Psychology. Journal of the Higher School of Economics. 2026. Vol. 23. N 2. P. 263–274.
Психология. Журнал Высшей школы экономики. 2026. Т. 23. № 2. С. 263–274.
DOI: 10.17323/1813-8918-2026-2-263-274

FROM MEASUREMENT TO UNDERSTANDING: AN ARCHITECTURE OF EMULATED EXPERTISE FOR LLM-DIAGNOSTICS IN EDUCATION

O.E. KLEPIKOV[a], E.M. POZDNYAKOVA[a], YA.O. SIZOV[a]

[a] *MGIMO University, 76 Vernadskogo Ave., Moscow, 119454, Russian Federation*

От измерения к пониманию: архитектура эмулированной экспертизы для LLM-диагностики в образовании

О.Е. Клепиков[a], Е.М. Позднякова[a], Я.О. Сизов[a]

[a] *Московский государственный институт международных отношений (университет) Министерства иностранных дел Российской Федерации (МГИМО МИД России), 119454, Россия, Москва, проспект Вернадского, д. 76*

Abstract

The demand for developmental diagnostics in personalized education reveals a fundamental crisis in classical psychometrics. This paper presents a conceptual blueprint for a new generation of LLM-based diagnostic systems engineered to ensure methodological validity in high-stakes educational contexts. The project rests on five interconnected principles: methodological grounding in evidence-centered design (ECD), a hybrid 'centaur' architecture, echeloned quality control, role-oriented utility, and data sovereignty. Its core innovation is an 'emulated expertise architecture' that emulates not the expert's consciousness but their idealized professional workflow, decomposing the act of judgment into a network of verifiable procedures – systematic falsification, independent peer review (supervision), and dialectical synthesis – executed by a multi-agent system of functionally specialized "prompt archetypes" ("researcher",

Резюме

Запрос на развивающую диагностику в персонализированном образовании обнажает фундаментальный кризис классической психометрики. Настоящая работа представляет концептуальный проект (blueprint) нового поколения LLM-диагностических систем, спроектированных для обеспечения методологической валидности в высокорисковых образовательных контекстах. Проект базируется на пяти взаимосвязанных принципах: методологическая обоснованность (evidence-centered design, ECD), гибридная «кентавр-архитектура», эшелонированный контроль качества, ролевая ориентированность и суверенитет данных. Ключевой инновационный элемент — «архитектура эмулированной экспертизы», которая эмулирует не сознание эксперта, а его идеализированный профессиональный рабочий процесс, структурируя акт суждения в виде сети верифицируемых процедур: систематической фальсификации, независимого рецензирования (супервизии) и диалектического синтеза, исполняемых мультиагентной системой функционально специализированных «промпт-архетипов» («исследователь»,

"auditor", "communicator"). The entire pipeline is governed by the 'human-over-the-loop' (HOTL) ethical imperative, positioning the system as a tool for expert augmentation rather than replacement. This approach facilitates a paradigm shift from a "psychometrics of measurement" to a "psychometrics of understanding", generating not merely a score but a dynamic, contextualized model of the learner's cognitive processes. Supplementary materials detail the five-phase methodology for translating psychometric instruments into behavioral analysis protocols and the complete microarchitectures of all agent archetypes. By integrating computational psychometrics with educational psychology, the proposed architecture opens a pathway toward valid, reliable, and interpretable tools, marks a shift from static measurement to a dynamic understanding of students' educational preferences and outcomes, and offers a practical instrument for advancing student-centered university education.

Keywords: computational psychodiagnostics, AI in education, large language models (LLM), computational psychometrics, construct validity, formative assessment, multi-agent systems, prompt engineering, explainable AI (XAI).

Oleg E. Klepikov — Head of Laboratory, IME MGIMO Laboratory, MGIMO University, Associate Professor.
Research Area: applied cognitive sciences, instrumental psychodiagnostics, stylometry in psychodiagnostic research, cognitive ergonomics.
E-mail: oleg.e.klepikov@gmail.com

Elena M. Pozdnyakova — Professor, Head of Neurolinguistics Research Center for Cognitive Potential Studies, MGIMO University, DSc in Philology, Professor.

«аудитор», «коммуникатор»). Весь процесс подчинен этическому императиву «Человек-над-циклом» (Human-over-the-Loop, HOTL), позиционирующему систему как инструмент аугментации эксперта, а не его замены. Такой подход обеспечивает парадигмальный сдвиг от «психометрии измерения» к «психометрии понимания», генерируя не просто балл, а динамическую, контекстуализированную модель когнитивных процессов учащегося. Для обеспечения прозрачности и воспроизводимости работа сопровождается развернутыми дополнительными материалами, детализирующими пятифазную методологию трансляции психометрических инструментов в протоколы поведенческого анализа, а также полные микроархитектуры всех агентных архетипов. Интегрируя вычислительную психометрию с педагогической психологией, предложенная архитектура открывает путь к созданию валидных, надежных и интерпретируемых инструментов, обозначает переход от статичного измерения к динамическому пониманию образовательных предпочтений и результатов обучающихся и предоставляет практический инструмент для развития личностно-ориентированного университетского образования.

Ключевые слова: вычислительная психодиагностика, искусственный интеллект в образовании, большие языковые модели (LLM), вычислительная психометрика, конструктная валидность, формирующее оценивание, мультиагентные системы, промпт-инжиниринг, объяснимый ИИ (XAI).

Клепиков Олег Евгеньевич — руководитель лаборатории, нейролаборатория ИМИП МГИМО, Московский государственный институт международных отношений (университет) Министерства иностранных дел Российской Федерации, доцент.
Сфера научных интересов: прикладные когнитивные науки, инструментальная психодиагностика, стилометрия в психодиагностических исследованиях, когнитивная эргономика.
Контакты: oleg.e.klepikov@gmail.com

Позднякова Елена Михайловна — профессор, руководитель центра, Центр нейролингвистических исследований когнитивного потенциала, Московский государственный институт международных отношений (университет) Министер-

Research Area: cognitive sciences (linguistics and education), semiotics, neurolinguistics.
E-mail: helenpozdnyakova@yandex.ru

Yaroslav O. Sizov — Research Assistant, MGIMO University.
Research Area: social anthropology, human ethology, applied cognitive sciences.
E-mail: sizoff.yaroslav@yandex.ru

ства иностранных дел Российской Федерации, доктор филологических наук, профессор.
Сфера научных интересов: когнитивные науки (лингвистика и образование), семиотика, нейролингвистика.
Контакты: helenpozdnyakova@yandex.ru

Сизов Ярослав Олегович — лаборант-исследователь, Московский государственный институт международных отношений (университет) Министерства иностранных дел Российской Федерации.
Сфера научных интересов: социальная антропология, этология человека, прикладные когнитивные науки.
Контакты: sizoff.yaroslav@yandex.ru

The demand for developmental diagnostics exceeds classical psychometrics' limits, creating a methodological impasse that requires a new, LLM-catalyzed paradigm.

Classical psychodiagnostics assumes accurate individual access to cognitive processes. However, research shows people often lack awareness of the true stimuli for their judgments, offering plausible post-hoc causal theories instead of accurate introspection (Nisbett & Wilson, 1977). Self-report validity is undermined by social desirability (Paulhus, 2002), response construction biases (Schwarz, 1999), referential shifts (Duckworth & Yeager, 2015), and retrospective memory distortions (Credé et al., 2005). Introspective methods thus measure the self-concept – an internal, socially constructed model, not behavioral dispositions.

Furthermore, the psychometric foundation of traditional diagnostics, particularly item response theory (IRT), is designed for stable, unidimensional latent traits (Embretson & Reise, 2013). This approach measures stable traits but fails to capture the dynamic cognitive processes essential for education, creating a significant methodological gap. This gap reflects the split between the "two disciplines of scientific psychology" (Cronbach, 1957) and echoes critiques of testology's detachment from real activity (Gurevich, 1998; Leontiev, 2011). The resulting low predictive power (Mischel, 1968) is now termed the "generalizability crisis" (Yarkoni, 2022).

Classical psychodiagnostics faces practical barriers: developing and standardizing tests is too resource-intensive for mass personalization (Baturin et al., 2015), and tools from WEIRD (Educated, Industrialized, Rich, and Democratic) samples (Henrich et al., 2010) risk cultural bias, requiring costly re-validation (Kornilov et al., 2009). This creates a methodological zugzwang, as reliance on standardized testing can degrade the educational process (Koretz, 2017).

This paper presents a framework for valid, reliable, and interpretable next-generation psychodiagnostic systems.

The Observational Paradigm: a Technological Catalyst and New Challenges

The introspective crisis requires shifting from self-reports to objective behavior. LLMs catalyze this shift but introduce new methodological challenges.

The observational paradigm measures what individuals do, not who they are. This approach, called "stealth assessment" (Rahimi & Shute, 2023), embeds assessment into naturalistic activity. Shifting from predefined answers to "digital footprints" (Lazer, 2015), it enables direct computational analysis of complex psychological phenomena with superior predictive power over self-reports (Kosinski et al., 2013). Natural language is a significant digital footprint, as word choice signals psychological states (Pennebaker, 2011) and narratives form identity (McAdams, 2015).

The transformer architecture (Vaswani et al., 2017) ensured the observational paradigm at scale. Modern "foundation models" (Bommasani et al., 2021) exhibit "emergent abilities" for reasoning (Wei et al., 2022) activated by techniques like chain-of-thought (CoT) prompting (Kojima et al., 2022).

A 'black box' application of foundation models poses validity and ethical risks:

• The 'black box' problem: neural network complexity renders LLMs internal logic uninterpretable (Burrell, 2016), making diagnostics opaque. High-stakes decisions require interpretable-by-design models, not post-hoc "explainable" black boxes (Rudin, 2019).

• Algorithmic bias: LLMs reproduce and amplify societal stereotypes (Caliskan et al., 2017; O'Neil, 2016). Even after explicit debiasing, large language models continue to show systematic group-related preference distortions in applied decision tasks (Bai et al., 2025). Autonomous, unaudited use in high-stakes settings is therefore not acceptable and requires multilayer auditing and human-over-the-loop control.

• Construct validity (a new form): LLMs may learn to rely on superficial linguistic correlations rather than deep psychological patterns, creating "measurement phantoms" (Lin, 2025). They can also simulate desired responses (Argyle et al., 2023), a computational analog to social desirability. Without an external, theoretically grounded methodology, their construct validity is unprovable.

Principles for Designing Methodologically Sound Systems

A sound methodology uses an LLM to execute an external, explicit, scientific protocol. Five interconnected principles form the system's conceptual framework:

• *Methodological grounding (evidence-centered design).* Externalizes diagnostic logic into an ECD protocol (Mislevy et al., 2003), requires interpretable, human-auditable reasoning (Rudin, 2019) and out-of-sample predictive validity; following Yarkoni & Westfall (2017), predictive performance on new data is treated as a necessary criterion for scientific credibility.

• *Hybrid 'centaur' architecture.* Rejects full automation for 'centaur' systems of AI-augmented experts (Brynjolfsson, 2022; Saghafian & Idan, 2024), using a deterministic orchestrator to manage specialized LLM agent networks.

• *Echeloned quality control.* A multi-level system institutionalizing "organized skepticism" (Merton, 1973) over unreliable model self-critique (Madaan et al., 2023) through systematic falsification and peer review.

• *Role-oriented and didactic utility.* The system must generate role-oriented outputs, not a single generic report. This includes an adaptive, formative report for

the learner to develop "feedback literacy" (Wiliam, 2018; Carless & Boud, 2018) and a professional report for the specialist.

• *Data sovereignty and security.* Mandates 'security by design,' ensuring strict compliance with legal norms like 152-FZ. Technical solutions ensure trust (e.g., local deployment, encryption, RBAC (Role-Based Access Control), auditable logging) and compliance with data governance (the General Data Protection Regulation (GDPR)) and quality (GOST R ISO/IEC 25010-2015) standards (Federal Agency for Technical Regulation and Metrology, 2015).

The Emulated Expertise Architecture: a System Blueprint

The 'emulated expertise architecture' implements these principles by decomposing an expert workflow into formalized, verifiable procedures (see Table 1, Figure 1).

1. Macroarchitecture: Decomposition and Control Flows

The macroarchitecture comprises functional units: "orchestrators", a "domain expert", a "domain supervisor", and a "simplifier". The workflow is an orchestrated pipeline of collection, analysis, synthesis, and persistence, with iterative feedback loops implementing echeloned quality control.

2. Microarchitecture: Functional Specialization of Prompts

The microarchitecture defines each agent's cognitive toolkit via prompt engineering, codifying methodology into executable instructions. It uses differentiated prompt assemblies for three archetypes: the "researcher", the "auditor", and the "communicator".

The "prompt researcher" (a domain expert) performs a three-stage hierarchical synthesis: 1) hypothesis generation and verification (heuristic core), 2) reliability self-audit, and 3) finalization, packaging verified knowledge into qualitative and quantitative outputs. The heuristic core uses hierarchical synthesis to manage complexity, systematic falsification via perspective shifting to combat confirmation bias, and formalized arbitration to objectify hypothesis selection.

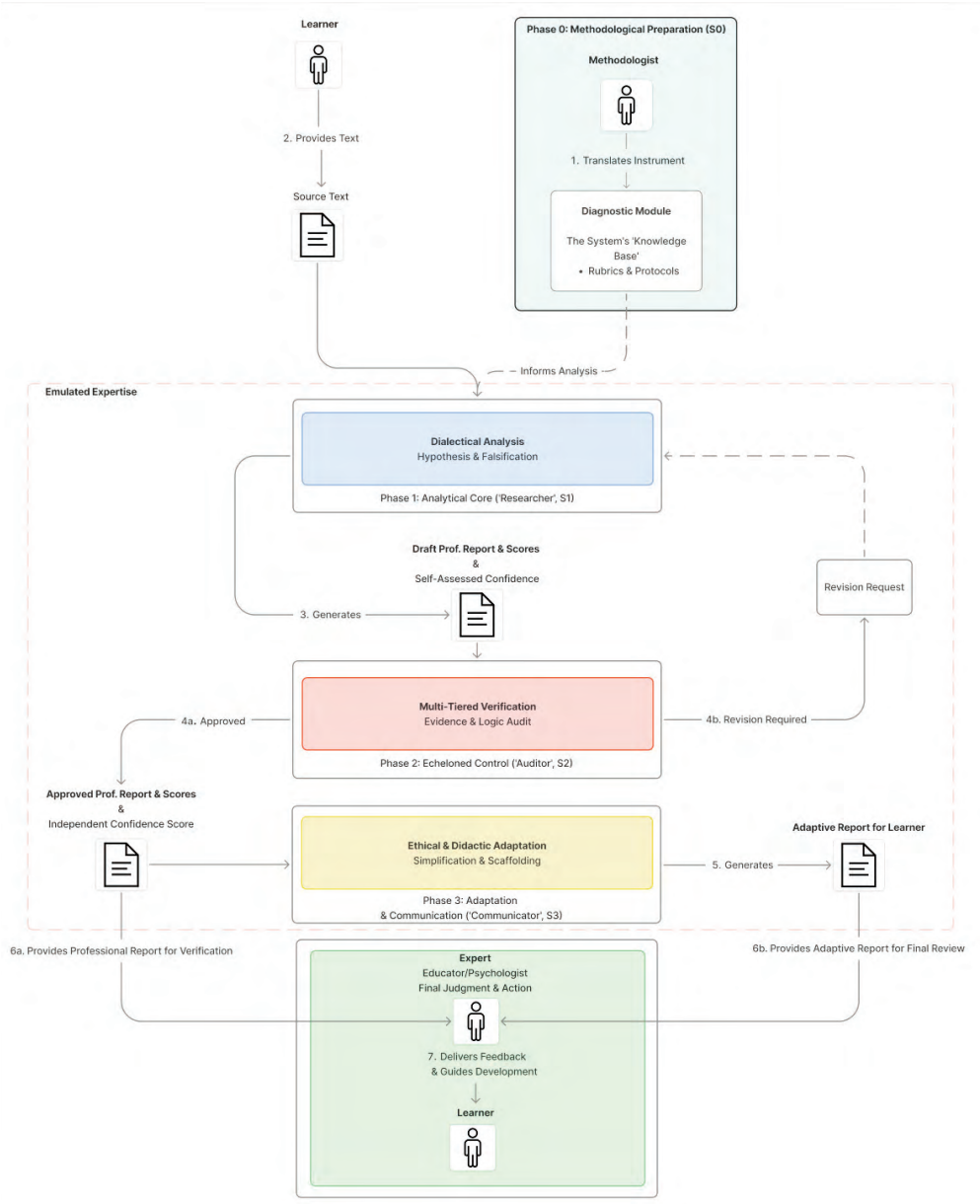
Table 1

Mapping of Evidence-Centered Design (ECD) Methodology to Architectural Components

ECD model	Architectural component	Generated artifact
Competency model	Prompt library, assessment rubrics	Diagnostic construct, scales
Evidence model	Interpreter agents, arbiter agent	Local findings, confidence score
Task model	Orchestrator, input data (text)	Professional and adaptive reports

Figure 1

High-Level Representation of the Conceptual Blueprint of The Emulated Expertise Architecture



The "prompt auditor" (a supervisor, validators) implements organized skepticism by verifying the 'researcher's' findings against predefined standards via a multi-tiered audit of the evidence, methodology, and reliability.

The "prompt communicator" (a simplifier) translates verified findings into safe, valuable end-user products through validation, stylistic adaptation, and generating auxiliary teaching materials.

Full specifications for the "researcher" (S1), "auditor" (S2), and "communicator" (S3) are in the supplementary materials (Klepikov et al., 2025).

3. Governing principle: human-over-the-loop (HOTL)

The HOTL principle defines roles and responsibility, positioning the machine as an expert augmentation tool under "meaningful human control" (IEEE, 2019). The system, as a "cognitive artifact" (Norman, 1993), provides the expert with a professional report, confidence scores for attention-prioritization, and full data traceability for independent verification (see Table 2). However, the HOTL model introduces risks like automation bias (Parasuraman & Riley, 1997; Goh et al., 2024) and the responsibility gap (Matthias, 2004).

Table 2

Correspondence of Quality Control Echelons to Identified Risks

Identified risk	Quality control echelon	Architectural implementation
Confirmation bias	Systematic falsification	"Skeptic" agent (in "researcher" pipeline)
Algorithmic bias	Independent peer review	"Auditor" agent pipeline
"Black box" opacity	Traceability & audit	Logging, confidence score, HOTL
Low finding reliability	Hierarchical synthesis	Two-tier agent architecture

Discussion: Theoretical and Practical Implications

This blueprint shifts from a "psychometrics of measurement" to a "psychometrics of understanding". While classical psychometrics, facing a validity crisis (Borsboom et al., 2004; Schimmack, 2021), focuses on scores, the proposed system inverts this: a quantitative score projects a contextualized qualitative model.

This entails an ontological shift from viewing personality as a set of static 'traits' to understanding it through dynamic cognitive-affective "repertoires". Such a perspective aligns with established theoretical frameworks, including the CAPS (cognitive-affective processing system) model (Mischel & Shoda, 1995) and the view of traits as state density distributions (Fleeson, 2001). This approach also resonates with the domestic psychological tradition of viewing personality as self-determination (Leontiev, 2011). It allows for transforming diagnostics from a controlling procedure into a developmental tool. The system's ontological grounding is detailed in supplementary materials (Klepikov et al., 2025, S5).

This shift has direct educational implications: it offers students a tool for metacognition (Schneider & Preckel, 2017); educators a basis for personalization (a "map of the ZPD" (Vygotsky, 2005)); and methodologists a research tool for learning analytics (Siemens, 2013).

Limitations and Future Directions

The blueprint requires full-scale validation. A detailed protocol for this is provided in the supplementary materials (Klepikov et al., 2025, S4), outlining a multistage process that includes psychometric testing, didactic evaluation via RCTs, and algorithmic bias audits.

Rubrics create a tension between depth and standardization, risking construct underrepresentation (Messick, 1989). At scale, this challenge is amplified, as cultural adaptation of rubrics becomes a critical question of cross-cultural validity. The system's effectiveness also hinges on human factors, demanding new competencies for educators to interpret its outputs. Technical limitations include the opacity of semantic synthesis nodes, partially mitigated by RAG (Retrieval-Augmented Generation) (Lewis et al., 2020), and the inherent quality-cost tradeoff of the complex multi-agent workflow, a deliberate choice favoring validity.

Conclusion

The solution is not human replacement but integrated, grounded "centaur" systems (Saghafian & Idan, 2024). Success utterly depends on an "architecture of trust": methodological grounding (ECD), hybrid design, echeloned quality control for biases (Bai et al., 2025), role-oriented utility, and data sovereignty. The protocol creation methodology is detailed in supplementary materials (Klepikov et al., 2025, S0).

This shift from measurement to understanding is consistent with computational psychometrics' evolution toward process modeling (von Davier et al., 2022), enabling the construction of verifiable, dynamic personality models. LLM-driven observational diagnostics operationalize this process view by modeling cognitive–affective repertoires rather than static traits, while the ECD mapping to the emulated expertise architecture transforms competency models into testable hypotheses-creating a platform for methodological research that integrates psychometrics, psychology, and machine learning. For educational institutions, incorporating such diagnostics into digital ecosystems facilitates longitudinal analytics: tracing repertoire evolution, measuring intervention effects, and linking process-level indicators to outcomes such as grades or retention-a pathway toward evidence-based quality assurance focused on developmental trajectories.

Such an approach, consonant with ideas of personality as self-determination (Leontiev, 2011), transforms diagnostics into a developmental tool. These tools are elements of a modern educational environment responding to the challenges of digital socialization (Soldatova, 2018) and the future of work (National Academies of Sciences, Engineering, and Medicine, 2022). Thus, psychodiagnostics can become a humanistic instrument, realizing AI's promise to augment, not imitate, humans (Brynjolfsson, 2022).

References

Argyle, L. P., Busby, E. C., Fulda, N., Gubler, J. R., Rytting, C., & Wingate, D. (2023). Out of one, many: Using language models to simulate human samples. *Political Analysis*, *31*(3), 337–351. https://doi.org/10.1017/pan.2023.2

Bai, X., Wang, A., Sucholutsky, I., & Griffiths, T. L. (2025). Explicitly unbiased large language models still form biased associations. *PNAS, 122*(8), Article e2416228122. https://doi.org/10.1073/pnas.2416228122

Baturin, N. A., Vuchetich, E. V., Kostromina, S. N., Kukarkin, B. A., Kupriyanov, E. A., Lurie, E. V., Miting, O. V., Naumenko, A. S., Orel, E. A., & Poletaeva, Yu. S. (2015). Russian standard for personal testing (interim version for a discussion). *Organizatsionnaya Psikhologiya [Organizational Psychology]*, *5*(2), 67–138. (in Russian)

Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., von Arx, S., Bernstein, M. S., Bohg, J., Bosselut, A., Brunskill, E., Brynjolfsson, E., Buch, S., Card, D., Castellon, R., Chatterji, N., Chen, A., Creel, K., Davis, J. Q., Demszky, D., ... Liang, P. (2021). *On the opportunities and risks of foundation models* [Preprint]. arXiv. https://doi.org/10.48550/arXiv.2108.07258

Borsboom, D., Mellenbergh, G. J., & van Heerden, J. (2004). The concept of validity. *Psychological Review*, *111*(4), 1061–1071. https://doi.org/10.1037/0033-295X.111.4.1061

Brynjolfsson, E. (2022). The Turing trap: The promise and peril of human-like artificial intelligence. *Daedalus*, *151*(2), 272–287.

Burrell, J. (2016). How the machine "thinks": Understanding opacity in machine learning algorithms. *Big Data & Society*, *3*(1). https://doi.org/10.1177/2053951715622512

Caliskan, A., Bryson, J. J., & Narayanan, A. (2017). Semantics derived automatically from language corpora contain human-like biases. *Science*, *356*(6334), 183–186. https://doi.org/10.1126/science.aal4230

Carless, D., & Boud, D. (2018). The development of student feedback literacy: Enabling uptake of feedback. *Assessment & Evaluation in Higher Education*, *43*(8), 1315–1325. https://doi.org/10.1080/02602938.2018.1463354

Credé, M., Kuncel, N. R., & Thomas, L. L. (2005). The validity of self-reported grade point averages, class ranks, and test scores: A meta-analysis and review of the literature. *Review of Educational Research*, *75*(1), 63–82. https://doi.org/10.3102/00346543075001063

Cronbach, L. J. (1957). The two disciplines of scientific psychology. *American Psychologist*, *12*(11), 671–684. https://doi.org/10.1037/h0043943

Duckworth, A. L., & Yeager, D. S. (2015). Measurement matters: Assessing personal qualities other than cognitive ability for educational purposes. *Educational Researcher*, *44*(4), 237–251. https://doi.org/10.3102/0013189X15584327

Embretson, S. E., & Reise, S. P. (2013). *Item response theory for psychologists*. Psychology Press.

Federal Agency for Technical Regulation and Metrology. (2015). *GOST R ISO/IEC 25010–2015. Informatsionnye tekhnologii. Sistemnaya i programmnaya inzheneriya. Trebovaniya i otsenka kachestva sistem i programmnogo obespecheniya (SQuaRE)* [Information technology. Systems and software engineering. Requirements and quality assessment of systems and software (SQuaRE)]. Moscow: Standartinform.

Fleeson, W. (2001). Toward a structure- and process-integrated view of personality: Traits as density distributions of states. *Journal of Personality and Social Psychology*, *80*(6), 1011–1027. https://doi.org/10.1037/0022-3514.80.6.1011

Goh, E., Gallo, R., Hom, J., Strong, E., Weng, Y., Kerman, H., Cool, J. A., Kanjee, Z., Parsons, A. S., Ahuja, N., Horvitz, E., Yang, D., Milstein, A., Olson, A. P. J., Rodman, A., & Chen, J. H. (2024).

Large language model influence on diagnostic reasoning: a randomized clinical trial. *JAMA Network Open, 7*(10), Article e2440969. https://doi.org/10.1001/jamanetworkopen.2024.40969

Gurevich, K. M. (1998). *Problemy differentsial'noi psikhologii* [Problems of differential psychology]. Moscow; Voronezh: Institut prakticheskoi psikhologii; NPO "MODEK".

Henrich, J., Heine, S. J., & Norenzayan, A. (2010). The weirdest people in the world? *Behavioral and Brain Sciences*, *33*(2–3), 61–83. https://doi.org/10.1017/S0140525X0999152X

IEEE. (2019). *Ethically aligned design: A vision for prioritizing human well-being with autonomous and intelligent systems*. https://ethicsinaction.ieee.org

Klepikov, O., Pozdnyakova, E., & Sizov, Y. (2025, September 29). *Supplementary materials for "From measurement to understanding: an architecture of emulated expertise for LLM-diagnostics in education"*. Zenodo. https://doi.org/10.5281/zenodo.17450229

Kojima, T., Gu, S. S., Reid, M., Matsuo, Y., & Iwasawa, Y. (2022). Large language models are zero-shot reasoners. In *Advances in Neural Information Processing Systems* (Vol. 35, pp. 22199–22213). https://proceedings.neurips.cc/paper_files/paper/2022/hash/8bb0d291acd4acf06ef112099c16f326-Abstract-Conference.html

Koretz, D. (2017). *The testing charade: Pretending to make schools better*. University of Chicago Press.

Kornilov, S. A., Smirnov, S. D., & Grigorenko, E. L. (2009). Contemporary assessments of the intellectual potential: a cross-cultural adaptation of foreign diagnostic instruments. *Lomonosov Psychology Journal*, *4*, 55–66. (in Russian)

Kosinski, M., Stillwell, D., & Graepel, T. (2013). Private traits and attributes are predictable from digital records of human behavior. *PNAS*, *110*(15), 5802–5805. https://doi.org/10.1073/pnas.1218772110

Lazer, D. (2015). The rise of the social algorithm. *Science*, *348*(6239), 1090–1091. https://doi.org/10.1126/science.aab1422

Leontiev, D. A. (2011). Novye orientiry ponimaniya lichnosti v psikhologii: ot neobkhodimogo k vozmozhnomu [New guidelines for understanding personality in psychology: From the necessary to the possible]. *Voprosy Psikhologii*, *1*, 3–27.

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W., Rocktäschel, T., Riedel, S., & Kiela, D. (2020). Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Advances in Neural Information Processing Systems* (Vol. 33, pp. 9459–9474). https://proceedings.neurips.cc/paper/2020/hash/6b493230205f780e1bc26945df7481e5-Abstract.html

Lin, Z. (2025). *A validity-guided workflow for robust large language model research in psychology* [Preprint]. arXiv. https://doi.org/10.48550/arXiv.2507.04491

Madaan, A., Tandon, N., Gupta, P., Hallinan, K., Gao, L., Wiegreffe, S., Alon, U., Dziri, N., Prabhumoye, S., Yang, Y., Gupta, S., Majumder, B. P., Hermann, K., Welleck, S., Yazdanbakhsh, A., & Clark, P. (2023). *Self-refine: Iterative refinement with self-feedback* [Preprint]. arXiv. https://doi.org/10.48550/arXiv.2303.17651

Matthias, A. (2004). The responsibility gap: Ascribing responsibility for the actions of learning automata. *Ethics and Information Technology*, *6*(3), 175–183. https://doi.org/10.1007/s10676-004-3422-1

McAdams, D. P. (2015). *The art and science of personality development*. The Guilford Press.

Merton, R. K. (1973). The normative structure of science. In N. W. Storer (Ed.), *The sociology of science: Theoretical and empirical investigations* (pp. 267–278). University of Chicago Press.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13–104). American Council on Education; Macmillan.

Mischel, W. (1968). *Personality and assessment*. Wiley.

Mischel, W., & Shoda, Y. (1995). A cognitive-affective system theory of personality: Reconceptualizing situations, dispositions, dynamics, and invariance in personality structure. *Psychological Review*, *102*(2), 246–268. https://doi.org/10.1037/0033-295X.102.2.246

Mislevy, R. J., Steinberg, L. S., & Almond, R. G. (2003). On the structure of educational assessments. *Measurement: Interdisciplinary Research and Perspectives*, *1*(1), 3–62.

National Academies of Sciences, Engineering, and Medicine. (2022). *Artificial intelligence and the future of work*. National Academies Press.

Nisbett, R. E., & Wilson, T. D. (1977). Telling more than we can know: Verbal reports on mental processes. *Psychological Review*, *84*(3), 231–259. https://doi.org/10.1037/0033-295X.84.3.231

Norman, D. A. (1993). *Things that make us smart: Defending human attributes in the age of the machine*. Addison-Wesley.

O'Neil, C. (2016). *Weapons of math destruction: How big data increases inequality and threatens democracy*. Crown.

Parasuraman, R., & Riley, V. (1997). Humans and automation: Use, misuse, disuse, and abuse. *Human Factors*, *39*(2), 230–253. https://doi.org/10.1518/001872097778543886

Paulhus, D. L. (2002). Socially desirable responding: The evolution of a construct. In H. I. Braun, D. N. Jackson, & D. E. Wiley (Eds.), *The role of constructs in psychological and educational measurement* (pp. 49–69). Lawrence Erlbaum Associates.

Pennebaker, J. W. (2011). The secret life of pronouns. *New Scientist, 211*(2828), 42–45.

Rahimi, S., & Shute, V. J. (2023). Stealth assessment: A theoretically grounded and psychometrically sound method to assess, support, and investigate learning in technology-rich environments. *Educational Technology Research and Development*, *72*, 1–25. https://doi.org/10.1007/s11423-023-10232-1

Rudin, C. (2019). Stop explaining black box machine learning models for high-stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, *1*(5), 206–215. https://doi.org/10.1038/s42256-019-0048-x

Saghafian, S., & Idan, L. (2024). *Effective generative AI: The human-algorithm centaur* [Preprint]. arXiv. https://doi.org/10.48550/arXiv.2406.10942

Siemens, G. (2013). Learning analytics: The emergence of a discipline. *American Behavioral Scientist*, *57*(10), 1380–1400. https://doi.org/10.1177/0002764213498851

Schimmack, U. (2021). The validation crisis in psychology. *Meta-Psychology, 5*. https://doi.org/10.15626/MP.2019.1645

Schneider, M., & Preckel, F. (2017). Variables associated with achievement in higher education: A systematic review of meta-analyses. *Psychological Bulletin*, *143*(6), 565–600. https://doi.org/10.1037/bul0000098

Schwarz, N. (1999). Self-reports: How the questions shape the answers. *American Psychologist*, *54*(2), 93–105. https://doi.org/10.1037/0003-066X.54.2.93

Soldatova, G. U. (2018). Digital socialization in the cultural-historical paradigm: A changing child in a changing world. *Sotsial'naya Psikhologiya i Obshchestvo [Social Psychology and Society]*, *9*(3), 8–18. https://doi.org/10.17759/sps.2018090308 (in Russian)

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). Attention is all you need. In I. Guyon et al. (Eds.), *31st Conference on Neural Information Processing Systems (NIPS)*. Curran Associates, Inc. https://arxiv.org/abs/1706.03762

Von Davier, A. A., Mislevy, R. J., & Hao, J. (2022). Introduction to computational psychometrics: towards a principled integration of data science and machine learning techniques into psychometrics.

In A. A. von Davier, R. J. Mislevy, & J. Hao (Eds.), *Computational psychometrics: New methodologies for a new generation of digital learning and assessment: With examples in R and Python* (pp. 1–6). Cham: Springer International Publishing. https://doi.org/10.1007/978-3-030-74394-9_1

Vygotsky, L. S. (2005). *Myshlenie i rech'* [Thinking and speech]. Moscow: Labirint.

Wei, J., Tay, Y., Bommasani, R., Raffel, C., Zoph, B., Borgeaud, S., Yogatama, D., Bosma, M., Zhou, D., Metzler, D., Chi, Ed H., Hashimoto, T., Vinyals, O., Liang, P., Dean, J., & Fedus, W. (2022). *Emergent abilities of large language models.* [Preprint]. arXiv. https://doi.org/10.48550/arXiv.2206.07682

Wiliam, D. (2018). *Embedded formative assessment* (2nd ed.). Solution Tree Press.

Yarkoni, T. (2022). The generalizability crisis. *Behavioral and Brain Sciences*, *45*, Article e1. https://doi.org/10.1017/S0140525X20001685

Yarkoni, T., & Westfall, J. (2017). Choosing prediction over explanation in psychology: Lessons from machine learning. *Perspectives on Psychological Science*, *12*(6), 1100–1122. https://doi.org/10.1177/1745691617693393