσουμε όσα αναφέρθηκαν παραπάνω με τα ακόλουθα.

Σύμφωνα με την κλασική θεωρία, η δυσκολία μιας ερώτησης καθορίζεται από την αναλογία εκείνων που απάντησαν σωστά σ' αυτήν στο σύνολο των εξετασθέντων. Κατά την ίδια θεωρία, ο δείκτης διακριτικότητας ενός ερωτήματος προσδιορίζεται με βάση τη διαφορά της συχνότητας των ορθών απαντήσεων της ανώτερης και της κατώτερης, ως προς τις επιδόσεις, ομάδας των εξετασθέντων. Σύμφωνα με τη *Θ.Μ.Ι.Α.Ε.*, η δυσκολία ενός ερωτήματος *«ορίζεται ως το σημείο της κλίμακας μέτρησης της δεξιότητας των εξεταζόμενων στο οποίο η πιθανότητα σωστής απάντησης είναι ίση με 0.5 (50%) για το μοντέλο Rasch, καθώς και για τα διπαραμετρικά μοντέλα και (1 + c)/2 για τα τριπαραμετρικά.»* (Baker, 2001: 35). Κατά συνέπεια, τα δεδομένα που αφορούν στις εκτιμώμενες παραμέτρους ερμηνεύονται μόνο σε συνάρτηση με την κλίμακα του χαρακτηριστικού που θεωρητικά μετρά ένα ερώτημα. Με άλλα λόγια, η δυσκολία είναι παράμετρος θέσης στον άξονα ικανότητας και δεν επηρεάζεται από τα χαρακτηριστικά των ερωτημάτων. Αυτό αναφέρεται ως **αρχή της ανεξαρτησίας** της ικανότητας των εξεταζομένων από τα ερωτήματα του τεστ, στην οποία έχουμε αναφερθεί. Συμπληρώνοντας τα ήδη μνημονευθέντα, σημειώνουμε ότι η αρχή αυτή ισχύει, υπό τις ακόλουθες παραδοχές: α) όλα τα ερωτήματα ενός τεστ εκτιμούν διαφορετικές όψεις του ίδιου λανθάνοντος χαρακτηριστικού και β) οι τιμές όλων των παραμέτρων εκφράζονται στην ίδια μετρική κλίμακα. Τούτο σημαίνει ότι αν ένα άτομο εξεταστεί με δύο τεστ που έχουν ισοπληθή ερωτήματα αλλά διαφορετικής δυσκολίας, η εκτίμηση της ικανότητάς του θα είναι η ίδια (όπ. παρ.: 91-92).

Υπενθυμίζουμε στον αναγνώστη ότι ένα άλλο σημαντικό στοιχείο της Θ.Μ.Ι.Α.Ε. είναι η ανεξαρτησία των τιμών των εκτιμώμενων παραμέτρων από το επίπεδο ικανότητας των εξεταζόμενων (group invariance of item

parameters). Αυτό σημαίνει ότι οι τιμές των υπό εξέταση παραμέτρων εκφράζουν ιδιότητες των ερωτήσεων και όχι των εξεταζομένων, στοιχείο που διευκολύνει την ερμηνεία τους.

Η δυνατότητα διάταξης των ερωτημάτων σε ένα ενιαίο άξονα επιτρέπει τη διαβάθμισή τους, αλλά και την ανεύρεση ισοδύναμων ερωτημάτων (ήτοι ερωτημάτων με τις ίδιες παραμέτρους), στοιχείο σημαντικό για την οργάνωση των τραπεζών ερωτήσεων, καθώς και για την κατασκευή πολλαπλών μορφών του ίδιου τεστ.

## 10.8. Επιλογή μοντέλου

Σύμφωνα με όσα έχουν ήδη αναφερθεί, τα μοντέλα, που εφαρμόζονται στον προσδιορισμό των ψυχομετρικών χαρακτηριστικών των ερωτημάτων ενός τεστ ποικίλουν, ακόμη κι αν περιοριστούμε σε διχοτομικού τύπου ερωτήματα. Η ποικιλία αυτή θέτει το ερώτημα: ποιο είναι το πλέον κατάλληλο μοντέλο σε κάθε περίπτωση, το μονοπαραμετρικό, το διπαραμετρικό ή το τριπαραμετρικό; Η απάντηση στο ερώτημα αυτό εξαρτάται από πολλούς παράγοντες, όπως υπογραμμίζουν οι Embretson και Reise (2000: 72). Μεταξύ αυτών περιλαμβάνονται: α) οι παράμετροι που ο ερευνητής επιθυμεί να συνεκτιμήσει (διακριτικότητα, δυσκολία ερωτημάτων, πιθανότητα τυχαίας ορθής απάντησης), β) η βαρύτητα που έχουν τα ερωτήματα στη βαθμολογία, γ) οι ιδιότητες της χρησιμοποιούμενης κλίμακας μέτρησης και δ) ο βαθμός προσαρμογής του μοντέλου στα δεδομένα.

Αν τα ερωτήματα έχουν ίση βαρύτητα και η ελεγχόμενη παράμετρος είναι η δυσκολία τους, τότε είναι προφανές ότι ενδείκνυται το μοντέλο του Rasch. Αν επιθυμούμε να εξετάσουμε και άλλες παραμέτρους για μεγαλύτερη ακρίβεια, τότε μπορούμε να επιλέξουμε το διπαραμετρικό ή το τριπαραμετρικό μοντέλο.

Βοηθητικό στοιχείο για την αξιολόγηση του μοντέλου που ενδείκνυται σε κάθε περίπτωση είναι ο υπολογισμός του σχετικού δείκτη προσαρμογής του στα δεδομένα. Στην εκτίμηση της μέγιστης πιθανότητας το αρνητικό διπλάσιο του λογαρίθμου (-2log(p)) των δεδομένων εκφράζει το μέγιστο βαθμό απόκλισης τους από το μοντέλο (Demars, 2010: 57). Για τη σύγκριση δύο διαφορετικών μοντέλων χρησιμοποιείται το κριτήριο της λογαριθμικής πιθανοφάνειας $\chi^2$ (log-likelihood $x^2$),[23] το οποίο προκύπτει με βάση τον ακόλουθο τύπο:

23. Το $x^2$ γράφεται και ως $G^2$, για να αντιδιαστέλλεται από το Pearson $x^2$, το οποίο χρησιμοποιείται στη Στατιστική για το έλεγχο της σημαντικότητας των διαφορών μεταξύ συχνοτήτων που έχουν διαμορφωθεί με βάση κατηγορικά δεδομένα.





$$\chi^2 = 2\log(p_A) - 2\log(p_B)$$

όπου το Α αντιστοιχεί στο ένα μοντέλο και το Β στο δεύτερο. Αν η μεταξύ τους διαφορά (απόλυτη τιμή) είναι στατιστικά σημαντική, επιλέγεται εκείνο με το μικρότερο δείκτη απόκλισης. Βαθμοί ελευθερίας για την εύρεση της στατιστικής σημαντικότητας της παραπάνω διαφοράς είναι το σύνολο των ερωτημάτων που περιέχει ένα τεστ.

## 10.9. Η χαρακτηριστική καμπύλη ενός τεστ, η αληθής βαθμολογία και η εκτίμηση της ικανότητας των εξεταζομένων

Μέχρι τώρα εξετάσαμε τα χαρακτηριστικά της καμπύλης των παραμέτρων των επιμέρους ερωτημάτων ενός τεστ. Ας δούμε τι ισχύει, όταν αναφερόμαστε σε ολόκληρο το τεστ. Ας πάρουμε ως παράδειγμα ένα τεστ που αποτελείται από ορισμένο πλήθος (N) ερωτημάτων πολλαπλής επιλογής, τα οποία βαθμολογούνται με 1 μονάδα, όταν η απάντηση είναι σωστή, και με 0, όταν είναι εσφαλμένη. Αν ένας εξεταζόμενος απαντήσει σε όλες τις ερωτήσεις σωστά, τότε η επίδοσή του θα είναι ίση με το N. Αν δεν απαντήσει σε καμία σωστά, τότε θα πάρει 0. Στην πράξη, ο βαθμός (x) των εξεταζομένων ικανοποιεί τη σχέση: $0 \leq x \leq N$. Αν θεωρήσουμε ότι ένας εξεταζόμενος συμμετέχει σε μια εξέταση με το ίδιο τεστ περισσότερες από μια φορές και δεν θυμάται τα ερωτήματα από τις προηγούμενες συμμετοχές του, αυτός μπορεί κάθε φορά να έχει διαφορετικές τελικές επιδόσεις. Η μέση τιμή των επιδόσεων αυτών αντιστοιχεί, σύμφωνα με την κλασική θεωρία των μετρήσεων, σ' αυτό που ονομάζεται **αληθής ή πραγματική επίδοση** (true score). Στην περίπτωση, όμως, της Θ.Μ.Ι.Α.Ε. ο αληθής βαθμός (AB) δίνεται από τον ακόλουθο τύπο του Lawley (αναφέρεται στο Baker, 2001: 66).

$$AB_j = \sum_{i=1}^{N} p_i(\theta_j)$$

όπου $AB_j$ είναι η βαθμολογία, την οποία μπορεί να πάρει ένα άτομο, το οποίο έχει επίπεδο ικανότητας θj και συμμετέχει σε μια εξέταση με ένα τεστ που περιλαμβάνει N ερωτήσεις, το $p_i(\theta_j)$ εκφράζει την πιθανότητά του να απαντήσει σωστά στην ερώτηση i (το i παίρνει τιμές από 1 μέχρι N).

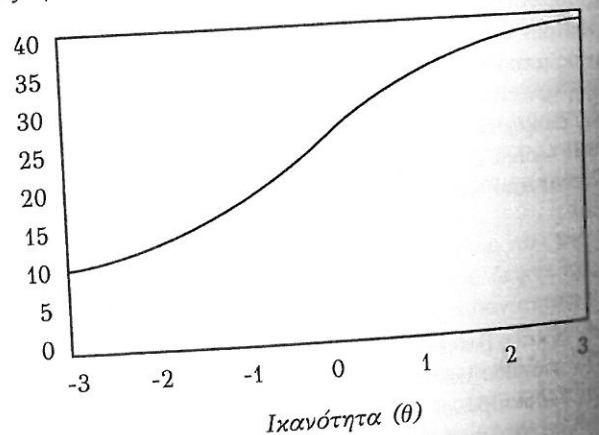
Πιο συγκεκριμένα, για να βρούμε τον αληθή βαθμό ενός ατόμου που εξετάζεται π.χ. με ένα τεστ δέκα ερωτήσεων πολλαπλής επιλογής και έχει προκαθορισμένο επίπεδο ικανότητας (π.χ. 1.0), υπολογίζουμε με βάση ένα

μοντέλο που επιλέγουμε, την πιθανότητα σωστής απάντησης σε κάθε ερώτημα, σύμφωνα με τη διαδικασία που έχουμε περιγράψει προηγουμένως (βλ. ενότητα 10.3) και αθροίζουμε τις υπολογιζόμενες πιθανότητες, όπως φαίνεται στο ακόλουθο υποθετικό παράδειγμα[24]

$$AB = .51 + .62 + .56 + .37 + .41 + .73 + .46 + .59 + .69 + .72 = 5.66$$

Για την ευκολότερη κατανόηση από το ευρύ κοινό των αποτελεσμάτων των τεστ στα οποία εφαρμόζεται η Θ.Μ.Ι.Α.Ε. είναι δυνατόν τα σχετικά δεδομένα να ανάγονται σε άλλη κατανομή με προκαθορισμένο μέσο όρο και συγκεκριμένη τυπική απόκλιση (γραμμική μετατροπή) (Lin, n.d.: 7, βλ. επίσης, το Κεφάλαιο ΙΕ΄).

Αν για κάθε επίπεδο ικανότητας υπολογίσουμε την αντίστοιχη αληθή βαθμολογία ($AB_j$), τότε προκύπτει η συνάρτηση $AB_j = AB_j(\theta_i)$, η γραφική παράσταση της οποίας αποτελεί τη **Χαρακτηριστική Καμπύλη του Τεστ** (Χ.Κ.Τ., Test Characteristic Curve - T.C.C.). Ένα ενδεικτικό παράδειγμα παριστάνεται στο σχήμα 20.




*Σχήμα 20. Ενδεικτικό παράδειγμα καμπύλης της αληθούς βαθμολογίας σε ένα τεστ*

24. Σχετικά παραδείγματα αναφέρονται στο Baker (2001: 67-69).





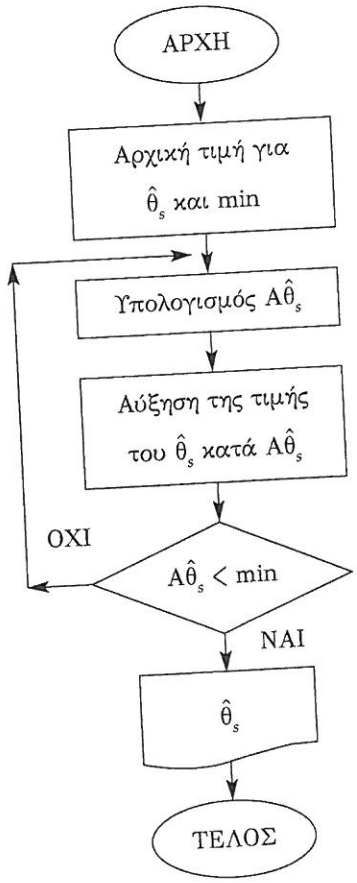
Σύμφωνα με τον Baker (2001: 71), «*ο κύριος ρόλος της χαρακτηριστικής καμπύλης ενός τεστ είναι να παράσχει ένα μέσο, με το οποίο θα μετασχηματίζεται η βαθμολογία ικανότητας σε αληθή βαθμολογία*». Ο μετασχηματισμός αυτός διευκολύνει την ερμηνεία των αποτελεσμάτων ενός τεστ. Χρησιμοποιείται, επίσης, ως κριτήριο για τον έλεγχο της ισοτιμίας πολλαπλών μορφών του ιδίου τεστ.

Ο υπολογισμός της αληθούς βαθμολογίας υπηρετεί αποτελεσματικότερα το σκοπό για τον οποίο χρησιμοποιείται ένα τεστ, ήτοι τον προσδιορισμό της πραγματικής ικανότητας των εξεταζομένων ως προς το λανθάνον στοιχείο που μετρά, σύμφωνα με το θεωρητικό πλαίσιο με βάση το οποίο κατασκευάστηκε. Εξυπηρετεί, επίσης, την πιο έγκυρη σύγκριση μεταξύ των εξεταζομένων, η οποία καθίσταται αναγκαία σε περιπτώσεις επιλογής.

Κατά την εξέταση με ένα τεστ, ο κάθε εξεταζόμενος απαντά σε ένα αριθμό ερωτημάτων (N) που, ας θεωρήσουμε, ότι βαθμολογούνται το καθένα με 1, όταν η απάντηση είναι ορθή και 0, όταν είναι λανθασμένη. Καθένα από τα ερωτήματα αξιολογεί κάποια πτυχή του λανθάνοντος χαρακτηριστικού. Ο κατάλογος των απαντήσεων 1/0 ονομάζεται **διάνυσμα απόκρισης** του εξεταζόμενου (examinee's item response vector) (Baker, 2001: 86).

Αυτό που ζητείται είναι να χρησιμοποιηθεί το παραπάνω διάνυσμα, για να εκτιμηθεί η αρχικά άγνωστη παράμετρος της ικανότητας ενός εξεταζομένου. Για να ευρεθεί αυτό, χρησιμοποιείται η διαδικασία, στην οποία αναφερθήκαμε συνοπτικά προηγουμένως. Τη διαδικασία αυτή απεικονίζει αδρομερώς το ακόλουθο λογικό διάγραμμα 21.[25]

25. Το διάγραμμα εκπονήθηκε, με βάση την εργασία του Baker (2001: 86-90).




*Σχήμα 21. Λογικό διάγραμμα διαδικασίας εκτίμησης της ικανότητας ενός εξεταζόμενου*

όπου $\hat{\theta}s = \hat{\theta}(s)$ η εκτιμώμενη ικανότητα κατά το s βήμα επανάληψης και min μια τιμή που μόλις η ποσότητα $A\hat{\theta}_s = A\hat{\theta}(s)$ γίνει μικρότερη από αυτήν, τότε η επαναληπτική διαδικασία σταματά. Στο παράρτημα παρατίθεται άλλο διάγραμμα, το οποίο στηρίζεται σε διαφορετική προσέγγιση της υπό εξέταση διαδικασίας.





Ο υπολογισμός του Αθ(s) έχει ως εξής:

$$S_1 = \sum_{i=1}^{N} (-a_i[u_i - P_i(\hat{\theta}s)])$$

$$S_2 = \sum_{i=1}^{N} ((a_i)^2 P_i(\hat{\theta}s)\, Q_i(\hat{\theta}s))$$

$$A\hat{\theta}s = S_1/S_2$$

όπου:

- $a_i$ είναι η παράμετρος της διακριτικής ισχύος για το στοιχείο i, (i = 1, 2 ... N),
- $u_i$ η απάντηση του εξεταζομένου στο στοιχείο i ($u^i = 1$ για σωστή απάντηση και $u_i = 0$ για εσφαλμένη),
- $P_i(\hat{\theta}s)]$ η πιθανότητα σωστής απάντησης στο στοιχείο i σε επίπεδο ικανότητας $\hat{\theta}s$ κατά το βήμα s της επαναληπτικής διαδικασίας για ορισμένο μοντέλο χαρακτηριστικής καμπύλης,
- $Q_i = 1 - P_i(\hat{\theta}s)]$ είναι η πιθανότητα λανθασμένης απάντησης στο στοιχείο i σε επίπεδο ικανότητας θs, κατά το βήμα s της επαναληπτικής διαδικασίας για ορισμένο μοντέλο χαρακτηριστικής καμπύλης και
- min ισούται με την τιμή εκείνη που μόλις η ποσότητα $A\theta_s = A\theta(s)$ γίνει μικρότερη από αυτήν, τότε η επαναληπτική διαδικασία σταματά (βλ. και την ενότητα 10.6).

Σημειώνουμε, τέλος, ότι η Θ.Μ.Ι.Α.Ε. αντικατέστησε την παλαιά αντίληψη για την αξιοπιστία των τεστ από την πληροφόρηση για τα ερωτήματα (I(θ) και για το τεστ (item and test information), η οποία προκύπτει ως συνάρτηση ενός μοντέλου παραμέτρων π.χ. Σύμφωνα με τη θεωρία της πληροφορίας του Fisher (Fisher Information Theory), ο υπό εξέταση δείκτης προκύπτει για τις διχοτομικής μορφής ερωτήσεις από τον πολλαπλασιασμό της πιθανότητας $p_k(\theta)$ να δοθεί σωστή απάντηση σε μια ερώτηση επί την πιθανότητα $q_k(\theta)$ να δοθεί λάθος απάντηση σ' αυτή, ήτοι:

$$I_k(\theta) = p_k(\theta)\, q_k(\theta)$$

Η συνάρτηση της πληροφορίας ενός τεστ που περιέχει N στοιχεία δίδεται από τη σχέση:

$$I(\theta) = \sum_{k=1}^{N} I_k(\theta)$$

Οι σχέσεις που δίνει το $I_x(\theta)$ εξαρτώνται από το μοντέλο της χαρακτηριστικής καμπύλης που εφαρμόζεται. Για το μοντέλο Rasch ισχύει:

$$I_x(\theta) = p_k(\theta)\, q_k(\theta)$$

Για το μοντέλο των δύο παραμέτρων η σχέση αυτή έχει ως ακολούθως:

$$I_x(\theta) = (a_k)^2 p_k(\theta)\, q_k(\theta)$$

Για το τριπαραμετρικό μοντέλο ισχύει:

$$I_k(\theta) = a_k^2 \left[\frac{q_k(\theta)}{p_k(\theta)}\right]\left[\frac{p_k(\theta)-c^2}{(1-c^2)}\right]$$

όπου $q_k(\theta) = 1 - p_k(\theta)$ και οι ποσότητες στα δεξιά ορίζονται στα αντίστοιχα μοντέλα k = 1, ... N.

Το τυπικό σφάλμα εκτίμησης (SE) για ένα δεδομένο επίπεδο χαρακτηριστικού (θ) είναι:

$$SE(\theta) = \frac{1}{\sqrt{I(\theta)}}$$

## 10.10. Η περίπτωση των ερωτημάτων με βαθμολογικά πολυτεμνόμενες (ή βαθμολογικά διατάξιμες) απαντήσεις

Στην εκπαιδευτική πράξη χρησιμοποιούνται συχνά ερωτήματα που δεν βαθμολογούνται με 1/0 (σωστό/λάθος), αλλά με βάση μια κλίμακα που μπορεί να λάβει διαφορετικές τιμές. Η νέα θεωρία των μετρήσεων μπορεί να εφαρμοστεί και στις περιπτώσεις αυτές. Τα μοντέλα που εφαρμόζονται στις παραπάνω περιπτώσεις είναι ποικίλα.[26] Η μεθοδολογική αρχή που τα διέπει είναι, κατά τους Van der Linden και Hambleton (1997: 4), η ύπαρξη μια χωριστής παραμέτρου για κάθε παράγοντα, ο οποίος επηρεάζει χωριστά τις απαντήσεις των εξεταζομένων.

Ωστόσο, εξαρτάται από την τεχνική κάθε μοντέλου ο προσδιορισμός ενός μαθηματικού σχεδίου, το οποίο μπορεί να εκφράζει όλες τις αλληλεπιδράσεις που ασκούν, ταυτόχρονα, διάφοροι παράγοντες. Τα μοντέλα της

26. Αναλυτική παρουσίασή τους βλ. στο Van der Linden & Hambleton (1997).





Θ.Μ.Ι.Α.Ε., τα οποία αναφέρονται σε ερωτήματα που επιδέχονται βαθμολογικές διαβαθμίσεις είναι δύσκολο να περιγραφούν στο πλαίσιο του παρόντος διδακτικού εγχειριδίου. Για το λόγο αυτό θα περιοριστούμε σε μια μόνο ενδεικτική κατηγορία και συγκεκριμένα στο Μοντέλο Μερικής Απόδοσης (M.M.A. - Partial Credit Model - P.C.M.) (Maters, 1982, Masters & Wright, 1996). Το μοντέλο αυτό εντάσσεται στην κατηγορία εκείνων που ακολουθούν τη μεθοδολογία του Rasch και θεωρούνται ως τα πιο απλά από τα μοντέλα που εφαρμόζονται σε περιπτώσεις ερωτημάτων, οι απαντήσεις των οποίων αξιολογούνται με κλίμακα δύο ή περισσότερων ιεραρχικών διαβαθμίσεων (π.χ. σε μετρήσεις στάσεων με κλίμακες τύπου Likert, σε βαθμολογίες ερωτήσεων ανάπτυξης, σε ασκήσεις επίλυσης προβλημάτων με σταδιακά βήματα κτλ.). Εξυπακούεται, ότι η διαβάθμιση ανταποκρίνεται σε προοδευτική αύξηση της λανθάνουσας ικανότητας. Το συγκεκριμένο μοντέλο έχει ευρύτατες εφαρμογές τόσο στην ψυχολογία όσο και στην εκπαίδευση, καθώς και σε άλλους επιστημονικούς τομείς.

Ευνόητο είναι ότι η εφαρμογή του στηρίζεται στην παραδοχή ότι η επίτευξη υψηλότερης βαθμολογίας υποδηλώνει αύξησης της ικανότητας την οποία μετρά το κάθε ερώτημα.

Τα λογισμικά που έχουν εκπονηθεί για την εφαρμογή του μοντέλου του Rasch περιλαμβάνουν τόσο τις περιπτώσεις των ερωτημάτων με διχοτομικές απαντήσεις όσο και εκείνες των ερωτημάτων με βαθμολογικά πολυτεμνόμενες απαντήσεις.

Σύμφωνα με τους Wu & Adams, (2007: 41), η παράγωγος ενός M.M.A. προσδιορίζει τη «**συμβατική πιθανότητα**» δύο παρακείμενων βαθμολογικών κατηγοριών και δεν υποδηλώνει ότι ένας εξεταζόμενος θα απαντά επιτυχώς σε όλα τα ερωτήματα.

Για λόγους ευκολίας, ας πάρουμε ως παράδειγμα μια ερώτηση στην οποία η απάντηση μπορεί να βαθμολογηθεί με μια κλίμακα μικρού εύρους π.χ. από 0 έως 2 (0,1,2). Με βάση το μοντέλο του Rasch για τα διχοτομούμενα ερωτήματα μπορούμε να υπολογίσουμε την πιθανότητα κάθε τιμής της παραπάνω κλίμακας ως ακολούθως:

$$p_0 = \frac{1}{1 + e^{(\theta - b1)} + e^{(2\theta - (b1+b2))}}$$

$$p_1 = \frac{e^{(\theta - b1)}}{1 + e^{(\theta - b1)} + e^{2(\theta - (b1+b2))}}$$

$$p_2 = \frac{e^{(2\theta - (b1+b2))}}{1 + e^{(\theta - b1)} + e^{(2\theta - (b1+b2))}}$$

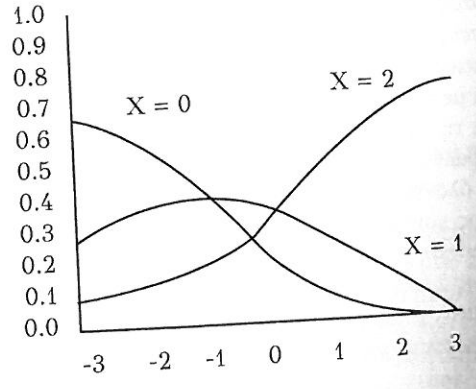
όπου τα σύμβολα έχουν την ίδια σημασία με αυτήν που αναφέρθηκε στις προηγούμενες περιπτώσεις.

Γενικεύοντας την εφαρμογή του συγκεκριμένου μοντέλου στις υπό εξέταση περιπτώσεις, σημειώνουμε ότι, αν η απάντηση σε ένα ερώτημα (i) λαμβάνει μία από τις τιμές 0,1,2...mi, τότε η πιθανότητα ενός ατόμου (n) να λάβει ως βαθμό (x) μία από αυτές δίνεται από τον ακόλουθο τύπο (Wu & Adams (2007: 40).

$$p(x_{ni} = x) = \frac{e^{\sum_{k=0}^{x}(\theta_n - b_{ik})}}{\sum_{h=0}^{m_i} e^{\sum_{k=0}^{h}(\theta_n - b_{ik})}}$$

όπου $\sum_{k=0}^{0}(\theta_n - b_{ik}) = 1$.

Οι χαρακτηριστικές καμπύλες για ένα μοντέλο M.M.A. με πολυτεμνόμενες απαντήσεις παρουσιάζονται στο σχήμα 22.




*Σχήμα 22. Καμπύλες κατανομής της βαθμολογίας σε τριβάθμιο (X = 0 ή 1 ή 2) ερώτημα*[27]

27. Ανάλογη είναι και η λογική της ανάλυσης απαντήσεων που στηρίζονται στην κλίμακα Likert. Για περισσότερες πληροφορίες παραπέμπουμε στις εργασίες Masters (1982) και Wua Adams (2007).





## 10.11. Κριτική της θεωρίας μέτρησης της ικανότητας απάντησης σε ερωτήσεις

Η εν λόγω θεωρία έχει πολλούς υποστηρικτές αλλά και ορισμένους επικριτές (Kline, 1993· Murhpy & Davidshofer, 1994· Chohen et al., 1996· Stocking, 1997· Αλεξόπουλος, 1998). Στα πλεονεκτήματά της κατατάσσονται, μεταξύ άλλων, τα εξής: α) ο προσδιορισμός της δυσκολίας και της διακριτικότητας των ερωτημάτων θεωρείται πληρέστερος σε σύγκριση με αυτόν που στηρίζεται στην κλασική θεωρία, β) τα τεστ που βασίζονται στη Θ.Μ.Ι.Α.Ε. είναι συντομότερα σε σχέση με τα παραδοσιακά τεστ, ενώ οι επιτυγχανόμενες εκτιμήσεις της ικανότητας των εξεταζομένων θεωρούνται ακριβέστερες, γ) οι μετρήσεις δεν εξαρτώνται από τα χαρακτηριστικά του δείγματος και δ) τα χαρακτηριστικά ενός τεστ είναι δυνατόν να προσδιοριστούν πριν από τη χορήγησή του.

Στις επικρίσεις της θεωρίας αυτής, που προέρχονται, ως επί το πλείστον από τους θιασώτες της κλασικής θεωρίας, περιλαμβάνονται: α) η δυσκολία εφαρμογής σε μη διχοτομούμενα ερωτήματα, ιδιαίτερα, β) η πολυπλοκότητα των μαθηματικών υπολογισμών και λοιπών σύνθετων αναλύσεων που δυσκολεύουν την κατανόησή της από τους μη ειδικούς και γ) η δυσκολία απομόνωσης ορισμένης λανθάνουσας ικανότητας των εξεταζομένων από άλλες παρεμφερείς που επηρεάζουν πιθανόν τις απαντήσεις τους στα ερωτήματα. Η τελευταία παρατήρηση αφορά στη δυσκολία επίτευξης μονοδιάστατων δοκιμασιών, όρος πάνω στον οποίο βασίζονται ορισμένα από τα μοντέλα της Θ.Μ.Ι.Α.Ε.

Στα μειονεκτήματα της υπό εξέταση θεωρίας εντάσσουν ορισμένοι την ανάγκη ύπαρξης μεγάλων δειγμάτων, επειδή οι διαφεύσεις των εφαρμοζόμενων μοντέλων είναι, όπως ισχυρίζονται (Ingenkamp, 1993: 163), δύσκολες σε περιπτώσεις μικρών δειγμάτων. Υποστηρίζεται, ακόμη, (όπ. παρ.) ότι η συμμόρφωση των δεδομένων προς την προσδοκία συγκεκριμένου μοντέλου δεν λύνει ικανοποιητικά το πρόβλημα της εγκυρότητας των μετρήσεων.

Αρκετές, τέλος, κριτικές έχουν ασκηθεί ειδικά κατά του μοντέλου του Rasch (για περισσότερες λεπτομέρειες βλ. Goldstein & Blinkhorn, 1977· Goldstein, 1979· Goldstein & Blinkhorn, 1982).