PISA – ΜΑΘΗΜΑΤΙΚΑ: ΠΟΣΟ ΑΞΙΟΠΙΣΤΑ ΕΙΝΑΙ ΤΑ ΑΠΟΤΕΛΕΣΜΑΤΑ;

Δρ. Παναγιώτης Παναγίδης
Καθηγητής Μαθηματικών
Απεήτειο Γυμνάσιο Αγρού
Λεμεσός-Κύπρος

Abstract

The aim of this study was to investigate the degree of reliability of the results of the PISA 2009 mathematics test. For the purposes of this study the mathematics results of the 2009 program were downloaded and analysed with the use of the Rasch Partial Credit Model. Analyses indicated that (a) reliability indices were very low, (b) the validity of a large proportion of students responses was questionable and (c) there was evidence of item bias in many items, against almost all countries. These results lead to the conclusion that the degree of reliability of the PISA mathematics exams is low.

Λέξεις κλειδιά

PISA, Μαθηματικά, Αξιοπιστία, μοντέλα Rasch.

0. Εισαγωγή

Το πρόγραμμα PISA (Programme for International Student Assessment) είναι μια διεθνής έρευνα που διοργανώνεται κάθε τρία χρόνια, με αφετηρία το 2000, από τον Οργανισμό Οικονομικής Συνεργασίας και Ανάπτυξης (ΟΟΣΑ). Το 2012 ήταν η πέμπτη φορά που διοργανώθηκε και έλαβαν μέρος 65 χώρες. Για πρώτη φορά στην έρευνα αυτή συμμετείχε και η Κύπρος.

Η έρευνα απευθύνεται σε μαθητές 15 χρονών (η ηλικία μέχρι την οποία τελειώνει η υποχρεωτική εκπαίδευση) και αξιολογεί την επίδοσή τους στα Μαθηματικά, στις Φυσικές Επιστήμες και στη Γλώσσα (κατανόηση κειμένου). Εκτός από την επίδοση των μαθητών σε δοκίμια, το πρόγραμμα αξιολογεί με τη χρήση ειδικών ερωτηματολογίων την οργάνωση των σχολικών ομάδων, τη διαδικασία μάθησης και την επίδραση διάφορων παραγόντων στην επίδοση των μαθητών, όπως για παράδειγμα το φύλο. Επίσης αξιολογεί την ποιότητα του εκπαιδευτικού συστήματος, συγκρίνοντας το με το εκπαιδευτικό σύστημα άλλων χωρών.

Τα οφέλη της Κύπρου από τη συμμετοχή στο πρόγραμμα είναι:

- Η μέτρηση και αξιολόγηση των εκπαιδευτικών αποτελεσμάτων με μέτρο σύγκρισης τις επιδόσεις πολλών άλλων χωρών.
- Η εξαγωγή γενικεύσιμων συμπερασμάτων και η επιστημονική υποστήριξη μεταρρυθμίσεων και θέσπισης πολιτικής σε σχέση με μετρήσιμους στόχους.
- Η διαχρονική παρακολούθηση των εκπαιδευτικών αποτελεσμάτων και η μέτρηση ρυθμού εκπαιδευτικής προόδου

Υπεύθυνο για τη διοργάνωση του προγράμματος στην Κύπρο είναι το Εθνικό Κέντρο PISA, το οποίο υπάγεται στο Κέντρο Εκπαιδευτικής Έρευνας και Αξιολόγησης.

(Οι πιο πάνω πληροφορίες πάρθηκαν από τη παρουσίαση με τίτλο «PISA» της κυρίας Αντιγόνης Μουγή, Συντονίστριας του Εθνικού Κέντρου PISA 2012)

0.1. Σκοπός της εργασίας

Ο Kreiner (2012) αμφισβητεί την αξιοπιστία των δοκιμίων κατανόησης κειμένου που χρησιμοποιούνται στο πρόγραμμα PISA και κατ' επέκταση τη χρησιμότητα του προγράμματος.

Έχοντας μελετήσει την παρουσίαση του Καθηγητή Kreiner που αφορούσε τα δοκίμια κατανόησης κειμένου που χρησιμοποιήθηκαν για το πρόγραμμα το 2006, ο ερευνητής ανέλαβε να διεξαγάγει μια παρόμοια εργασία, για τη διερεύνηση του βαθμού αξιοπιστίας των αποτελεσμάτων των δοκιμίων στα Μαθηματικά για το PISA 2009, με τη χρήση των μοντέλων Rasch.

1. Μοντέλα Rasch

Τα μοντέλα Rasch βασίζονται στο ότι ένα άτομο με μεγαλύτερη ικανότητα σε ένα αντικείμενο έχει πάντα μεγαλύτερη πιθανότητα να απαντήσει μια ερώτηση οποιασδήποτε δυσκολίας από ένα άτομο με χαμηλότερη ικανότητα. Ταυτόχρονα, μια ερώτηση μεγαλύτερης δυσκολίας έχει πάντα μικρότερη πιθανότητα να απαντηθεί από άτομο οποιασδήποτε ικανότητας, παρά μια ευκολότερη ερώτηση.

Για να μπορούν να χρησιμοποιηθούν τα μοντέλα Rasch πρέπει τα δεδομένα να πληρούν κάποιες βασικές προϋποθέσεις των μετρήσεων. Πρώτα πρέπει οι ερωτήσεις του δοκιμίου να μετρούν μόνο μια ικανότητα (unidimensionality). Επίσης, πρέπει να υπάρχει τοπική ανεξαρτησία (local independence), δηλαδή, με απλά λόγια, οι απαντήσεις των εξεταζομένων σε μια ερώτηση να μην επηρεάζουν ή να υποδεικνύουν τις απαντήσεις άλλων ερωτήσεων. Τέλος πρέπει όλοι οι εξεταζόμενοι να έχουν αρκετό χρόνο για να δοκιμάσουν να απαντήσουν όλες τις ερωτήσεις, οι απαντήσεις να μην επηρεάζονται από τον παράγοντα τύχη και οι ερωτήσεις να έχουν περίπου την ίδια διάκριση.

Το παραδοσιακό μοντέλο (Rasch, 1960) αξιολογεί τη πιθανότητα ενός ατόμου να απαντήσει σωστά σε κάθε ερώτηση ενός δοκιμίου ως συνάρτηση της ικανότητάς

του (B), η οποία εκτιμάται από τη συνολική του βαθμολογία στο δοκίμιο, και της δυσκολίας (D) της συγκεκριμένης ερώτησης, η οποία εκτιμάται από τον αριθμό των εξεταζόμενων που απάντησαν σωστά στην ερώτηση. Το μοντέλο, όπως αρχικά αναπτύχθηκε για διχοτομικά δεδομένα (τύπου σωστό/λάθος ή πολλαπλής επιλογής), δίνεται από τον τύπο:

$$P_{ni} = \frac{e^{Bn-Di}}{1+e^{Bn-Di}}$$

Η παράμετρος P_{ni} εκφράζει την πιθανότητα το άτομο n να απαντήσει σωστά στην ερώτηση i, δεδομένης της ικανότητας του Bn και της δυσκολίας της ερώτησης Di μετρημένες και οι δύο σε λογαριθμικές μονάδες (logits). Άρα η πιθανότητα ενός ατόμου να απαντήσει σωστά μια ερώτηση είναι συνάρτηση της διαφοράς ικανότητας και δυσκολίας. Στην περίπτωση που η ικανότητα ενός εξεταζομένου ισούται με τη δυσκολία της ερώτησης (B = D), η πιθανότητα να απαντήσει σωστά την ερώτηση ο εξεταζόμενος είναι 0.5.

Η αρχική επανάσταση στη Σύγχρονη Θεωρία Μέτρησης (Item Response Theory) που έφερε ο Rasch (1960) για διχοτομικά δεδομένα έχει εξελιχθεί και επεκταθεί, για να απευθύνεται σε οποιοδήποτε τύπο δεδομένων, όπως για παράδειγμα τις κλίμακες Likert (Andrich, 1978) ή σε δοκίμια στα οποία οι ερωτήσεις βαθμολογούνται με διαφορετικό αριθμό μονάδων η κάθε μια (Masters, 1982). Οι Panayides, Collins και Tymms (2010) απαντούν στις πιο σημαντικές κριτικές εναντίον των μοντέλων αυτών και αναφέρουν μια σειρά εφαρμογών των μοντέλων Rasch δείχνοντας έτσι το μεγάλο εύρος των περιπτώσεων που μπορούν να χρησιμοποιηθούν αυτά στις κοινωνικές επιστήμες.

2. Μεθοδολογία

Για την ανάλυση των δεδομένων χρησιμοποιήθηκε το Partial Credit Model (PCM) της οικογένειας των μοντέλων Rasch (Masters, 1982), το οποίο χρησιμοποιείται όταν οι ερωτήσεις του δοκιμίου βαθμολογούνται με περισσότερες από μια μονάδες και κάποιες από τις μονάδες δίνονται για μερικώς σωστές απαντήσεις στις ερωτήσεις αυτές.

Για τη διερεύνηση του εάν οι ερωτήσεις και οι εξεταζόμενοι ικανοποιούν τις προϋποθέσεις του μοντέλου χρησιμοποιήθηκαν δύο δείκτες, το infit mean square και το outfit mean square, γνωστοί στη βιβλιογραφία ως "fit statistics". Οι δείκτες αυτοί ακολουθούν κατά προσέγγιση την κατανομή X^2 και παίρνουν τιμές από το 0 μέχρι το άπειρο με αναμενόμενη τιμή τη μονάδα. Ο δείκτης outfit είναι ευαίσθητος σε ακραίες τιμές (δηλ. όταν η ικανότητα του εξεταζόμενου βρίσκεται μακριά από τη δυσκολία της ερώτησης). Έτσι, όταν ένας μικρός αριθμός απαντήσεων δε συμφωνεί με την πρόβλεψη του μοντέλου, η τιμή του outfit είναι ψηλή και η σειρά απαντήσεων του εξεταζόμενου θεωρείται απρόβλεπτη. Ο δείκτης infit δείχνει την απόκλιση από το αναμενόμενο από το μοντέλο, για περιπτώσεις στις οποίες η ικανότητα του εξετα-

ζόμενου βρίσκεται κοντά στη δυσκολία των ερωτήσεων. Σύμφωνα με τον Linacre (2006) η τιμή του δείκτη outfit δεν έχει τόσο μεγάλη επίδραση στην εγκυρότητα των μετρήσεων όσο αυτή του infit.

Οι αναλύσεις έγιναν με τη χρήση του λογισμικού WINSTEPS 3.65 (Linacre, 2005).

2.1. Μεροληψία Ερωτήσεων

Ο όρος μεροληψία ερωτήσεων (item bias) αναφέρεται στην πιθανότητα κάποια ερώτηση να είναι στατιστικά πιο δύσκολη για κάποιο πληθυσμό σε σύγκριση με κάποιο άλλο. Η συνήθης πρακτική είναι η διερεύνηση μέσω της διαδικασίας ανίχνευσης μεροληψίας, ή της ανάλυσης διαφορικής λειτουργίας DIF (Differential Item Functioning) με την οποία εκτιμάται αν συγκεκριμένες ερωτήσεις είναι στατιστικά σημαντικά πιο δύσκολες για τους εξεταζόμενους μιας ομάδας σε σύγκριση με τους εξεταζόμενους μιας άλλης ομάδας. Πιο συγκεκριμένα, μια μέθοδος διερεύνησης της πιθανότητας ύπαρξης μεροληψίας είναι με τη χρήση του δείκτη που προτάθηκε από τους Wright and Stone (1977) ο οποίος ακολουθεί την κατανομή t και δίνεται από τον τύπο

$$t_i = \frac{d_{i1} - d_{i2}}{\sqrt{s_{i1}^2 + s_{i2}^2}}$$

όπου d_{i1} και d_{i2} οι δείκτες δυσκολίας της ερώτησης i και S_{i1} και S_{i2} το τυπικό σφάλμα της εκτίμησης, όπως υπολογίζονται ξεχωριστά από τους δύο πληθυσμούς. Άλλη μέθοδος γνωστή στην ψυχομετρία είναι η Mantel-Haenszel η οποία, όπως εξηγούν οι Linacre and Wright (1989), υστερεί από τη μέθοδο μέσω των μοντέλων Rasch λόγω του ότι η δεύτερη δίνει πιο εύκολους και καλύτερα ορισμένους στατιστικούς δείκτες. Το μειονέκτημα αυτών των μεθόδων είναι ότι όπως σε κάθε στατιστικό έλεγχο ο δείκτης επηρεάζεται από δείγματα με πολύ μεγάλα μεγέθη (όπως στην προκειμένη περίπτωση) καθιστώντας τη διερεύνηση στατιστικά σημαντική για ακόμα και πολύ μικρές διαφορές.

Μια πολύ πιο απλή μέθοδος που προτάθηκε από τον Draba (1977) με τη χρήση των μοντέλων Rasch, εισηγείται απλά την εκτίμηση της δυσκολίας των ερωτήσεων ξεχωριστά για τους δύο υπό διερεύνηση πληθυσμούς και αν η διαφορά των δύο εκτιμήσεων είναι μεγαλύτερη του 0,5 logits για κάποια ερώτηση τότε η συγκεκριμένη ερώτηση μεροληπτεί κατά του πληθυσμού του οποίου η εκτιμημένη δυσκολία της ερώτησης είναι μεγαλύτερη. Ενώ αρχικά, η μέθοδος αυτή αντιμετωπίστηκε αρνητικά, στη συνέχεια αποδείχτηκε ότι δίνει περίπου τα ίδια αποτελέσματα με τη μέθοδο Mantel-Haenszel (Scheuneman & Subhiyah, 1998). Δεδομένης της απλότητας της μεθόδου του Draba (1977) και του επηρεασμού των στατιστικών ελέγχων από τα τεράστιου μεγέθους δείγματα, στην εργασία αυτή ο ερευνητής αποφάσισε να τη χρησιμοποιήσει σαν (πολύ πιθανή) ένδειξη ύπαρξης μεροληψίας ερωτήσεων.

3. Αποτελέσματα

Το δοκίμιο των Μαθηματικών αποτελείτο από 24 ερωτήσεις, έξι εκ των οποίων είχαν δύο υπό-ερωτήματα (ερωτήσεις 6, 8, 12, 17, 19 και 21), μία είχε τρία (ερώτηση 24) και μία τέσσερα (ερώτηση 3). Δηλαδή, συνολικά υπήρχαν 35 ερωτήσεις, από τις οποίες οι 32 έπαιρναν μια μονάδα ενώ τρεις δύο μονάδες. Οι ερωτήσεις κατανέμονταν σε 27 βιβλιάρια-δοκίμια τα οποία περιείχαν από 7 μέχρι 23 ερωτήσεις στα Μαθηματικά.

Ο αριθμός των εξεταζόμενων ήταν 515958 από 75 χώρες, από τους οποίους 357642 απάντησαν ερωτήσεις στα Μαθηματικά. Κάθε ερώτηση δόθηκε κατά μέσο όρο σε 158024 εξεταζόμενους (περίπου 30% του δείγματος, ελάχιστος αριθμός 153981, μέγιστος αριθμός 159562). Η εκτίμηση των ικανοτήτων των μαθητών έγινε με τη χρήση του Partial Credit Model (PCM) της οικογένειας των μοντέλων Rasch (Masters, 1982).

Ο Πίνακας 1 δείχνει την κατάταξη των 10 χωρών των οποίων οι μαθητές είχαν τη μεγαλύτερη διακύμανση στην κατάταξη βάσει των απαντήσεων σε τρία από τα βιβλιάρια-δοκίμια. Αυτά τα τρία βιβλιάρια-δοκίμια χορηγήθηκαν σε δείγματα μαθητών από 45 χώρες, τα ίδια δοκίμια και στις 45 από τις 75 χώρες. Φαίνεται ότι το Λουξεμβούργο είχε τη μεγαλύτερη διακύμανση αφού έχει καταταχθεί στην 23η θέση στο βιβλιάριο-δοκίμιο B1 και την 36η θέση στο B3, μια διακύμανση 13 θέσεων σε σύνολο 45 χωρών. Η Πολωνία ακολουθεί με διακύμανση 11 θέσεων (30η στο B1 και 19η στο B3), η Πορτογαλία με διακύμανση 9 θέσεων μέχρι τις ΗΠΑ με διακύμανση 6 θέσεων.

Πίνακας 1: Κατάταξη 10 χωρών βάσει αποτελεσμάτων σε τρία βιβλιάρια-δοκίμια

Country	B1(23)	B3(12)	B5(24)	Range	Max	min
LUX	23	36	31	13	36	23
POL	30	19	20	11	30	19
PRT	33	26	35	9	35	26
IRL	24	32	25	8	32	24
NZL	15	8	11	7	15	8
FRA	20	22	26	6	26	20
JPN	7	13	13	6	13	7
LVA	28	34	33	6	34	28
NOR	27	25	21	6	27	21
USA	36	31	30	6	36	30

Το πρόβλημα γίνεται πιο έντονο αν αντί για ολόκληρο δοκίμιο διερευνήσουμε τις θέσεις κατάταξης ξεχωριστά στην κάθε ερώτηση. Ο Πίνακας 2 δείχνει την γενική κατάταξη τεσσάρων χωρών (στήλη 2), την κατάταξη στις τρεις πιο εύκολες ερωτήσεις (στήλες 3, 4 και 5) και στις τρεις πιο δύσκολες ερωτήσεις (στήλες 6, 7 και 8) και τη διακύμανση (εύρος) στην κατάταξη των χωρών αυτών (τελευταία στήλη).

Πίνακας 2: Κατάταξη τεσσάρων χωρών σε έξι συγκεκριμένες ερωτήσεις

	Total	E22	E10	E2	E14	E6b	E12b	Range of
Country	Rank	Rank	Rank	Rank	Rank	Rank	Rank	ranks
AZE	46	60	58	30	4	26	24	56
LTU	36	6	34	34	56	35	35	50
TUR	43	21	19	53	5	53	20	48
MEX	52	50	27	48	74	58	63	47

Το Αζερμπαϊτζάν, έχει καταταχτεί τέταρτο στην ερώτηση E14 και εξηκοστό στην ερώτηση E22, με μια διακύμανση 56 θέσεων σε σύνολο 75 χωρών. Επίσης, η Λιθουανία σε μια ερώτηση κατατάσσεται 6η και σε άλλη 56η, διακύμανση 50 θέσεων. Άρα, οι δύο πρώτοι πίνακες δείχνουν τη μεγάλη διακύμανση στην κατάταξη κάποιων χωρών σε συγκεκριμένα δοκίμια και σε συγκεκριμένες ερωτήσεις.

3.1. Αξιοπιστία

Η επεξεργασία των δεδομένων με τα μοντέλα Rasch έδειξε ότι ο δείκτης αξιοπιστίας Person Reliability ήταν μόλις 0.73 και ο Person Separation μόλις 1.64. Ο πρώτος δείκτης είναι αντίστοιχος του δείκτη άλφα (Cronbach, 1952) και δείχνει πόσο καλά το δοκίμιο μπορεί να διακρίνει-ξεχωρίζει τους εξεταζόμενους. Ο δεύτερος δείκτης δείχνει την έκταση (το εύρος) των εκτιμήσεων των ικανοτήτων των εξεταζόμενων ως προς το τυπικό σφάλμα. Η τιμή 0,73 είναι πολύ χαμηλή για δοκίμια μαθηματικών και θα ήταν οριακά ικανοποιητική για κλίμακες μέτρησης ψυχολογικών χαρακτηριστικών (Nunnally, 1978). Επίσης, ο Linacre (2006) εισηγείται τιμές για τον πρώτο δείκτη μεγαλύτερες του 0,80 και για το δεύτερο μεγαλύτερες του 2,0 ως ένδειξη ικανοποιητικής αξιοπιστίας.

3.2. Διερεύνηση των fit statistics, infit και outfit, των ερωτήσεων

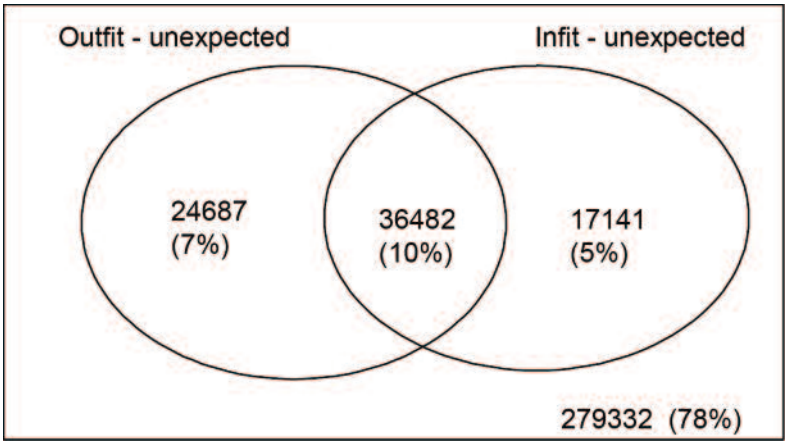
Ο δείκτης outfit έδειξε ότι πέντε από τις 33 ερωτήσεις δεν απαντήθηκαν με τον αναμενόμενο τρόπο από τους εξεταζόμενους (σύμφωνα με την πιθανότητα να απαντήσουν σωστά ή λάθος την ερώτηση). Οι ερωτήσεις ήταν οι E22, E10, E14, E8b

και E19a οι οποίες είχαν outfit > 1,3, τιμή που θεωρείται το όριο, σύμφωνα με τους Wright, Linacre, Gustafson και Martin-Lof (1994) και τους Bond και Fox (2001, 2007) για τέτοιου είδους δοκίμια. Οι τιμές του outfit ήταν 1,97, 1,64, 1,44, 1,35 και 1,33 αντίστοιχα. Οι τελευταίες τρεις μπορεί να θεωρηθούν οριακές.

3.3. Διερεύνηση των fit statistics, infit και outfit, των απαντήσεων των εξεταζόμενων

Το διάγραμμα 1 δείχνει τον αριθμό και ποσοστό των εξεταζόμενων που διαγνώστηκαν ότι δεν απάντησαν με τον αναμενόμενο τρόπο (σύμφωνα με την πιθανότητα να απαντήσουν σωστά ή λάθος τις ερωτήσεις) με τη βοήθεια των δύο fit statistics.

Διάγραμμα 1: Κατανομή τιμών δεικτών infit και outfit με όριο το 1,3



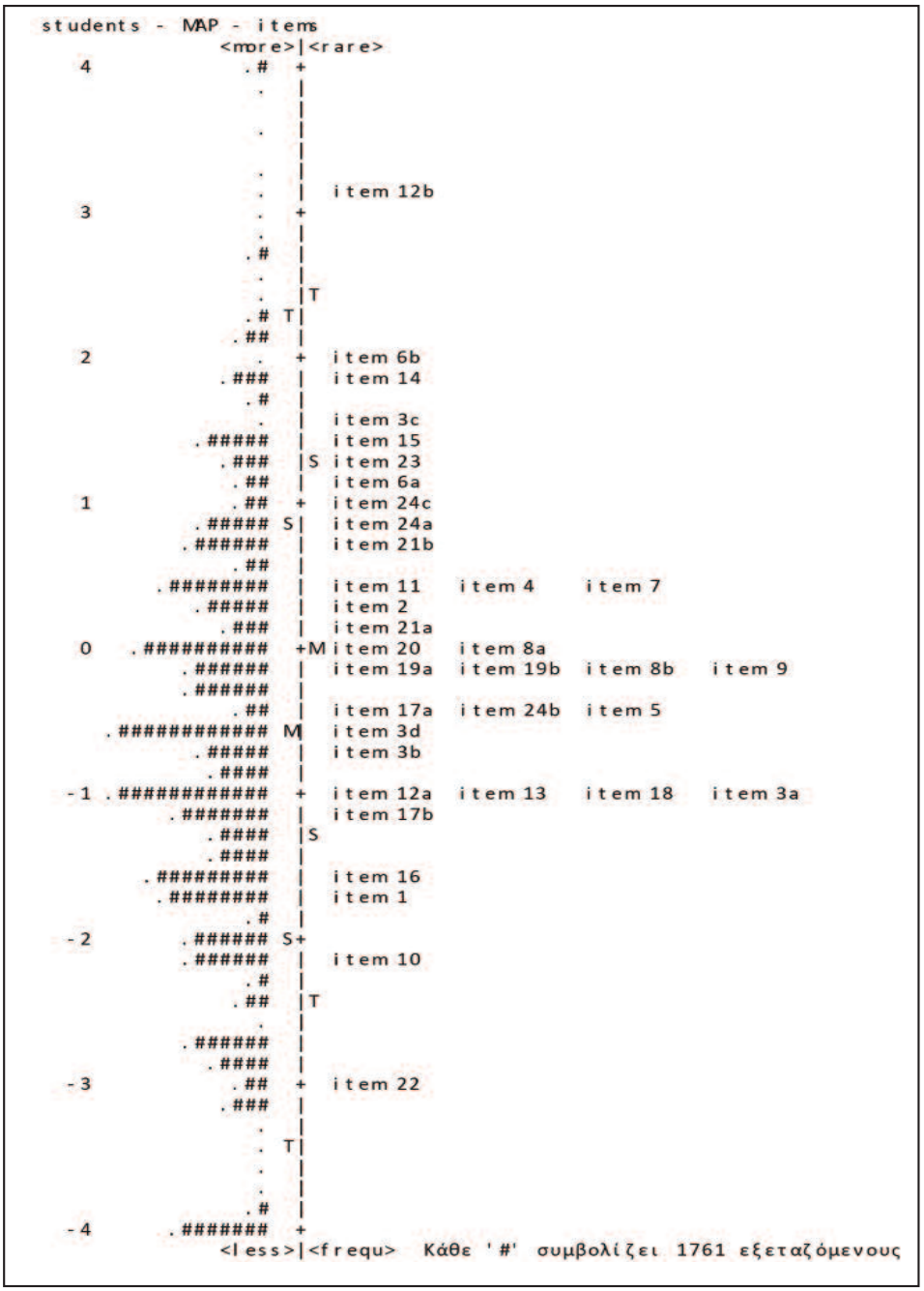
Ένα ποσοστό περίπου 22% διαγνώστηκε ότι απάντησε με απρόβλεπτο τρόπο, δηλαδή ένας στους πέντε εξεταζόμενους. Ειδικά το 15% διαγνώστηκε από το δείκτη infit ο οποίος σύμφωνα με το Linacre (2006) δείχνει κίνδυνο για την εγκυρότητα των μετρήσεων των ικανοτήτων των εξεταζομένων αυτών.

3.4. Πόσο καλά στοχευμένες είναι οι ερωτήσεις;

Το διάγραμμα 2 δείχνει την τοποθέτηση των εξεταζομένων (βάσει της ικανότητας τους), στα αριστερά και των ερωτήσεων (βάσει της δυσκολίας τους), στα δεξιά στον θεωρητικό άξονα Μαθηματικής ικανότητας. Αυτό είναι ένα από τα πλεονεκτήματα και μοναδικότητα των μοντέλων Rasch σε σχέση με άλλα μοντέλα της Σύγχρονης Θεωρίας Μέτρησης: η απεικόνιση της κατανομής των ικανοτήτων των εξεταζομένων και της κατανομής των δυσκολιών των ερωτήσεων στο ίδιο άξονα. Οι ερωτήσεις φαίνονται να είναι καλά στοχευμένες για τον πληθυσμό των εξεταζομένων καλύπτοντας

όλα τα επίπεδα ικανοτήτων και οι δύο κατανομές (ικανοτήτων και δυσκολιών) είναι περίπου συμμετρικές γύρω από τον άξονα μέτρησης.

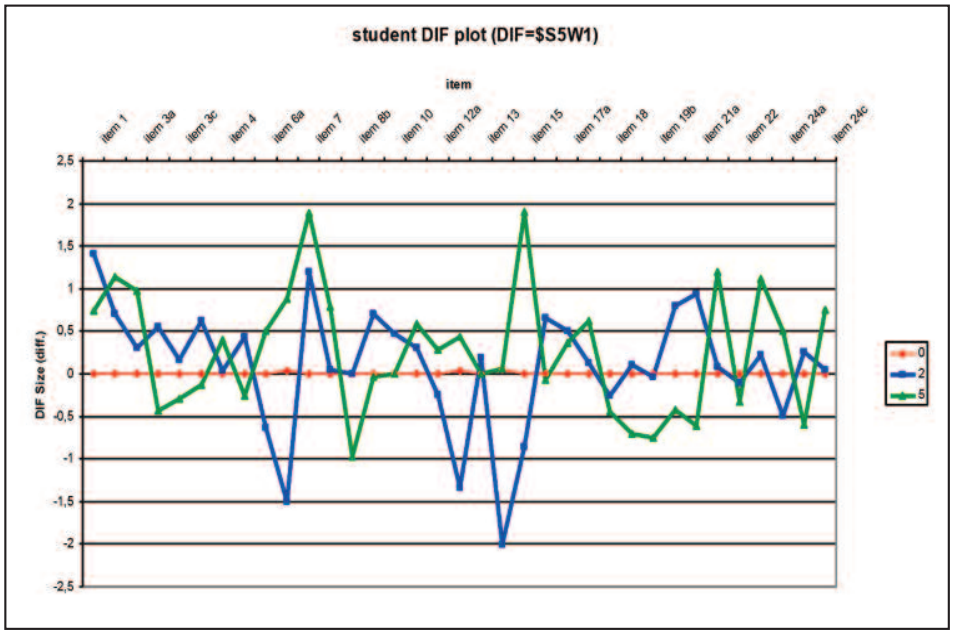
Διάγραμμα 2: Διάγραμμα εξεταζομένων - ερωτήσεων



3.5. Διερεύνηση μεροληψίας ερωτήσεων

Το διάγραμμα 3 δείχνει τη διαφορά στην εκτίμηση της δυσκολίας των ερωτήσεων από δύο δείγματα δύο διαφορετικών εξεταζόμενων χωρών, από τη γενική εκτίμηση από όλο τον πληθυσμό. Οι δύο χώρες είναι η Κίνα που κατετάγη πρώτη και η Tamil Nadu-India (QTN) που κατετάγη 73η. Ο οριζόντιος άξονας (στο 0) αντιπροσωπεύει τη δυσκολία των ερωτήσεων χωρίς μεροληψία. Χρησιμοποιώντας το κριτήριο που εισηγήθηκε ο Draba (1977), δηλαδή ότι διαφορά πάνω από 0,5 logit είναι ενδεικτική μεροληψίας υπέρ ή κατά του υπό διερεύνηση δείγματος, φαίνεται ότι αρκετές από τις ερωτήσεις επιδεικνύουν μεροληψία. Συγκεκριμένα, εννέα ερωτήσεις (27,3% του συνόλου των ερωτήσεων) φαίνεται ότι είναι σημαντικά πιο δύσκολες για τη Κίνα και πέντε (15,2%) σημαντικά πιο εύκολες. Όσο αφορά στη QTN 12 ερωτήσεις (36,4%) ήταν σημαντικά πιο δύσκολες, ενώ 5 (15,2%) σημαντικά πιο εύκολες. Οι διαφορές είναι αρκετά μεγάλες.

Διάγραμμα 3: Διάγραμμα διερεύνησης μεροληψίας ερωτήσεων



Λόγω χώρου παρουσιάζεται μόνο ένα διάγραμμα. Τα αποτελέσματα είναι όμως παρόμοια και για πολλές χώρες, δείχνοντας ότι στα δοκίμια Μαθηματικών του PISA υπάρχει μεροληψία υπέρ, αλλά πιο σημαντικό, κατά του πληθυσμού σχεδόν όλων των χωρών. Η μεροληψία αυτή μπορεί να οφείλεται σε ερωτήσεις προσβλητικές ή εκτός κουλτούρας ομάδας εξεταζομένων ή το πιο πιθανό, ερωτήσεων εκτός των

αναλυτικών προγραμμάτων χωρών άρα ερωτήσεις σε θέματα που δεν έχουν διδαχθεί ή με τα οποία έχουν εξοικειωθεί οι μαθητές της χώρας.

Στην περίπτωση της Ελλάδας, μόνο δύο ερωτήσεις παρουσιάζουν μεροληψία κατά των Ελλήνων εξεταζομένων, και το ίδιο ισχύει και για τη Σιγκαπούρη που κατετάγη δεύτερη.

4. Συμπεράσματα

Ο κύριος στόχος της εργασίας αυτής ήταν η αξιολόγηση της αξιοπιστίας των αποτελεσμάτων του προγράμματος PISA για τα δοκίμια στα Μαθηματικά. Για την διερεύνηση αυτή χρησιμοποιήθηκαν τα αποτελέσματα των δοκιμίων του προγράμματος από το 2009, τα οποία δημοσιεύονται στο διαδίκτυο στην ιστοσελίδα του ΟΟΣΑ.

Μια θεμελιώδης αρχή της θεωρίας των μετρήσεων είναι αυτή της σταθερότητας. Αυτή αναφέρει ότι: 'Η δυσκολία κάθε ερώτησης είναι σταθερή και ανεξάρτητη από τα άτομα που αποτελούν το δείγμα. Ταυτόχρονα, η ικανότητα των ατόμων είναι σταθερή και ανεξάρτητη από τις ερωτήσεις που απαρτίζουν το δοκίμιο που θα χρησιμοποιηθεί'. Μια πρώτη ανάλυση στην κατάταξη των εξεταζόμενων των διάφορων χωρών σε έξι από τις ερωτήσεις έδειξε ότι υπάρχει πολύ μεγάλη διακύμανση στην κατάταξη των χωρών σε κάθε ερώτηση ξεχωριστά. Επίσης, διερευνήθηκε η κατάταξη των χωρών σε τρία βιβλιάρια-δοκίμια δείχνοντας επίσης μεγάλες διακυμάνσεις. Προφανώς η αρχή της σταθερότητας δεν ισχύει για τις ερωτήσεις που χρησιμοποιήθηκαν και κατ' επέκταση και για το δοκίμιο των Μαθηματικών που χρησιμοποιήθηκε για τους σκοπούς του προγράμματος.

Με τη βοήθεια των μοντέλων Rasch έγιναν περαιτέρω εξειδικευμένες αναλύσεις που έδειξαν τα πιο κάτω:

- Οι ερωτήσεις που χρησιμοποιήθηκαν στο δοκίμιο των Μαθηματικών ήταν καλά στοχευμένες στην κατανομή των ικανοτήτων των εξεταζόμενων.
- Παρόλ' αυτά, μεγάλο ποσοστό των εξεταζομένων (22%) αναγνωρίστηκε από τους δείκτες infit και outfit mean square statistics ως άτομα που απάντησαν στο δοκίμιο με απρόβλεπτο τρόπο, αμφισβητώντας έτσι την εγκυρότητα των μετρήσεων των ικανοτήτων τους. Ένα ποσοστό των εξεταζομένων αναμένεται (από τα μοντέλα Rasch) ότι θα απαντά με απρόβλεπτο τρόπο, αλλά κατά τη γνώμη του ερευνητή το ποσοστό που παρατηρήθηκε σ' αυτή την περίπτωση είναι μεγάλο.
- Οι δύο δείκτες μέτρησης της αξιοπιστίας των αποτελεσμάτων που χρησιμοποιήθηκαν (Person Reliability και Person Separation) είχαν πολύ χαμηλές τιμές, δεικνύοντας έτσι πολύ χαμηλό βαθμό αξιοπιστίας στα αποτελέσματα των εξεταζομένων στο δοκίμιο.

- Τέλος, η διερεύνηση μεροληψίας των ερωτήσεων έδειξε ότι αρκετές ερωτήσεις μεροληπτούσαν κατά σχεδόν όλων των χωρών. Για παράδειγμα, περίπου μια στις τρεις ερωτήσεις επιδείκνυε μεροληψία κατά των εξεταζομένων από την Tamil Nadu-India που κατατάγηκε 73η σε σύνολο 75 χωρών. Ταυτόχρονα, περίπου μια στις τέσσερις ερωτήσεις επιδείκνυε μεροληψία κατά των εξεταζομένων από την Κίνα, που κατατάγηκε πρώτη στη γενική κατάταξη.

4.1. Καταληκτικά σχόλια

Τα αποτελέσματα της εργασίας αυτής δείχνουν ότι η αξιοπιστία του δοκιμίου που χρησιμοποιήθηκε για το πρόγραμμα PISA ήταν πολύ χαμηλή, με την εγκυρότητα των αποτελεσμάτων μεγάλου ποσοστού των εξεταζομένων να είναι αμφισβητούμενη. Επίσης πολλές από τις ερωτήσεις του δοκιμίου επιδεικνύουν μεροληψία εις βάρος αρκετών χωρών, οδηγώντας στο συμπέρασμα ότι η επίδοση της κάθε χώρας στο δοκίμιο εξαρτάται από το ποιες ερωτήσεις θα κληθεί να απαντήσει το μεγαλύτερο μέρος του πληθυσμού των εξεταζομένων της. Πιθανοί λόγοι για τη μεροληψία των ερωτήσεων είναι, κατά τη γνώμη του ερευνητή, η διαφορετικότητα ανάμεσα στα αναλυτικά προγράμματα και στην έμφαση που αυτά δίνουν στα διάφορα κεφάλαια όπως επίσης και στη διαφορά κουλτούρας ανάμεσα στις διάφορες χώρες ως προς την αντιμετώπιση τέτοιων διαγωνισμών.

Φυσικά το ερώτημα που τίθεται είναι πόση βαρύτητα πρέπει να δίνουμε και πως να αξιοποιούμε τα αποτελέσματα αυτά. Λόγω της χαμηλής αξιοπιστίας των δοκιμίων, τουλάχιστο στα μαθηματικά, δεν πρέπει να δίνεται μεγάλη σημασία στην κατάταξη των χωρών σε διαδοχικές χορηγήσεις των δοκιμίων αυτών. Δηλαδή αν η θέση της Κύπρου ή της Ελλάδας βελτιωθεί κατά 10 θέσεις από το 2012 στο 2015, αυτό δε σημαίνει κατ' ανάγκη βελτίωση. Μπορεί απλά να σημαίνει ότι χορηγήθηκαν πιο ευνοϊκές ερωτήσεις στους εξεταζομένους των χωρών αυτών. Ο ερευνητής πιστεύει ότι πρέπει να μας προβληματίσει η άσχημη κατάταξη μας (Κύπρου και Ελλάδας) στο διαγωνισμό του 2012, να δούμε αναλυτικά την επίδοση των μαθητών μας ανά ερώτηση και να αποφασίσουμε αν θέλουμε να βελτιώσουμε τους τομείς στους οποίους υστερούμε, αν φυσικά θεωρούμε ότι αυτοί οι τομείς είναι σημαντικοί στη μαθηματική παιδεία.

Ο ερευνητής εισηγείται όπως διεκπεραιωθούν παρόμοιες εργασίες με τα αποτελέσματα του διαγωνισμού για το 2012 έτσι ώστε να διαπιστωθεί ο βαθμός εγκυρότητας και αξιοπιστίας των δοκιμίων σ' όλες της θεματικές ενότητες. Αυτό είναι απαραίτητο έτσι ώστε να μην παρθούν βεβιασμένες αποφάσεις για το μέλλον του εκπαιδευτικού συστήματος στην Κύπρο που να βασίζονται σε λανθασμένο εργαλείο μέτρησης.

Τέλος, χρησιμοποιώντας τα λόγια του καθηγητή Linacre: "International educational League Tables make for exciting politics but not much for education". (Linacre, personal communication, April 24, 2012).

Βιβλιογραφικές πηγές

Andrich, D. (1978) A rating formulation for ordered response categories. *Psychometrika, 43,* 561-573.

Bond, T. G. & Fox, C. M. (2001) *Applying the Rasch model: Fundamental measurement in the social sciences* (2nd edition): Erlbaum Associates.

Bond, T. G. & Fox, C. M. (2007). *Applying the Rasch model: Fundamental measurement in the social sciences*: Erlbaum Associates.

Draba (1977) The identification and interpretation of item bias. Memorandum No. 25. Chicago, Press. Available: http://www.rasch.org/memo25.htm

Kreiner, S. (2012) Is the foundation under solid? A critical look at the scaling model underlying international comparisons of student attainment. Study presented at the *6th Annual UK Rasch Users Group Meeting,* 20-21 March.

Linacre, J. M. (2006) WINSTEPS (3.61.2) [Computer Software]: Winsteps.com.

Linacre, J. M. (2005) *WINSTEPS Rasch measurement computer program*: Winsteps.com.

Masters, G. N. (1982) A Rasch model for partial credit scoring. *Psychometrika*, 47(2), 149-174.

Nunnally, J. C. (1978) *Psychometric theory* (2nd ed.): McGraw-Hill.

Panayides, P., Robinson, C. & Tymms, P. (2010) The assessment revolution that has passed by: Rasch measurement. *British Educational Research Journal, 36*(4), 611-626.

Rasch, G. (1960) *Probabilistic models for some intelligence and attainment tests.* (Reprinted in 1980 with a forward and afterward by Benjamin D. Wright) (Chicago, Press).

Scheuneman, J. D. & Subhiyah, R. G. (1998) Evidence for the validity of a Rasch technique for identifying Differential Item Functioning. *Journal of Outcome Measurement. 2*, 33-42.

Wright, B. D., Linacre, J. M., Gustafson, J-E. & Martin-Lof, P. (1994) Reasonable mean square fit values. *Rasch measurement transactions*, *8*(3), 370. Retrieved July 2011 from http://www.rasch.org/rmt/rmt83b.htm.

Wright, B. D. & Stone, M. H. (1979) Best Test Design: Press.