ΓΕΩΡΓΙΟΣ ΤΣΙΤΑΣ
ΠΜΣ «ΠΟΛΙΤΙΚΗ, ΔΙΟΙΚΗΣΗ ΚΑΙ ΑΞΙΟΛΟΓΗΣΗ ΣΤΗΝ ΕΚΠΑΙΔΕΥΣΗ» (ΠΔΑΕ)
ΑΞΙΟΛΟΓΗΣΗ ΣΤΗΝ ΕΚΠΑΙΔΕΥΣΗ

Item Response Theory

(Σύγχρονη Θεωρία Μέτρησης της Ικανότητας Απάντησης σε Ερωτήματα ή Θεωρία των Λανθανόντων Χαρακτηριστικών ή Θεωρία της Ισχυρής Πραγματικής Επίδοσης)

Έχει συμβεί ποτέ;

- Να «αγνοήσετε» το χέρι ενός άριστου μαθητή στην τάξη θεωρώντας δεδομένο ότι ξέρει την απάντηση λόγω της ευκολίας της;
- Να απαντήσει κάποια μαθήτρια σε μία πολύ δύσκολη ερώτηση και να συμπεράνετε ότι γνωρίζει και όλα τα άλλα ερωτήματα που θέσατε κατόπιν;
- Να εντυπωσιαστείτε τόσο πολύ από την απάντηση ενός παιδιού στο γραπτό του, που να παραβλέψετε κάποιο άλλο μικρό λάθος ως ασήμαντο, και να μην το συνυπολογίσετε στο τελικό αποτέλεσμα;
- Να νιώσετε πως υπάρχουν κάποια ερωτήματα που μας οδηγούν, ίσως αυθαίρετα, στο συμπέρασμα για το πόσο πολλά και πόσο καλά γνωρίζει κάποιος;

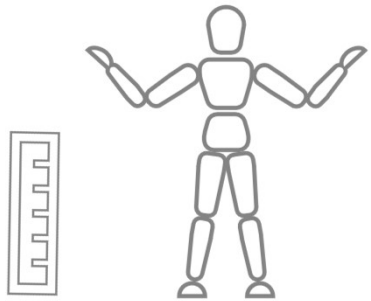
Αν ναι, τότε μπορείτε να καταλάβετε τη λογική «πίσω» από την IRT!

Για ποιον λόγο IRT;

- Ας υποθέσουμε ότι θέλουμε να μετρήσουμε το ύψος κάποιων ανθρώπων, που δεν τους βλέπουμε, μέσω π.χ. ερωτήσεων που μετράνε το πόσο εύκολα φτάνουν κάτι.
- Ιδού η κλίμακα των απαντήσεων:

σκύβοντας, απλώνοντας το χέρι, σηκώνοντας το χέρι, πηδώντας

- Ας υποθέσουμε ότι έχουμε λάβει από 2 διαφορετικούς ανθρώπους παρόμοιες απαντήσεις: 5 (σκύβοντας), 2 (σηκώνοντας το χέρι) 1 (πηδώντας).
- Μπορούμε να βγάλουμε κάποιο πραγματικό συμπέρασμα για τους δύο ανθρώπους ή μας λείπει κάτι; Τι δεν ξέρουμε;
- Τις ερωτήσεις και τον βαθμό δυσκολίας τους.



1. Φτάνεις το πόμολο;
2. Φτάνεις τη χειρολαβή των πάνω ντουλαπιών της κουζίνας;
3. Φτάνεις το πάνω μέρος από το κάσωμα της πόρτας;
4. Φτάνεις το στεφάνι της μπασκέτας;

- Τώρα που ξέρουμε κάποιες ενδεικτικές απαντήσεις, θα αλλάζαμε κάτι, αν μπορούσαμε, στις ερωτήσεις;

Τι είναι η Θεωρία Απόκρισης σε Ερωτήματα;

Σύγχρονη ψυχομετρική προσέγγιση. Η αφετηρία της σε μαθηματικούς (Rasch),κοινωνιολόγους (Lazarfeld) κ.α., από το 1960 και μετά.

Χρησιμοποιείται στην εκπαιδευτική αξιολόγηση και την ψυχολογία.

Στοχεύει στην ανάλυση της απόκρισης ενός ατόμου σε συγκεκριμένα "αντικείμενα" ή "ερωτήσεις" ενός τεστ και όχι στο τεστ συνολικά.

Βρίσκεται στον αντίποδα της κλασικής θεωρίας των τεστ (Classical Test Theory).

Επικεντρώνεται στην αξιολόγηση των επιμέρους ερωτήσεων ενός τεστ. Ως εκ τούτου, η αξιολόγηση της καταλληλότητας των ερωτημάτων είναι κρίσιμη για την IRT.

Λαμβάνει υπόψη τρία βασικά χαρακτηριστικά:

την ικανότητα του εξεταζόμενου,	τη δυσκολία των ερωτήσεων,	την ευαισθησία του τεστ στο να διακρίνει διαφορετικά επίπεδα απόδοσης.

Βασικές Αρχές της IRT

Επίπεδο Ικανότητας του Εξεταζόμενου: Η IRT θεωρεί ότι κάθε εξεταζόμενος έχει μια «κρυφή, λανθάνουσα ικανότητα» ή «ικανότητα χαρακτηριστικό», η οποία επηρεάζει την πιθανότητα να απαντήσει σωστά σε κάθε ερώτηση. Η ικανότητα αυτή μπορεί να είναι κάποια γνώση, δεξιότητα ή χαρακτηριστικό προσωπικότητας.

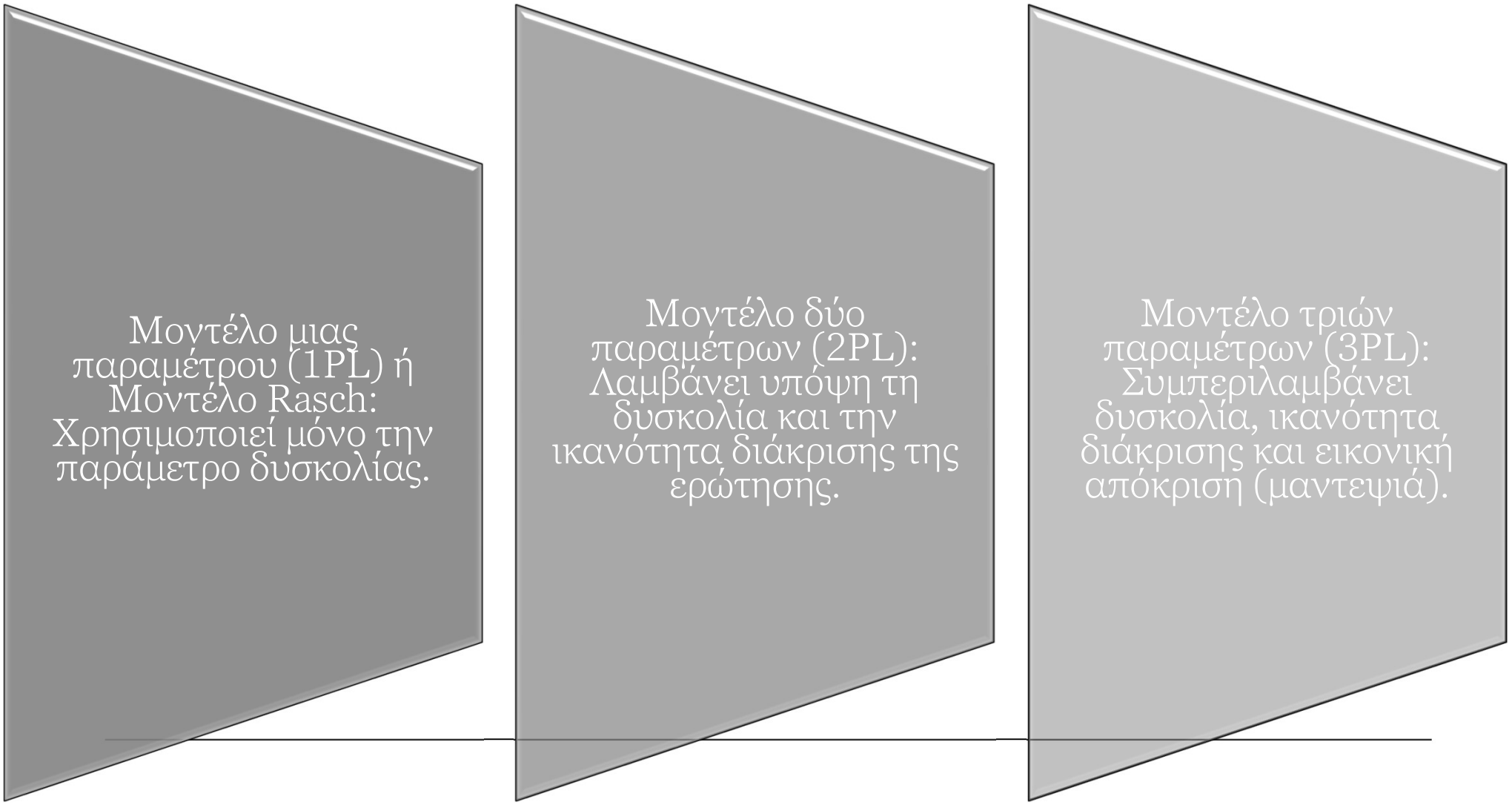
Προϋποθέτει την ανεξαρτησία μεταξύ των διαφόρων ερωτημάτων (local independence).

Χαρακτηριστικά των Ερωτήσεων (Αντικείμενα)

Δυσκολία: Αντιπροσωπεύει το επίπεδο στο οποίο μια ερώτηση διαχωρίζει τους εξεταζόμενους ανάλογα με την ικανότητά τους. Οι ερωτήσεις με υψηλή δυσκολία απευθύνονται σε άτομα με υψηλότερη ικανότητα.	Ικανότητα Διάκρισης: Το χαρακτηριστικό που αναφέρεται στην ικανότητα της ερώτησης να διαφοροποιεί σωστά τα άτομα με υψηλή και χαμηλή ικανότητα.	Εικονική Απόκριση ή Μαντεψιά: Αφορά την πιθανότητα να απαντήσει κάποιος σωστά τυχαία (συχνά σχετίζεται με πολυεπιλογικές ερωτήσεις).

Λογιστική Συνάρτηση (Logistic Function): Η IRT χρησιμοποιεί τη λογιστική συνάρτηση για να περιγράψει την πιθανότητα σωστής απάντησης σε μια ερώτηση, η οποία αυξάνεται όσο μεγαλύτερη είναι η ικανότητα του εξεταζόμενου. Αυτή η συνάρτηση απεικονίζει την πιθανότητα μιας σωστής απάντησης ως συνάρτηση της ικανότητας του εξεταζόμενου.

Μοντέλα IRT


Μοντέλο μιας παραμέτρου (1PL) ή Μοντέλο Rasch: Χρησιμοποιεί μόνο την παράμετρο δυσκολίας.
Μοντέλο δύο παραμέτρων (2PL): Λαμβάνει υπόψη τη δυσκολία και την ικανότητα διάκρισης της ερώτησης.
Μοντέλο τριών παραμέτρων (3PL): Συμπεριλαμβάνει δυσκολία, ικανότητα διάκρισης και εικονική απόκριση (μαντεψιά).

Πλεονεκτήματα της IRT

Ακρίβεια στην Αξιολόγηση

Κάθε ερώτηση βαθμολογείται με βάση την ικανότητα του εξεταζόμενου, προσφέροντας μια πιο ακριβή εικόνα της απόδοσής του.

Σχεδίαση και Ανάπτυξη Τεστ

Οι δημιουργοί τεστ μπορούν να επιλέξουν αντικείμενα που ταιριάζουν καλύτερα στο επίπεδο των εξεταζόμενων.

Προσαρμοστική Δοκιμασία (Adaptive Testing)

Με τη χρήση της IRT είναι δυνατή η δημιουργία προσαρμοστικών τεστ, όπου οι ερωτήσεις προσαρμόζονται δυναμικά στην ικανότητα του εξεταζόμενου, βελτιώνοντας την ακρίβεια και μειώνοντας τον χρόνο εξέτασης.

Εφαρμογές της IRT

- Εκπαιδευτικές αξιολογήσεις και τεστ εισαγωγής.
- Ψυχολογικές και προσωπικές αξιολογήσεις.
- Δοκιμασίες αξιολόγησης προσωπικών δεξιοτήτων, όπως οι αξιολογήσεις των δεξιοτήτων στην εργασία και τα ερωτηματολόγια προσωπικότητας.

Χρήση της IRT στην Ελλάδα

Στην Ελλάδα, η Item Response Theory (IRT) δεν χρησιμοποιείται ευρέως σε επίσημες εκπαιδευτικές αξιολογήσεις, όπως οι πανελλαδικές εξετάσεις. Η αξιολόγηση των μαθητών και η σχεδίαση των τεστ στην ελληνική εκπαίδευση βασίζεται κυρίως στην κλασική θεωρία των τεστ (Classical Test Theory - CTT), όπου οι ερωτήσεις έχουν παρόμοια βαρύτητα, και οι βαθμολογίες αναλύονται με βάση το συνολικό σκορ και την κατανομή βαθμολογιών.

Ερευνητικά Προγράμματα και Μελέτες σε ελληνικά πανεπιστήμια εξετάζουν τη χρήση της IRT για την ανάλυση ψυχομετρικών εργαλείων και την αξιολόγηση των ερωτήσεων.

Τομείς: ψυχολογία, κοινωνιολογία, εκπαίδευση.

Σκοπός: οι ερευνητές χρησιμοποιούν την IRT για την κατασκευή και τον έλεγχο των ψυχομετρικών χαρακτηριστικών δοκιμασιών και ερωτηματολογίων.

Πιλοτικές Εφαρμογές σε Διαγνωστικά Τεστ σε ορισμένα προγράμματα επαγγελματικής κατάρτισης και διαγνωστικών αξιολογήσεων, χρησιμοποιείται η IRT για την ανάπτυξη διαγνωστικών εργαλείων.

Σκοπός: ο εντοπισμός συγκεκριμένων δεξιοτήτων και γνώσεων των εξεταζόμενων.

Σημείωση: αυτές οι περιπτώσεις δεν είναι ακόμα συστηματικές και περιορίζονται κυρίως στον ιδιωτικό τομέα ή σε πιλοτικά προγράμματα.

Προοπτική για Προσαρμοστικά Τεστ (Computerized Adaptive Testing) μελλοντικά και στην Ελλάδα, (ανάπτυξη σχετικής τεχνογνωσίας) μολονότι δεν είναι ακόμη δημοφιλή στην Ελλάδα.

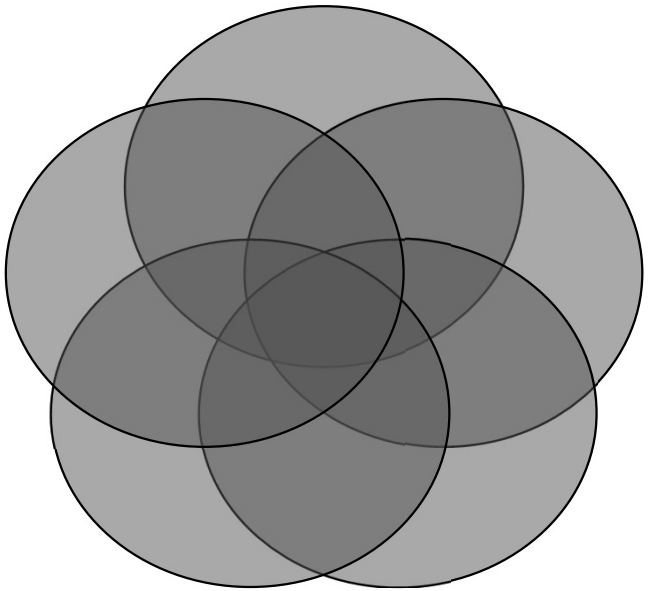
Ψηφιακή ή αναλογική χρήση της IRT;

Η εφαρμογή της Item Response Theory (IRT) δεν απαιτεί απαραίτητα ψηφιακή μορφή εξέτασης, αλλά η ψηφιακή εκδοχή την καθιστά σαφώς πιο αποτελεσματική και ακριβή. Η IRT επικεντρώνεται στην ανάλυση των αντικειμένων (ερωτήσεων) και την εξατομίκευση της αξιολόγησης, η οποία μπορεί να γίνει και με χαρτί, αλλά η ψηφιακή διεξαγωγή επιτρέπει περισσότερα από τα βασικά πλεονεκτήματά της, ειδικά στα προσαρμοστικά τεστ.

Γιατί η IRT είναι πιο αποτελεσματική ψηφιακά:

Ευκολία στην Προσθήκη Παραμέτρων και Δεικτών

Στα ψηφιακά συστήματα είναι εύκολο να ενσωματωθούν παράμετροι, όπως η πιθανότητα τυχαίας απάντησης, η ικανότητα διάκρισης και η δυσκολία, χαρακτηριστικά που είναι βασικά στην IRT.



Προσαρμοστική Εξέταση (Computerized Adaptive Testing - CAT)

Ένα από τα κύρια πλεονεκτήματα της IRT είναι η ικανότητά της να προσαρμόζει τη δυσκολία των ερωτήσεων ανάλογα με τις απαντήσεις του εξεταζόμενου. Αυτό είναι πολύ πιο εύκολο και αποδοτικό ψηφιακά, όπου το σύστημα μπορεί αυτόματα να επιλέξει την κατάλληλη ερώτηση βάσει των προηγούμενων απαντήσεων. Με χαρτί, η διαδικασία αυτή δεν είναι εφικτή.

Ακρίβεια στην Καταγραφή Απαντήσεων και Αποτελεσμάτων

Τα ψηφιακά συστήματα βοηθούν στην καταγραφή ακριβών δεδομένων για κάθε ερώτηση, διασφαλίζοντας αμεσότητα στη συλλογή και ανάλυση της απόκρισης χωρίς ανθρώπινα λάθη που μπορεί να προκύψουν στο χαρτί.

Ταχεία Ανάλυση και Διαχείριση Δεδομένων

Η ψηφιακή εξέταση επιτρέπει την άμεση ανάλυση των δεδομένων και των αποτελεσμάτων για κάθε ερώτηση, κάνοντας τη βαθμολόγηση πιο γρήγορη και ακριβή, κάτι που θα απαιτούσε περισσότερο χρόνο και κόπο σε χειρόγραφη μορφή.

Μη Ψηφιακή (Χειρόγραφη) Εφαρμογή της IRT

Η IRT μπορεί να εφαρμοστεί και σε εξετάσεις με χαρτί και μολύβι, εφόσον δεν απαιτείται προσαρμοστικότητα της εξέτασης. Στις περιπτώσεις αυτές, η IRT χρησιμοποιείται μόνο για την εκ των υστέρων ανάλυση των απαντήσεων, π.χ., για να υπολογιστεί η δυσκολία ή η ικανότητα διάκρισης των ερωτήσεων, παράμετροι που βελτιώνουν το σχεδιασμό μελλοντικών τεστ. Στην περίπτωση αυτή, ωστόσο, χάνεται η δυνατότητα προσαρμοστικότητας και η εξέταση δεν αξιοποιεί πλήρως τις δυνατότητες της IRT.

Είναι αλήθεια ότι με βάση την IRT η ερώτηση που ακολουθεί βασίζεται στο αν απάντησες σωστά ή λάθος την ερώτηση που σου δίνεται;

1 › 2 › 3 › 4 › 5

Ναι, αυτό ισχύει κυρίως στα **Προσαρμοστικά Τεστ (Computerized Adaptive Testing - CAT)** που βασίζονται στην **Item Response Theory (IRT)**. Στα προσαρμοστικά τεστ, η επόμενη ερώτηση επιλέγεται με βάση την απάντηση του εξεταζόμενου στην προηγούμενη:

Αν ο εξεταζόμενος απαντήσει σωστά, τότε το σύστημα θα επιλέξει μια ερώτηση με **υψηλότερο επίπεδο δυσκολίας**, για να δει εάν ο εξεταζόμενος μπορεί να ανταποκριθεί σε πιο απαιτητικές ερωτήσεις.

Αν απαντήσει λανθασμένα, το σύστημα επιλέγει μια **ευκολότερη ερώτηση** για να παρακολουθήσει καλύτερα το επίπεδο ικανότητάς του.

Αυτή η διαδικασία προσαρμογής βελτιώνει την ακρίβεια στην αξιολόγηση της ικανότητας κάθε εξεταζόμενου, μειώνοντας παράλληλα τον συνολικό αριθμό ερωτήσεων που απαιτούνται για αξιόπιστα αποτελέσματα.

Στα παραδοσιακά τεστ, οι εξεταζόμενοι απαντούν όλες τις ερωτήσεις ανεξάρτητα από τις απαντήσεις τους, κάτι που δεν εκμεταλλεύεται τις προσαρμοστικές δυνατότητες της IRT.

Στην παραδοσιακή αξιολόγηση, για να πετύχουμε αντίστοιχο αποτέλεσμα με την IRT θα χρειαζόμασταν πολύ μεγάλο αριθμό ερωτήσεων. Ισχύει;

Ναι, ισχύει! Στην **παραδοσιακή αξιολόγηση** (με την Κλασική Θεωρία Τεστ, ή CTT), για να φτάσουμε την ίδια ακρίβεια που παρέχει η **IRT** θα χρειαζόταν να χρησιμοποιήσουμε **πολύ μεγαλύτερο αριθμό ερωτήσεων**. Αυτό συμβαίνει γιατί η CTT βασίζεται κυρίως στο **συνολικό σκορ** των εξεταζόμενων και όχι στη δυσκολία ή διακριτική ικανότητα κάθε ερώτησης.

Αντίθετα, στην **IRT** κάθε ερώτηση αναλύεται ατομικά, με βάση:

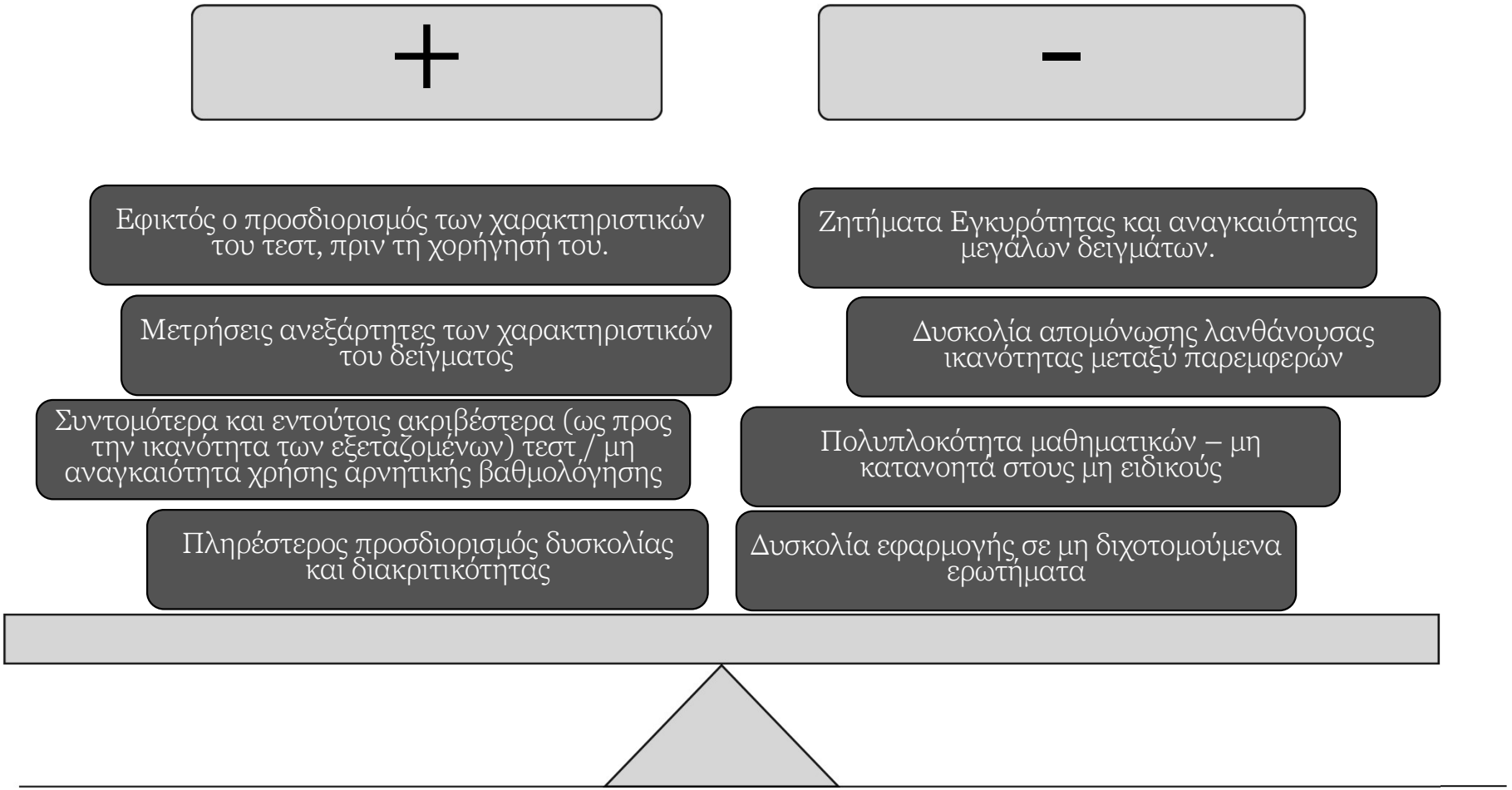
α) τη δυσκολία της, που δεν έχει να κάνει με το πόσοι απάντησαν ορθά αλλά με το σημείο της κλίμακας μέτρησης της δεξιότητας των εξεταζομένων στο οποίο η πιθανότητα ορθής απάντησης είναι π.χ. 0,5 (50%) για το μοντέλο Rasch,

β) την ικανότητά της να διακρίνει διαφορετικά επίπεδα γνώσης ή δεξιοτήτων,

γ) και (σε κάποια μοντέλα) την πιθανότητα τυχαίας απάντησης.

Γιατί η IRT χρειάζεται λιγότερες ερωτήσεις;

Πλεονεκτήματα και μειονεκτήματα της IRT.



Παράδειγμα με βάση την Κλασική Θεωρία Μέτρησης (CTT)

Στην Κλασική Θεωρία Μέτρησης, οι ερωτήσεις έχουν την ίδια βαρύτητα για όλους τους εξεταζόμενους, ανεξαρτήτως του επιπέδου τους.

Ερώτηση: Ποια είναι η πρωτεύουσα της Γαλλίας;

Σωστή απάντηση: Παρίσι

Εκτίμηση: Όλοι οι εξεταζόμενοι που απαντούν σωστά, λαμβάνουν την ίδια βαθμολογία για την ερώτηση, ανεξάρτητα από το επίπεδο γνώσεων τους (π.χ., κάποιος αρχάριος ή ειδικός στη γεωγραφία). Στην CTT, η δυσκολία της ερώτησης δεν λαμβάνεται υπόψη κατά άμεσο τρόπο. Μπορεί μόνο να αποδίδεται διαφορετική βαθμολογία σε μία δύσκολη ερώτηση σε σχέση με μία εύκολη στο συνολικό τεστ. Η αξιοπιστία και εγκυρότητα βασίζονται στο συνολικό τεστ, όχι σε κάθε μεμονωμένη ερώτηση. Ως εκ τούτου, όλοι οι εξεταζόμενοι που συμμετέχουν στη συγκεκριμένη αξιολόγηση θα πρέπει να απαντήσουν σε αυτή την ερώτηση.

Παράδειγμα με βάση τη Θεωρία Ανταπόκρισης σε Θέμα (IRT):

Στην IRT, κάθε ερώτηση έχει συγκεκριμένα χαρακτηριστικά (π.χ., βαθμός δυσκολίας, διακριτική ικανότητα) και η ανάλυση λαμβάνει υπόψη την πιθανότητα σωστής απάντησης ανάλογα με την ικανότητα του εξεταζόμενου.

Ερώτηση: Ποια είναι η πρωτεύουσα της Γαλλίας;

Χαρακτηριστικά ερώτησης (παράμετροι):Δυσκολία: Χαμηλή (εύκολη ερώτηση)

Διακριτική ικανότητα: Μέτρια (ξεχωρίζει καλά τους εξεταζόμενους χαμηλής από υψηλής ικανότητας)

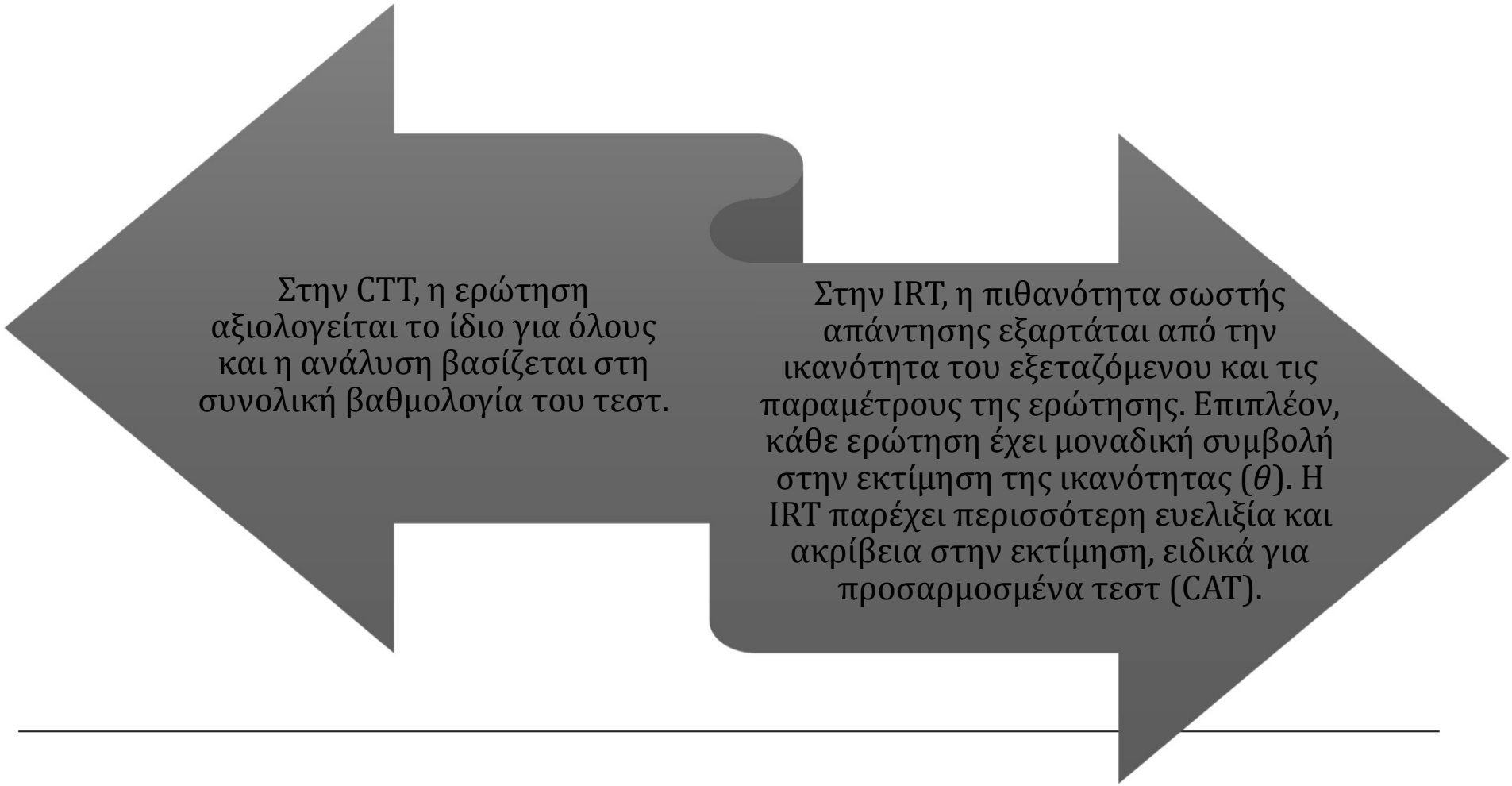
Ανάλυση IRT:

Ένας εξεταζόμενος με χαμηλό επίπεδο γνώσεων (π.χ., $\theta=-2$) έχει χαμηλή πιθανότητα να απαντήσει σωστά.

Ένας εξεταζόμενος με μέσο επίπεδο γνώσεων (π.χ., $\theta=0$) έχει μέτρια πιθανότητα να απαντήσει σωστά.

Ένας εξεταζόμενος με υψηλό επίπεδο γνώσεων (π.χ., $\theta=+2$) έχει σχεδόν σίγουρη πιθανότητα να απαντήσει σωστά.

Διαφορές


Στην CTT, η ερώτηση αξιολογείται το ίδιο για όλους και η ανάλυση βασίζεται στη συνολική βαθμολογία του τεστ.
Στην IRT, η πιθανότητα σωστής απάντησης εξαρτάται από την ικανότητα του εξεταζόμενου και τις παραμέτρους της ερώτησης. Επιπλέον, κάθε ερώτηση έχει μοναδική συμβολή στην εκτίμηση της ικανότητας (θ). Η IRT παρέχει περισσότερη ευελιξία και ακρίβεια στην εκτίμηση, ειδικά για προσαρμοσμένα τεστ (CAT).

Τι τιμές μπορεί να λάβει το θ;

- Στη Θεωρία Ανταπόκρισης σε Θέμα (IRT), η τιμή του θ, που αντιπροσωπεύει την **ικανότητα του εξεταζόμενου**, μπορεί να λάβει οποιαδήποτε πραγματική τιμή στον αριθμητικό άξονα, θεωρητικά από $-\infty$ έως $+\infty$ Ωστόσο, στην πράξη:
- **Πρακτικό Εύρος Τιμών του θ:**

Συνήθως, το θ κυμαίνεται στο διάστημα περίπου από −3 έως +3, όπου:

1. −3: Πολύ χαμηλή ικανότητα.
2. 0: Μέση ικανότητα (είναι η τιμή αναφοράς, καθώς η κατανομή θεωρείται συνήθως κανονική με μέσο όρο 0).
3. +3: Πολύ υψηλή ικανότητα.

- **Λογική Ερμηνεία:**

1. *Οι τιμές κοντά στο −3 αντιπροσωπεύουν εξεταζόμενους με πολύ χαμηλή επίδοση, που πιθανώς απαντούν σωστά μόνο στις πιο εύκολες ερωτήσεις.*
2. *Οι τιμές κοντά στο +3 αντιπροσωπεύουν εξεταζόμενους με πολύ υψηλή επίδοση, που πιθανώς απαντούν σωστά ακόμα και στις πιο δύσκολες ερωτήσεις.*
3. *Το 0 αντιπροσωπεύει έναν εξεταζόμενο με μέση ικανότητα στον πληθυσμό.*

Εξήγηση των τιμών του θ με βάση την κατανομή

Στις περισσότερες εφαρμογές της IRT, η **ικανότητα (θ)** θεωρείται ότι ακολουθεί την **κανονική κατανομή** (N 0,1), όπου: Το μέσο είναι 0.

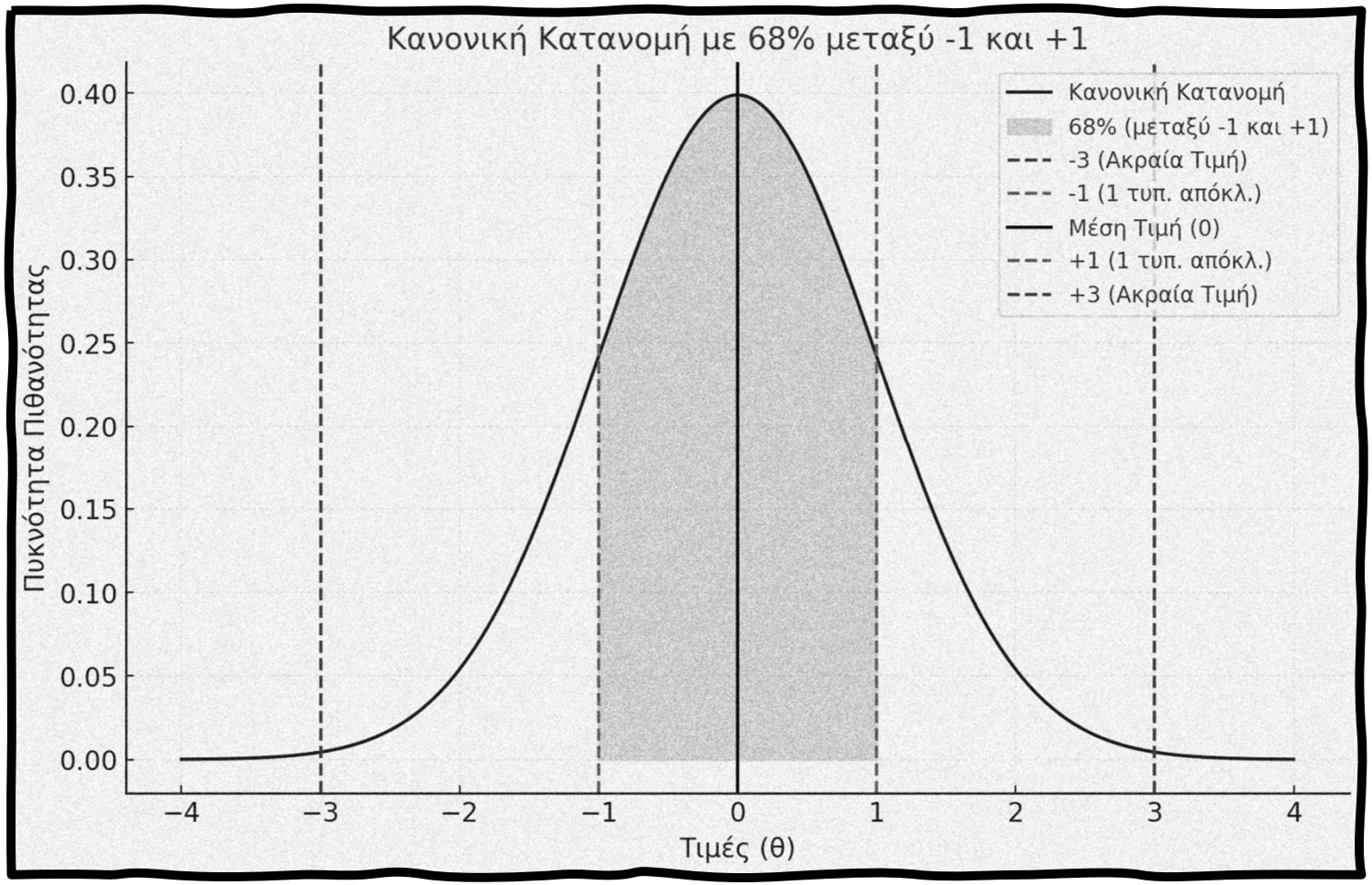
Η τυπική απόκλιση είναι 1, δηλαδή η πλειοψηφία των εξεταζομένων (περίπου 68%) έχει ικανότητα εντός του διαστήματος [−1,+1]. (βλ. εικόνα)

Πολύ λίγοι έχουν τιμές εκτός του διαστήματος [−3,+3].

Άρα

Θεωρητικά: $\theta \in (-\infty, +\infty)$.

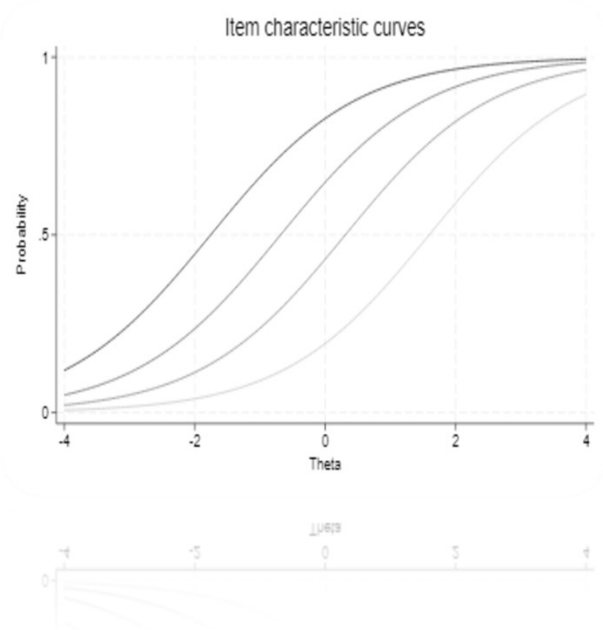
Πρακτικά: Κυμαίνεται κυρίως μεταξύ −3 και +3, με τις περισσότερες τιμές να είναι κοντά στο 0.


Κανονική Κατανομή με 68% μεταξύ -1 και +1
Κανονική Κατανομή
68% (μεταξύ -1 και +1)
-3 (Ακραία Τιμή)
-1 (1 τυπ. απόκλ.)
Μέση Τιμή (0)
+1 (1 τυπ. απόκλ.)
+3 (Ακραία Τιμή)
Πυκνότητα Πιθανότητας
0.40
0.35
0.30
0.25
0.20
0.15
0.10
0.05
0.00
−4
−3
−2
−1
0
1
2
3
4
Τιμές (θ)

Αυτή η εικόνα αναπαριστά τις καμπύλες χαρακτηριστικών ερωτήσεων (Item Characteristic Curves - ICCs) στη Θεωρία Ανταπόκρισης σε Θέμα (IRT). Ας δούμε τι σημαίνουν τα στοιχεία της:

Τι είναι οι καμπύλες χαρακτηριστικών ερωτήσεων;

• Οι καμπύλες αυτές δείχνουν τη πιθανότητα (P) ένας εξεταζόμενος με συγκεκριμένη ικανότητα (θ) να απαντήσει σωστά σε μια ερώτηση.
• Στον οριζόντιο άξονα (x-άξονας) είναι η ικανότητα (θ), η οποία κυμαίνεται από περίπου −4 (πολύ χαμηλή ικανότητα) έως +4 (πολύ υψηλή ικανότητα).
• Στον κάθετο άξονα (y-άξονας) είναι η πιθανότητα σωστής απάντησης που κυμαίνεται από 0 έως 1.

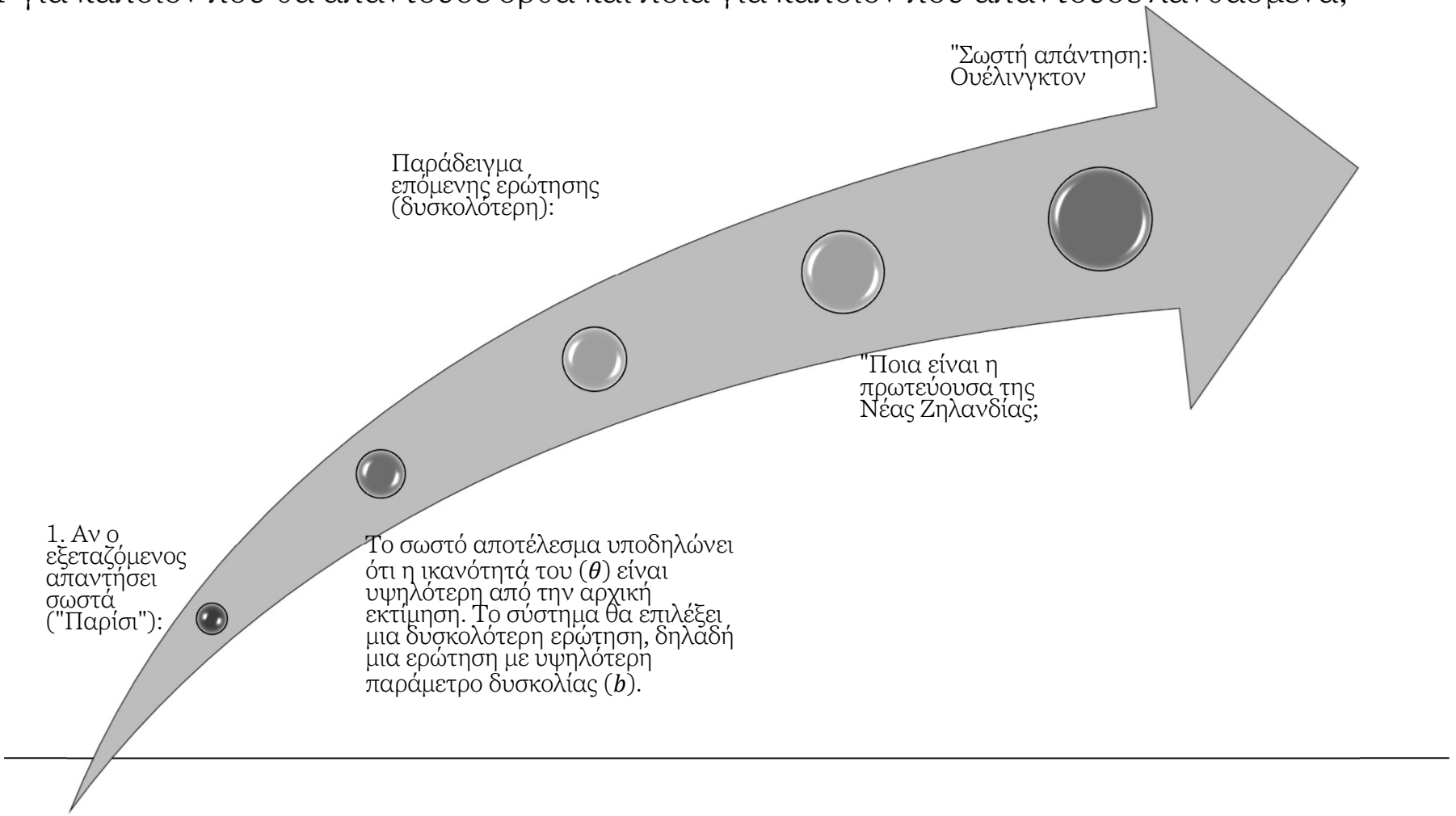


- Για μια ερώτηση με υψηλή δυσκολία (b>0), όπως η κίτρινη καμπύλη, μόνο οι εξεταζόμενοι με υψηλή ικανότητα (θ>1) έχουν μεγάλη πιθανότητα σωστής απάντησης.
- Αντίθετα, μια ερώτηση με χαμηλή δυσκολία (b<0), όπως η κόκκινη καμπύλη, είναι εύκολη και ακόμα και εξεταζόμενοι με χαμηλή ικανότητα (θ<0) μπορούν να την απαντήσουν σωστά .
- Οι πιο απότομες καμπύλες (π.χ., η μπλε) είναι καλύτερες για τη διάκριση ανάμεσα σε εξεταζόμενους κοντά στο επίπεδο δυσκολίας τους.

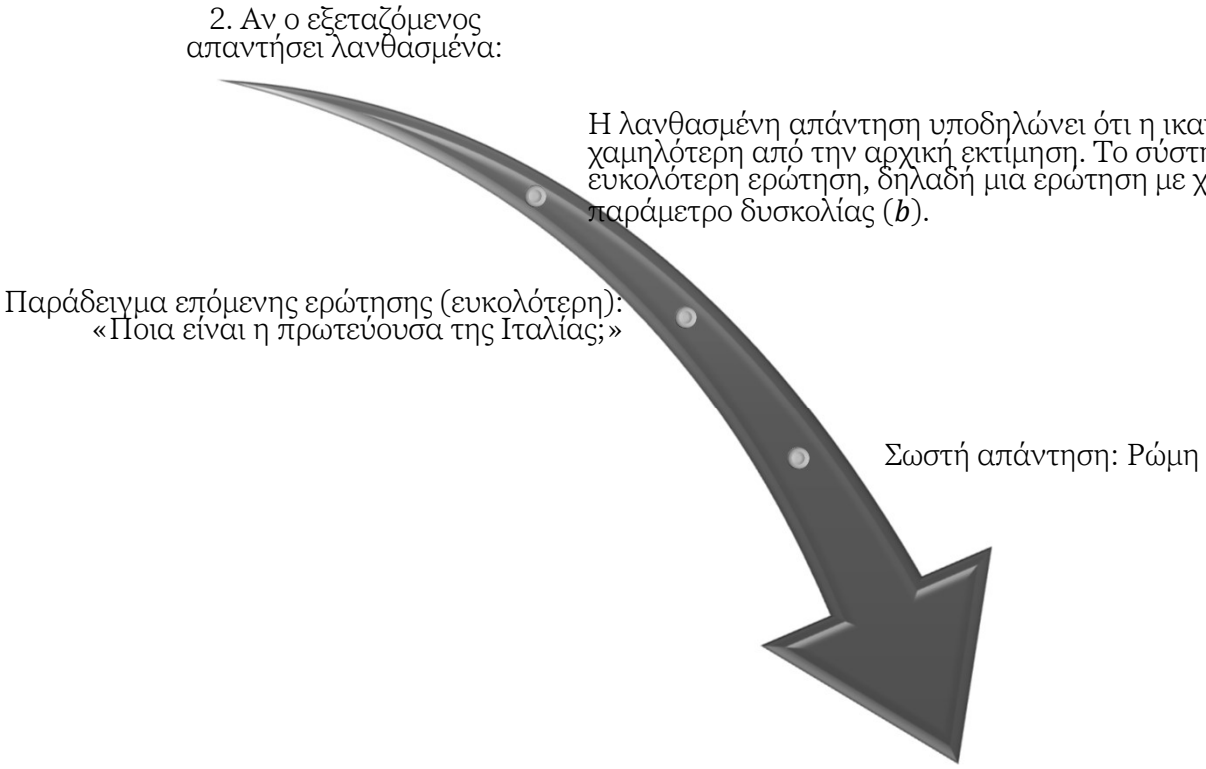
Ποια θα ήταν, λοιπόν, η επόμενη ερώτηση που θα δινόταν μετά την πρωτεύουσα της Γαλλίας σε ένα τεστ με IRT για κάποιον που θα απαντούσε ορθά και ποια για κάποιον που απαντούσε λανθασμένα;

- Σε ένα τεστ που βασίζεται στη **Θεωρία Ανταπόκρισης σε Θέμα (IRT)** και χρησιμοποιεί προσαρμοστική εξέταση (Computerized Adaptive Testing - CAT), η επιλογή της επόμενης ερώτησης εξαρτάται από την απάντηση που δίνει ο εξεταζόμενος στην προηγούμενη ερώτηση. Ο στόχος είναι να εκτιμηθεί η ικανότητά του (θ) με τη μέγιστη δυνατή ακρίβεια, επιλέγοντας ερωτήσεις που είναι κατάλληλες για το επίπεδό του.
- Ας δούμε πώς αυτό επηρεάζεται στις δύο περιπτώσεις:

Ποια θα ήταν, λοιπόν, η επόμενη ερώτηση που θα δινόταν μετά την πρωτεύουσα της Γαλλίας σε ένα τεστ με IRT για κάποιον που θα απαντούσε ορθά και ποια για κάποιον που απαντούσε λανθασμένα;



Ποια θα ήταν, λοιπόν, η επόμενη ερώτηση που θα δινόταν μετά την πρωτεύουσα της Γαλλίας σε ένα τεστ με IRT για κάποιον που θα απαντούσε ορθά και ποια για κάποιον που απαντούσε λανθασμένα;



Γενική Λογική - Πλεονεκτήματα

- Η επιλογή της επόμενης ερώτησης γίνεται με βάση την τρέχουσα εκτίμηση της ικανότητας (θ) του εξεταζόμενου:
  - *Αν απαντήσει σωστά → Το τεστ αυξάνει τη δυσκολία για να προσδιορίσει με ακρίβεια το επίπεδό του.*
  - *Αν απαντήσει λάθος → Το τεστ μειώνει τη δυσκολία για να εντοπίσει ερωτήσεις που αντιστοιχούν στο επίπεδό του.*
- Το τεστ προσαρμόζεται δυναμικά στις δυνατότητες του εξεταζόμενου. Περιορίζεται ο αριθμός των ερωτήσεων, αφού δεν απαιτείται να απαντηθούν όλες οι ερωτήσεις, αλλά μόνο όσες χρειάζονται για να εκτιμηθεί με ακρίβεια το θ.

Σε πόσες ερωτήσεις θα σταματούσε το τεστ για τις πρωτεύουσες εάν κάποιος απαντούσε σωστά και σε πόσες εάν απαντούσε συνεχώς λάθος;

- *Ο αριθμός των ερωτήσεων που απαιτούνται σε ένα* ***προσαρμοστικό τεστ (CAT)*** *βασισμένο στη* ***Θεωρία Ανταπόκρισης σε Θέμα (IRT)*** *εξαρτάται από την* ***ακρίβεια*** *που θέλουμε να επιτύχουμε στην εκτίμηση της ικανότητας (θ) του εξεταζόμενου. Το τεστ συνεχίζεται μέχρι η* ***εκτίμηση της ικανότητας (θ)*** *να σταθεροποιηθεί (δηλαδή, η αβεβαιότητα της εκτίμησης να είναι μικρότερη από ένα προκαθορισμένο όριο).*
- Ας δούμε πώς αυτό επηρεάζεται στις δύο περιπτώσεις:

1. Αν ο εξεταζόμενος απαντά συνεχώς σωστά:

- Κάθε σωστή απάντηση αυξάνει την εκτίμηση της ικανότητας (θ).
- Το τεστ θα συνεχίζει να δίνει όλο και δυσκολότερες ερωτήσεις.
- Καθώς οι ερωτήσεις γίνονται δυσκολότερες, η πιθανότητα να απαντήσει σωστά αρχίζει να μειώνεται (ανάλογα με την πραγματική ικανότητά του).
- Πότε σταματά;
    1. *Το τεστ σταματά όταν το σύστημα δεν μπορεί πλέον να βελτιώσει σημαντικά την ακρίβεια της εκτίμησης (θ).*
    2. *Συνήθως, αυτό συμβαίνει σε **10-15 ερωτήσεις**, αν οι ερωτήσεις έχουν καλή διακριτική ικανότητα (παράμετρο α) και καλύπτουν όλο το εύρος δυσκολιών.*

2. Αν ο εξεταζόμενος απαντά συνεχώς λάθος:

- Κάθε λανθασμένη απάντηση μειώνει την εκτίμηση της ικανότητας (θ).
- Το τεστ θα συνεχίζει να δίνει όλο και ευκολότερες ερωτήσεις.
- Τελικά, το τεστ φτάνει σε ερωτήσεις με πολύ χαμηλή δυσκολία ($b<-2$), δηλαδή πολύ εύκολες). Αν ο εξεταζόμενος συνεχίζει να απαντά λανθασμένα, η εκτίμηση της ικανότητας (θ) σταθεροποιείται σε χαμηλή τιμή ($\theta\approx-3$).
- Πότε σταματά;
  1. Και εδώ, το τεστ σταματά μόλις η ακρίβεια στην εκτίμηση (θ) φτάσει το προκαθορισμένο επίπεδο.
  2. Συνήθως, αυτό συμβαίνει επίσης σε περίπου 10-15 ερωτήσεις, αν και μπορεί να χρειαστούν λιγότερες αν η χαμηλή ικανότητα είναι προφανής από τις πρώτες απαντήσεις.

Συμπερασματικά, ως γενικός κανόνας:

- Ο αριθμός των ερωτήσεων εξαρτάται από:

1. *Το επίπεδο της απαιτούμενης ακρίβειας (π.χ., τυπικό σφάλμα < 0.3).*
2. *Τις παραμέτρους των ερωτήσεων (δυσκολία **b**, διακριτική ικανότητα **a**, πιθανότητα τυχαίας απάντησης **c**).*
3. *Το εύρος των διαθέσιμων ερωτήσεων (π.χ., αν το τεστ περιλαμβάνει ερωτήσεις με ευρύ φάσμα δυσκολίας).*

- Συνήθως, για τεστ με ερωτήσεις σαν τις παραπάνω, που έχει καλά διατυπωμένες και σωστά επιλεγμένες ερωτήσεις καθώς και ξεκάθαρες σωστές απαντήσεις, το τεστ ολοκληρώνεται σε 10-15 ερωτήσεις, ανεξαρτήτως αν ο εξεταζόμενος απαντά συνεχώς σωστά ή λάθος.

Έτσι…

Στην κλασική θεωρία… όλοι οι μαθητές με 15/20 σωστές απαντήσεις, θα έπαιρναν βαθμό 15 ή 75/100, απαντώντας σε ερωτήσεις κοινής δυσκολίας για όλους τους εξεταζόμενους.

Στην IRT, μπορεί και πάλι να απαντήσουν αρκετοί μαθητές σε 15 ερωτήσεις… λόγω όμως του διαφορετικού επιπέδου δυσκολίας, ο ένας μπορεί να είχε σκορ: 72/10 και ο άλλος 79/100.

Ενδεικτική βιβλιογραφία

Baker, F. B., & Kim, S.-H. (2004). Item Response Theory: Parameter Estimation Techniques. New York, NY: CRC Press.

De Ayala, R. J. (2009). The Theory and Practice of Item Response Theory. New York, NY: Guilford Press.

Embretson, S. E. (1996). "The new rules of measurement." Psychological Assessment, 8(4), 341.

Embretson, S. E., & Reise, S. P. (2000). Item Response Theory for Psychologists. Mahwah, NJ: Lawrence Erlbaum Associates.

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). Fundamentals of Item Response Theory. Newbury Park, CA: Sage Publications.

Lord, F. M. (1980). Applications of Item Response Theory to Practical Testing Problems. Hillsdale, NJ: Lawrence Erlbaum Associates.

Reise, S. P., & Revicki, D. A. (2015). "Handbook of Item Response Theory Modeling: Applications to Typical Performance Assessment." Multivariate Applications Series.

Samejima, F. (1969). "Estimation of latent ability using a response pattern of graded scores." Psychometrika, 34(1), 1-97.

Thissen, D., & Wainer, H. (2001). Test Scoring. Mahwah, NJ: Lawrence Erlbaum Associates.

van der Linden, W. J. (2006). "A lognormal model for response times on test items." Journal of Educational and Behavioral Statistics, 31(2), 181-204.

van der Linden, W. J., & Hambleton, R. K. (Eds.). (2013). Handbook of Modern Item Response Theory. New York, NY: Springer.