



The Same Voice, Different Scores: An Autoethnography of Mood, Examiners, and Real-Time Performance in IELTS Speaking and Listening

Muhammad Isa* 

English Study Program (D3), State Polytechnic of Madiun, East Java 61312, Indonesia

ARTICLE INFO	ABSTRACT
Received: 21 Mei 2026	The research examines why the IELTS Speaking and Listening scores of a single candidate differed when he took the tests in 2021 and 2025, even though his English proficiency did not change. The author is an English lecturer in Indonesia with bipolar disorder. The research adopts the methodology and stance of autoethnography and is based on six full IELTS exams and one single-skill Listening retake, supplemented by the autobiographical memory of the day the test was taken and the artifacts of the preparation prepared by the author. It reduces to speaking and listening, since the two skills are practiced on the spot and are time-constrained, where immediate focus and impact are most strongly exerted on performance. The results show that speaking and listening were between band 6 and 7.5, and between 6 and 8, and in most instances, just months apart. It is most easily isolated when listening, with no examiner to be taken into account, and with the same skill obtaining 6 at the close of a long testing day and 8 two weeks later when alone. The human element is added through speaking, as the oral interview is co-constructed with an examiner. The author had thought he was rated more severely by non-native examiners, but his self-ratings do not bear this out, which he reads as internalized native-speakerism that turned self-fulfilling on bad-mood days. The research concludes that single-sitting Speaking and Listening bands can indicate the condition and state of a test day as much as ability with implications of test fairness, preparation, and teacher consciousness when candidates have a mood disorder.
Accepted: 01 July 2026	
Published: 06 August 2026	
Keywords: Autoethnography, IELTS Speaking and Listening, Foreign Language Anxiety, Examiner Co- construction, Native- speakerism	

Copyright © 2026 The Author(s)

Published by CV Wahana Publikasi.

This article is an open access article under the [CC BY-NC-SA 4.0 License](https://creativecommons.org/licenses/by-nc-sa/4.0/)

1. Introduction

IELTS is a high-stakes assessment: Speaking and Listening bands are read as a stable representation of a candidate's English proficiency, and decisions about study, migration, and employment can follow from a single test day. The author is an English lecturer who has taken the IELTS five or six times, including a single-skill retake, and scored between Band 6 and 7.5 in Speaking and between 6 and 8 in Listening across sittings spread unevenly, from weeks to years apart, between 2021 and 2025. By 2021, the author had already met his general proficiency requirement, yet the scores kept moving. This variation across repeated sittings by the same candidate, without a corresponding change in underlying proficiency, is the central problem this research addresses. Speaking and Listening are demonstrated at the moment, and there is no time to reread a passage or plan and make changes to a draft, so immediate attention and mood are among the most pressing factors affecting performance (Fu et al., 2025; MacIntyre & Gardner, 1994; Teimouri et al., 2019). In Speaking, an examiner is also present, and the interaction between candidate and examiner can shape the score that is finally recorded.

The author is diagnosed with bipolar disorder; it was initially diagnosed in 2018. Ordinary arguments about testing believe that a score is a form of approximation of a constant ability, and the error is viewed as a chance error (Kane, 2013). This record does not say that reading: the variation was systematic, and, according to mood, the circumstances of every test day, and, in Speaking, the face of the examiner. The language assessment literature has little to say about within-person variability in IELTS Speaking and Listening performance, particularly for a candidate whose test-day condition includes a mood disorder; this is the gap this research addresses. Because this variability sits within one person, it is best captured from the inside, through the author's own autobiographical memory and preparation artifacts, treated here as autoethnographic data. It is informed by three questions: (1) How did the Speaking and Listening scores vary across repeated sittings, and how did that relate to mood state? (2) Why might these two real-time skills, more than Reading and Writing, be sensitive to mood and momentary attention? (3) What does this within-person volatility imply for the fair interpretation of single-sitting scores, and for the preparation of candidates who live with a mood disorder?

2. Literature Review

2.1 *Real-Time Skills, Mood, and Accent*

Speaking and listening processing are online, time-limited and without re-reading and editing. This imbibes attention, working memory and in the moment affect more than the self-paced reading of a passage or the plan and revise writing of an essay. A meta-analysis has repeatedly found that anxiety is negatively related to achievement (Teimouri et al., 2019). Idiodynamic studies followed the dynamics of anxiety minute-by-minute during the IELTS speaking test (Fu et al., 2025). MacIntyre & Gardner (1994) It demonstrated that anxiety reduces performance at the time of use as opposed to underlying ability and that emotion in language work is now seen as context-structured as opposed to being privative (Benesch, 2017; Benesch & Prior, 2023). This is escalated to a clinical mood disorder, since the respective condition is not just a scenario of situational nervousness, but an internal undersetting that is taken into the room before the first inquiry (MacIntyre et al., 2020; Tao et al., 2025). Accent familiarity and speech rate are also sensitive to listening, as unfamiliar or fast accents increase the processing load and may reduce the comprehension scores independent of ability (Kang et al., 2024, 2025; Miao, 2023).

2.2 Examiner Co-Construction and Native-Speakerism

Unlike the listening test, the Speaking test involves live interlocutor and a lengthy line of assessment research considers the oral interview as collaborative production and not a performance and the interaction of the interlocutor and the raters is included within the score (Brown, 2003; Clapham & McNamara, 2000). Knowledge of the accent of a speaker affects rater judgements in themselves (Carey & Szocs, 2024). Here ideology intrudes, as well. The perception that a native-speaking test-taker is the more rightful or more fair one is a type of native-speakerism (Holliday, 2006), and new assessment scholarship has made the same argument by revealing the existence of two sides to the story of native-speakerism (Isaacs & Rose, 2022). This belief can affect the performance of a candidate as shown in this research even when it is not described as being the behavior of the examiner.

2.3 Validity and Fairness

The test has no validity but there is inferences based on scores (Kane, 2013), and the Standards of Educational and Psychological Testing defines a bias as a construct-irrelevant variance that differs among the test-takers (American Educational Research Association et al., 2014). The recent scholarship on language assessment considers fairness and accommodation to be a design issue instead of an afterthought (Lovett & Harrison, 2023; Taylor & Banerjee, 2023), and critical scholarship has called the ableism that the assessment regimes may entail (Nieminen, 2023). The idiodynamic test anxiety indicates the same as construct-irrelevant variance (Arai et al., 2026). In the case of a candidate whose performance is dependent on a clinical condition, the day rather than the ability can be reported by a single band.

2.4 Conceptual Framework

Combined, these strands set up the conceptual framework that drives this research: a band score in a real-time skill is a combined product of three layers, and only the deepest is the proficiency of the candidate that is stable. The first one is the affective state of the candidate that may be altered by a mood disorder to the level lower than that of normal nerves prior to the exam itself (MacIntyre & Gardner, 1994; Tao et al., 2025). The second is real-time task demand, which entails temporary attention and processing demands, e.g., the demand due to a foreign or fast accent, direct path to the score (Teimouri et al., 2019; Miao, 2023). The third one that is present in Speaking but absent in Listening is the interactional state of the interaction with an examiner, in which co-construction and beliefs of the candidate about such an examiner affect performance (Brown, 2003; Holliday, 2006; Isaacs & Rose, 2022). The theory of validity offers the lenses that identifies the outcome: a change in a score is construct-irrelevant variance when it changes with these layers rather than proficiency (American Educational Research Association, 2011; Kane, 2013). In the study, the three-layer framework is used as it reads a decade of one candidate Speaking and Listening scores.

3. Method

It is a hybrid evocative and analytic autoethnography (Adams et al., 2021; L. Anderson, 2006; Chang, 2016; Ellis et al., 2011), which has a long-standing tradition in the field of applied linguistics and TESOL, and a history of investigating the concept of teacher identity and emotion using single situated narratives (R. W. Anderson, 2024; Ghanbar & Kamali, 2026; Keles, 2022; Liu et al., 2025). The evocative strand is realized through first-person narrative reconstruction of the mood, sensation, and interaction of each test day, while the analytic strand is realized through systematic comparison of scores, dates, and testing conditions

across sittings. The data consists of six complete IELTS tests and a single-skill retake of Listening in between 2021 and 2025 (Table 1) that forms the score reports and the autobiographical memory of all the test days and study artifacts of the author. It is analyzed in two ways: chronologically, a narrative of the sittings, and analytically, a reading that is mindful of the skills in motion with the state and conditions, in which the isolated Listening retake is viewed as a within-person comparison that keeps the candidate constant but varies the testing condition. The chronological reading orders the score reports and remembered detail of each sitting into a single narrative, from which the pattern of rises and falls across 2021 to 2025 emerges. The analytical reading then cuts across that narrative to compare sittings against one another on mood, processing demand, and examiner presence, treating the Listening retake taken two weeks after the June 2024 sitting as a within-person comparison in which the candidate and his underlying proficiency are held constant while the testing condition varies; the two readings together generate the findings on mood, real-time processing, examiner interaction, and score variability reported in Section 4. Examiner details are mentioned as what the author perceives at the time and not a fact.

3.1 Researcher Context and Ethics

The researcher occupies both the researcher and participant positions. He is an English teacher at one of the state vocational higher education institutions in East Java, Indonesia, where he teaches English as a specific purpose, and has a master degree in English language teaching. He is bipolar (was diagnosed in 2018), and in 2023 experienced an episode of psychotherapy, which is reflected in the story. The study is an autoethnographic self-study, so there were no other human subjects and they required consent to participate in the study; the author concurs to utilize his own data, even disclosing a mental health problem. Memories of mood, and of examiners, are re-created and not recorded at the moment, and are subjected to treatment as impressions all the way through.

4. Results and Discussion

4.1 How the Scores Varied Across the Sittings

Speaking and Listening bands moved up and down across the seven sittings, without a steady rise or fall, and the perceived examiner and testing centre changed from one sitting to the next. The narrative below follows these sittings in order. Table 1 summarizes the pattern:

Table 1. IELTS Speaking and Listening Scores, 2021 to 2025

Date	Centre	Examiner (perceived)	Speaking	Listening
2021	IDP Malang	Elderly Australian man	7	7.5
Jul/Aug 2023	IDP Jogja	Asian man	6	7
Dec 2023	British Council Jogja	Asian woman	7	7
Jun 2024	IALF Malang	Australian	7.5	6
~2 weeks later, 2024	IDP Surabaya	Listening retake	-	8
Late 2024	IDP Jogja	Indian	6	7
Jul 2025	IDP Malang	Native-speaking man	7	7

Note. Examiner descriptions are the author's impression at the time and were not verified. IELTS examiners are trained and certified against a common rubric, and many are not native speakers by birth.

As of 2021, I was not studying the test in the same manner that I used to. At this point my preparation was a daily routine: picking up the language spoken by the native speakers, the collocations and informal phrases that I heard on YouTube and on the British and Australian podcasts and forcing myself to use them rather than just writing them down in a notebook. At IDP Malang, the old Australian man was the one to conduct Speaking exam. In Part 1 I have employed the connector, *mind you*, instead of using *but* to join two thoughts. He lifted his eyes, and the features were drawn apart in what may be termed astonishment, that he had not expected that little British introducer of an Indonesian candidate. I took the glance to mean acceptance and it relieved something in me. My Speaking, which I had always thought impossible, and my Listening, rose together, 6 to 7.5.

In mid 2023, my earlier score had lapsed, so I was back at IDP Jogja, and my Speaking was down to 6. I thought to myself it was the examiner, an Asian man, in the room. But that is the year when my bipolar disorder and anxiety entered the scene, a psychologist informed me that I was shaken by a new work culture, and I began a few months of psychotherapy. I had a Listening of 7 the day before. I had a woman, whom I thought to be Asian, as my examiner in the British Council, Jogja, in December 2023, when I had been very hasty, and had the two mountain hikes that evened out my mood, and Speaking was 7 and Listening was 7. I achieved my peak of 7.5 in Speaking, with an Australian examiner in June 2024, in IALF Malang, but only 6 in Listening, against fast, India-accented audio. Two weeks later, during one of my Listening retakes at IDP Surabaya, I got up to 8 after two hours of train practice. A few months later at the IDP, Jogja where I had a Scorer who I believed was an Indian, my Speaking was 6 and my Listening was 7. With a native speaking man, both of them finally reached the study abroad threshold in July 2025 in IDP Malang. Read across the sittings, the scores move without a matching change in preparation or ability, tracking mood and the condition of each test day instead; this is consistent with the wider finding that anxiety and momentary state, not underlying competence, drive performance on the day (MacIntyre & Gardner, 1994; Teimouri et al., 2019).

4.2 *Why the Real-Time Skills Move*

No point is cleaned so well as through listening, as there is neither examiner nor interlocutor. It ranked a 6 on a long test day of fast, foreign accents in June 2024; two weeks later, it alone after practice, this skill rated an 8. It also had something to do with accent, as accents that are unfamiliar and fast have been known to decrease listening scores due to processing load and not due to proficiency (Kang et al., 2024; Miao, 2023). There was nothing in fourteen days that altered my ability; only the circumstances in which the art was tested. This is the construct-irrelevant variance that current scholarship has come to refer to (Arai et al., 2026; Nieminen, 2023) and it is the most powerful single evidence in the record (Kane, 2013). Reading and Writing are external to this story in the sense that they support it: Writing did not vary at all over the same years or even on a particular retake, and Reading varied less. The two live and time pressure skills are the ones that are volatile.

Mood and attention also help explain sittings beyond the isolated Listening retake, which speaks directly to RQ2. In mid-2023, arriving anxious and newly navigating a mood episode, my Speaking fell to 6; by December 2023, after two hikes that had settled my mood, both Speaking and Listening returned to 7. The June 2024 sitting on which the retake comparison rests was also the end of a long testing day, when sustained attention across four skills had already been drawn down before Listening began, whereas the retake was sat alone, rested, and after two hours of focused practice. Read together, these sittings suggest that mood state

and momentary attention, and not only accent and processing load, tracked the movement in scores (Fu et al., 2025; MacIntyre & Gardner, 1994; Teimouri et al., 2019).

4.3 Examiners, Native-Speakerism, and Fairness

There is human layer added by speaking to that state. The oral test is designed in such a way that it involves two people and assessment research takes into account the interviewer behavior and rater judgment as an element of what the score is (Brown, 2003; Clapham & McNamara, 2000). A little event in 2021 was the turning point: when an examiner gave a slight shudder at one of my phrases, it raised my spirits, and these spirits accompanied me through the rest of the interview. I went into this study with the belief that it was non-native examiners who lowered my marks. To such a belief my data do not live. In December 2023, my examiner whom I thought was Asian scored 7 level by my Australian examiners, and in 2024, my examiner whom I thought was Indian scored 6. Examiner origin does not track my scores; my state does. Even the supposition that a native examiner is the fairer one is a sort of internalized native-speakerism (Holliday, 2006; Isaacs & Rose, 2022), and on a bad-mood day it became a self-fulfilling prophecy: I was anticipating a tougher test, so I tightened my muscles and behaved like it was tougher. The recoveries, theirs, were following a goal-oriented recovering, a series of psychotherapy, and gains that refilled my mood, which, as a candidate with a mood disorder, cannot be considered a wellness in itself, but, instead, a component of test preparation. These conditions, when combined into a single pattern, would mean that a single Speaking or Listening band would report the conditions and state of a test day as many times as an ability does (American Educational Research Association, 2011; Taylor & Banerjee, 2023).

5. Conclusion

Ten body scores indicate that single sitting Speaking and Listening bands can convey the state and the situation of a test day as much as aptitude of a candidate, and that it is sharpest in situations where the output of a candidate is related to mood disorder. Listening, 6 in a full test day, 8 in isolation, two weeks apart is the case that makes the case almost experimental, the within-person contrast. This translates to be careful in interpreting one band as a fact and be accommodative to the test users. Real goals in teaching and preparation in listening are accent familiarity and speed of processing, mood self-regulation is a goal in preparation, the native-speaker assumption is a question to be explored, rather than blindly trusted, in that it can affect performance within. It is a study that is interpretive, not statistically generalizable, its mood states are recreated, not recorded, and its examiner description are unconfirmed impressions, as an isolated case. These are the right limits of the method and to where larger studies can continue.

Funding

This research received no external funding.

Conflict of Interest

The author declares that there is no conflict of interest.

Authors' Contribution

The author made substantial contributions to the conception and design of the study. The author took responsibility for data collection, analysis, interpretation, and discussion of results, and read and approved the final manuscript.

Acknowledgments

The author used a generative AI assistant (Claude, Anthropic) to help locate literature, organize and draft portions of the text, and format the manuscript; the author conceived the study, provided all data, wrote the narrative, verified all cited sources, and is responsible for the final content.

References

- Adams, T. E., Jones, S. H., & Ellis, C. (Eds.). (2021). *Handbook of autoethnography* (2nd ed.). Routledge. <https://doi.org/10.4324/9780429431760>
- American Educational Research Association. (2011). *Report and recommendations for the reauthorization of the Institute of Education Sciences*. American Educational Research Association.
- Anderson, L. (2006). Analytic autoethnography. *Journal of Contemporary Ethnography*, 35(4), 373–395. <https://doi.org/10.1177/0891241605280449>
- Anderson, R. W. (2024). My most intense years: An autoethnographic analysis of language teacher identities and emotions of one dissertating (adjunct) language instructor. *TESOL Journal*, 15(4), e845. <https://doi.org/10.1002/tesj.845>
- Arai, Y., Takizawa, K., & Paterson, R. (2026). Anxiety in an artificial intelligence (AI)-based English conversation test: An idiodynamic approach to explore potential construct-irrelevant variance. *Language Testing*, 43(3), 290–317. <https://doi.org/10.1177/02655322251397040>
- Benesch, S. (2017). *Emotions and English language teaching: Exploring teachers' emotion labor*. Routledge.
- Benesch, S., & Prior, M. T. (2023). Rescuing “emotion labor” from (and for) language teacher emotion research. *System*, 113, Article 102995. <https://doi.org/10.1016/j.system.2023.102995>
- Brown, A. (2003). Interviewer variation and the co-construction of speaking proficiency. *Language Testing*, 20(1), 1–25. <https://doi.org/10.1191/0265532203lt242oa>
- Carey, M. D., & Szocs, S. (2024). Revisiting raters' accent familiarity in speaking tests: Evidence that presentation mode interacts with accent familiarity to variably affect comprehensibility ratings. *Language Testing*, 41(2), 290–315. <https://doi.org/10.1177/02655322231200808>
- Chang, H. (2016). *Autoethnography as method*. Routledge. <https://doi.org/10.4324/9781315433370>
- Clapham, C. (2000). Measuring second language performance [Review of the book *Measuring second language performance*, by T. McNamara]. *TESOL Quarterly*, 34(2), 376–377. <https://doi.org/10.2307/3587967>
- Ellis, C., Adams, T. E., & Bochner, A. P. (2011). Autoethnography: An overview. *Historical Social Research*, 36(4), 273–290. <https://doi.org/10.12759/hsr.36.2011.4.273-290>
- Fu, L. J., Ducker, N. T., & Thomas, N. (2025). Have you prepared them psychologically? An idiodynamic investigation into foreign language anxiety in IELTS video-call speaking tests. *The Language Learning Journal*. Advance online publication. <https://doi.org/10.1080/09571736.2025.2578720>
- Ghanbar, H., & Kamali, J. (2026). Autoethnography in TESOL teacher education: A methodological synthesis. *TESOL Quarterly*. Advance online publication. <https://doi.org/10.1002/tesq.70058>

- Holliday, A. (2006). Native-speakerism. *ELT Journal*, 60(4), 385–387. <https://doi.org/10.1093/elt/ccl030>
- Isaacs, T., & Rose, H. (2022). Redressing the balance in the native speaker debate: Assessment standards, standard language, and exposing double standards. *TESOL Quarterly*, 56(1), 401–412. <https://doi.org/10.1002/tesq.3041>
- Kane, M. T. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement*, 50(1), 1–73. <https://doi.org/10.1111/jedm.12000>
- Kang, O., Miao, Y., Hirschi, K., & Ahn, H. (2025). Test-takers' perceptions of the inclusion of different English accent varieties in the IELTS listening tests. *RELC Journal*, 56(2), 362–376. <https://doi.org/10.1177/00336882241231310>
- Kang, O., Yan, X., Kostromitina, M., Thomson, R., & Isaacs, T. (2024). Fairness of using different English accents: The effect of shared L1s in listening tasks of the Duolingo English test. *Language Testing*, 41(2), 263–289. <https://doi.org/10.1177/02655322231179134>
- Keleş, U. (2022). Autoethnography as a recent methodology in applied linguistics: A methodological review. *The Qualitative Report*, 27(2), 448–474. <https://doi.org/10.46743/2160-3715/2022.5131>
- Liu, S., Yuan, R., & Wang, C. (2025). 'Let emotion ring': An autoethnographic self-study of an EFL instructor in Wuhan during COVID-19. *Language Teaching Research*, 29(1), 221–241. <https://doi.org/10.1177/13621688211053498>
- Lovett, B. J., & Harrison, J. R. (2023). Ensuring test access for students with disabilities: A special issue of *Assessment for Effective Intervention*. *Assessment for Effective Intervention*, 49(1), 3–6. <https://doi.org/10.1177/15345084231205314>
- MacIntyre, P. D., & Gardner, R. C. (1994). The subtle effects of language anxiety on cognitive processing in the second language. *Language Learning*, 44(2), 283–305. <https://doi.org/10.1111/j.1467-1770.1994.tb01103.x>
- MacIntyre, P. D., Gregersen, T., & Mercer, S. (2020). Language teachers' coping strategies during the Covid-19 conversion to online teaching: Correlations with stress, wellbeing and negative emotions. *System*, 94, Article 102352. <https://doi.org/10.1016/j.system.2020.102352>
- Miao, Y. (2023). The relationship among accent familiarity, shared L1, and comprehensibility: A path analysis perspective. *Language Testing*, 40(3), 723–747. <https://doi.org/10.1177/02655322231156105>
- Nieminen, J. H. (2023). Unveiling ableism and disablism in assessment: A critical analysis of disabled students' experiences of assessment and assessment accommodations. *Higher Education*, 85(3), 613–636. <https://doi.org/10.1007/s10734-022-00857-1>
- Tao, J., Xu, Y., Wang, S., & Gao, X. (2025). The emotions of language teachers: A systematic review of studies between 2015 and 2024. *Language Teaching*, 58(4), 450–473. <https://doi.org/10.1017/S0261444825000205>
- Taylor, L., & Banerjee, J. (2023). Accommodations in language testing and assessment: Safeguarding equity, access, and inclusion. *Language Testing*, 40(4), 847–855. <https://doi.org/10.1177/02655322231186221>
- Teimouri, Y., Goetze, J., & Plonsky, L. (2019). Second language anxiety and achievement: A meta-analysis. *Studies in Second Language Acquisition*, 41(2), 363–387. <https://doi.org/10.1017/S0272263118000311>