**Vilniaus universitetas**
**Matematikos ir informatikos**
**institutas**
**LIETUVA**

INFORMATIKA (09 P)

# INFORMATIKOS EDUKACINIŲ TESTŲ GENERAVIMAS IR VALIDUMO TYRIMAS

**Lina Vinikienė**

2015 m. spalis

Mokslinė ataskaita MII-DS-09P-15-12

VU Matematikos ir informatikos institutas, Akademijos g. 4, Vilnius LT-08663

www.mii.lt

# Santrauka

Ataskaitoje pateikiama trumpa klasikinės ir moderniosios testų teorijos apžvalga. Tolimesniems tyrimams modernioji testų teorija bus naudojama kaip metodas įvertinti testo patikimumą ir sudaryti modelį, kuris aprašytų kaip teisingai turi būti parengtas testas ir duomenys, leidžiantys mokytojui analizuoti mokinio pasiekimus, įvertinti testo atitikimą ugdymo programai. Tyrimu siekiama išspręsti testų validumo problemą.

**Reikšminiai žodžiai: modernioji testų teorija, įvertinimas, patikimumas**

# Turinys

# 1 Įvadas

Apibrėžiant bendruosius ugdymo tikslus pabrėžiami reikalavimai turinio formavimui ir ugdymo procesui organizuoti. Kiekvieno švietimo sistemos dalyvio siekis, kad mokinys šiame ugdymo procese įgytų bendrąsias ir dalykines kompetencijas. Rengiamos įvairios konferencijos, seminarai, rašomi straipsniai apie tai, kokia turi tūi formuojama mokymosi aplinka. Populiaru tirti mokymo ir mokymosi metodus, ieškoti pagrįstų modelių, kurie užtikrintų efektyvų mokymą/mokymąsi, įvertintų besimokančiojo įgūdžius, kompetencijas, žinias. Kiekvienas modelis yra sudėtingas ir reikalaujantis „pamatuojamos“ sąveikos tarp pažinimo įvedimo (angl. Cognitive inputs), pažinimo savybių (angl. Cognitive attributes) ir pasiekiamų tikslų (Lamb, Richard L.et al. 2014). Tačiau svarbiausia šių dienų problema, kaip įvertinti mokinius. Dažnai mokytojai mokinių kompetencijų įvertinimui pasirenka testavimo sistemas, patys kuria testų klausimus naudojantis sistemoje pateikiamais šablonais. Tačiau kyla klausimas, ar šie testai yra patikimi (validūs), kokias kompetencijas galima įvertinti ir kaip testas susijęs su ugdymo tikslais, rezultatais, ar pasirinkta vertinimo forma atitinka bendruosius ugdymo reikalavimus, ar korektiškai įvertinami besimokančiųjų pasiekimai.

Literatūroje plačiai aprašomas kompiuterinis testavimas kaip metodas patikrinti bendruosius gebėjimus, tačiau mokslinių tyrimų, ar testai yra validūs (patikimi) nėra daug, todėl testų generavimas ir jų patikimumas yra aktuali tema moksline ir praktine prasme ypač informatikos ir edukologijos moksluose.

# 2 Testų teorija

Švietimo sistemoje naudojami testai turi apibūdinti mokinio žinių lygį, pasiekimus ir ugdymo programos atitikimą mokymo tikslams. Klausimų generavimas yra esminis procesas adaptyviuose testuose, tačiau statistinė analizė padeda mokytojams parinkti testo klausimus ir įvertinti, koks mokinių žinių lygis. Statistinių duomenų analizei naudojamos įvairios testų teorijos, kuriose pagrindinės testo charakteristikos yra šios:

- Klausimo sudėtingumas. Jo analizė padeda pasirinkti klausimus pagal sudėtingumą ir eilę teste. Klausimo sudėtingumas taip pat lemia testo validumą ir patikimumą. Laikoma, kad tam tikrą klausimą atsakiusių mokinių dažnis yra atsitiktinis dydis. Dažnai į testą įtraukiami tie klausimai, kurių sudėtingumo įvertis yra 0,16 arba 0,84.
- Klausimo skiriamoji geba (diskriminacija) – klausimo savybė išskirti mokinius pagal visų klausimų rezultatus.
- Testo ir klausimo koreliacija.
- Patikimumo koeficientas (reliability coefficient).
- Standartinė matavimų paklaida (Bodoff, D., Li, P., 2007).

## 2.1 *Klasikinė testų teorija*

Paprasčiausias testų teorijos pavyzdys yra klasikinė testų teorija. Klasikinėje testų teorijoje nėra apibrėžiama, kaip atskiri asmenys ar grupės egaminuojamųjų atsakys į specifinius klausimus (Hambleton, Swaminathan, Rogers, 1991). Egzaminuojamojo testo statistika priklauso nuo klausimų, įtrauktų į testą, imties ir klausimų charakteristikų tokių, kaip klausimų sudėtingumas, diskriminacija (angl. Item discrimination) (Aesaert, K. at al., 2014).

Klasikinė testų teorija aprašoma lygtimi:

$$X=T+e \qquad (1)$$

T – tikrasis mokinio rezultatas;

e – atsitikinė paklaida, kuri priklauso nuo subjekto (emocinės būklės, streso ir pan.), testo, aplinkos sąlygų.

Patikimumo įvertinimo metodai.

- *alternatyvios ar analogiškos formos metodas. Tai reiškia, jog du testai galėtų būti pakeičiami vienas kitu ir visi kandidatai iš abiejų testų gautų identiškus rezultatus.*
- *perskeltos dalies metodas. Testas padalijamas į dvi tapačias vienodo ilgio ir vienodo sudėtingumo dalis. Apskaičiuojami kiekvieno laikiusiojo testą galutiniai taškai – kiekvienos dalies atskirai.*
- *vidinio nuoseklumo metodas. Kiekviena atskira testo užduotis gali būti traktuojama kaip atskiras testas.*

Testo patikimumą galima įvertinti apskaičiuojant Cronbacho alfa koeficientą:

$$\acute{\alpha} = \left(\frac{k}{k-1}\right)\left(1 - \frac{\sum_{i=1}^{N} S_i^2}{S_X^2}\right) \qquad (2)$$

k-testo užduočių skaičius, $S_x^2$ - galutinių testo rezultatų dispersija, $S_i^2$ - klausimo i dispersija.

Kai žinomas šis koeficientas, galima paskaičiuoti standartinę paklaidą:

$$S_E = S_X\sqrt{1 - \acute{\alpha}} \qquad (3)$$

Sx –teso rezultatų, gautų mokinių populiacijos, standartinis nuokrypis.

Formulė parodo, kad kuo aukštesnis testo patikimumas, tuo mažesnė standartinė matavimų paklaida.

Testo patikimumą lemia užduočių patikimumas, testo rezultatų pasiskirstymas pagal egaminuojamąjį, užduočių skiriamoji geba, užduočių sudėtingumas.

Pagrindinė klasikinės testų teorijos problema yra ta, kad daroma prielaida, jog visų egzaminuojamųjų matavimo paklaida yra vienoda. Šiuo atveju turi būti atsižvelgta į skirtingų gebėjimų mokinius. Klasikinėje testų teorijoje atmetami netinkami klausimai ir iš naujo įvertinami visi testo parametrai.

Privalumai klasikinės testų teorijos yra paremti silpnai pagrįstomis prielaidomis ir žiniomis (Peeraer, J., Van Petegem, P., 2012). Gera testų teorija ar modelis padeda parengti testo struktūrą.

Naudojantis klasikine testų teorija galima sudaryti testo matricą, kurioje siejamos informacinių komunikacinių technologijų kompetencijos ir tai, ką mokiniai turi mokėti. Baker & Kim (2004) tyrime naudojamą matricą įvertino ekpertai, mokytojai, specialistai. Tyrimo metu 560 mokiniams (6 klasės) pateiktos užduotys, kuriomis jie turėjo pademonstruoti informacinių ir komunikacinių technologijų kompetencijas naudojant kompiuterines aplikacijas ir programinę įrangą (naudojamos modeliavimu paremtos užduotys). Užduotyse pateikiami aprašymai, ką mokinys turi atikti. Šiame tyrime buvo panaudota ir moderni testų teorija. Siekiama įvertinti latentines tiriamųjų savybes (angl. Latent trait) (Baker & Kim, 2004), nors šia savybe negalima tiesiogiai įvertinti informacinių ir komunikacinių technologijų kompetencijų.

Šiuo tyrimu buvo siekiama parodyti, kuris modelis teisingas norint įvertinti kompetencijas. Kaip geriausias modelis pasirinktas 2 parametrų modelis (2PLM), tačiau reikalingi tolimesni tyrimai, kurie parodytų kaip vertinimas susijęs su mokymo programa, jos tikslais. Klasikinėje testų teorijoje pastebima problema, kad sudėtinga įvertinti atsakymus, kai testo bandymų parametrai priklauso nuo pagrįstumo. Klasikinė testų toerija turi būti vertinama iš naujo, kai antrojo testo bandymo atsakymai skiriasi nuo pirmojo.

## 2.2 Modernioji testų teorija

Dažnai kyla klausimas, kodėl mokinys atsakė klausimą neteisingai. Gali būti įvardijamos tokios priežastys, kaip mokymosi medžiagos nesupratimas, klausimo nesuvokimas, testo klausimo sudėtingumas arba klausimas paremtas kitos temos žiniomis (Sudol, L. A., Studer, C., 2010). Modernioji testų teorija (angl. Item Response Theory, IRT) modeliuoja klausimų sudėtingumą, testo statistiką, kuri nepriklauso nuo egzaminuojamojo ir jo balų (Hambletonet al., 1991, Lamb, R. et al. 2014) ir gali apibrėžti minėtas priežastis.

IRT modeliai matuoja skalės tikslumą per latentinį kintamąjį θ. Kuo kintamasis didenis, tuo egaminuojamasis turi didesnę tikimybę atsakyti klausimą teisingai.

Moderniosios testų teorijos modelis naudoja aibę klausimų, kurie vertinami gebėjimų skalėje, ir taip leidžia palyginti asmens latentines savybes ir klausimų charakteristikas. (Hambleton et al., 1991). Remiantis šia teorija apskaičiuojama tikimybė atlikti konkretų testą, todėl egzaminuojamasis ir testas yra nepriklausomi. Santykis tarp mokinio ir klausimo vaizduojamas charakteringąja užduoties kreive (angl. Item Characteristic Curve) (Conejo, R., et al., 2014). Klausimų charakteringoji kreivė parodo tikimybę, kad mokinys su gebėjimais θ atsakys klausimą teisingai (Conejo, R., et al., 2014).

Tiek klasikinėje, tiek modernioje testų teorijoje apibrėžiama sudėtingumo reikšmė, kuri įvertinama statistiniais arba matematiniais metodais (Conejo, R., et al., 2014).

Naudojami vieno, dviejų, trijų parametrų logistiniai modeliai (1 PLM, 2 PLM, 3 PLM) Jų pasirinkimas priklauso nuo klausimo charakteristikų. Kiekviename modelyje apibrėžiamas sudėtingumo parametras, kuris atspinti klausimų išdėstymą gebėjimų skalėje, kur teisingo atsakymo tikimybė lygi 0.5. Vadinasi, kuo didesnė sudėtingumo parametro reikšmė, tuo daugiau gebėjimų reikia atsakyti klausimą teisingai 50% (Hambletonet al., 1991).

**Rasch modelis**

Rasch matavimų modelis ir metodologija yra moderni testų teorija, kuri apima tikslią ir išsamią duomenų analizę ir teikia psichometrinę informaciją, kuri negali būti parodoma remiantis klasikine testų teorija (Peeraer, J., Van Petegem, P., 2012).

Rasch modelio savybės:

- *Asmens ir pasirinkimo įverčiai yra teisingi intervale nuo − ∞ iki + ∞.*
- *Dviejų lyginamų subjektų sudėtingumo lygiai nepriklauso nuo klausimų aibės sudarymo.*

Rasch teorija paremta lygtimi, kurią sukūrė George Rasch (Rasch, 1960). Jo teigimu yra matematinis ryšys, kuris parodo, kaip asmuo atsako į klausimą. Anot Rasch, asmuo žinantis daugiau turi didesnę tikimybę išspręsti tą patį klausimą, nei žinodamas mažiau ir atsakyti klausimą teisingai antrą kartą tikimybė yra didesnė ( Wei, S., Liu, X., Jia, Y., 2013). Toks matematinis santykis aprašomas lygtimi:

$$P(X=1|\theta_n, b_i) = \frac{e^{(\theta_n - b_i)}}{1+e^{(\theta_n - b_i)}} \quad (4)$$

Lygtis aprašo tikimybę $P_{ni}$ asmens n su gebėjimais $\theta_n$ teisingai atsakyti klausimą i su sudėtingumu $b_i$, kuris vertinamas teisinga (X=1) arba klaidinga (X=0). (Wheadon, Ch., 2013). Kad mokinys atsakytų klausimą teisingai su tikimybe 0,5, tai jis turi mokėti tiek, kad $\theta_n = b_i$.

Lygtis gali būti perrašyta taip:

$$ln\left(\frac{P_{ni}}{1-P_{ni}}\right) = \theta_n - b_i$$

Matoma, kad tikimybė atsakyti klausimą teisingai yra nustatoma skirtumo tarp asmens gebėjimų ir klausimo sudėtingumo. Kuo didesnis skirtumas, tuo labiau tikėtina, kad asmuo klausimą atsakys teisingai.

Rasch modelyje vyrauja vienas parametras – klausimo skiriamoji geba.

Modelis turi tik vieną laisvąjį parametrą, kuris yra sudėtingumo parametras. Jei pastebėta didelė variacija tarp diskriminacijos indekso skirtingų klausimų, tai 1PML nerekomenduojamas. Mažų gebėjimų mokiniai turi didesnį šansą nei didesnių gebėjimų mokiniai atsakant klausimus teisingai (Osterlind, 2002).

**Dviejų ir trijų parametrų logistiniai modeliai.**

Trijų parametrų logistinis modelis aprašomas lygtimi:

$$P(u_i = 1|\theta) = (1 - c_i)\frac{1}{1+e^{-1,7a_i(\theta-b_i)}} \qquad (6)$$

$u_i = 1$ - tikimybė atsakyti klausimą i teisingai;

$\Theta$ – mokinio gebėjimas. Jis įgija tikrąsias reikšmes intervale [1,1], bet praktiniuose skaičiavimuose pasirenkami intervalai [4.0, 4.0] arba [3.0, 3.0].

$a_i$ – diskriminavimo faktorius;

$b_i$ – testo klausimo sudėtingumas;

$c_i$ – spėjimo faktorius (tikimybė teisingai atsakyti į testo u i-ajį klausimą, net jei egzaminuojamasis nežino atsakymo į klausimą).

Į lygtį įtraukta konstanta 1,7, kad modelis artėtų prie normaliosios parabolės modelio.

Kai $c_i$=0, tai gaunamas 2 parametrų logistinis modelis:

$$p_i(\theta) = \frac{1}{1 + e^{-a_i(\theta-b_i)}}$$

$p_i(\theta)$ - tikimybė atsakyti klausimą teisingai;

$a_i$ – diskriminavimo faktorius;

$b_i$ – sudėtingumas.

Klausimų arba testo informacinė funkcija (angl. Item and Test Information Functions):

$$I(\theta) = a_i^2 p_i(\theta)(1 - p_i(\theta))$$

$p_i(\theta)$ - tikimybė atsakyti klausimą teisingai, kai žinių lygis $\theta$;

$a_i$ – diskriminavimo faktorius.

Dviejų parametrų logistinis modelis naudojmas, kad paskaičiuoti populiacijos tikimybę remiantis tikraisiais testo taškais (angl. True Score method) (Dimitrov, 2007, Peeraer, J., Van Petegem, P., 2012) Naudojant dviejų parametrų modelį galima pastebėti problemas susijusias su klausimo turinio formulavimu, konstruktu, ir leidžia pastebėti, kaip klausimas priklauso nuo „žinau ar nežinau“ spėjimo ir fakto (Sudol, L. A., Studer, C, 2010). Taip pat šis modelis parodo, kaip mokiniai gali atsakyti tipiškus klausimus.

Taigi, 2 PLM ir 3PLM turi antrą parametrą - klausimo diskriminacijos parametrą. Šis parametras leidžia 2 PLM ir 3PLM skirtingai diskriminuojamiems klausimams su aukštesne diskriminacijos parametro reikšme atskirti skirtingų gebėjimų mokinių lygį (Hambleton et al.,1991).

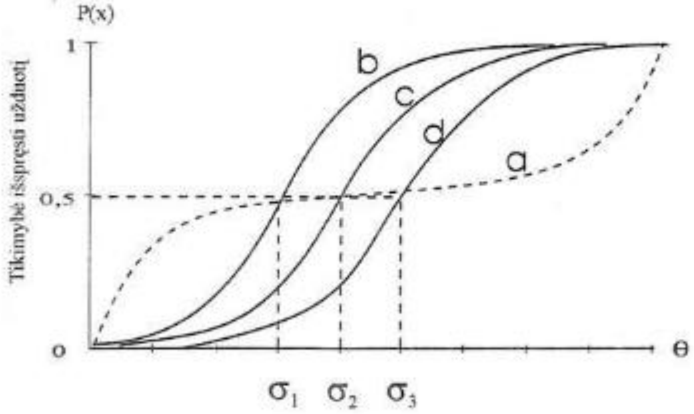
3PLM turi ir trečią parametrą žinomą, kaip spėjimo parametrą arba faktorių. Trijų parametrų modelio pavyzdys yra testavimo sistema SIETLE (Barla, M. et al., 2010).

Moderniosios testų teorijos modelis taikomas su kitais statistiniais modeliais įgalina dinaminių ir adaptyviųjų klausimų pasirinkimą. Pavyzdžiui, Youngseok, L., Jungwon, Ch. (2015) straipsnyje remiantis šia teorija minimas klausimų banko sudarymas testavimo sistemoje, kurioje klausimai parenkami atsižvelgiant į temą, potėmę ir siūlomas mokinio gebėjimų nustatymo modelis, kuris parodo mokėjimo lygį.

**Klausimų ir testo charakteringosios kreivės** (angl. Item and Test charakteristics curves).

Klausimo sudėtingumo kreivė parodo tikimybę atsakyti klausimus teisingai. Kreivė pereinanti į kairę (dešinę) pusę vaizduoja lengvesnius (sudėtingesnius) klausimus. Atitinkamai, diskriminavimo įvertis vaizduojamas iki kreivės nuolydžio.

Kuo arčiau ir daugiau stebimi balai pasiskirsto apie klausimų charakteringąją kreivę (ICC), tuo geriau modelis atitinka duomenis. ICC forma yra S:



1 pav. Rasch modelio ICC. Skirtingo sudėtingmo uždaviniai: b – lengvas, c – vidutinis, d- sunkus, a – klausimas, netenkinantis modelio sąlygos.

Klausimų atsakymų funkcija (angl. Item Response Function) vaizduoja tikimybę, kaip teisingai mokinys atsako klausimą.

Testo informacinė funkcija yra suma testo klausimų funkcijų. Ši funkcija parodo priklausomybę nuo diskriminacijos parametro, t. y. kuo didesnė diskriminacija, tuo didesnis informacijos kiekis teikiamas apie egzaminuojamuosius.

Pagrindiniai IRT principai:

- *egzaminuojamojo veikla atsakant klausimą nusakoma latentinėmis savybėmis.*
- *ryšys tarp egzaminuojamojo ir klausimo gali būti išreiškiamas charakteringaja kreive arba charakteringaja funkcija. (Huang, Y., Lin, Y., Cheng, S., 2009)*

Taigi moderniosios testų teorijos modelis naudojamas analizuoti testo patikimumą gali būti naudojamas mokinių vertinimo sistemos pasirinkimui. Tačiau remiantis modeliu negalime analizuoti klausimų pagal visus testo rezultatus, bet galime – pagal klausimų charakteristinę kreivę su unikaliomis nekintamomis klausimų charakteristikomis. Vieno klausimo charakteristikos neparodo informacijos apie visą egzaminuojamųjų grupę (Youngseok, L., Jungwon, Ch., 2015).

# 3 Literatūra

1. Adams, R. J., Wilson, M. & Wang, W. (1997). The multidimensional random coefficients multinomial logit model. *Applied Psychological Measurement,* 21(1), 1–23.
2. Asert,K. et al. (2014). Direct measures of digital information processing and communication skills in primary education: Using item response theory for the development and validation of an ICT competence scale in *Computers & education*, Volume: 76 Pages: 168-181
3. Baker, F. B., & Kim, S. H. (2004). Item response theory. Parameter estimation techniques (2nd ed.). New York: Marcel Dekker.
4. Barla, M. et al. (2010). On the impact of adaptive test question selection for learning efficiency in *Computers & Education*, Volume: 55, Issue: 2, Pages: 847-857, doi:10.1016/j.compedu.2010.03.016
5. Bodoff, D., Li, P. (2007). Test Theory for Assessing IR Test Collections
6. Conejo, R.et al.(2014). An empirical study on the quantitative notion of task difficulty in *Expert Systems with Applications*
7. Dimitrov, D. M. (2007). Least squares distance method of cognitive validation and analysis for binary items using their item response theory parameters in *Applied Psychological Measurement*, 31(5), 367e387.
8. Huang, Y. et al. (2009). An adaptive testing system for supporting versatile educational assessment in *Computers & education* Volume: 52, Issue: 1, Pages: 53-67
9. Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). Fundamentals of item response theory. New Bury, California: Sage Publications.
10. Kuo, Ch.; Wu, H. (2013). Toward an integrated model for designing assessment systems: An analysis of the current status of computer-based assessments in science in *Computers & education*, Volume: 68, Pages: 388- 403
11. Lamb R. et al. (2014). Cognitive diagnostic like approaches using neural-network analysis of serious educational videogames in *Computers & education*, Volume: 70, Pages: 92-104
12. Lamb, R. et al. (2014). A computational modeling of student cognitive processes in science education in *Computers & education*, Volume: 79, Pages: 116-125
13. Rasch, G. (1960). Probabilistic models for some intelligence and attainment tests. Copenhagen: Danmarks Paedagogiske Institut.
14. Peeraer, J., Van Petegem, P. (2012). Measuring integration of information and communication technology in education: An item response modeling approach in *Computers & education*, Volume: 58, Issue: 4, Pages: 1247-1259
15. Sudol, L. A., Studer, C. (2010) Analyzing Test Items: Using Item Response Theory to Validate Assessments in *SIGCSE '10 Proceedings of the 41st ACM technical symposium on Computer science education*, 436-440
16. Wei, S., Liu, X., Jia, Y (2014) Using rasch measurement to validate the instrument of students' understanding of models in science (sums) in *International journal of science and mathematics education*, Volume: 12, Issue: 5, Pages:1067-1082.

17. Wheadon, Ch. (2013). Using modern test theory to maintain standards in public qualifications in England in *Research papers in education*, Volume: 28, Issue: 5 Pages: 628-647
18. Wise, S. L. (1997). Overview of practical issues in a CAT program. In Paper presented at the annual meeting of the national council on measurement in education.
19. Youngseok, L., Jungwon, Ch, 2015. Personalizes item generation method for adaptive testing systems in *Multimedia Tools and Applicatio*ns. Volume: 74, Issue: 19, Pages: 8571-8591.