Qabul qilindi: 12-sentabr 2026-yil • Ma’qullandi: 18-sentabr 2026-yil • Nashr qilindi: 24-sentabr 2026-yil

MEZONLI DIAGNOSTIKANING ILMIY-METODIK ASOSLARI: ANALITIK TAFAKKURNI BAHOLASH TOPSHIRIQLARI, SIFAT DARAJALARI VA BAHOLOVCHILAR MUVOFIQLIGINI TEKSHIRISH

НАУЧНО-МЕТОДИЧЕСКИЕ ОСНОВЫ КРИТЕРИАЛЬНОЙ ДИАГНОСТИКИ: ЗАДАНИЯ ДЛЯ ОЦЕНКИ АНАЛИТИЧЕСКОГО МЫШЛЕНИЯ, УРОВНИ КАЧЕСТВА И ПРОВЕРКА СОГЛАСОВАННОСТИ ЭКСПЕРТОВ

SCIENTIFIC AND METHODOLOGICAL FOUNDATIONS OF CRITERION-REFERENCED DIAGNOSTICS: ANALYTICAL THINKING ASSESSMENT TASKS, PERFORMANCE LEVELS AND INTER-RATER AGREEMENT

Aliyev Nizomiddin Maxmasaitovich
Farg‘ona davlat universiteti mustaqil izlanuvchisi
pedagogika fanlari bo‘yicha falsafa doktori (PhD)
Orcid: 0009-0005-2609-9453

Annotatsiya: Maqolada analitik tafakkurni mezonli baholashning ilmiy-metodik asoslari yoritiladi. Manbalar tahlili asosida topshiriq, mezon va sifat darajalarining bog‘lanishi loyihalanadi. Olti mezonli rubrika hamda mustaqil baholash va moderatsiya protokoli taklif etiladi. Muvofiqlikni tekshirish usullari farqlanib, vaznli kappa sun’iy misolda hisoblanadi.

Kalit so‘zlar: mezonli diagnostika, analitik tafakkur, diagnostik topshiriq, rubrika, sifat darajasi, validlik, baholovchilar muvofiqligi, vaznli kappa.

Аннотация: Раскрыты научно-методические основы критериальной оценки аналитического мышления. На основе анализа источников согласованы задания, критерии и уровни качества ответов. Предложены шесть критериев и протокол

независимого оценивания и модерации. Показатели согласия разграничены; взвешенная каппа рассчитана на искусственных данных.

Ключевые слова: критериальная диагностика, аналитическое мышление, диагностическое задание, рубрика, уровни качества, валидность, согласие экспертов, взвешенная каппа.

Abstract: The article develops a criterion-referenced approach to assessing analytical thinking. A literature-based synthesis links assessment tasks, criteria and performance levels. A six-criterion rubric and an independent scoring and moderation protocol are proposed. Agreement measures are distinguished; weighted kappa is calculated using synthetic data.

Keywords: criterion-referenced diagnostics, analytical thinking, diagnostic task, rubric, performance levels, validity, inter-rater agreement, weighted kappa.

KIRISH

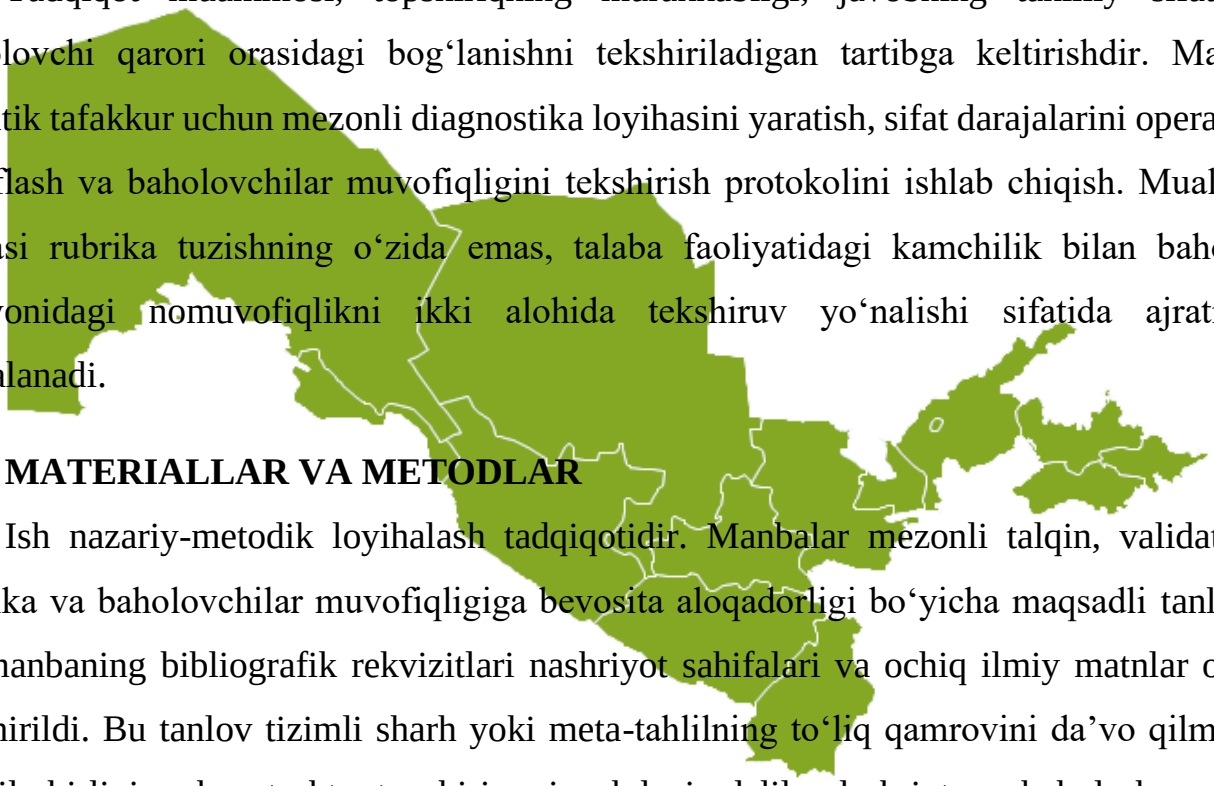
Bo‘lajak pedagogning analitik tafakkurini baholashda asosiy masala javobga ball qo‘yishdan ko‘ra kengroq: kuzatilgan faoliyatdan qaysi tahliliy qobiliyat haqida va qanday chegarada xulosa chiqarish mumkinligini asoslash zarur. Mezonli yondashuv natijani guruhdagi o‘rin bilan emas, oldindan ta’riflangan faoliyat talablari bilan bog‘laydi. Shuning uchun bir xil jami ball dalilni tekshirish, sabablarni farqlash va qarorni asoslash bo‘yicha bir xil tayyorgarlikni anglatishi shart emas [1].

Validlikni instrumentning o‘zgarmas xususiyati sifatida talqin qilish yetarli emas. M. Kane ballardan chiqariladigan talqin va ulardan foydalanishning daliliy asosini tekshirishni taklif etadi [2]. Dalillarga asoslangan loyihalash esa o‘lchanadigan qobiliyat, uni namoyon qiladigan javob va shu javobni yuzaga keltiradigan topshiriqni o‘zaro muvofiqlashtiradi [3]. Mazkur maqolada bu bog‘lanish pedagogik muammo tahlilini tashxislash uchun operatsionlashtiriladi.

S. Brookhart sifat rubrikasini oddiy tekshiruv ro‘yxati yoki «yaxshi–qoniqarli» kabi umumiy baholardan farqlaydi: mezon faoliyatning muhim tomonini, deskriptor esa uning kuzatiladigan sifatini ifodalashi kerak [4]. F. N. Yılmazning 2024-yilgi tadqiqotida reyting

shkalasi orqali olingan ballar analitik rubrika ballariga nisbatan yuqoriroq ishonchlilik ko‘rsatgan [5]. J. Chamberlain va hammualliflarining 2026-yilgi biologiya yo‘nalishidagi yozma ishlar bo‘yicha tadqiqoti ham rubrika turini avtomatik ustun deb bo‘lmasligini, baholovchi tajribasini hisobga olish zarurligini ko‘rsatadi [6]. Bu natijalar O‘zbekiston pedagogika talabalari uchun bevosita tasdiq emas, balki mahalliy tekshiruvni loyihalash uchun dalildir.

Tadqiqot muammosi, topshiriqning murakkabligi, javobning tahliliy sifati va baholovchi qarori orasidagi bog‘lanishni tekshiriladigan tartibga keltirishdir. Maqsad analitik tafakkur uchun mezonli diagnostika loyihasini yaratish, sifat darajalarini operatsion ta’riflash va baholovchilar muvofiqligini tekshirish protokolini ishlab chiqish. Mualliflik hissasi rubrika tuzishning o‘zida emas, talaba faoliyatidagi kamchilik bilan baholash jarayonidagi nomuvofiqlikni ikki alohida tekshiruv yo‘nalishi sifatida ajratishda ifodalanadi.



MATERIALLAR VA METODLAR

Ish nazariy-metodik loyihalash tadqiqotidir. Manbalar mezonli talqin, validatsiya, rubrika va baholovchilar muvofiqligiga bevosita aloqadorligi bo‘yicha maqsadli tanlandi; 12 manbaning bibliografik rekvizitlari nashriyot sahifalari va ochiq ilmiy matnlar orqali tekshirildi. Bu tanlov tizimli sharh yoki meta-tahlilning to‘liq qamrovini da’vo qilmaydi. Tahlil birligi, «konstrukt, topshiriq, javobdagi dalil, deskriptor, baholash qarori» bog‘lanishidir [2; 3].

Konseptual qiyoslash, mezonlarni operatsionlashtirish va mantiqiy modellashtirish orqali to‘rtta topshiriq turi hamda olti mezonli rubrika loyihalandi. Hisoblashning takrorlanuvchanligini ko‘rsatish uchun sakkiz juft sun’iy baho asosida vaznli kappa hisoblandi. Talabalar jalb qilinmadi; ekspertiza, pilot sinov va empirik validatsiya o‘tkazilgani da’vo qilinmaydi. Natijalar bo‘limidagi model va tashxis qoidalari tekshirilishi lozim bo‘lgan metodik takliflardir.

NATIJALAR

Analitik tafakkurning diagnostik modeli. Ushbu ishlanmada analitik tafakkur

pedagogik vaziyatni tarkiblarga ajratish, dalilning sifatini tekshirish, munosabatlarni qiyoslash, muqobil izohlarni baholash va xulosani yangi axborotga muvofiq qayta asoslash faoliyati sifatida ta’riflanadi. Konstrukt umumiy bilim hajmi, nutqning bezakdorligi yoki javob uzunligi bilan tenglashtirilmaydi. M1–M6 mezonlari mos ravishda muammoni tarkiblash, dalilni baholash, sababiy tahlil, qiyoslash, qarorni dalillash va refleksiv tuzatishni qamrab oladi. Bular empirik aniqlangan faktorlar emas, mazmuniy modelning tekshiriladigan qismlaridir.

Topshiriqlar uchun yagona loyiha formati taklif etiladi: qisqa kasbiy vaziyat, manbasi ko‘rsatilgan ma’lumot paketi, oltita tahliliy savol va yangi axborot berilgandan keyingi qayta xulosa. Har bir vaziyatda barcha mezonlarga oid dalil olish ko‘zda tutiladi. To‘rtta vaziyatning mavzusi farqlanadi; rubrika esa saqlanadi. Pilot reja uchun 10 daqiqalik yo‘riqnoma va to‘rtta 20 daqiqalik topshiriq, jami 90 daqiqa ajratiladi. Bu vaqt me’yor emas: pilotda tezlik omilining natijaga ta’siri tekshirilib, zarur bo‘lsa muddat o‘zgartiriladi [1].

Diagnostik topshiriqlarning mazmuniy spetsifikatsiyasi.

T1. Pedagogik ma’lumotni tahlil qilish. Talabaga ikki sinfning boshlang‘ich va yakuniy natijalari, ishtirokchilar soni va topshiriq sharoiti beriladi. U muammoni ajratadi, hisoblarning qiyoslanishini tekshiradi, natijadagi farqqa kamida ikki izoh taklif qiladi va ehtiyotkor pedagogik qaror chiqaradi. Keyin guruhlardan birida ayrim o‘quvchilar yakuniy tekshiruvda qatnashmagani ma’lum qilinadi; xulosani qayta ko‘rish talab etiladi.

T2. Manbalar ishonchliligini baholash. Bir metod haqida tadqiqot parchasi, o‘qituvchi kuzatuvi va reklama mazmunidagi xabar taqdim etiladi. Talaba da’volarning doirasini, dalil bilan fikrning farqini va ma’lumot yig‘ish usulini qiyoslaydi. Bir xil xabarni qayta tarqatgan manbalarni mustaqil tasdiq deb hisoblamasligi tekshiriladi. Yangi cheklov axboroti berilgach, tavsiya va ishonch darajasi qayta asoslanadi.

T3. Muqobil pedagogik qarorlarni qiyoslash. Past ishtirok muammosiga uchta yechim taklif qilinadi. Talaba vaqt, resurs, o‘quv maqsadi va ehtimoliy zarar mezonlari asosida variantlarni taqqoslaydi; o‘z tanloviga qarshi dalilni ham ko‘rsatadi. O‘quvchining intizomini uning motivatsiyasi bilan avtomatik tenglashtirishga yo‘l qo‘ymasligi kuzatiladi. Qo‘shimcha kontekst qaysi qarorni va nima sababdan o‘zgartirishini yozadi.

T4. Tahlilni yangi vaziyatga ko‘chirish. Oldingi vaziyatlarning sonlari yoki nomlari almashtirilmaydi; boshqa pedagogik mazmundagi muammo beriladi. Talaba tanish tahlil usulini moslashtiradi, uning qo‘llanish chegarasini belgilaydi va qo‘shimcha dalil so‘raydi. Yakuniy javob «qaysi dalil xulosamni o‘zgartiradi?» savolini qamrab oladi. Shu tariqa andazani takrorlash bilan asosli transferni farqlashga sharoit yaratiladi.

Har bir javobdagi baholash birligi, fikrning o‘zi emas, «da’vo, tegishli dalil, bog‘lanish izohi» bo‘lagidir. To‘g‘ri yakuniy qarorga tasodifan kelish yuqori tahliliy sifat deb qabul qilinmaydi. Barcha vaziyatlar mualliflik o‘quv loyihalari bo‘lib, real ta’lim muassasasida qayd etilgan holatlar sifatida berilmaydi.

Mezonlar va sifat darajalarini operatsionlashtirish. Darajalar javob sifatini bildiradi, shaxsga doimiy yorliq qo‘ymaydi. 0-3 kodlari tartibli kategoriyalardir: darajalar orasidagi masofa tengligi oldindan isbotlangan deb olinmaydi. Jadvaldagi mezonlar vaziyatga xos tayanch javoblar bilan to‘ldirilishi kerak; umumiy deskriptorning o‘zi baholovchi tayyorgarligini almashtirmaydi [4].

1-jadval. Analitik tafakkurni baholash uchun olti mezonli rubrika loyihasi

Mezon	0- dalil namoyon emas	1- fragmentar tahlil	2- asoslangan tahlil	3- integrativ- refleksiv tahlil
M1. Muammoni tarkiblash	Muammo o‘rniga aloqasiz fikr yoki vaziyat nusxasini beradi.	Muammoni aytadi; muhim shartlar va tarkiblar aralash qoladi.	Muammo, tarkib va shartlarni ajratib, aloqalarini ko‘rsatadi.	Tarkiblar o‘zaro ta’sirini va muammo chegarasini asoslaydi.
M2. Dalil sifatini baholash	Da’voni dalilsiz qabul qiladi yoki manbani noto‘g‘ri ko‘rsatadi.	Manbani ataydi; uning mosligi va cheklovini tekshirmaydi.	Dalilning manbasi, mosligi va asosiy cheklovini tekshiradi.	Zid dalillarni solishtiradi; manbalar mustaqilligi va noaniqlikni baholaydi.
M3. Sababiy bog‘lanish	Birga kuzatilgan hodisalarni asossiz sabab–oqibat deb oladi.	Bitta sababni taxmin qiladi, ammo tekshirish yo‘lini bermaydi.	Muqobil sabablarni ajratib, ularni tekshiradigan dalilni belgilaydi.	Aralashuvchi omil va qarshi dalilni hisobga olib, sababiy da’voni chegaralaydi.
M4. Asosli qiyoslash	Mos kelmaydigan ko‘rsatkichlardan noto‘g‘ri xulosa chiqaradi.	Farqni qayd etadi; qiyoslash sharti yoki mezoni noaniq.	Yagona mezon va mos ko‘rsatkichlar asosida to‘g‘ri qiyoslaydi.	Qiyos natijasini sharoit, muqobil mezon va noaniqlik bilan tekshiradi.

M5. Qarorni dalillash	Qaror dalilga zid yoki daliliy asosga ega emas.	Qaror va dalil bor, ammo ular orasidagi bog‘lanish zaif.	Qarorni dalil va mantiqiy izoh bilan asoslaydi.	Muqobil qaror, qarshi dalil va qo‘llanish shartlarini asosli muvofiqlashtiradi.
M6. Refleksiv tuzatish	Yangi zid axborotni e’tiborsiz qoldiradi.	Javobni o‘zgartiradi, lekin o‘zgarish sababini izohlamaydi.	Yangi dalilga ko‘ra xulosa va uning asosini tuzatadi.	O‘zgargan taxminni ochadi; qayta xulosani yangi vaziyatda tekshirish yo‘lini belgilaydi.

Kodlash qoidasi. Baholovchi har bir mezon uchun javobdan dalil bo‘lagini qayd etib, eng mos deskriptorni tanlaydi. Qo‘shni darajalar orasidagi noaniqlik alohida belgilanadi. Texnik sababli olinmagan yoki topshirilmagan javob «NA» sifatida ajratiladi; bajarilgan javobda zarur dalil yo‘qligi esa 0 bilan kodlanadi. Bu ikki holat avtomatik birlashtirilmaydi.

Talaba bo‘yicha to‘rtta vaziyatdagi M1–M6 profili saqlanadi. Jami ballni teng oraliqlarga bo‘lib «past, o‘rta, yuqori» chegaralarini belgilash taklif etilmaydi. Chegaraviy qarorlar uchun alohida standart belgilash, mustaqil tekshiruv va noto‘g‘ri tasniflash oqibatlarini baholash kerak [1; 2].

Baholovchilar muvofiqligini tekshirish protokoli. Tayyorgarlik bosqichida baholovchilar topshiriq maqsadi va har bir mezonning chegarasini muhokama qiladi; turli sifatdagi tayanch javoblar bilan mashq bajaradi. Ikki baholovchi pilotdagi barcha ishlarni bir-birining bahosini ko‘rmasdan kodlashi taklif etiladi. Trening javoblari audit namunasiga qo‘shilmaydi. Har bir yozuv talaba kodi, topshiriq, mezon, baholovchi, ball, dalil parchasi va rubrika versiyasini saqlaydi [9].

Nomuvofiqlik aniqlanganda u «deskriptor noaniqligi», «dalilni turlicha talqin qilish», «topshiriq kamchiligi» yoki «kodlash xatosi» kabi ishchi toifalarda qayd etiladi. Avval mustaqil baholar bo‘yicha muvofiqlik hisoblanadi, undan keyingina moderatsiya qilinadi. Kelishilgan yakuniy ball dastlabki baholarni o‘chirmaydi. Rubrika o‘zgartirilsa, uning yangi versiyasi boshqa javoblarda qayta tekshiriladi; baholarni oldindan kelishtirib, so‘ng yuqori muvofiqlik e’lon qilishga yo‘l qo‘yilmaydi.

Ordinal deskriptorlar uchun ikki baholovchining aynan mos kelish ulushi, bir pog‘ona ichidagi moslik va Cohenning vaznli kappasi birga beriladi [8; 9]. Dastlabki ko‘rsatkichlar

farqning amaliy hajmini tavsiflaydi, kappa esa marginal taqsimotlardan kutiladigan moslikni hisobga oladi. Ular bir-birining o‘rnini bosmaydi. Vaznlar tahlildan oldin belgilanadi; mazkur loyiha darajalar orasidagi qadamlar uchun chiziqli kelishuv vaznini qo‘llaydi.

$w_{ij} = 1 - |i - j|/3$; $\kappa w = (Pw - Pe)/(1 - Pe)$, $i, j \in \{0, 1, 2, 3\}$.

Bu yerda Pw- kuzatilgan juft baholarning vaznli moslik ulushi; Pe- baholovchilarning alohida toifa ulushlari ko‘paytmalaridan hisoblangan kutiladigan vaznli moslik. To‘liq moslikka 1, bir pog‘ona farqqa 2/3, ikki pog‘ona farqqa 1/3, uch pog‘ona farqqa 0 vazn beriladi. Bu vazn tanlovi hisoblash qoidasi bo‘lib, konstruktning teng intervalli shkala ekanini isbotlamaydi [8].

Sun’iy ma’lumotlar bilan takrorlanadigan hisoblash.

Faqat bitta mezon va bitta topshiriq bo‘yicha sakkizta sun’iy javob baholangan deb olinadi: A = (0; 1; 1; 2; 2; 2; 3; 3), B = (0; 1; 2; 2; 2; 3; 3; 3). Aynan mos baholar 6 ta: 6/8 × 100 = 75%. Qolgan ikki farq bir pog‘onadan, shuning uchun bir pog‘ona ichidagi moslik 100%. A ning 0–3 toifalar bo‘yicha sonlari (1; 2; 3; 2), B niki (1; 1; 3; 3).

Kuzatilgan vaznli moslik: $Pw = [6 \times 1 + 2 \times (2/3)]/8 = 11/12$. Marginal ulushlar bo‘yicha $Pe = \Sigma_i\Sigma_j w_{ij}(a_i/8)(b_j/8) = 61/96$. Demak, $\kappa w = [(88 - 61)/96]/[(96 - 61)/96] = 27/35 \approx 0{,}771$. B baholovchi ikkita javobga bir pog‘ona yuqori ball bergan; shu sababli 100% yaqin moslik mutlaq teng baholash degani emas.

Mazkur sakkiz juftlik usulni tushuntirish uchungina yaratilgan; 0,771 mahalliy rubrikaning ishonchlilik natijasi emas. Haqiqiy hisobotda har bir mezon, topshiriq kesimida juftlar soni, toifalar taqsimoti, yo‘qolgan ma’lumotlar, kelishmovchilik yo‘nalishi va 95% ishonch oralig‘i keltirilishi kerak. Bir talabaning bir nechta javobi birgalikda tahlil qilinsa, ularning bog‘liqligi saqlanadigan klasterli baholash usuli qo‘llanadi; satrlarni mustaqil kuzatuv deb ko‘paytirishdan saqlaniladi.

MUHOKAMA

Mezonli diagnostikada statistik muvofiqlik zarur, ammo mazmuniy to‘g‘rilikning o‘rnini bosa olmaydi: ikki baholovchi bir xil noto‘g‘ri deskriptorni izchil qo‘llashi mumkin

[2; 9]. Ishlanmaning afzalligi, baho bilan uni asoslagan javob bo‘lagini saqlash; cheklovi esa tahliliy tafakkurning murakkabligini olti mezon bilan to‘liq qamrab olish hozircha tekshirilmaganidir. Baholovchilar kelishmasa, talabaning «past darajasi» haqida shoshilinch xulosa chiqarish o‘rniga topshiriq va rubrika ham audit qilinishi lozim.

Umumlashtirilgan ballning mazmuniy tuzilishi va uni sonli shkala sifatida qo‘llash asoslangandagina qo‘shimcha ICC tahlili rejalashtiriladi. Baholovchilarni kengroq o‘xshash baholovchilar populyatsiyasiga umumlashtirish ko‘zlansa, ikki yo‘nalishli tasodifiy effektli, mutlaq kelishuvga asoslangan ICC(2,1) bitta baholovchi balli uchun; ICC(2,k) esa k baholovchi o‘rtachasi uchun mos keladi. Faqat muayyan hay’at bilan cheklangan xulosa uchun aralash effektli model va umumlashtirish chegarasi ko‘rsatiladi. Model, birlik, mutlaq kelishuv yoki izchillik tanlovi hamda 95% ishonch oralig‘i ochiq yoziladi [10].

Kappa ordinal profil, ICC esa asoslangan sonli umumlashtirish uchun xizmat qiladi; yuqori Pearson korrelyatsiyasi aynan bir xil ball qo‘yilishini bildirmaydi [9; 10]. Ko‘p topshiriqli dizaynda «talaba, topshiriq, baholovchi» omillarini ajratadigan umumlashtiriluvchanlik nazariyasi qo‘shimcha imkoniyat beradi. Mezonli, mutlaq qaror uchun Φ koeffitsiyenti nisbiy tartiblashga xizmat qiluvchi G koeffitsiyentidan farqlanadi. Keyingi D-tadqiqot topshiriqlar yoki baholovchilar sonini o‘zgartirishning aniqlikka ta’sirini baholashi mumkin [11]. Bu maqolada mazkur koeffitsiyentlar hisoblanmagan.

Validatsiya rejasi to‘rtta bog‘langan yo‘nalishni qamrab oladi: mustaqil ekspertlarning mazmuniy qamrov va deskriptor chegaralarini tekshirishi; talabalarning ovoz chiqarib fikrlashi orqali kutilgan tahliliy amal yuzaga kelishini kuzatish; mustaqil baholar va vaziyatlararo natijalar barqarorligini baholash; yangi pedagogik vaziyatdagi faoliyat bilan bog‘lanishni tekshirish. Tanlanma hajmi qulay son bilan emas, rejalashtirilgan ko‘rsatkichlarning aniqligi, model murakkabligi va maqsadli populyatsiya xilma-xilligi asosida belgilanadi.

Diagnostik profil asosida teskari aloqa «qaysi mezon, qaysi dalil, qanday tuzatish» shaklida tuziladi. D. Sadler sifat mezonini tushunish, o‘z ishini u bilan qiyoslash va tuzatish usullarini o‘zlashtirishni bog‘laydi [12]. E. Panadero va hammualliflarining 2025-yilgi universitet talabalari ishtirokidagi tadqiqoti rubrikaga e’tibor hamda fikr-mulohaza turi

faoliyat bilan bog‘liqligini ko‘rsatgan [7]. Shu bois mahalliy sinovda faqat ball emas, talabaga tavsiya tushunarli va bajariladigan bo‘lishi ham tekshiriladi.

Mahalliy qo‘llashda fan bo‘yicha bilim, topshiriq tili, raqamli savodxonlik va javob yozish tezligi natijani chalkashtirishi mumkin; zarur moslashtirishlar hujjatlashtiriladi. Rubrikani takror mashq qilish natijasi umumiy analitik rivojlanish bilan tenglashtirilmaydi. Ishlanma mahalliy pilot va tashqi tekshiruvsiz saralash, attestatsiya yoki boshqa yuqori mas’uliyatli qarorlar uchun tayyor instrument deb taqdim etilmaydi.

XULOSA

Taklif etilgan model analitik tafakkurni to‘rtta kasbiy vaziyat, olti mezon va to‘rt tartibli sifat darajasi orqali tekshirishni loyihalaydi. Uning asosiy metodik talabi, talaba javobidagi dalil, baholovchining mustaqil qarori va keyingi moderatsiya natijasini alohida saqlashdir. Ishonchli mezonli diagnostika bitta koeffitsiyentga emas, mazmuniy validatsiya, mos statistik tahlil va tashxisga muvofiq pedagogik tuzatishning birligiga tayanishi lozim. Ushbu birlik empirik tasdiq emas, keyingi tadqiqot uchun tekshiriladigan loyiha sifatida taklif etiladi.

FOYDALANILGAN ADABIYOTLAR

1. American Educational Research Association; American Psychological Association; National Council on Measurement in Education. Standards for Educational and Psychological Testing. Washington, DC: AERA, 2014. https://www.testingstandards.net/open-access-files.html

2. Kane M. T. Validating the Interpretations and Uses of Test Scores. Journal of Educational Measurement. 2013; 50(1): 1–73. https://doi.org/10.1111/jedm.12000

3. Mislevy R. J., Almond R. G., Lukas J. F. A Brief Introduction to Evidence-Centered Design. ETS Research Report Series. 2003; 2003(1): i–29. https://doi.org/10.1002/j.2333-8504.2003.tb01908.x

4. Brookhart S. M. Appropriate Criteria: Key to Effective Rubrics. Frontiers in Education. 2018; 3: 22. https://doi.org/10.3389/feduc.2018.00022

5. Yılmaz F. N. Comparing the reliability of performance task scores obtained from

rating scale and analytic rubric using the generalizability theory. Studies in Educational Evaluation. 2024; 83: 101413. https://doi.org/10.1016/j.stueduc.2024.101413

6. Chamberlain J., Francis S., Herrick T. Assessor experience, not rubric type, determines grading reliability in biosciences coursework. Frontiers in Education. 2026; 11: 1729644. https://doi.org/10.3389/feduc.2026.1729644

7. Panadero E., Delgado P., Zamorano D., Pinedo L., Fernández-Ortube A., Barrenetxea-Mínguez L. Putting excellence first: How rubric performance level order and feedback type influence students’ reading patterns and task performance. Learning and Instruction. 2025; 99: 102168. https://doi.org/10.1016/j.learninstruc.2025.102168

8. Cohen J. Weighted kappa: Nominal scale agreement with provision for scaled disagreement or partial credit. Psychological Bulletin. 1968; 70(4): 213–220. https://doi.org/10.1037/h0026256

9. Hallgren K. A. Computing Inter-Rater Reliability for Observational Data: An Overview and Tutorial. Tutorials in Quantitative Methods for Psychology. 2012; 8(1): 23–34. https://doi.org/10.20982/tqmp.08.1.p023

10. Koo T. K., Li M. Y. A Guideline of Selecting and Reporting Intraclass Correlation Coefficients for Reliability Research. Journal of Chiropractic Medicine. 2016; 15(2): 155–163. https://doi.org/10.1016/j.jcm.2016.02.012

11. Brennan R. L. Generalizability Theory. New York: Springer, 2001. https://doi.org/10.1007/978-1-4757-3456-0

12. Sadler D. R. Formative assessment and the design of instructional systems. Instructional Science. 1989; 18: 119–144. https://doi.org/10.1007/BF00117714