Psychometric Evaluation of Multimodal AI Assessment of University Students' Well-being

Xingyu Wang, Dongbo Che, Xutang Tong and Wah Cheung Wong
The Hong Kong Polytechnic University, Hong Kong SAR, China

maxime.wang@polyu.edu.hk
daniel.che@connect.polyu.hk
mindy.tong@polyu.edu.hk
henrywc.wong@polyu.edu.hk

Abstract: Universities need credible evidence to support student well-being before difficulties disrupt learning. Self-report questionnaires remain essential, but fixed responses may omit context that students express in their own words. A multimodal model can process language, voice and video from a short interview, although agreement with a benchmark alone cannot establish that its outputs measure well-being. This work in progress sets out an evaluation of My Well-being My Zone, a voluntary, non-diagnostic system intended for private student reflection. The central question is whether its model-generated scores and feedback can be interpreted reliably and usefully for that purpose. We plan to recruit 1,300 university students to answer five recorded questions on self-awareness, emotions, stress and behavioural tendencies and to complete matched self-report measures. Independent experts will score responses and provide formative feedback using predefined criteria. Data from 1,000 students will support parameter-efficient fine-tuning of an open-source video-language model. In a separate 300-student hold-out sample, we will compare the fine-tuned model with its base model on identical inputs. We will examine score stability, generalisability across questions and raters, agreement with independent experts, theoretically expected relationships with self-report measures, and added agreement with experts beyond self-report. Independent raters will assess feedback relevance and actionability; subgroup analyses will examine uneven errors. A related 101-person employee pilot and a working prototype establish technical feasibility only. We expect fine-tuning to improve expert agreement and feedback quality, but student results are not yet available. The evaluation will distinguish agreement with expert ratings from evidence about the broader meaning or benefits of an assessment. By separating expert agreement, score meaning, feedback quality and subgroup errors, the study contributes an assessment-oriented validation protocol for multimodal AI in higher education, rather than treating predictive accuracy alone as evidence of well-being measurement. It will identify which claims the data support, where uncertainty remains and what further evidence would be needed before the system is used beyond voluntary developmental reflection.

Keywords: Multimodal AI assessment, Student well-being, Psychometrics, Formative assessment, Parameter-efficient fine-tuning, Automated video interview, Higher education

1. Introduction

Student well-being is associated with academic engagement and institutional support (Chaudhry et al., 2024). In Hong Kong, a trend study also found substantial mental health concerns among university students, although its cohort design does not establish causes (Pan et al., 2025). Universities therefore have a reason to offer opportunities for reflection before stress or disengagement interferes with study. The proposed assessment is intended for that preventive, non-clinical purpose.

Self-report questionnaires give students a direct way to describe experiences that an observer cannot infer from appearance or speech. Yet fixed-response items may leave out the circumstances and meaning of a recent difficulty. When measures share a respondent and format, common method effects may also influence their associations (Podsakoff et al., 2003). A short interview could offer another source of information, but the fact that a model can process video does not show that visual or vocal features measure well-being.

Automated video-interview research provides a useful methodological precedent, but it largely concerns employment rather than students. Hickman et al. (2022) examined reliability, validity and generalisability in personality assessment. In a later study of cognitive ability, verbal behaviour was more informative than nonverbal or paraverbal behaviour, and validity evidence was mixed (Hickman et al., 2025). Neither finding can be transferred directly to student well-being. The gap is therefore not merely whether a model predicts expert scores, but whether its scores and feedback support the particular interpretation and use proposed here.

Our central research question is whether a fine-tuned multimodal model can produce scores and feedback that support private student reflection with adequate psychometric evidence. RQ1 asks how stable the scores are across repeated assessments and questions, and how closely they agree with independent expert ratings. RQ2 asks whether the scores show the expected relationships with matched self-report measures and add agreement with expert ratings beyond those measures. RQ3 asks whether fine-tuning improves feedback relevance and

actionability without concealing uneven errors across student groups. These questions separate evidence about scoring, score meaning and feedback rather than treating any single accuracy estimate as validation.

2. Related Work

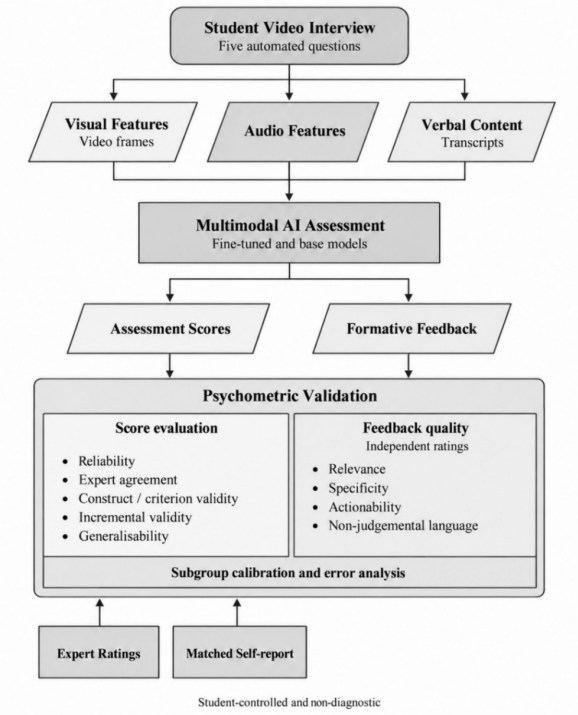
Well-being includes positive functioning as well as distress (Ryan and Deci, 2001). Flourishing and affect measures provide established ways to examine positive functioning (Diener et al., 2010). A recent systematic review found that higher-education studies often rely on clinical symptom instruments while positive well-being and educational experiences receive less attention (Baik et al., 2026). This supports a developmental assessment that considers students' accounts of strengths and difficulties, but it does not justify inferring an internal state directly from a facial or vocal cue.

Automated interview studies show why technical performance must be examined together with measurement quality. Models trained on recorded responses can produce scores related to human ratings, yet their reliability and validity vary by construct and input mode (Hickman et al., 2022; Hickman et al., 2025). Parameter-efficient fine-tuning can adapt a pretrained model with fewer trainable parameters than full fine-tuning (Hu et al., 2022). Whether that adaptation improves the interpretation of student interviews is an empirical question, not a consequence of the training method.

Validity concerns whether scores reflect the intended attribute and whether evidence supports their proposed interpretation and use, not simply their correlation with one benchmark (Borsboom et al., 2004; Messick, 1995; Kane, 2013). The Standards for Educational and Psychological Testing similarly require evidence and theory relevant to an intended use (American Educational Research Association et al., 2014). For this system, expert ratings provide a comparison for interview-based scoring, while self-report provides a different account of the student's experience. Neither is error-free ground truth. The proposed evaluation therefore examines consistency, expected relationships and additional information, and assesses feedback as a separate student-facing output.

3. Methodology

This prospective study evaluates the assessment component of My Well-being My Zone. Students will answer five recorded questions concerning self-awareness, emotions, stress and behavioural tendencies and complete matched self-report measures. Before recruitment, the team will specify what each rating is meant to represent and align the interview prompts, expert rubric and analyses with those interpretations. The system is intended to support voluntary reflection, not diagnosis or decisions about academic progression. A working prototype processes recorded responses into structured ratings and student-facing reports. This demonstrates that the proposed assessment workflow can be implemented; it does not establish the reliability or validity of its outputs.



We plan to recruit 1,300 students. In the 1,000-person development sample, trained experts will independently score the five responses per student and write brief formative feedback using predefined criteria. We will estimate agreement between experts rather than treat their ratings as error-free. The separate 300-person hold-out sample will remain outside all model training and selection. A subset of students will repeat the interview to estimate short-interval score stability.

With informed consent, recordings will be processed as video frames, audio features and transcripts. Parameter-efficient fine-tuning will adapt an open-source video-language model to the development data (Hu et al., 2022). The fine-tuned and base models will receive identical inputs from each hold-out participant and will produce ratings and concise formative feedback. The participant-level split prevents responses from the same student appearing in both development and evaluation data.

For RQ1, we will report test-retest estimates, variation across questions, expert inter-rater agreement and model-expert agreement, with uncertainty intervals. For RQ2, prespecified correlations will test theoretically expected convergence and separation among model scores and matched self-report measures. Hierarchical analyses will then ask whether model scores explain additional variation in hold-out expert ratings after self-report scores are entered. Because the experts rate the same interviews, this analysis tests added agreement with expert interpretation; it is not evidence that the model predicts an independent well-being outcome.

For RQ3, raters who did not generate the feedback will judge its relevance, specificity, actionability and non-judgemental language. We will compare the two models on the same hold-out interviews and inspect unsupported or overly certain statements. Subgroup estimates and error cases will be reported with uncertainty, since a 300-person hold-out sample may not support precise estimates for small groups. The evaluation plan will be fixed before examining hold-out results.

Table 1: Planned sources of validation evidence

Domain	Evidence	Interpretation tested
Reliability	Repeat interviews; five questions	Stability and question dependence
Score meaning	Expert ratings; self-report	Agreement and expected relationships
Incremental value	Model and self-report scores	Added agreement with experts
Feedback and fairness	Independent ratings; subgroup errors	Actionability and uneven performance

4. Preliminary Evidence and Planned Evaluation

Student data collection and validation are pending. A related pilot with 101 employees tested a feature-engineering pipeline using five recorded responses, human annotations and manager-reported performance. It found acceptable reliability and associations with the external criterion in that employment setting. The pilot supports the feasibility of collecting and processing interview data, but neither its population nor its target construct establishes validity for student well-being. The working student-facing prototype likewise shows that ratings can be converted into reports, not that the reports are accurate or helpful.

The student hold-out study will report the three research questions separately. We expect fine-tuning to improve agreement with experts and feedback quality relative to the base model, but these are hypotheses. An improvement in average agreement could coexist with poor test-retest stability, little added information beyond self-report or uneven subgroup errors. Null and adverse findings will therefore be retained rather than folded into a single performance score.

5. Discussion

The first interpretation concerns score consistency. If ratings change markedly when a student repeats a comparable interview, or depend heavily on the particular question, close average agreement with experts will not be enough for individual reflection. Hickman et al. (2022) show the value of examining reliability and generalisability alongside model-expert relationships. Our hold-out design applies that standard to a different construct and population. Since expert agreement itself can be limited, the analysis will report the human benchmark's variability before judging whether a model approaches it.

The second interpretation concerns what the scores mean. Theoretical relationships with self-report can support a proposed construct interpretation, but agreement with expert ratings of the same videos cannot establish that the scores capture well-being outside the interview. This distinction follows the view that validity evidence must

support the intended interpretation and use (Messick, 1995; Kane, 2013). If model scores add little agreement with experts after self-report is considered, the interview may not justify its additional burden. If they add agreement, the result would still call for independent student outcomes and longitudinal evidence before wider claims are made.

The third interpretation concerns the report a student actually receives. A score that resembles an expert rating may still be translated into generic or overconfident feedback. The planned independent feedback ratings therefore test relevance and actionability directly. The same caution applies to fairness. Hickman et al. (2025) found that the usefulness of behavioural modes and validity evidence varied in a different automated-interview setting. We will not infer that visual or vocal inputs improve this assessment merely because the complete model uses them, and subgroup analyses will identify possible disparities rather than certify their absence.

These results would guide a limited practical decision. Consistent scores, theoretically coherent relationships, useful feedback and no clear pattern of unequal errors would support further evaluation of voluntary student reflection. Weak evidence in any of these areas would narrow the interpretation or warrant redesign. This is especially important because self-report remains the student's account of an internal experience and cannot simply be replaced by an algorithmic rating (Baik et al., 2026).

The proposed study has several limits. A single-institution sample may not represent other languages or educational settings. Interview comfort, disability, recording quality and expert assumptions may affect ratings. The design does not isolate the added value of audio or video over transcripts, and it does not test whether using the reports improves well-being. Future work should compare input modes, test independent and longitudinal outcomes, and replicate the evaluation across institutions before extending the system's use.

6. Conclusion

This work in progress proposes a hold-out evaluation of whether a multimodal model can generate consistent scores and useful feedback for private student reflection. Its contribution is to make the limits of that claim testable: expert agreement, relationships with self-report, added information, feedback quality and subgroup errors will be reported separately. Until student results and independent outcome evidence are available, technical feasibility and employee-pilot findings cannot justify broader assessment use.

Ethics Declaration: University ethics approval has been obtained. Participation will require informed consent and allow withdrawal. Recordings and individual outputs will not be used for academic, disciplinary or clinical decisions. Students reporting serious distress will be directed to qualified human support.

AI Declaration: OpenAI's GPT-6 Sol (high reasoning effort) assisted with initial drafting, language editing and manuscript organisation. It was not used to generate or analyse study data, conduct statistical analyses or make research decisions. The authors reviewed and revised AI-assisted text, checked factual claims and references, and accept responsibility for the final manuscript.

References

American Educational Research Association, American Psychological Association and National Council on Measurement in Education (2014) *Standards for Educational and Psychological Testing*. Washington, DC: American Educational Research Association.

Baik, C., Ryan, T. and Lynch-Wells, S. (2026) 'Assessing student mental health and wellbeing in higher education: A systematic review of measurement approaches', *Educational Research Review*, 52, 100823. doi:10.1016/j.edurev.2026.100823.

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

Chaudhry, S., Tandon, A., Shinde, S. and Bhattacharya, A. (2024) 'Student psychological well-being in higher education: The role of internal team environment, institutional, friends and family support and academic engagement', *PLOS ONE*, 19(1), e0297508. doi:10.1371/journal.pone.0297508.

Diener, E., Wirtz, D., Tov, W., Kim-Prieto, C., Choi, D. W., Oishi, S. and Biswas-Diener, R. (2010) 'New well-being measures: Short scales to assess flourishing and positive and negative feelings', *Social Indicators Research*, 97(2), pp. 143-156. doi:10.1007/s11205-009-9493-y.

Hickman, L., Bosch, N., Ng, V., Saef, R., Tay, L. and Woo, S. E. (2022) 'Automated video interview personality assessments: Reliability, validity, and generalizability investigations', *Journal of Applied Psychology*, 107(8), pp. 1323-1351. doi:10.1037/apl0000695.

Hickman, L., Tay, L. and Woo, S. E. (2025) 'Are automated video interviews smart enough? Behavioral modes, reliability, validity, and bias of machine learning cognitive ability assessments', *Journal of Applied Psychology*, 110(3), pp. 314-335. doi:10.1037/apl0001236.

Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L. and Chen, W. (2022) 'LoRA: Low-rank adaptation of large language models', in *International Conference on Learning Representations*. Available at: https://openreview.net/forum?id=nZeVKeeFYf9.

Kane, M. T. (2013) 'Validating the interpretations and uses of test scores', *Journal of Educational Measurement*, 50(1), pp. 1-73. doi:10.1111/jedm.12000.

Messick, S. (1995) 'Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning', *American Psychologist*, 50(9), pp. 741-749. doi:10.1037/0003-066X.50.9.741.

Pan, J.-Y., Wang, N. and Shum, M. H. Y. (2025) 'Mental health of Hong Kong university students: A 5.5-year trend study across the COVID-19 pandemic', *International Journal of Adolescence and Youth*, 30(1), 2496445. doi:10.1080/02673843.2025.2496445.

Podsakoff, P. M., MacKenzie, S. B., Lee, J.-Y. and Podsakoff, N. P. (2003) 'Common method biases in behavioral research: A critical review of the literature and recommended remedies', *Journal of Applied Psychology*, 88(5), pp. 879-903. doi:10.1037/0021-9010.88.5.879.

Ryan, R. M. and Deci, E. L. (2001) 'On happiness and human potentials: A review of research on hedonic and eudaimonic well-being', *Annual Review of Psychology*, 52, pp. 141-166. doi:10.1146/annurev.psych.52.1.141.