Correspondence

J Educ Eval Health Prof 2026;23:33. https://doi.org/10.3352/jeehp.2026.23.33

Validity and fairness of Ukraine's Krok examinations in the national medical qualification system during war

Iurii Kuchyn[1], Lesya Lymar[2], Kateryna Bielka[1], Olha Oseredchuk[3], Anna Blagaia[4*], Iryna Kostiuk[5]

[1]Postgraduate Department of Surgery, Anesthesiology and Intensive Care, Bogomolets National Medical University, Kyiv, Ukraine

[2]Academic CPE Center of Institute of Postgraduate Education, Bogomolets National Medical University, Kyiv, Ukraine

[3]Department of Practical Psychology and Mental Health, Lviv Polytechnic National University, Lviv, Ukraine

[4] Department of Hygiene and Ecology, Bogomolets National Medical University, Kyiv, Ukraine;

[5]Department of Organization and Economics of Pharmacy, Bogomolets National Medical University, Kyiv, Ukraine

Epub ahead of print

***Corresponding e-mail:** anna.blagaia@nmu.ua

Editor: Sun Huh, Hallym University, Korea

Received: September 9, 2026; **Accepted:** October 15, 2026; **Published:** October 26, 2026

This article is available from http://jeehp.org.

© 2026 Korea Health Personnel Licensing Examination Institute.

This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

High-stakes medical examinations fulfill their public-protection function only when their scores support defensible decisions about candidates. Examination difficulty alone does not establish assessment quality. The central questions are whether an observed score represents the intended medical knowledge with sufficient consistency and whether the resulting pass–fail decision is supported by a defensible validity argument [1-3]. These questions are particularly important when examinations take place during prolonged armed conflict.

In Ukraine, Krok 1 assesses fundamental biomedical sciences, and Krok 2 assesses professionally oriented clinical disciplines; both are standardized, test-based components of the Unified State Qualification Examination. The current official minimum passing score is 64% for both examinations [4]. Since Russia's full-scale invasion on February 24, 2022, candidates have prepared for and taken these examinations amid air-raid alerts, displacement, disrupted sleep, infrastructure damage, power outages, and unequal access to clinical education. The national Krok examination system has continued to operate despite these disruptions. Its continuation during prolonged armed conflict raises a broader question: under what conditions can high-stakes examination results remain valid, interpretable, and fair?

Wartime influences on assessment should not be subsumed under a single, undifferentiated concept of "stress." We propose distinguishing three related but conceptually distinct dimensions: the competence expected of candidates, their opportunity to acquire it, and the conditions under which they demonstrate it. The first dimension concerns the knowledge and performance required for safe professional practice. War does not, in itself, make essential biomedical or clinical knowledge less necessary. The second concerns candidates' access to teaching, clinical placements, supervision, electricity, digital resources, and protected learning time. Although Ukrainian medical schools have sustained education through remote and hybrid approaches, substantial educational and psychological burdens have persisted [5,6]. The third concerns circumstances immediately before and during the examination, when acute events may affect candidates' ability to demonstrate acquired knowledge under comparable conditions.

These distinctions matter because each dimension requires a different response. Prolonged disruption of teaching and clinical exposure may leave genuine gaps in acquired knowledge by reducing candidates' opportunities to learn. Addressing such educational loss primarily requires restoring access to essential learning experiences, identifying affected learners, providing targeted remediation, and monitoring educational outcomes. Acute examination-day disruption presents a different problem. Meta-analytic evidence indicates that acute stress can impair working memory and cognitive flexibility, particularly during cognitively demanding tasks [7]. Direct evidence linking wartime examination interruptions to Krok performance remains limited; nevertheless, experimental evidence on acute stress suggests a plausible mechanism by which severe examination-day disruption could introduce construct-irrelevant variance. An air-raid alert, evacuation, nearby attack, acute illness, power failure, or major technical interruption could therefore plausibly affect an observed score for reasons unrelated to stable medical knowledge. The appropriate response in these circumstances is to protect measurement integrity, rather than lower the

Epub ahead of print

knowledge standard assessed by Krok, through predefined procedures for interruption, documentation, preservation of testing time, resumption under equivalent conditions, or rescheduling. Table 1 summarizes these pathways, their implications for score interpretation, and the proportionate assessment responses.

Table 1. A framework for interpreting threats to high-stakes assessment during prolonged crisis

Challenge	Primary locus of validity threat	**Potential assessment threat**	**Appropriate response**
Ordinary examination anxiety	Conditions of demonstration	Usually limited evidence that the score is invalid	Preparation, practice, mentoring, psychological support
Chronic wartime educational disruption	Opportunity to acquire competence	Unequal opportunity to learn	Restore learning opportunities, provide remediation and targeted support, and monitor subgroup outcomes
Acute attack, air raid, or evacuation during the examination	Examination administration and conditions for demonstrating competence	Construct-irrelevant variance	Interrupt testing, preserve testing time, and resume under equivalent conditions or reschedule according to predefined rules
Poor item quality/translation	Assessment instrument and scoring	Construct-irrelevant variance or construct underrepresentation	Blueprinting, expert review, item analysis
Inadequately justified pass standard	Standard setting and classification	Misclassification around the pass threshold	Evidence-based standard setting and periodic review
Reliance on Krok alone	Interpretation, extrapolation, and construct coverage	Construct underrepresentation	Combine Krok with objective structured clinical examinations, simulation, workplace-based assessment, and longitudinal assessment

Epub ahead of print

This distinction also prevents ordinary examination anxiety from being conflated with threats to assessment validity. Distress is relevant to candidate well-being and warrants appropriate support, but anxiety alone does not establish that an examination score is invalid. Conversely, the need for medical professionals to function under stress does not justify avoidable construct-irrelevant influences during a licensing examination. The question is whether examination conditions introduce score variation unrelated to the intended construct, rather than simply whether candidates experience stress.

The limited published evidence on Krok during the full-scale invasion illustrates the importance of these distinctions. The most directly relevant peer-reviewed study compared aggregate Krok 1 outcomes in medicine and pediatrics from 2019 to 2022. Selected national indicators did not deteriorate uniformly, and several pediatric outcomes improved in 2022 [8]. These findings do not support a simple account of system-wide performance collapse after the invasion. Aggregate outcomes, however, cannot rule out effects

on individuals or subgroups. They cannot establish links between examination performance and displacement, psychological symptoms, interrupted clinical education, proximity to attacks, or acute examination-day events, nor can they show whether effects varied by region, language, testing location, or other candidate characteristics. The evidence thus supports neither the automatic rejection of wartime Krok scores nor the assumption that stable aggregate outcomes demonstrate individual fairness.

Beyond candidates' learning opportunities and conditions for demonstrating competence, the assessment system must safeguard each inference from item response to professional decision. Examination content should remain aligned with the intended blueprint and clinically relevant knowledge. Scoring should be protected against ambiguous wording, translation problems, poorly functioning items, technical errors, and irregular administration. Test forms should yield reliable, comparable evidence. Interpretation must also respect the examination's limits: a standardized written examination can provide valuable evidence of medical knowledge but cannot, by itself, represent the full range of professional competence. Krok results should therefore inform high-stakes decisions within a broader assessment program that includes, where appropriate, objective structured clinical examinations, simulation, workplace-based assessment, supervised clinical performance, and longitudinal evidence [9].

Fairness in this context requires more than nominally identical procedures for every candidate. Identical procedures do not necessarily yield equivalent measurement when one group completes an examination without interruption while another loses testing time because of an air raid, evacuation, or infrastructure failure. Conversely, unequal educational opportunities before an examination do not automatically mean that it inaccurately measures candidates' current knowledge. Assessing fairness therefore requires distinguishing inequity in opportunities to acquire competence from construct-irrelevant influences on its demonstration. Here, fairness is considered primarily in terms of the defensibility and comparability of score interpretation. Broader questions of educational and distributive equity are also important, but cannot be reduced to psychometric validity. The objective should be equivalent score interpretation, rather than procedural uniformity for its own sake.

The same distinction applies to competence standards and numerical pass thresholds. The knowledge construct assessed, the classification cut score, and the difficulty of a particular test form are related but not interchangeable. The official 64% threshold is a decision rule [4]; its defensibility depends on the definition of minimally acceptable performance on the knowledge construct, the method used to establish the threshold, and the comparability of evidence across examination forms. Lowering the threshold solely because war persists would risk conflating the minimum acceptable performance standard for the knowledge assessed by Krok with the circumstances in which candidates learned or were assessed. The Krok pass standard should not be treated as a stand-alone threshold for overall professional competence or safe clinical practice. Yet retaining the same percentage does not, by itself, demonstrate fairness if blueprint coverage, item functioning, form difficulty, or administration conditions differ. For candidates whose scores are close to the pass threshold, transparent standard setting, evidence of form

Epub ahead of print

comparability, consideration of measurement error, and attention to classification consistency are particularly important.

Wartime contingency procedures should likewise be established before an incident occurs, rather than replaced by retrospective leniency. Ukraine has already introduced important wartime adaptations to examination delivery, including requirements for safe testing conditions and access to emergency medical care [10]. The question is therefore whether procedures for substantial interruptions are sufficiently predefined, documented, consistently applied, and able to preserve comparable score interpretation. Discretionary score adjustments made after an interrupted examination without a prespecified psychometric rationale would weaken comparability and obscure the resulting score's interpretation. A documented air raid, evacuation, power failure, acute medical event, or substantial loss of testing time should instead trigger a predefined procedure specifying when testing is paused, how remaining time is preserved, whether equivalent conditions can be restored, when rescheduling is required, and how the incident is documented. These adaptations do not weaken the knowledge standard assessed by Krok; they protect the conditions needed for the score to represent that standard.

Psychometric and procedural transparency is therefore an important component of assessment that can withstand crises. Examination authorities can preserve item security while reporting blueprint coverage, reliability, form comparability, standard-setting procedures, classification consistency, item review, and relevant subgroup performance. Differential item functioning analysis may identify items that behave differently across candidate groups after overall ability is taken into account [11], although psychometric analysis alone cannot establish whether learning opportunities or examination-day conditions were comparable. Incident reports could also document air-raid alerts, evacuations, infrastructure failures, resumed sessions, and rescheduled examinations. Clear procedures for reviewing scores, addressing complaints about item content, and hearing appeals concerning examination administration would further strengthen confidence in high-stakes decisions.

These implications extend beyond Ukraine. Armed conflict, natural disasters, epidemics, mass displacement, cyberattacks, and prolonged infrastructure failures can all create a gap between nominal standardization and genuine comparability. Systems designed only for stable conditions may place the burden of resilience on individual candidates precisely when external circumstances are least controllable. Resilience should instead be built into assessment systems through alternative testing arrangements, secure contingency procedures, incident documentation, psychometric monitoring, and complementary evidence of competence.

Ukraine's experience challenges a common assumption about standardization in high-stakes assessment. Standardization should protect comparability of score interpretation, rather than impose rigid procedural uniformity when external circumstances make equivalent measurement impossible. Medical knowledge can be assessed during war, but defensible decisions require a clear distinction among the competence expected of candidates, their opportunity to acquire it, and the conditions under which they

Epub ahead of print

demonstrate it. Krok can provide meaningful standardized evidence of medical knowledge, but should not be treated as a complete representation of professional competence. Maintaining clinically justified standards and adapting assessment procedures are compatible objectives. During a prolonged crisis, appropriate adaptation may be necessary to preserve validity, fairness, and public confidence in high-stakes assessment.

ORCID

Iurii Kuchyn: https://orcid.org/0000-0002-9667-1911

Lesya Lymar: https://orcid.org/0000-0002-9407-1066

Kateryna Bielka: https://orcid.org/0000-0003-1185-6835

Olha Oseredchuk: https://orcid.org/0000-0002-3192-3129

Anna Blagaia: https://orcid.org/0000-0002-2451-9689

Iryna Kostiuk: https://orcid.org/0000-0002-3689-3379

Authors' contributions

Conceptualization: LL (ideas and formulation or evolution of overarching research goals and aims).
Methodology: KB, OO, AB, IKo (development and refinement of the conceptual framework).
Project administration: IKu.
Writing – original draft: LL.
Writing – review & editing: LL, IKu, KB, OO, AB, IKo

Conflict of interest

No potential conflict of interest relevant to this article was reported.

Funding

None.

Data availability

Not applicable.

Acknowledgments

None.

Supplementary materials

None.

Epub ahead of print

References

1. Downing SM. Validity: on the meaningful interpretation of assessment data. Med Educ. 2003;37:830-837. https://doi.org/10.1046/j.1365-2923.2003.01594.x

2. Cook DA, Brydges R, Ginsburg S, Hatala R. A contemporary approach to validity arguments: a practical guide to Kane's framework. Med Educ. 2015;49:560-575. https://doi.org/10.1111/medu.12678

3. Norcini J, Anderson MB, Bollela V, Burch V, Costa MJ, Duvivier R, Hays R, Palacios Mackay MF, Roberts T, Swanson D. 2018 consensus framework for good assessment. Med Teach. 2018;40:1102-1109. https://doi.org/10.1080/0142159X.2018.1500016

4. State Non-Profit Center of Education Quality of the Ministry of Health of Ukraine. Passing criterion [Internet]. Kyiv: State Non-Profit Center of Education Quality of the Ministry of Health of Ukraine; [cited 2026 Aug 19]. Available from: https://www.testcentr.org.ua/uk/ispyty/kriteriji-sklav-ne-sklav.html

5. Mayer A, Yaremko O, Shchudrova T, Korotun O, Dospil K, Hege I. Medical education in times of war: a mixed-methods needs analysis at Ukrainian medical schools. BMC Med Educ. 2023;23:804. https://doi.org/10.1186/s12909-023-04768-2

6. Korda M, Shulhai A, Shevchuk O, Shulhai O, Shulhai AM. Psychological well-being and academic performance of Ukrainian medical students under the burden of war: a cross-sectional study. Front Public Health. 2025;12:1457026. https://doi.org/10.3389/fpubh.2024.1457026

7. Shields GS, Sazma MA, Yonelinas AP. The effects of acute stress on core executive functions: a meta-analysis and comparison with cortisol. Neurosci Biobehav Rev. 2016;68:651-668. https://doi.org/10.1016/j.neubiorev.2016.06.038

8. Obernikhina NV, Yanitska LV, Vygovska OV. The unified state qualification exam Step 1 as a marker of the success of the education of pediatric doctors in Ukraine and Bogomolets National Medical University. BMC Med Educ. 2024;24:263. https://doi.org/10.1186/s12909-024-05261-0

9. Heeneman S, de Jong LH, Dawson LJ, Wilkinson TJ, Ryan A, Tait GR, Rice N, Torre D, Freeman A, van der Vleuten CPM. Ottawa 2020 consensus statement for programmatic assessment—1. Agreement on the principles. Med Teach. 2021;43:1139-1148. https://doi.org/10.1080/0142159X.2021.1957088

10. State Non-Profit Center of Education Quality of the Ministry of Health of Ukraine. Procedure for conducting the Unified State Qualification Examination [Internet]. Kyiv: State Non-Profit Center of Education Quality of the Ministry of Health of Ukraine; [cited 2026 Sep 28]. Available from: https://www.testcentr.org.ua/uk/component/tbpb/page-587.html

11. Tavakol M, Stewart C, Sharpe CC. Ensuring fairness in assessment in health professions education: rapid analysis tools to detect differential item functioning across groups. Int J Med Educ. 2024;15:80-83. https://doi.org/10.5116/ijme.6694.de69

Epub ahead of print