Psychological and pedagogical aspects of objectivity in assessing productive speech activities in large-scale standardized testing

Iryna Vasylivna Boichuk [1]

Received: 2025-09-10
Accepted: 2025-10-22
DOI: https://doi.org/10.5281/zenodo.19765800

Abstract. The study introduces a customized methodological model for evaluating the professional effectiveness of foreign language teachers. The model is connected to the dynamics of students' academic achievement and teachers' professional effectiveness. The findings from the literature review and presented in the paper support the notion that Value-Added Modeling (VAM) combined with communicative-oriented assessment practice forms a valid, rigorous, and context-sensitive method of measuring teacher results. Research evidence supports this fact, as standardized examination scores increased from 156 to 178 points. This is evidence that communicative and formative assessment strategies may not only lead to improved academic performance, but also greater consistency in learning outcomes. Through blending quantitative elements of student progress with qualitative aspects of instructional practice, the suggested model overcomes the main constraints in conventional approaches to evaluation. It also conforms to current models of instructional practice rooted in competency-based and data-driven teaching and learning. This paradigm-making process has implications for policymakers and schools that want to better account for their policies that affect accountability, transparency, and instructional quality in the context of ongoing educational reforms. In conclusion, the research enhances the literature on teacher effectiveness and is theoretically and methodologically relevant to teachers; and lays the groundwork for further refinement and application of VAM-based approaches to teaching in a variety of contexts. The proposed framework may also serve as a methodological basis for improving examiner training and strengthening quality assurance in large-scale language assessment. In this regard, the study highlights practical pathways for enhancing fairness and consistency in productive speech evaluation without diminishing the professional role of expert raters.

Key words: rater bias; Psychological determinants; Pedagogical regulation; Productive speech assessment; Many-Facet Rasch Model; CEFR descriptors; High-stakes testing; Ukrainian standardized assessment.

[1] English Teacher of the Highest Category, Methodologist, Researcher at Lesia Ukrainka Lyceum No. 75 of the Lviv City Council,
ORCID ID: 0009-0008-2072-1135

Introduction

Standardized tests play a major role in education today because tests evaluate what students learn, making it more challenging for students to develop skills that enhance their social advancement. In Ukraine, the External Independent Evaluation (ZNO) has been identified as a high-stakes assessment, requiring very careful consideration to ensure that validity, reliability, and fairness are maintained. Although the ZNO uses an objective test format to enable automatic scoring procedures for multiple choice tests, the evaluation of productive speech activities—writing and speaking—continues to require human-mediated assessment, in other words rater mediation.

There are several qualitative differences between evaluating productive language performance and evaluating receptive language performance. Productive language performance evaluations involve raters' interpretations of complex linguistic output, application of multi-dimensional criteria, and the need for qualitative judgments that must be made under time pressure.

Consequently, while there are certainly technical aspects to the process of scoring, the process also contains psychological components. When raters interpret descriptors, apply priorities to criteria, manage their cognitive load, and make decisions about the performance of students, they do so using internal cognitive frameworks that will inherently affect their decision-making processes.

Contemporary research shows that rater severity, halo effect, variation in how raters interpret criteria, and changes in rater criteria over time are not anomalies or chance occurrences but rather structural elements of performance-based assessments [10; 12; 19]. However, the objectivity of assessments cannot be limited to just the psychological variance of raters. Pedagogy can play a critical role in establishing the degree to which scoring variability is stabilized or exacerbated.

Rating descriptor clarity, scale alignment with CEFR performance levels, the development of assessment literacy, and rater training are all factors that contribute to whether raters' subjective variability will be minimized or maximized [4; 11; 16]. Thus, the objectivity of assessments of productive language skills is constructed not as a natural attribute of the scoring process but as a structured pedagogical construction.

The Ukrainian context illustrates this dualism. Studies in assessment literacy show teachers' inconsistent preparedness and varying degrees of methodological readiness for their assessment tasks [16; 17]. Reviews of national proficiency testing suggest developing systemic alignments and clear scoring guidelines [13]. These findings illustrate that psychological and pedagogical determinants are interdependent within an institutional framework.

Although significant advancements in psychometric modeling have been developed, specifically through the use of the Many-Facet Rasch Model and Generalizability Theory [5; 18], the theoretical connection between the psychological and pedagogical dimensions of objectivity has not been well articulated. Measurement models quantify variability; however, they do not provide the means for creating the pedagogical conditions required to minimize it.

Therefore, this article develops a conceptual framework that connects the psychological influences of rater behavior with the pedagogical influences of assessment design in the context of large-scale standardized testing. This article draws upon the current body of literature and utilizes the author's experience as an expert examiner to demonstrate that objectivity in assessing productive speech is the result of both intentional psychological stabilization and pedagogical design within standardized testing systems.

Literature review

2.1 Psychological Determinants of Rater-Mediated Assessment

In contrast to selected response items, the human factor in grading writing and speaking assessments results in an interpretative judgment being required for the scoring of

those assessments. Research has consistently demonstrated that the interpretative judgment of human raters in assessing writing and/or speaking is subject to both cognitive and affective mechanisms that guide rater behavior. One of the most well-documented psychological phenomena that has been demonstrated to impact rater behavior is the halo effect [10]. Using the many-facet rasch model, Kim (2020) demonstrated that how the rating criteria are presented to the rater impacts how they will score the candidates. When raters receive either a strong or weak feature first, that first impression may have an unconscious impact on their subsequent judgments [10]. As such, the halo effect illustrates that the scoring decision of a rater is not isolated but is cognitively linked to other features assessed.

Another aspect of the psychological stability of rater behavior is related to the concept of severity and leniency variability. Mohd Noh and Mohd Matore (2022) showed that raters are severity biased in a systematic way depending upon their field, training, and teaching experience [12]. These biases are not random but are systematic. Likewise, Khodi (2021) utilized generalizability theory to demonstrate that rater variance is one of the primary sources of measurement error in the assessment of student writing. The results indicated that raters contributed independently to the variability of scores independent of the task and candidate effects [9].

Jia and Zhang (2023) studied raters' thinking in integrated writing tasks and found that raters engaged in problem solving and created mental models of performance quality that do not simply involve comparing answers to a rubric. These findings support the idea that rating is an active, interpretative process that is influenced by mental heuristics [8].

On a related note, Heidari et al. (2022) built on this by focusing on how raters perceive the criteria for the different levels on the scoring scale [6]. The findings showed that raters considered the scale criteria differently, which led to different scoring results. Even though the rating scales used were the same for all raters, there was still some psychological uncertainty regarding how the criteria for ratings were interpreted.

Sureeyatanapas et al. (2024) further researched the topic of marking reliability by utilizing gauge repeatability and reproducibility analysis to investigate speaking assessment [15]. Their findings confirmed that variability in rater judgments can be statistically decomposed and quantified. Additionally, Rezai et al. (2022) noted that potential demographic biases in assessment fairness that supported the idea that subjective influences occur at multiple levels of the assessment process [14].

As a whole, these investigations support the idea that rater-mediated assessments involve a degree of psychology. Raters' variability arises from their cognitive processing, flexible interpretations of their role in the assessment, and their individual behavioral tendencies [7]. As such, the understanding of objectivity cannot be defined as the removal of human variables; it can only be defined as controlling the level of psychological variability present.

2 Measurement Models as Tools for Regulating Subjectivity

Psychological variables introduce variability; however, current measurement models offer tools to measure and manage this variability. The Many-Facet Rasch Model (MFRM) is now being used to model the impact of raters on assessments [5].

Govindasamy et al. (2019) illustrated the ability to separate and measure the component variances due to raters, tasks and candidates using MFRM. The model estimates the severity parameters for raters and, therefore, identifies raters who are consistently more severe or lenient than other raters [5]. Such a bias might be safer to measure than other biases, which tend to be more hypothetical.

Uto (2021) built on the use of MFRM with linked performance assessments, demonstrating how the various facets affected score comparability [18]. In a later publication, Uto (2023) used Bayesian methods to identify shifts in the severity of raters over different periods [19]. It is in large scale assessment situations (i.e., large scale, long-term standardized

assessments) that shifts in rater severity commonly occur, as raters tend to engage in different behaviors over time. Incorporating Neural Automated Essay Scoring with Many-Facet Modeling, Uto and Aramaki (2024) also made an addition to the MFRM framework which could be a starting point in the development of hybrid systems for rater monitoring and scoring systems [20].

Moreover, G-Theory gave clues to teachers on how to divide variability into subcomponents of evaluators, tasks, and scoring techniques, and how to systematically hone in on which of these elements are responsible for the greatest degree of errors [9]. Additionally, Burchell et al. (2022) illustrated that the use of structured assessment frameworks along with ongoing supervision of rater activity improved the consistency of assessment results [2].

Dong et al. (2022) noted in their bibliometric analysis of research trends in language testing that there was growing interest in assessing performance, rater effects and the ways to reduce subjectivity in testing, indicating a larger disciplinary movement toward reducing subjectivity in assessments [3]. However, although statistical models can quantify variability, they cannot eliminate it. Therefore, statistical models must be situated in a larger pedagogical framework.

2.3 Pedagogical Determinants of Assessment Objectivity

The Pedagogical structure attempts to limit or control the amount of subjective judgement from assessors in the rater-mediated evaluation process. Effective assessment literacy may be complemented by preventive approaches such as thorough development of scoring rubrics, straightforward scoring rubric descriptors, and systematic assessment literacy development.

The systematic construction and validation of analytic rating scales have resulted in increased scoring consistency if developed and validated through the analytical method [4]. The use of well-defined constructs mitigates ambiguity and leads to the clear and unambiguous interpretation of the rating scales. Likewise, Li and Wang (2021) found that rating scales that fit the assessed constructs improve reliability in a construct integrated assessment [11].

Heidari et al. (2022) further established that the way raters view the criteria targeted by the rating scale influences their scoring [6]. This also emphasizes that descriptors should be unambiguous and clear to the raters to achieve the intended purpose.

Ukrainian researchers have found that while many pre-service teachers develop assessment literacy in Ukraine, it is largely due to the inclusion of assessment training into university programs. Ukrayinska (2024) studied the development of assessment literacy in Ukrainian language assessment and found that although the majority of pre-service teachers develop some level of assessment literacy, the quality and quantity of the training varies significantly [16]. Systematic inclusion of assessment training could lead to consistent assessment competence. The research also showed that teachers' assessment practices are a function of the pedagogical culture in which they taught; assessment practice is contextualized in the wider educational philosophy and traditions.

Finally, Ozernyi and Suvorov (2023) reviewed language proficiency testing in Ukraine and highlighted the need to provide transparency in the assessment tasks being aligned and described at the CEFR performance level as well as in the provision of marking guides [13]. The authors are in favor of establishing national system-wide guidelines for test development and implementation. In total, the results of these studies indicate that the educational design constrains the level of psychological variability, leaving room for the assessor's subjective judgment.

Well-defined descriptors aligned to the CEFR performance levels serve as a reference point (or cognitive anchor) for the assessor. Greater inclusion of training ultimately raises

assessment literacy [7]. Finally, organized calibration sessions within an institution are designed to reduce the degree of variability in assessment interpretation by the assessor.

Therefore, objectivity does not arise out of the complete elimination of subjectivity but rather from the pedagogical regulation of its manifestation.

2.4 Integrative Perspective: The Need for Conceptual Synthesis

The reviewed literature reveals three interconnected domains (See Table 1):

1. Psychological variability in rater behavior,
2. Measurement-based quantification of variability,
3. Pedagogical structuring aimed at minimizing interpretative divergence.

Table 1. Theoretical synthesis of determinants influencing objectivity in productive speech assessment.

Dimension	Key Determinants of Variability / Regulation	Representative Sources
Psychological variability	Rater severity and leniency; halo effect; criterion interpretation variability; cognitive heuristics; rater drift over time	Kim (2020); Mohd Noh & Mohd Matore (2022); Heidari et al. (2022); Jia & Zhang (2023); Uto (2023)
Measurement Monitoring	Many-Facet Rasch modeling; Generalizability Theory; severity index estimation; variance partitioning; drift detection mechanisms	Govindasamy et al. (2019); Khodi (2021); Uto (2021, 2023); Uto & Aramaki (2024); Burchell et al. (2022)
Pedagogical Structuring	Descriptor clarity; CEFR alignment; analytic rubric design; calibration procedures; development of assessment literacy	Ghanbari & Barati (2020); Li & Wang (2021); Ukrayinska (2024, 2025)

While psychological, measurement, and pedagogical research into objectivity may have some overlap, they can also tend to operate independently. For example, psychological researchers may investigate bias while measurement researchers may develop models for assessing test performance, and pedagogical researchers will primarily explore how to train teachers to use those tests effectively. Therefore, there has been limited work to date developing a complete conceptual model for objectivity.

An important reason why this is so, is that, in Ukraine, with its numerous large scale standardized assessments, which carry heavy consequences for students, educators, and schools (i.e., productive speech activities), a complete conceptual model of objectivity is particularly important. In addition, because objective scores on standardized assessments need to be seen as a result of both psychological characteristics, pedagogical features of the test instrument, and measurement monitoring of the assessment process, it is necessary to formulate a theoretical model that captures this interplay.

As a result, this study introduces a theoretical model for objectivity as an intended product of purposeful interactions among psychological, pedagogical, and measurement

factors. It does not aim at eliminating subjectivity, but at structuring and regulating it through standardized assessment processes.

Methodology

The current research utilizes a theoretical-analytic research strategy for the development of the theoretical and pedagogical parameters of objective evaluation of students' productive speech abilities as well as other productive speech abilities in the context of large-scale standardized assessments. This research is based on an analytical methodology of conceptual integration as opposed to empirical collection of data. Thus, its primary function is to create an integrated model using established theoretical models and contemporary scholarly ideas that can be used to explain and regulate the subjective nature of raters when evaluating students' productive speech abilities.

Methodologically, this research was conducted through three related phases.

1 Theoretical Content Analysis of Scholarly Discourse

The initial stage entailed a systematic theoretical content analysis of recent research regarding the assessment of language in general, with special consideration for research that has investigated rater cognition, severity and leniency in scoring, rater interpretations of rating scales, models of measurement, and the development of assessment literacy [6; 9; 10; 12; 19]. In contrast to aggregating empirically-derived findings through quantitative methods, the content analysis examined the conceptual patterns and recurring explanatory constructs in the literature as a whole.

In terms of content, the primary analytical categories were derived inductively from the literature and included:

- cognitive processes underlying raters' scoring behaviors,
- structural factors leading to inter-rater variation,
- ambiguities in rater interpretation resulting from descriptors on rating scales,
- psychometric approaches to decomposing the variance associated with raters' scores,
- pedagogical techniques for improving consistency in assessment practices.

By comparing the categories identified above, theoretical intersections between psychological aspects of assessing language and pedagogical aspects of assessment were identified.

2 Conceptual Modeling Through Psychometric Frameworks

The second phase included the use of theoretical models for the evaluation of rater-mediated measurement error by employing established theories on measurement —the Many-Facet Rasch Model (MFRM) and Generalizability Theory (G-theory) —as conceptual tools [5; 18]. Although these two theories have been used in empirical studies, they were utilized here to represent theoretically how rater-induced variability may be structured, identified and monitored.

A model-based approach enabled the theoretical construction of objectivity as an interactive process among three facets: Candidate Performance, Task Characteristics, and Rater Behavior. This model-based approach especially addressed theoretical mechanisms of severity drift [19] and the interactions between facets that illustrate the non-stationary nature of stable scores.

3 Reflective Professional Integration

The third is a methodological component that is represented by the use of structured professional reflection based on the author's experience as an examiner for large-scale standardized English assessments. It provides a means for bridging the gap between the theoretical concepts presented and the practical application of those concepts in assessment settings.

A second way in which the author utilized their experience in order to inform their evaluation of the theoretical models discussed was to reflect upon how well they could be applied to actual scoring process; specifically regarding calibration techniques, descriptors' interpretations, and the various constraints associated with each institution. As such, this type of reflective integration exemplifies the tradition of theoretically-informed professional inquiry in which practitioners' insights are able to enhance conceptual clarity.

4 Methodological Logic of the Study

The methodological architecture of the study is presented via a triadic interaction model between:

1. Psychological determinants (rater cognition and bias),
2. Pedagogical structuring mechanisms (descriptor design and training),
3. Measurement-based monitoring (variance modeling and drift detection).

Through the integration of these components, the study develops a theoretically grounded model of engineered objectivity in the assessment of productive speech activities.

Results and Discussion

Beginning with an examination of the theoretical-analytical integration (synthesis) developed as part of this research, it has been found that the objective evaluation of productive speech activities on a large scale during standardized testing, is a product of the engineering of structure through the interaction of the three domains of psychological determinants, pedagogical structural elements, and measurement-based monitoring systems. The results have been organized at three levels of conceptualization.

1 Psychological Determinants of Scoring Variability

There have been many studies on the structural nature of psychological variability within the context of rater-mediated assessments, and the findings suggest that subjectivity is a fundamental aspect of human judgment as opposed to simply a flaw in human judgment. Research has shown that rater severity and leniency can be viewed as consistent behavioral characteristics and not as random deviations from some ideal or normative level [12]. Furthermore, research using Generalizability Theory [9] demonstrates that raters create an additional source of measurement error as part of the assessment process, which suggests that psychological variability needs to be understood as a structural component of performance assessments.

Additionally, research has identified several cognitive mechanisms that also contribute to the increased variability associated with rater-based assessments. For example, Kim (2020) demonstrated the halo effect as a mechanism through which initial impressions are used to inform subsequent evaluations of the criterion [10]. While raters claim to evaluate each criterion independently, in practice, they tend to form a global impression based on their experience of the overall performance of the candidate. This global impression will then likely influence their analytic judgments for each criterion. Additionally, Jia and Zhang (2023) demonstrated that raters are actively engaged in problem-solving during the assessment process and are not merely mechanically comparing the candidate's performance to the descriptors listed on the rubric [8]. As such, the interpretive act of evaluating a candidate's performance against the rubric introduces a variety of cognitive heuristics and internal weightings into the assessment process.

Furthermore, descriptor interpretation variability is another psychological variable that contributes to increased variability in the ratings assigned by raters. Research by Heidari et al. (2022) indicated that while raters have formalized rubrics with standardized descriptors, their interpretations of those descriptors differed. These findings suggest that rubrics serve as interpretative prompts as opposed to mechanical scoring templates [6].

Finally, the psychological instability caused by drift over time affects the evaluation process. Uto (2021) showed that rater severity is not a constant, but it changes over time

depending on how long a rater has been evaluating candidates [18]. This type of drift is particularly relevant when evaluating large-scale standardized tests, since the test takers have a long time to work on the tests.

These findings seem to suggest that for the purpose of these evaluation processes, the concept of objectivity should not be limited to the complete absence of subjective influence. Rather, objectivity should be defined as controlled variability, as it is impossible to eliminate all subjective influences on an assessment.

2 Pedagogical Structuring as a Stabilizing Mechanism

Pedagogical structuring is an important method to control subjectivity in scoring. As psychological variables can be unstable, pedagogical design can provide methods to stabilize scores.

Rating scale development is one of the most important stabilizers. For example, Ghanbari and Barati (2020) found that using analytically developed scales (which are validated), improves the stability of scoring (by reducing ambiguities) [4]. Likewise, Li and Wang (2021) indicated that using rating scales based on clearly identified constructs, will improve the reliability of scores when students complete an integrated assessment task [11].

According to the research literature, descriptors, which are aligned with CEFR bands serve as cognitive anchors to help clarify what students' scores mean. Performance descriptors have been shown to influence raters' subjective assessments of student responses. Furthermore, Heidari et al. (2022) noted that raters' perceptions of criteria influence the scores given to students. Hence, the descriptors have to be clearer [6].

Two other important factors that influence teachers' assessment practices are pedagogical culture and pedagocial literacy. In particular, Ukrayinska (2024) states that systematic training of pre-service teachers, is preparation for assessment practice that is equitable and draws on structure [16]. Furthermore, Ukrayinska (2025) demonstrated that pedagogical culture has an effect on how assessment practices occur and how criteria are measured; indicating that teachers' scoring behaviors are influenced by the broader educational tradition [17].

Therefore, in the context of standardized testing in Ukraine, this means that objective scores depend on coordinated pedagogical structure at the institutional level. Without systematic calibration, common understanding of descriptors, or ongoing professional development, the psychological variability of ratings increases instead of being controlled.

Therefore, pedagogical structuring functions as a prevention mechanism by limiting uncontrolled subjectivity through specified criteria, calibration processes, and structured training of teachers.

3 Measurement-Based Monitoring and Quantification of Variability

The third conceptual layer deals with the use of psychometric modeling to ensure objectivity. In addition to pedagogically-based design providing stability in the way that students' performances are interpreted, measurement models provide ways to measure performance variability.

The Many-Facet Rasch Model (MFRM) measures rater severity as a separate parameter from the other facets [5]. Thus, instead of regarding subjectivity as an invisible variable, the MFRM provides a structure for measuring it statistically. Facet interaction influences both linking and comparability of scores [18], therefore emphasizing the necessity of systematic modeling in large-scale environments. Also, Uto (2023) demonstrates that Bayesian modeling can identify rater severity drifts, which can be used to identify potential instability in scoring [19].

Thus, integrating automated scoring with many-facet models may create new hybrid monitoring opportunities that will add to the level of objectivity available while still retaining some level of human judgment [20]. Structured monitoring protocols were shown to greatly increase the reliability of discourse assessments [2]. Increasing scholarly attention has been

focused on rater effects by Dong et al. (2022), thus establishing that the field acknowledges subjectivity as a significant concern [3].

The conceptual outcome of this synthesis is that, while measurement modeling is not a substitute for psychological or pedagogical interventions, it is a complementary approach to those interventions. Monitoring systems serve as corrective systems to identify deviations that are inevitable due to the limitations of pedagogically based design.

4 Toward a Model of Engineered Objectivity

Using all three of these areas creates the main idea of the study that shows how to engineer objectivity for assessing productive speech activity by the interplay between psychological regulation, pedagogical structuring and measurement monitoring.

As shown in Figure 1, the conceptual framework illustrates a triadic interaction model with the following elements:

- Psychological variability (inherent),
- Pedagogical structuring (preventative) and,
- Measurement monitoring (corrective).

In this model, objectivity has been redefined as the equilibrium of controlled influence of the abovementioned interactive elements.

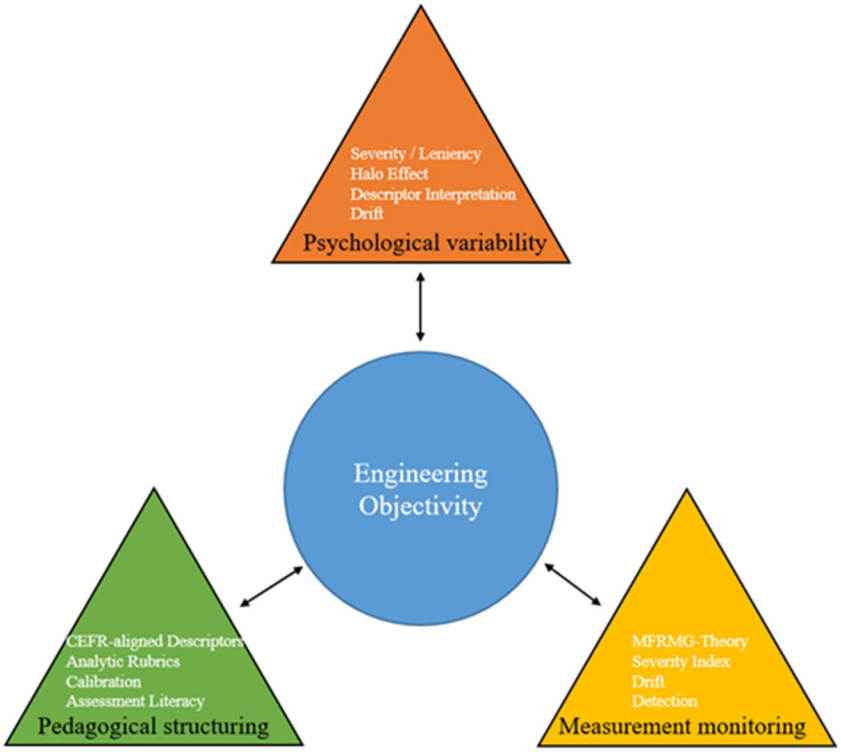

Figure 1. Conceptual triadic model illustrating objectivity as the engineered interaction of psychological, pedagogical, and measurement determinants.

The findings suggest that objectivity in large-scale standardized testing is not an intrinsic characteristic of scoring procedures. It is the outcome of deliberate system design aimed at stabilizing human judgment within a structured institutional framework.

This research is designed to conceptualize how objectivity can be achieved in the evaluation of productive speech (the end product) of students when they participate in structured interactions that are based on psychological variables, instructional strategies, and data-driven monitoring in a large scale, standardized test environment. The results of the analysis support the notion that objectivity is not a matter of eliminating subjectivity; it is instead about establishing the parameters of human judgment within an institutional context.

The remainder of the Discussion will follow the conceptual levels of the results.

1 Psychological Variability as an Inherent Structural Condition

Rater variability is inherent in evaluation, because of the tendency for assessors to rate based upon severity or leniency [12], to judge all aspects of performance with the same lens (halo effect) [10], to interpret performance along different criteria (criterion interpretation difference) [6], and to increase their rating over time due to familiarity with the test-taker (dynamic severity drift) [19]. Therefore, evaluating performance is not a strictly neutral technical task, but an evaluative one, which can be affected by an assessor's psychology.

Therefore, there are important theoretical implications to this observation. In order to develop methods to reduce variability in performance evaluations, researchers and practitioners need to recognize that variability is an inherent aspect of the evaluation process and, therefore, is an aspect of the evaluation process that needs to be regulated through the design of assessment systems.

The author's professional experiences as an employee of a major testing company also support these conclusions. Although the author worked under well-structured rules governing how test scores were assigned by examiners, the author observed variability in how the examiner interpreted the test-taker's responses, with the variability being influenced by the examiner's subjective internal weighing of the importance of each criterion and their cognitive impression of the test-taker. This experience led the author to conclude that objectivity in the evaluation of performance cannot be achieved by simply relying upon an individual's professionalism and good intentions, but that objectivity requires systemic support.

2 Pedagogical Structuring as Preventive Regulation

The second conceptual area is how teachers use pedagogical structure to help students stay on track. The research showed that when teachers clearly explain what they want their students to do, define what they mean by specific terms, and provide clear, structured ways for students to learn, it helps prevent them from drifting away from the goals.

Using rating scales that were based on the CEFR Performance Descriptors [4; 11] helped reduce ambiguity for raters as to how to evaluate student performance. Raters were able to develop more structured, analytic evaluations of student performance, and therefore used fewer overall impressions or heuristics.

Teacher education programs that include the systematic integration of assessment literacy also enhance teachers' ability to conduct structured assessments, according to Ukrayinska (2024) [16]. In addition, more research [17] indicates that the criteria weighting process is influenced by the pedagogical culture of the school, which suggests that unless schools are willing to critically examine the nature of their culture, then they will continue to experience variability in their assessments.

Pedagogical structure can act as a type of cognitive scaffold for teachers. Pedagogical structure limits the amount of latitude teachers have to make judgments about student performance, but does not eliminate teachers' professional judgments. Calibration sessions, standard exemplars, and shared interpretations of descriptors, all serve to help turn individual teacher evaluations into an institutionalized practice of evaluation.

Thus, in the context of Ukraine's large-scale testing program, the degree to which the testing is objective will depend upon the degree of coordination among institutions. If training is fragmented, or there is no consistent interpretation of the descriptors, then the likelihood of psychological variability occurring will increase. Therefore, the need for pedagogical engineering to be continuous is evident.

3 Measurement Monitoring as Corrective Regulation

Pedagogical structuring prevents errors while measurement based monitoring corrects them. The Many-Facet Rasch Model (MFRM), with its emphasis on raters as facets in a structured framework, is capable of quantifying what previously was considered to be subjective [1].

The use of Drift Detection Mechanisms illustrates that rater behavior can and will change over time [19]. If rater behavior changes over time it is likely to go unnoticed without measurement. Therefore, Measurement Modeling serves as an Institutional Safeguard for detecting rater deviations which could not be identified solely by training.

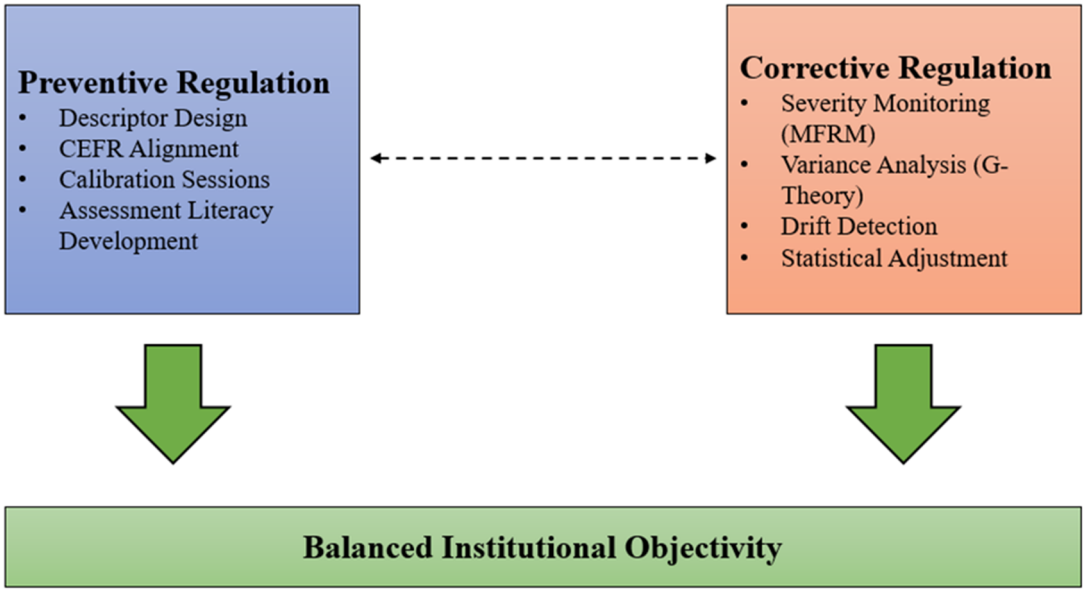

Figure 2. Preventive (pedagogical) and corrective (measurement-based) mechanisms regulating objectivity in large-scale standardized testing.

Statistical adjustment is a powerful method to provide an objective measure of the variability in student performance; however, it has limitations. Specifically, statistical adjustment does not absolve educators from their responsibility to develop appropriate learning environments for students. While measurement tools may demonstrate where the variability lies, they are unable to eliminate or diminish the root causes of that variability psychologically. Objectivity is thus established by the inter-relationship of the preventive instructional design and the corrective use of measurement to monitor student performance (see Fig. 2).

The development of hybrid methods such as automated scoring combined with many-facet models expands the possibilities of regulatory actions [20]. However, no matter how sophisticated the model used to assess communicative ability is, it will always be limited in its ability to assess the full range of this ability. Automated assessment models will only serve to establish the conditions under which the assessment of communicative ability occurs.

4 Objectivity as Engineered Equilibrium

A fundamental theoretical contribution of this research is the thri-factoral concept of engineered equilibrium for objectivity in the evaluation of productive skills:

- Psychological variability is acknowledged rather than denied,
- Pedagogical structuring limits interpretative divergence,
- Measurement monitoring identifies and corrects systematic bias.

When all three elements interact, they produce a state of balance (equilibrium), which becomes unstable if one or more of its components are weakened.

There are two potential pitfalls to the achievement of stability in the system. The first occurs when there is too much emphasis placed on psychology at the expense of the measurement process (i.e., a failure to monitor for bias). This could lead to unobserved drift with respect to the definition of the construct being measured. Conversely, there is also the risk that over-reliance on statistical corrections may lead to mechanical adjustments in scores that have little relationship to the intended construct.

Therefore, the practical application of the equilibrium model in the Ukrainian context for standardized testing will require that such items as severity tracking, systematic

calibration of descriptors, and continued refinement of descriptors, be incorporated into the overall design of large-scale assessments. Furthermore, the assessment literacy training provided to teachers and other assessors will need to extend beyond procedural knowledge of the assessment processes, and provide them with knowledge of the psychological factors that influence their judgments.

Thus, the theory developed in this dissertation provides a new perspective on objectivity as a systemic and dynamic construct that develops through the coordination of psychological processes, pedagogical processes, and measurement processes rather than as an inherent property of the scoring process.

5 Implications for Large-Scale Standardized Testing

Within high-stakes environments such as ZNO, the consequences of uncontrolled variability extend beyond technical measurement concerns; they affect fairness, trust, and institutional legitimacy. The engineered equilibrium model provides a conceptual blueprint for strengthening standardized testing systems.

Specifically, large-scale assessment frameworks should:

1. Embed CEFR-aligned analytic descriptors as cognitive anchors.
2. Institutionalize regular calibration and reflective training.
3. Implement systematic severity monitoring and drift detection.
4. Integrate feedback loops connecting measurement results to pedagogical **revision.**

Such systemic integration transforms objectivity from a declared principle into an operationalized practice.

Conclusion

The current research defines objectivity in assessing students' productive speech as a dynamic product of three elements: psychological factors (i.e., rater-mediated scoring), pedagogical factors (i.e., training and calibration of raters), and measurement-based factors (i.e., quality and quantity of data), which are all embedded in large-scale standardized assessments. While there is evidence to support that rater-mediated scoring is influenced by both subjective cognition and temporal drift, these factors are inevitable in human judgment. However, the pedagogical structure of assessments (e.g., use of CEFR-aligned rubrics, calibration of raters, etc.) acts as a regulatory mechanism to limit the potential of unstructured rater bias.

Measurement models (e.g., Rasch, G-Theory) can serve as mechanisms to identify and monitor rater-related variability, however, the application of these models does not establish objectivity on their own, they do aid in establishing and maintaining it [1].

A key theoretical contribution of this research includes the reframing of objectivity as not being an inherent property of rating procedures, but rather as an emergent product of the coordination of psychological stability through rater mediation, pedagogical design through the calibration of raters, and the establishment of assessment literacy, and ongoing measurement-based monitoring through the application of measurement models. In addition, this integrated approach provides a basis for building increased fairness, reliability, and institutional trust in Ukraine's large-scale assessments of productive speech. Productive speech assessment objectivity is created, it is not an innate characteristic of rating processes.

References:

1. Bijani, H., Hashempour, B., Ibrahim, K. A. A. A., Orabah, S. S. B., & Heydarnejad, T. (2022). Investigating the effect of classroom-based feedback on speaking assessment: a multifaceted Rasch analysis. Language Testing in Asia, 12(1), 26. https://doi.org/10.1186/s40468-022-00176-3

2. Burchell, D., Bourassa Bédard, V., Boyce, K., McLaren, J., Brandeker, M., Squires, B., & FrEnDS-CAN. (2022). Exploring the validity and reliability of online assessment for

conversational, narrative, and expository discourse measures in school-aged children. Frontiers in Communication, 7, 798196. https://doi.org/10.3389/fcomm.2022.798196

3. Dong, M., Gan, C., Zheng, Y., & Yang, R. (2022). Research trends and development patterns in language testing over the past three decades: A bibliometric study. Frontiers in Psychology, 13, 801604. https://doi.org/10.3389/fpsyg.2022.801604

4. Ghanbari, N., & Barati, H. (2020). Development and validation of a rating scale for Iranian EFL academic writing assessment: A mixed-methods study. Language Testing in Asia, 10(1), 17. https://doi.org/10.1186/s40468-020-00112-3

5. Govindasamy, P., del Carmen Salazar, M., Lerner, J., & Green, K. E. (2019). Assessing the reliability of the framework for equitable and effective teaching with the many-facet rasch model. Frontiers in psychology, 10, 1363. https://doi.org/10.3389/fpsyg.2019.01363

6. Heidari, N., Ghanbari, N., & Abbasi, A. (2022). Raters' perceptions of rating scales criteria and its effect on the process and outcome of their rating. Language Testing in Asia, 12(1), 20. https://doi.org/10.1186/s40468-022-00168-3

7. Jeong, H. (2019). Writing scale effects on raters: An exploratory study. Language Testing in Asia, 9(1), 20. https://doi.org/10.1186/s40468-019-0097-4

8. Jia, W., & Zhang, P. (2023). Rater cognitive processes in integrated writing tasks: from the perspective of problem-solving. Language Testing in Asia, 13(1), 50. https://doi.org/10.1186/s40468-023-00265-x

9. Khodi, A. (2021). The affectability of writing assessment scores: a G-theory analysis of rater, task, and scoring method contribution. Language Testing in Asia, 11(1), 30. https://doi.org/10.1186/s40468-021-00134-5

10. Kim, H. (2020). Effects of rating criteria order on the halo effect in L2 writing assessment: a many-facet Rasch measurement analysis. Language Testing in Asia, 10(1), 16. https://doi.org/10.1186/s40468-020-00115-0

11. Li, J., & Wang, Q. (2021). Development and validation of a rating scale for summarization as an integrated task. Asian-Pacific Journal of Second and Foreign Language Education, 6(1), 11. https://doi.org/10.1186/s40862-021-00113-6

12. Mohd Noh, M. F., & Mohd Matore, M. E. E. (2022). Rater severity differences in English language as a second language speaking assessment based on rating experience, training experience, and teaching experience through many-faceted Rasch measurement analysis. Frontiers in Psychology, 13, 941084. https://doi.org/10.3389/fpsyg.2022.941084

13. Ozernyi, D. M., & Suvorov, R. (2023). Ukrainian language proficiency test review. Language Testing, 40(3), 828-839. https://doi.org/10.1177/02655322231156819

14. Rezai, A., Namaziandost, E., Miri, M., & Kumar, T. (2022). Demographic biases and assessment fairness in classroom: insights from Iranian university teachers. Language Testing in Asia, 12(1), 8. https://doi.org/10.1186/s40468-022-00157-6

15. Sureeyatanapas, P., Sureeyatanapas, P., Panitanarak, U., Kraisriwattana, J., Sarootyanapat, P., & O'Connell, D. (2024). The analysis of marking reliability through the approach of gauge repeatability and reproducibility (GR&R) study: a case of English-speaking test. Language Testing in Asia, 14(1), 1. https://doi.org/10.1186/s40468-023-00271-z

16. Ukrayinska, O. (2024). Synergies in developing pre-service teachers' language assessment literacy in Ukrainian universities. Education Sciences, 14(3), 223. https://doi.org/10.3390/educsci14030223

17. Ukrayinska, O. (2025). English and French Teachers' Assessment Practices in the Ukrainian Context: Understanding Language Assessment Literacy Needs Across Two Languages. Education Sciences, 15(1), 42. https://doi.org/10.3390/educsci15010042

18. Uto, M. (2021). Accuracy of performance-test linking based on a many-facet Rasch model. Behavior Research Methods, 53(4), 1440-1454. https://doi.org/10.3758/s13428-020-01498-x

19. Uto, M. (2023). A Bayesian many-facet Rasch model with Markov modeling for rater severity drift. Behavior research methods, 55(7), 3910-3928. https://doi.org/10.3758/s13428-022-01997-z

20. Uto, M., & Aramaki, K. (2024). Linking essay-writing tests using many-facet models and neural automated essay scoring. Behavior Research Methods, 56(8), 8450-8479. https://doi.org/10.3758/s13428-024-02485-2