E-ISSN: 1116 - 3240 P-ISSN: 3026-9334
MPRI-JAIMR | INTERNATIONAL JOURNAL OF AFRICAN INNOVATION & MULTIDSCIPLINARY RES.
VOL. 14 NO. 2 – OCTOBER, 2026

MACHINE LEARNING AND TEST ANXIETY: A CASE FOR RESTRAINED, TRANSPARENT AND STUDENT-GOVERNED USE

MNGUSUUR JAM Ph.D
School of General Education, Federal College of Education, Odugbo, Benue State, Nigeria.
Corresponding Author: jammngusuur@fceodugbo.edu.ng
DOI: https://doi.org/10.70382/mejaimr.v14i2.0155

Abstract

Test anxiety is a well-documented barrier to student performance, and machine learning (ML) is increasingly proposed as a way to predict, detect and respond to it. This position paper argues that, in the current state of evidence and regulation, ML should be used for test anxiety with restraint. Five problems are identified: weak and noisy ground-truth labels; evaluation practices, such as student-level data leakage and the absence of simple baselines, that overstate accuracy; unequal performance and base rates across groups; reliance on emotion inference from biometric signals that is scientifically contested and, in education institutions, prohibited in the European Union's Artificial Intelligence Act with narrow exceptions; and the power asymmetries of assessment settings, where students cannot readily refuse and labels persist. The paper proposes a six-principle framework (support not sanction, student governance, simple models first, prospective and subgroup-aware validation, human in the loop, and no biometric emotion inference during assessment), and applies it to common proposals in a decision table. It addresses four objections and sets out a research agenda that includes trials of whether acting on predictions improves student outcomes, fairness analysis, and studies in low-resource contexts such as Nigerian schools.

Keywords: test anxiety; machine learning; educational assessment; algorithmic fairness; emotion recognition

Introduction

Test anxiety is among the most studied emotional problems in education. Hembree's (1988) meta-analysis drew on 562 studies, and von der Embse, Jester, Roy and Post (2018) updated the evidence with 238 further studies, again documenting a negative relationship between test anxiety and a range of educational outcomes. Alongside this long research tradition, a newer literature has appeared in which machine learning (ML) is used to predict, detect or manage test

EDITOR-IN-CHIEF
Prof. A.I. Mteiye [MALAWI]

anxiety. Proposed applications include models that flag at-risk students from questionnaires and records, systems that infer anxiety from heart rate, skin conductance, voice or facial expression, and adaptive tools that change the test or the support offered in response to a predicted emotional state. The appeal is obvious. Anxiety is common, often unreported, and costly for learners, and ML promises scale and objectivity that human observation cannot match. This paper does not argue that ML has no place in this area. It argues for **restraint:** that current ML approaches to test anxiety rest on weak ground truth, are frequently evaluated in ways that overstate their accuracy, can perform unequally across groups, depend on emotion inference that is scientifically contested and increasingly restricted by law, and are deployed in settings where students have little power to refuse. On this view, ML should be used for test anxiety only for low-stakes support that students control, that is built on transparent and validated methods, and that is shown to help. I call this position restrained, transparent and student-governed use, and I set out a framework to apply it.

Scope, Definitions and Approach

This is a conceptual position paper. It reports no new data and draws on three kinds of source: syntheses of the test-anxiety literature, methodological and fairness literature on machine learning, and regulatory material on emotion-inferring systems in education. Examples and numerical illustrations are hypothetical and are labelled as such.

Three definitions fix the scope. **Test anxiety** is the set of cognitive, affective, physiological and behavioural responses to concern about performing poorly in an evaluative situation. **Machine learning** refers to methods that learn a mapping from inputs to outputs from data, whether logistic regression, tree ensembles, support vector machines or neural networks; the argument does not assume that complex models are used, and indeed recommends that simple ones be preferred. **Emotion inference** means estimating a person\u2019s emotional state from observable data such as facial movements, voice, physiology or behaviour. The paper is concerned with educational use, and in particular with the assessment settings in which test anxiety arises, and not with clinical use under professional supervision, which has its own standards.

The Construct

Test anxiety is not a single thing. Liebert and Morris (1967) distinguished worry, the cognitive component, from emotionality, the physiological arousal component, and later work has treated the construct as multidimensional, with cognitive, affective, physiological and behavioural elements (Zeidner, 1998; Cassady & Johnson, 2002). Newer scales reflect this, such as the Multidimensional Test Anxiety Scale of Putwain, von der Embse, Rainbird and West (2021). Attentional control theory explains how anxiety harms performance: anxious individuals devote attentional and working-memory resources to threat and worry, which impairs the efficiency of processing on demanding tasks (Eysenck, Derakshan, Santos & Calvo, 2007). Three features of the construct matter for ML. First, anxiety has trait and state forms. A student can be dispositionally anxious about tests yet calm on a given day, or the reverse. Second, its expression varies across individuals, cultures and tasks, so the same physiological signal or behaviour may mean different things. Third, the relationship with performance is not simple. Anxiety is often viewed as causing

poor performance, but poor prior performance can also cause anxiety (Shi & Yuan, 2025), and a model that predicts anxiety from past grades may be capturing the second pathway.

Prevalence and Consequences

Estimates of prevalence depend on the instrument, the cut-off and the population, and should be read with care. The meta-analytic literature consistently reports negative associations between test anxiety and performance, and also associations with poorer self-concept, motivation and wellbeing (Hembree, 1988; von der Embse et al., 2018). Narrative reviews citing the 2018 meta-analysis report that a notable minority of students in the United States, of the order of 15 to 20 per cent, experience test anxiety at problematic levels (Shi & Yuan, 2025). There is much less systematic evidence for many other education systems, including Nigeria\u2019s, although high-stakes examinations at the end of secondary school make the issue plainly relevant. The prevalence matters for the argument because the usefulness of any predictive tool depends on it, as Section 4 shows.

How Test Anxiety Is Measured Today

The principal methods are self-report scales, most of which are brief and inexpensive. The Test Anxiety Inventory (Spielberger, 1980) separates worry and emotionality; more recent instruments, such as the Multidimensional Test Anxiety Scale, differentiate components such as worry, tension, and cognitive interference (Putwain et al., 2021). Physiological measures, including heart rate, heart-rate variability and electrodermal activity, can capture arousal, but they are influenced by many factors unrelated to anxiety, such as physical activity, caffeine and excitement. Behavioural observation is rarely systematic. Each method has limitations, none is a gold standard, and the correspondence between self-report and physiological indicators is typically modest. This matters for ML because the choice of measure fixes what the model learns to predict.

Established Interventions

A model of test anxiety needs to be seen against what is already known about helping students. Reviews of interventions report that cognitive-behavioural and skills-based approaches reduce anxiety, and that interventions combining study-skills training with anxiety-management techniques are particularly promising, although effects on performance are smaller and less consistent than effects on self-reported anxiety (Hembree, 1988; von der Embse et al., 2018; Shi & Yuan, 2025). Such interventions can be delivered to whole classes at low cost. This is important for the question of targeting: if effective support can be delivered universally at modest expense, the added value of an algorithm that identifies who should receive it is correspondingly reduced.

The Typical Pipeline

Studies in this area typically follow a common sequence. Data are collected from a sample of students, often a class, a school or a university cohort, at one or more points in time. Features are derived: questionnaire items, demographic variables, prior grades, platform interaction logs, or signals from sensors. A label is defined, usually by thresholding a self-report scale or by taking a student\u2019s own rating of state anxiety at the time of recording. A model is trained and

evaluated, most often by cross-validation, and results are reported as accuracy, area under the curve, F1-score or similar. The model is then proposed as the basis of a tool that can screen students or trigger support. Each step conceals decisions that affect validity: how the sample was drawn, how the features relate to the construct, how the threshold was chosen, how the evaluation was partitioned and what the performance figure means for the intended use. The five problems below correspond to the weakest points in this pipeline.

It helps to distinguish three tasks, because the case for and against ML differs among them.

1. **Risk prediction from existing data.** Models use questionnaires, demographic information, past grades, and sometimes learning-platform logs to predict who is likely to experience high test anxiety.
2. **Affect detection from signals.** Models infer momentary anxiety from physiological sensors, facial video, voice or keystroke and mouse behaviour, in the tradition of affective computing and affect detection (Calvo & D'Mello, 2010; D'Mello & Graesser, 2012).
3. **Adaptive response.** Systems alter the test, the feedback or the support offered on the basis of a predicted anxiety state or risk score.

Research on affect detection has shown that emotions can be recognised above chance in controlled settings (Calvo & D'Mello, 2010). The step from this to a reliable, fair, deployable tool for real examinations is large, and the sections that follow explain why.

What the Affect-Detection Literature Does and Does Not Show

Reviews of affect detection report that emotional states can be classified from physiological, facial, vocal and behavioural signals at levels above chance, particularly in laboratory conditions with posed or elicited emotion, and with models trained and tested on the same individuals (Calvo & D\u2019Mello, 2010). Studies of learners\u2019 affect during complex learning show that states such as confusion, frustration, boredom and engagement are common, dynamic and tied to the structure of the task (D\u2019Mello & Graesser, 2012). These are valuable findings for understanding learning. They do not by themselves show that anxiety during a real examination can be reliably identified in unfamiliar students, at the level of the individual, in sufficiently varied conditions to support decisions. Three gaps lie between the research findings and deployment. The first is ecological: examinations differ from laboratory tasks in stakes, setting, motion constraints and the presence of invigilators, all of which change signals. The second is generalisation across persons: models that work within persons often degrade for new persons, because baselines, expression styles and physiology vary. The third is the distinction between arousal and anxiety. Heightened heart rate can reflect excitement, effort or caffeine as readily as anxiety, and a model trained on one will mislabel the others. Careful research can address these gaps, but until it has done so, claims of readiness for use with students are premature.

The Chain from Signal to Construct

It helps to see why the task is harder than it appears by laying out the chain of inference that a physiological or behavioural detector depends on. A sensor records a signal, for instance a change in skin conductance. The signal is a response to some process in the body, which may be

sympathetic arousal. Arousal may be caused by anxiety about the test, but it may equally be caused by excitement, effort, temperature, movement or something unrelated to the examination. The anxious state, in turn, is only one component of test anxiety, which also includes worry and cognitive interference that may not be visible in physiology at all. Finally, the student\u2019s own report of anxiety, which is used as the label, is a further step removed, since it depends on what the student notices, how she interprets it and how willing she is to disclose it. At each link of the chain there is slippage, and the errors compound. Evidence that a signal discriminates between conditions in a controlled experiment is not evidence that it measures the target construct in an examination hall. This is the reason why the paper urges that tools be evaluated against outcomes that matter to students, and not only against a proxy label.

Ground truth is weak

A supervised model cannot exceed the quality of its labels. In test-anxiety research, labels almost always come from self-report scales, sometimes from a single administration. Scales differ in what they measure, are open to response bias and are imperfectly correlated with physiological and behavioural indicators. Dichotomising a continuous scale into anxious and non-anxious groups introduces arbitrary cut-offs that vary between studies. When a model is trained to reproduce such labels, the reported accuracy describes agreement with a noisy and partly arbitrary criterion, not accuracy in detecting a student's condition. Where the label is the student's later test result, the model is predicting performance and not anxiety, and may simply rediscover prior attainment. A simple calculation illustrates the consequences of imperfect labels. Suppose, hypothetically, that a self-report scale used as the label has a reliability of .70 for the underlying state. The correlation between the scale and the true state cannot exceed the square root of the reliability, which is about .84, and any predictor can therefore correlate with the label at most as well as the label correlates with truth if its own relationship to the truth is imperfect. A model that appears to predict the label with 80 per cent accuracy may be considerably less accurate against the true condition, and nothing in the standard evaluation will reveal this. The problem is compounded when thresholds are applied. A cut-off that divides students into high and low anxiety discards information, creates arbitrary classes and places many students whose scores lie near the threshold in the wrong category by chance. Models trained to reproduce such classes are asked to find a boundary that does not exist in nature. Continuous or ordinal targets, or probabilistic outputs with uncertainty, are more honest representations.

Evaluation practices overstate performance

ML studies in education frequently use small samples, multiple measurements per student, and many candidate features. Under these conditions, data leakage and over-fitting are easy to introduce and hard to notice. Kapoor and Narayanan (2023) documented how leakage has led to inflated claims across fields that use ML, and Varoquaux and Cheplygina (2022) described similar problems in medical imaging, including small test sets and the absence of comparison with simple baselines. The most common leak in the present area is splitting at the level of observations and not students, so that records from the same learner appear in both training and test sets and the model recognises people instead of anxiety. Other problems include selecting features or tuning

hyperparameters on the full dataset, and reporting accuracy without regard to class imbalance, when anxious students may be a minority. A credible study should therefore report group-aware cross-validation, a held-out test set from a different cohort or school, comparison with a simple baseline such as a short validated scale or logistic regression, calibration, and measures appropriate to class imbalance, with confidence intervals that reflect the true number of independent students.

Data leakage through repeated measures deserves a numerical illustration. When several observations are collected from each student, observations are not independent. The effective sample size can be approximated by n divided by (1 plus (m minus 1) times the intraclass correlation), where m is the number of observations per student. Suppose, hypothetically, that 60 students are each observed 10 times and that the intraclass correlation is .50. The effective sample size is then 600 divided by 5.5, or about 109, not 600. A model evaluated on randomly split observations will have seen most test students in training, and its apparent accuracy will reflect recognition of individuals. Confidence intervals computed as if there were 600 independent cases will be far too narrow. Another common failure is the absence of a meaningful baseline. If a model that predicts anxiety from prior grades and gender achieves a particular area under the curve, the relevant question is whether it improves on a three-item questionnaire or on simply asking students. Without such comparisons, reports of accuracy have no reference point and cannot guide decisions about whether the added complexity is warranted.

Fairness and generalisability

Educational ML has a well-documented problem of unequal performance across groups (Baker & Hawn, 2022; Kizilcec & Lee, 2022), and the sources of bias are numerous, ranging from unrepresentative training data to measurement differences and the way outcomes are defined (Mehrabi, Morstatter, Saxena, Lerman & Galstyan, 2021). Test anxiety offers specific reasons for concern. Reported levels differ by gender and across regions in some syntheses, so base rates differ, and with different base rates the same model will have different predictive values in different groups. Cultural norms of emotional display, differences in physiology and variation in device quality and connectivity can all affect signals. A model validated in one school type or country is not thereby validated in another, and few studies report performance by subgroup. A tool that under-detects anxiety in some groups or over-flags others would distribute support and stigma unequally. The difficulty is not only unequal error rates but a deeper tension among fairness criteria. Results in the fairness literature show that when the prevalence of the outcome differs between groups, a risk score cannot in general satisfy calibration within groups and equal false-positive and false-negative rates at the same time, except in special cases (Kleinberg, Mullainathan & Raghavan, 2017; Chouldechova, 2017). Developers must therefore choose which property to prioritise and justify the choice, and users should be told what was chosen.

The effect of base rates on usefulness can be shown with a hypothetical example. Suppose a classifier has sensitivity and specificity of .80 in every group. The probability that a student flagged as anxious really is anxious, the positive predictive value (PPV), then depends on prevalence, as Table 1 shows. At 10 per cent prevalence, fewer than a third of flagged students are truly anxious; at 30 per cent, nearly two thirds are. The same tool, with the same intrinsic accuracy, therefore

gives different groups very different experiences: in a low-prevalence group most flags will be false alarms, whereas in a high-prevalence group most flags will be correct. If a tool is applied across schools or groups with different prevalence, as is likely, then the meaning of a flag differs in each.

Table 1. Hypothetical positive predictive value of a classifier with sensitivity = specificity = .80 at different prevalence levels

Prevalence of high test anxiety	Flagged students who are truly anxious (PPV)	Flagged students who are not anxious
10%	31%	69%
20%	50%	50%
30%	63%	37%

Note. PPV = (sensitivity \u00d7 prevalence) / [(sensitivity \u00d7 prevalence) + ((1 \u2212 specificity) \u00d7 (1 \u2212 prevalence))]. Figures are illustrative and not drawn from any study.

Emotion inference is scientifically contested and legally restricted

Systems that read emotions from faces, voices or bodies assume that emotional states can be reliably inferred from such outward signals. Barrett, Adolphs, Marsella, Martinez and Pollak (2019), in a comprehensive review, concluded that the evidence does not support the claim that people can infer a person's emotional state reliably from facial movements alone, because expressions vary across cultures and situations and the same expression can reflect different states. If so, models trained on labelled expressions may simply learn the labelling conventions of the dataset. The law is moving in the same direction. The European Union's Artificial Intelligence Act (Regulation (EU) 2024/1689) prohibits AI systems that infer the emotions of natural persons in workplaces and education institutions, except for medical or safety reasons, and the prohibition has applied since 2 February 2025. The Commission's guidelines and commentary indicate that the prohibition is directed at inference from biometric data, and that inferring emotions from written text falls outside its scope (European Commission, 2025; see also Future of Privacy Forum, 2025). Whether a particular system falls within the ban will depend on its design and purpose, and the boundaries of the medical and safety exceptions remain debated. For the present argument, two consequences follow. A biometric anxiety detector used during assessment in a European education institution is likely to be unlawful, and designers elsewhere should treat that as a signal about the weakness of the science and the severity of the power imbalance, not as a regional peculiarity. Where Nigerian or other national data-protection law applies, such as the Nigeria Data Protection Act 2023, developers should also consider consent, purpose limitation and the sensitivity of data about emotional and health states.

Outside the European Union the prohibition does not apply, but the reasoning behind it is portable. Reasons given for it include the power imbalance between institutions and students, the weakness of the science, and the risk of discrimination. The Act also treats several education uses as high-risk, including systems used to evaluate learning outcomes and to monitor student behaviour in tests, which attract requirements for risk management, data quality, transparency and human oversight. A developer who follows these requirements voluntarily, wherever the product is sold, reduces legal exposure and improves quality. Data protection law adds a further layer. Information

about a person\u2019s emotional or mental state may be treated as health-related and therefore as a special category of personal data in many regimes, requiring a stronger justification for processing and, usually, explicit consent. Consent given by a student who cannot realistically refuse in an examination hall is of doubtful validity. These considerations apply to the collection of the data as well as to the use of the resulting model.

Power, consent and the persistence of labels

Assessment settings are asymmetric. Students cannot easily decline to be observed during an examination, and a predicted label may follow them in records. Learning analytics has long raised concerns about surveillance, consent and the effects of categorising learners (Slade & Prinsloo, 2013). In the case of test anxiety, a label such as high-risk may influence how teachers perceive a learner, may alter the test the learner receives, or may deter learners from disclosing difficulties. A tool intended to support students could in practice monitor and sort them. There is also a risk of self-fulfilling expectations. The classic experiment on teacher expectations reported that students identified to teachers as likely to show intellectual growth later showed greater gains (Rosenthal & Jacobson, 1968). Its findings and methods have been debated, but the broad principle that teachers\u2019 beliefs about students can influence their treatment of them is widely accepted. A label that predicts high anxiety could, if known to teachers, change how they interact with a student, whether through lowered expectations or through well-meant but unnecessary accommodations, and could affect how the student views herself. The potential harms of a mistaken label, including false positives, are therefore real even when no formal decision is taken on the basis of the label.

The Position: Restrained, Transparent and Student-Governed Use

The problems above do not mean that ML can never help. They justify a framework in which use is restrained to cases where benefits are plausible and harms controllable. The framework has six principles.

1. **Support, not sanction or selection.** Outputs may inform optional support. They should never influence grades, progression, placement or accusations of misconduct.
2. **Student governance.** Participation should be voluntary and informed, with the ability to withdraw and to see and challenge one's own data and results. Data should be minimised, held securely and deleted when no longer needed.
3. **Simple first.** A short validated scale, or logistic regression on a few well-chosen predictors, should be the baseline against which any complex model is judged. Complexity should be adopted only if it brings clear gains. Where decisions affect people, interpretable models are preferable to post hoc explanation of opaque ones (Rudin, 2019), and explanation tools such as SHAP (Lundberg & Lee, 2017) should be treated as aids to understanding and not as proof of validity.
4. **Rigorous, prospective validation.** Evaluation should use student-level splits, external cohorts, pre-registered analysis plans, and subgroup performance reporting, with documentation such as model cards that state intended use, data and limitations (Mitchell et al., 2019).

5. **Human in the loop.** Predictions should prompt a conversation with a counsellor or teacher, who retains judgement. No intervention should be triggered automatically.
6. **No biometric emotion inference in assessment settings.** Inference of anxiety from faces, voices or physiology should not be used on students during assessment, and any research with such signals should take place outside assessment, with ethical approval, informed consent and appropriate legal review.

Table 2. Applying the framework to common proposals

Proposed use	Position	Conditions
Questionnaire-based risk flag that invites students to opt in to support (text or survey data)	Acceptable	Voluntary; validated scale as baseline; student-level validation; subgroup performance reported; results seen first by student
Prediction of anxiety from grades and platform logs to target interventions	Acceptable with caution	Check that the model is not simply predicting prior attainment; protect against label persistence; compare with universal provision
Physiological or voice sensing in research outside assessment	Acceptable in research only	Ethics approval; informed consent; no consequences for participants; legal review of local rules
Facial, voice or physiological anxiety detection during examinations	Not acceptable	Weak inference, power imbalance, and likely illegal where the EU AI Act applies
Automatic adjustment of test content or scoring based on predicted anxiety	Not acceptable	Risks unequal treatment and distorts the construct being assessed
Use of anxiety predictions in decisions about progression or discipline	Not acceptable	High stakes, unvalidated labels, no meaningful consent

The Principles in Detail

The six principles listed above call for further explanation, because their application involves judgement.

Support, not sanction or selection. The distinction is between uses that offer something to students and uses that decide something about them. An optional workshop, a conversation with a counsellor, or a self-help resource are offers. Placement, grading adjustment, access to opportunities and disciplinary action are decisions. The principle also forbids secondary uses: data collected to offer support must not be repurposed for decisions, and agreements with vendors should say so.

Student governance. Governance is more than consent at the outset. It includes the ability to see what is held, to correct it, to withdraw and to have the data deleted, and it means involving students and parents in decisions about whether and how a tool is used. For minors, the consent

of parents or guardians is also needed, but the assent of the student should be sought and respected.

Simple first. Complexity should be justified by demonstrated gain over a transparent baseline, evaluated honestly. Where the same decision can be made as well by a short validated scale as by a model drawing on logs and sensors, the scale is preferable because it is cheaper, more transparent, easier to audit and less invasive.

Rigorous, prospective validation. Retrospective cross-validation on a convenience sample shows only that a model can fit that sample. Prospective validation in a new cohort, preferably in a different school, with the analysis plan registered in advance, shows whether it works when it matters. Subgroup performance, calibration and the cost of different kinds of error should be reported, and results should be published whether favourable or not.

Human in the loop. A trained person should interpret any output in context, talk with the student and decide what to offer. The person must have both the authority and the time to override the model. Where automation bias is likely, as when staff are busy and the model appears authoritative, procedures should require that the model\u2019s output be treated as a prompt for inquiry and not as a finding.

No biometric emotion inference in assessment settings. The principle follows from the combination of weak science, power imbalance and legal restriction. It does not forbid research on physiological correlates of anxiety conducted with consenting participants outside assessment, and it does not forbid the use of text or survey data with the safeguards above.

A Validation Checklist

Table 3 sets out items that a study or product claiming to predict test anxiety should be able to answer. A claim that cannot meet them should be treated as unproven.

Table 3. Minimum validation checklist for ML tools that predict test anxiety

Domain	Question
Construct and label	What exactly is predicted (trait, state, performance impact)? How was the label obtained, how reliable is it, and how was any threshold chosen?
Data	Who is in the sample? How were students recruited? How many independent students, and how many observations per student?
Evaluation	Were splits made at student level? Was there an external or prospective test set? Was the analysis pre-registered?
Baselines	How does the model compare with a short validated scale, a simple regression and the base rate?
Calibration and errors	Are probabilities calibrated? What are sensitivity, specificity and PPV at the expected prevalence?
Fairness	Are performance and calibration reported by gender, language, school type and other relevant groups? What was prioritised and why?
Interpretability	Can staff and students understand what drives a prediction? Are explanation tools validated or only illustrative?

Domain	Question
Governance	Is use voluntary? Who sees results? How long are data kept? Can students see, correct and delete data?
Impact	Is there evidence that acting on predictions improves outcomes compared with universal provision?
Documentation	Is there a model card stating intended use, limitations and known failure modes?

A Worked Example: A Responsible Questionnaire-Based Support System

To show that the framework is workable, consider a hypothetical design for a public senior secondary school in Makurdi that wants to help students who are anxious about the final examinations. The school does not need machine learning to begin. In the first stage, all students in the examination class are offered a short, validated test-anxiety questionnaire in the language of instruction, with an explanation that participation is voluntary, that responses are confidential to the counsellor and that no scores go to teachers or into records. Students who choose to share their results are invited to optional sessions on coping and study strategies, and all students receive a class-wide session on the same material, which is delivered whether or not they complete the questionnaire. In the second stage, if the school and a university partner wish to explore whether a model can add value, a research project is set up with ethical approval. Students consent to the use of their questionnaire responses, with grades from earlier terms, in a prospective study. A simple model such as regularised logistic regression is compared with the short questionnaire and with the base rate, using student-level cross-validation and a held-out cohort the following year. Performance is reported by gender and other relevant groups, along with PPV at the observed prevalence. A random half of the participants is offered targeted support when the model flags them, while the other half receives universal support only, so that the study can show whether targeting improves outcomes. Nothing in the process requires cameras, sensors or any inference from biometric signals. If the model does not beat the questionnaire or does not improve outcomes, the school continues with universal provision and the short scale.

Text-Based and Conversational Systems

Much of the recent interest in educational technology concerns language-based systems, including chatbots that offer coping tips or respond to students\u2019 written descriptions of their worries. These differ from biometric detectors in important ways. According to the Commission\u2019s guidelines as summarised in commentary, the prohibition in the EU Act concerns inference from biometric data, and inferring emotions from written text is outside its scope (Future of Privacy Forum, 2025). Conversational tools can also be offered voluntarily, used anonymously and consulted when the student chooses, which fits the framework better than sensing that runs in the background. The framework nevertheless applies to them. A system that reads a student\u2019s messages and classifies them as anxious has the same validity problems as other classifiers: unclear labels, unknown accuracy across languages and dialects, and the risk of unequal performance for students who write in non-standard English or in code-switched language. Generative systems may also produce advice that is inaccurate or inappropriate, may respond

poorly to signs of serious distress, and may collect sensitive disclosures that are stored by third parties. Minimum safeguards include clear statements that the tool is not a clinician, reliable routes to human help, protection of conversation data, evaluation of responses to distress and testing across the languages and registers that students use. The same question applies as elsewhere: what evidence shows that using the tool helps?

Governance over Time

A model that is valid when it is built can cease to be valid. Student populations change, examination formats change, and the meaning of the features can shift, for example when a learning platform is redesigned or when a pandemic changes patterns of study. Models should therefore be monitored after deployment, with regular checks of performance and calibration on new data and by subgroup, with defined thresholds that trigger review, and with a plan for retiring a model that no longer meets standards. Responsibility for this monitoring should be assigned to named people, and results should be reported to the body that oversees the programme and, in summary, to students and parents. Documentation should record changes, since a model that has been updated is a different model and may need fresh validation. Procurement contracts should require vendors to supply the information needed for such oversight and to allow independent audit.

Objections and Responses

Objection 1: Early detection at scale could help many students. The goal is sound, but the choice is not between ML and doing nothing. Because test anxiety is common, a universal approach such as teaching evidence-based coping and study strategies, adjusting classroom assessment practice and normalising help-seeking may reach more students at lower risk than targeted prediction. Targeting is most justified where resources are scarce and the predictor is accurate, fair and shown to improve outcomes, and that is the evidence that is currently missing.

Objection 2: ML can outperform simple questionnaires. It may, in some datasets, but the claim must be established with sound evaluation. Many reported advantages shrink under group-aware validation and against a strong simple baseline. Even where gains are real, the question is whether they are large enough to justify added opacity, data collection and risk.

Objection 3: Restrictions on emotion inference will stifle innovation. The restrictions apply to a narrow range of designs and exempt medical and safety uses. They leave open text and survey-based approaches, research outside assessment, and the development of better measures. Innovation that depends on covert inference from students under examination conditions is not one that education should aim to encourage.

Objection 4: Students may welcome tools that detect their stress. Some may, and a student-controlled tool that they choose to use for self-monitoring is consistent with the framework. The difficulty arises when institutions deploy such tools on students who cannot refuse.

Objection 5: Voluntary, opt-in systems exclude exactly the students who need help. This is a real limitation, since students who are most anxious may be least likely to opt in, or may fear the consequences of disclosure. The response is not to replace choice with compulsion but to make participation safe: guarantee confidentiality, separate the service from assessment and discipline, offer universal supports so that no one needs to identify themselves in order to benefit, and build

trust through the conduct of the service. Evidence that opt-in systems fail to reach those in need would be a reason to redesign them, and not to abandon voluntariness.

Objection 6: Other predictive systems in education are already in use, so this concern is overstated. The existence of risk scores for attrition or failure shows that predictive analytics is common, and it raises similar concerns about validity, fairness and the effects of labels. It is a reason to apply the same scrutiny widely and not to relax it. Predicting anxiety is arguably more sensitive, because it involves inference about internal states and health-related information, and the difficulty of verifying such states makes errors harder to detect.

A Research Agenda

1. **Prospective, pre-registered validation** of any anxiety-prediction model in new cohorts and settings, with student-level splits and comparison with simple baselines.
2. **Trials of support, not just models.** Randomised comparisons of ML-targeted support with universal support and with usual practice, using student wellbeing and performance as outcomes. A prediction is valuable only if acting on it improves something.
3. **Better ground truth**, combining validated multidimensional scales, repeated measurement across occasions, and, where ethical, physiological data collected outside assessment.
4. **Fairness analysis** by gender, language, school type, disability and region, with reporting of error rates and calibration for each group.
5. **Studies in low-resource contexts**, including Nigerian schools, where device access, connectivity, language and data-governance conditions differ, and where local data and local participation in design are essential.
6. **Governance research** on consent, data sovereignty and the way students and parents understand and use predictions.

Several of these priorities call for specific design features. Prospective validation should specify the intended use in advance, such as offering support to students in the top quartile of predicted risk, and should evaluate the model in terms of that use, including the capacity of counselling services to respond to the number of students flagged. A model that flags more students than can be helped creates frustration and no benefit. Trials of prediction-led support should randomise at the student or class level, depending on the risk of contamination, include a universal-support comparison arm, and measure outcomes such as self-reported anxiety, wellbeing, performance and help-seeking, with follow-up beyond the immediate examination period. Work on ground truth should combine validated self-report scales with repeated measurement across days and situations, so that trait and state components can be separated, and should assess the agreement between methods rather than assuming that one is correct. Fairness analysis should be planned in advance, with enough participants in each subgroup to estimate error rates with useful precision, and should report calibration and the cost of different errors for each group. Studies in Nigerian and other low-resource settings should be led or co-led by local researchers, should consider the availability of devices and connectivity, and should address language of instruction and cultural expression of distress, as well as local data-protection requirements.

Recommendations for Stakeholders

Developers and vendors should adopt the validation checklist in Table 3, publish model cards, avoid biometric emotion inference for students in assessment settings, design for student control of data, and make clear in marketing what has and has not been validated, including by group.

Schools and universities should begin with universal measures of support and with validated short questionnaires, require evidence of benefit before adopting predictive tools, ensure that use is voluntary and that results cannot reach decisions about grades or progression, and train staff to treat predictions as prompts for conversation.

Regulators and ministries should clarify rules on emotion-inferring technology in education, require impact assessments for tools used with children, and support independent evaluation. In jurisdictions that do not yet regulate this area, drawing on the logic of the EU approach, together with data-protection law, would be a prudent starting point.

Researchers and funders should prioritise prospective, pre-registered validation and trials of whether prediction-led support helps, report negative results, share code and data where consent allows, and include under-represented settings so that evidence does not come only from a few countries.

Students and parents should be told in plain language what is collected, by whom, for what purpose and for how long, and should be able to decline without penalty.

Conclusion

Test anxiety is a real and widespread educational problem, and ML tools offer an appealing way to address it. But the current state of the evidence calls for restraint. The labels are noisy, evaluation practices often inflate performance, fairness is rarely examined, emotion inference from biometric signals is scientifically contested and increasingly prohibited in education, and assessment settings give students little power to refuse. The position defended here is not a rejection of ML but a set of conditions under which it can be used responsibly: for voluntary, low-stakes support, with simple and transparent methods first, validated prospectively and by group, with humans in the loop and with students in control of their own data. Until the evidence shows that predicting test anxiety leads to better outcomes for students, the burden of proof lies with those who would deploy such systems. Two broader lessons emerge from this case. First, it is possible to take emotional and psychological wellbeing in education seriously without adopting every available technology for measuring it; indeed, taking wellbeing seriously argues for care about who holds information on students\u2019 inner states and what it is used for. Second, the discipline of asking what a prediction is for, and whether acting on it helps, is more important than any improvement in accuracy. A tool that predicts test anxiety with high accuracy but leads to no better outcome for students has no educational value, whereas a simple, voluntary and humane support system that students trust may help many of them.

Limitations of This Analysis

This paper is a position paper and not a systematic review. The literature it draws on is large and fast-moving, and the selection of sources reflects the judgement of the author. The regulatory discussion rests on the text of the EU Artificial Intelligence Act and on published commentary and

guidance whose interpretation is still developing, and it should not be read as legal advice; readers in particular jurisdictions should consult current law. The numerical examples are hypothetical and are included to illustrate principles. The recommendations are made on the evidence available now and would change if prospective, well-controlled studies showed that ML-based tools for test anxiety are valid, fair and beneficial in the conditions in which they are used.

References

Baker, R. S., & Hawn, A. (2022). Algorithmic bias in education. *International Journal of Artificial Intelligence in Education, 32*, 1052–1092.

Barrett, L. F., Adolphs, R., Marsella, S., Martinez, A. M., & Pollak, S. D. (2019). Emotional expressions reconsidered: Challenges to inferring emotion from human facial movements. *Psychological Science in the Public Interest, 20*(1), 1–68.

Calvo, R. A., & D'Mello, S. (2010). Affect detection: An interdisciplinary review of models, methods, and their applications. *IEEE Transactions on Affective Computing, 1*(1), 18–37.

Cassady, J. C., & Johnson, R. E. (2002). Cognitive test anxiety and academic performance. *Contemporary Educational Psychology, 27*(2), 270–295.

Chouldechova, A. (2017). Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big Data, 5*(2), 153–163.

D'Mello, S., & Graesser, A. (2012). Dynamics of affective states during complex learning. *Learning and Instruction, 22*(2), 145–157.

European Commission. (2025). *Guidelines on prohibited artificial intelligence practices established by Regulation (EU) 2024/1689 (AI Act)*. [Add the official reference number and URL from the Commission website.]

Eysenck, M. W., Derakshan, N., Santos, R., & Calvo, M. G. (2007). Anxiety and cognitive performance: Attentional control theory. *Emotion, 7*(2), 336–353.

Federal Republic of Nigeria. (2023). *Nigeria Data Protection Act 2023*. [Add official gazette citation.]

Future of Privacy Forum. (2025). *Red lines under the EU AI Act: Unpacking the prohibition of emotion recognition in the workplace and education institutions*. https://fpf.org/blog/red-lines-under-eu-ai-act-unpacking-the-prohibition-of-emotion-recognition-in-the-workplace-and-education-institutions/

Hembree, R. (1988). Correlates, causes, effects, and treatment of test anxiety. *Review of Educational Research, 58*(1), 47–77.

Kapoor, S., & Narayanan, A. (2023). Leakage and the reproducibility crisis in machine-learning-based science. *Patterns, 4*(9), 100804.

Kizilcec, R. F., & Lee, H. (2022). Algorithmic fairness in education. In W. Holmes & K. Porayska-Pomsta (Eds.), *The ethics of artificial intelligence in education* (pp. 174–202). Routledge.

Kleinberg, J., Mullainathan, S., & Raghavan, M. (2017). Inherent trade-offs in the fair determination of risk scores. In *8th Innovations in Theoretical Computer Science Conference (ITCS 2017)* (Article 43). Schloss Dagstuhl.

Liebert, R. M., & Morris, L. W. (1967). Cognitive and emotional components of test anxiety: A distinction and some initial data. *Psychological Reports, 20*(3), 975–978.

Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems, 30*, 4765–4774.

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2021). A survey on bias and fairness in machine learning. *ACM Computing Surveys, 54*(6), Article 115.

Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., Hutchinson, B., Spitzer, E., Raji, I. D., & Gebru, T. (2019). Model cards for model reporting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency* (pp. 220–229). ACM.

Putwain, D. W., von der Embse, N. P., Rainbird, E. C., & West, G. (2021). The development and validation of a new Multidimensional Test Anxiety Scale (MTAS). *European Journal of Psychological Assessment, 37*(3), 236–246.

Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act). *Official Journal of the European Union*.

Rosenthal, R., & Jacobson, L. (1968). *Pygmalion in the classroom: Teacher expectation and pupils' intellectual development*. Holt, Rinehart & Winston.

Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence, 1*, 206–215.

Shi, R., & Yuan, Y. (2025). Theoretical models, mechanisms, assessment methods, and intervention strategies for test anxiety: A narrative review. *Journal of Clinical and Basic Psychosomatics, 3*(2), 1–10. https://doi.org/10.36922/jcbp.3946

Slade, S., & Prinsloo, P. (2013). Learning analytics: Ethical issues and dilemmas. *American Behavioral Scientist, 57*(10), 1510–1529.

Spielberger, C. D. (1980). *Test Anxiety Inventory: Preliminary professional manual*. Consulting Psychologists Press.

Varoquaux, G., & Cheplygina, V. (2022). Machine learning for medical imaging: Methodological failures and recommendations for the future. *npj Digital Medicine, 5*, Article 48.

von der Embse, N., Jester, D., Roy, D., & Post, J. (2018). Test anxiety effects, predictors, and correlates: A 30-year meta-analytic review. *Journal of Affective Disorders, 227*, 483–493.

Zeidner, M. (1998). *Test anxiety: The state of the art*. Plenum.

EDITOR-IN-CHIEF
Prof. A.I. Mteiye [MALAWI]