Assessing Writing 61 (2024) 100848

Contents lists available at ScienceDirect

Assessing Writing

journal homepage: www.elsevier.com/locate/asw

A teacher's inquiry into diagnostic assessment in an EAP writing course

Rabail Qayyum

University of Hawaii at Mānoa, USA

ARTICLE INFO

Keywords:
Diagnostic assessment
Argument-based validation
English for Academic Purposes writing
International students

ABSTRACT

Research into diagnostic assessment of writing has largely ignored how diagnostic feedback information leads to differentiated instruction and learning. This case study research presents a teacher's account of validating an in-house diagnostic assessment procedure in an English for Academic Purposes writing course with a view to refining it. I developed a validity argument and gathered and interpreted related evidence, focusing on one student's performance in and perception of the assessment. The analysis revealed that to an extent the absence of proper feedback mechanisms limited the use of the test, somewhat weakened its impact, and reduced the potential for learning. I propose a modification to the assessment procedure involving a sample student feedback report.

1. Introduction

Studies on diagnostic assessment of writing have documented how to identify, operationalize, and measure the underlying constructs (e.g., Dolgova & Siczek, 2019; Kim, 2011; Urmston et al., 2013), yet there remains much to learn about how to use diagnostic instruments (Pellegrino et al., 2016). There is a broad consensus that validity is the most important consideration in evaluating the quality of the uses and interpretations of the results of tests and assessments. Widely cited validity theorists like Kane (2013) and Pellegrino et al. (2016) have conceptualized validity as a multifaceted element that is not an instrument's inherent property; rather it lies in the interpretation of an instrument's use and impact. Whereas some of the current approaches have attempted to link validity to issue of fairness (e.g., Randall et al., 2024, argue for a social justice orientation to validity), the argument-based validation approach is a popular one (Dursun & Li, 2021) that lays out the interpretation and uses of an assessment procedure in as clear and complete as possible terms, and is a relatively straightforward process (Kane, 2021). Its general framework and some of the terminologies originate from the work of Toulmin (2003). It is a useful approach in listing the criteria for an assessment instrument and clearly detailing its limitations, so that the purpose of the instrument does not stray.

Among the different approaches in diagnostic assessment such as cognitive diagnostic assessment (Kim, 2011) and automated feedback (Koltovskaia, 2020), the approach I adopt is based on Alderson (2005) in which the focus is not only on identifying strengths and weaknesses of learners, but also on following up on these aspects. Therefore, in this paper, the term *diagnostic assessment procedure* is used to characterize the complete cycle of diagnosis, comprising five recurring stages: 1) define what is to be assessed, 2) operationalize it, 3) conduct diagnostic assessment, 4) give feedback, and 5) implement feedback (Huhta et al., 2024).

Research into diagnostic assessment of writing (e.g., Kim, 2011; Knoch, 2011; Xie, 2017) has largely ignored insights generated after a test is administered and how those insights aid in adjusting learning goals, lesson plans, or assignments. Moreover, few studies

E-mail address: rqayyum@hawaii.edu.

https://doi.org/10.1016/j.asw.2024.100848
Received 5 January 2023; Received in revised form 25 April 2024; Accepted 7 May 2024
Available online 30 May 2024

are teacher-driven initiatives, situated within classroom contexts. To bridge these gaps, this case study research is a teacher's attempt at validating an in-house diagnostic assessment procedure for an English for Academic Purposes (EAP) course by placing a student's perspective at the center.

By and large, EAP programs are the first among the few opportunities for international students to receive university-level writing instruction. Indeed, these programs do some heavy lifting, and the instructors here have to meet substantial expectations (Fox, 2009). In this scenario, diagnostic assessment is one avenue to provide beneficial information to teachers, students, and program administrators about students' individual needs. This paper describes how I designed and implemented a validity argument to explore the value and impact of the existing diagnostic assessment procedure. My goal is to provide a unifying framework to classroom instructors for spearheading an effort into investigating the validity of their own diagnostic assessments.

In the next section, I describe the findings of key studies on diagnostic assessment of writing and the gaps in it, to build grounds for how the present study addresses those gaps.

2. Diagnostic assessment of EAP writing

One issue with scholarship on diagnostic assessment of writing is that it often carries a portrayal of diagnostic assessment in both theory and practice that renders it indistinguishable from other assessment forms. In this paper, *diagnostic assessment* refers to assessment with the function of making inferences about strengths or weaknesses in learners' writing abilities (encompassing linguistic abilities) in order to assign targeted learning activities to yield positive changes in learning (Jang & Sinclair, 2021).

Validity is the fundamental consideration in not only developing tests, but also in evaluating them. Two useful examples for exploring validity of diagnostic instruments are provided by Chapelle et al. (2015) and Knoch and Elder (2016). Chapelle et al. (2015) provide argument-based validation examples of two automated-scoring instruments for diagnostic assessment of undergraduate and graduate academic writing. However, these instruments were designed for a particular writing task, and it is unclear how the application of diagnostic feedback generated extended over a semester. On the other hand, Knoch and Elder (2016) analyzed two diagnostic instruments, namely *Diagnostic English Language Assessment* and *Measuring the Academic Skills of University Students*–termed post-entrance language assessments (PELAs)–using an argument-based validity framework. They determined that neither instrument fully satisfied their established criteria.

Among few studies that explore diagnostic feedback implementation and use, Doe's (2015) study is one notable example. The researcher examined how students interpreted diagnostic assessment in an EAP course at one Canadian university. Drawing on the assessment use argument, this study probed the extent to which diagnostic test feedback enabled students to beneficially adjust their study practices to improve their academic skills. The findings related to two assessment claims: interpretations and consequences. The interpretations claim addressed whether students' strengths and weaknesses in reading, listening, and reading were meaningful and relevant to student learning. The consequences claim examined whether using the diagnostic test in the EAP course was beneficial to the students. Students' responses to reflection tasks and interviews were used to support or refute the underlying warrants for these claims. Overall, the researcher found that the students interpreted the diagnostic feedback appropriately and benefited from it within a supportive content that enabled them to adjust their study practices.

Other studies that explore diagnostic feedback applications frequently involve implementation of large-scale interventions or introduction of campus-wide policies. Fox et al. (2016) reported on a multistage-evaluation mixed methods study, which evaluated the impact of a diagnostic assessment procedure on the first-year experience, student engagement, achievement, and retention in an undergraduate engineering program. Their study integrated EAP assessment with mathematical knowledge, and the participants included all students, not just international. In another study, Fox (2009) carried out a large-scale mixed methods study examining the role diagnostic assessment played in moderating policy impact and designing EAP curricular reform. She ascertained that diagnostic assessment was, to a degree, able to mitigate the negative impact of the admission policy.

Overall, there is a need for more small-scale qualitative studies that delve into applications of diagnostic assessment that is more closely tied to academic course settings. Such an exploration will also be useful because assessments designed locally to address local initiatives and contexts are more likely to portray those contexts accurately and treat the stakeholders fairly than are large-scale assessments (Conference on College Composition and Communication CCCC, 2022). Placing diagnostic assessment in classroom contexts and curricular goals will allow a connection of diagnostic feedback with actual actions that students and teachers can take to promote learning. In such micro-level applications of diagnostic feedback, decision making as well as the decision makers will be noteworthy factors in the course of the procedure.

Since teachers are key stakeholders in the diagnostic assessment procedure, a teacher-driven initiative can guide further thinking about and research into diagnostic assessment. A teacher's viewpoint merits examination because "when purposes are set and outcomes measured by those who are beyond the classroom, the teacher is reduced to an element in the equation" (Freeman, 1998, p. 58). In comparison to macro-level feedback applications, a classroom-based application can bring out the teacher as the driver and controller of the assessment processes, which is crucial for teacher empowerment (Crusan, 2010). For diagnostic assessment specifically, as proposed by Alderson et al. (2015a), the first principle states that it is not the instrument, but the user who diagnoses.

Furthermore, relegating assessment as the domain of testing experts only and not of writing teachers could be problematic because "testing companies often mysticize assessment through the use of statistical and measurement jargon, making assessment a scary proposition for both teachers and students" (Crusan, 2010, p. 257). Edelenbos and Kubanek-German (2004) emphasize fostering language teachers' diagnostic competence, which remains a neglected area in teacher education (Jang & Sinclair, 2021).

To address these concerns, this case study research is a teacher's attempt at validating an in-house developed diagnostic assessment procedure within an EAP course. Using an argument-based approach, the research integrates attempts I, the teacher, made to make

sense of diagnostic test information with a focal student's perspective of receiving, comprehending, valuing, and acting upon the feedback. Both these perspectives are vital because students' understandings may be at odds with those of their teachers (Carless, 2006), enabling different layers of the situation to emerge.

3. The present study

The approach of the present study is a case study (Creswell, 2013), which enables close investigation of dynamic social processes as they evolve in the research setting and can provide rich holistic data that contribute to the understanding of complex situations. The strengths of case study method are in its ability to gain an insider's viewpoint during the research process, more in-depth and nuanced findings based on that, and in its flexibility in using different methods.

In contrast to the existing diagnostic assessment research dominated by the quantitative paradigm, the case study design can help to answer how diagnostic feedback information leads to differentiated instruction and learning, how evidence can be sought and interpreted for this purpose, and why certain subskills are more valued. A rich and descriptive case of a classroom-based and course-embedded approach to diagnostic assessment can promote learning by giving a practical framework for investigating validity, elaborating on the nature of evidence required for this purpose and how to analyze it. Case study research is not only an ideal approach to researching diagnostic writing assessment, but it can also capture the richness of data necessary to understand the multifaceted aspects of the classroom environment. There is increasing recognition of the situated nature of classroom-based assessment, and a case study research can render that context open.

At the time of data collection, I was a doctoral student in applied linguistics and taught the EAP course as part of my graduate assistantship. I had substantial experience of teaching writing to multilingual learners, tutoring at the writing center, and delivering workshops and attending conferences related to my field. These experiences had given me an insight into the needs of graduate students in general and international students in particular (I happened to be one myself). I had designed university entrance examinations at my previous place of employment.

I taught the EAP course for the first time in fall 2021 and again in spring 2022. This study deals with spring 2022. As I started teaching the course, I became interested in exploring the value and impact of the existing diagnostic assessment procedure to make explicit its expected benefits, provide evidential basis of its expectations, and improve my practice. The point of this exercise was to illustrate an argument-based framework, in which the intended interpretation and use of the existing diagnostic assessment procedure is laid out in some detail, with an emphasis on the inferences being made and the assumptions required for these inferences to be plausible. I sought a coherent rationale for using the existing instrument. It was my understanding that this instrument had been used in past iterations of the course, but a validity investigation had not been carried out before. With these objectives, the study addresses the following research question: To what extent is the existing diagnostic assessment procedure valid for assessing international graduate students' EAP writing?

Since the present study takes a grassroots level of analysis and provides a glimpse inside a class, it is necessary to first outline the specific university and course context within which the diagnostic assessment procedure was situated.

4. Context

4.1. Institutional background

The study took place at a large state university in the Pacific region. 535 international graduate students were enrolled in fall 2021. All doctoral students must write a dissertation, and some master's students are expected to write a thesis or a scholarly paper. Incoming students who have English as an additional language and do not meet any of the university's criteria for automatic exemption (which include evidence of academic English proficiency e.g., a minimum score of 7 on International English Language Testing System, IELTS) take a placement test and are placed into, or exempted from, EAP courses. Although these courses are non-credit bearing, they are mandatory for completing degree requirements. They cover intermediate and advanced levels of three skill-based areas: listening and speaking, reading, and writing. The course under discussion is one such writing course.

Due to the constraints caused by the COVID-19 pandemic, students were not offered a placement test, which is otherwise held in-person. They were, instead, placed based on proficiency tests and schooling in English-medium contexts. In many instances, students were given the opportunity to provide an academic writing sample to assist with the placement decisions.

Ongoing writing support resides primarily at the writing center that provides free consultations to the university community. The consultants are trained to assist at all stages of the writing process. Additionally, there is sporadic writing support through departmental or campus-wide workshops and trainings. It is reasonable to conclude that international graduate students are not heavily scaffolded for their language or writing skills enhancement, and the provision for developing writing is delivered primarily through the EAP or discipline-specific writing-intensive courses.

4.2. Curricular context

The 16-week advanced writing course followed a genre analysis approach (Bhatia, 1997) to help graduate students learn the disciplinary conventions of their fields of study and gain linguistic competence (Swales & Feak, 2012). The main course aim was not to teach students how to compose the specific genres used in their major fields; rather how to observe and comprehend these genres by performing a linguistic and rhetorical analysis of the disciplinary texts. Students were assessed by six assignments as well as through

other ungraded written pieces. Consistent with the course aims, the assignments constituted short papers through which students investigated, collected, and analyzed models of the disciplinary conventions of their fields. Through these tasks, they learned to interpret task constraints, practice intellectual processes required in the task, create working drafts, receive peer and teacher response, and learn to revise effectively based on that response. They participated in teacher-student group conferences (Ching, 2014) before each assignment was due, in which they read each other's papers and offered feedback. They also discussed anonymous samples as models. Swales and Feak's *Academic writing for graduate students: Essential skills and tasks* was used as the textbook. The class met two days per week for seventy-five minutes. There was only one course section being offered, so I was the only one teaching this course.

Seven students were in class, hailing from countries like Nepal, Cambodia, and Japan, studying programs like computer science, applied linguistics, and Asian studies. Their ages ranged from 21 to early-30 s. All spoke English and at least one additional language.

4.3. Instrument

Diagnostic tests can be grounded in either theory or syllabus (Alderson, 2005; Huhta, 2008); the EAP course involved the latter. Students were administered a 70-minute diagnostic test in the first week of the semester through the university's learning management system (LMS). The test comprised a writing with sources task, an independent task that shows how a writer follows conventions in order to place their text within a network of other texts.[1] The prompt involved five quotations on the topic of time management, and students had to 1) identify one or more points on this topic that they found interesting or important; 2) analyze the point(s); and 3) support their analysis with information from at least two quotations, their own experience, observations, and/or background reading. In this way, the task primarily involved the cognitive processes of analyzing, evaluating, and synthesizing. No word limit was specified, and dictionary usage was permissible. Students were encouraged to draft and revise their responses within the 70-minute time frame. I assessed the responses through an evaluation criteria, which focused on subskills like content, organization, vocabulary, and grammar (including mechanics).

As a result, students were expected to demonstrate abilities such as what stance or perspective should be taken toward the topic, how to structure a text, what style to use, how to represent one's persona, and which lexical items were most appropriate to employ. I was provided the test material with other instructional resources, and the test material was accompanied by a brief note advising instructors that they "respond *briefly*" (emphasis original) and hold individual conferences with students if necessary.

5. Method

5.1. Participant

Emily[2] was the only student who volunteered as a participant. She was a motivated student and took great interest in class activities. She was a 27-year-old, first-year doctoral student in marine biology. She had completed an undergraduate degree in veterinary medicine in Thailand in 2019 and was admitted directly into the Ph.D. program in fall 2021, which was her first time both visiting the U.S. and studying abroad.

In Thailand Emily had used English for academic purposes including communicating with her (then potential) advisor over Zoom. To prepare for her Ph.D. admission application, she took an IELTS preparation course.[3] After completing the course, she continued to prepare for the exam by practicing writing, which she shared for feedback with her more proficient friends in Thailand. She spent two hours daily for three months to prepare and ended up with a 5.5 band score in writing.

When Emily started the Ph.D. program, she struggled to "catch up" with classmates and faculty orally discussing scientific concepts. Even though everyday communication was not a barrier, she found it challenging to communicate scientific knowledge. In the first interview, she highlighted the additional obstacle of switching to a new subject area: "It's really challenge me because first I have language barrier, and second I study in new field that I don't have any background knowledge before, so it's kinda like combination–two things."[4] This remark documents the dual challenge that she had to contend with: coping with the intellectual demands of a new discipline and mastering a second language. Moving from one social domain to another requires adjusting writing, learning new skills, and transforming the knowledge one brings from previous experience.

5.2. Data collection

The process for accumulating evidence spanned across the semester. The data sources included the diagnostic instrument, Emily's written response to it, my reflection journal, course materials (syllabus, lesson plans, and handouts), semi-structured interviews with Emily at mid-semester and at the end of the semester (see Appendix A for interview guides), Emily's six written assignments (including first drafts) in the EAP course, and two graded assignments produced for two separate marine biology courses (self-selected by Emily). Moss (2013) advocates such a data-driven approach where teachers employ multiple sources of evidence, which also allows to see connections and triangulate findings (Merriam & Tisdell, 2016). I believed that this collection of data adequately represented the

[1] I am unable to share the actual test because of program policies.
[2] This pseudonym was self-selected by the participant.
[3] GRE was not a requirement for Emily's program.
[4] I am replicating Emily's words verbatim out of respect of her language levels (Mangelsdorf & Ruecker, 2018).

content of instruction and its effects on student learning.

The first interview was held in the seventh week of instruction,[5] and the second was held one week after the course had ended. Both interviews were held over Zoom, audio recorded, conducted in English, and lasted 16 min and 27 min respectively. In both interviews, I shared my screen with Emily to show her the test prompt and her response, in order to stimulate her memory of the test. I acknowledge that one possible limitation of this data was that the teacher-student relationship that I shared with her was likely to have influenced her responses. Nevertheless, I made every effort to push her to critically reflect on her responses. I transcribed the interviews using Otter.ai. I exported the downloaded transcript in a Microsoft Word document and listened to the audio again to revise errors in automatic transcription. I aimed to get the transcription as accurate as possible.

Emily's two graded assignments written for her marine biology courses were a study question and a press release. The study question was a page-and-a-half manuscript review of an article meant for journal publication along with guiding questions (Table 2 includes its excerpt). The one-page press release related to sea turtle tumors. The study question was completed in week 7, and the press release in week 10. I used these samples to benchmark the disciplinary expectations Emily had to meet and to connect the writing she produced in the EAP classroom with the writing goals that were valued beyond the classroom. These samples provided a window into her larger experience of writing in her major, however brief or subjective. To me, such insight was vital to making informed judgments about her academic writing needs.

5.2.1. Developing the Validity Argument

My past experience in the EAP course helped me in designing and conducting the argument-based validation. This process commenced with Chapelle et al.'s (2015) framework. Adapting their framework, I operationalized validity as comprising six core inferences: domain definition, evaluation, feedback, utilization, extrapolation, and decisions. Table 1 presents the classification of the six inferences, their warrants and underlying assumptions, and their corresponding sources of evidence.

For the framework to serve the context in which the diagnostic test was employed, I made several modifications to Chapelle et al.'s (2015) model, and borrowed from Knoch and Elder (2016) for these changes. The warrants for utilization and decisions inferences were tailored to represent the study's aims, which was to make explicit the expected benefits of the assessment, provide evidential basis of its expectations, and improve my practice. *Domain definition* refers to the target construct or theory about which test developers want to draw conclusions and the decisions those conclusions inform (Moss, 2013). In my approach, the domain was *EAP writing*, which was operationalized as writing produced for both the EAP course and other discipline-specific writing-intensive courses and included both writing and language ability. Fig. 1 provides a Toulmin diagram to illustrate the evidentiary argument for the first assumption of domain definition inference, drawing on Oliveri et al. (2019). *Evaluation*, *feedback,* and *utilization* inferences were distinguished in the sense that the evaluation inference pertained to the diagnostic test and its response; feedback inference constituted feedback reporting, i.e., instructions or guidance that students receive or actionable information for teachers; and utilization inference was what both students and teachers decide to *do* with the information, and was consequently more action-oriented.

One point of departure from both Chapelle et al. (2015) and Knoch and Elder (2016), was the *generalization* inference, which I excluded since psychometric reliability and the notion of different test occasions were irrelevant to my purposes, nor did I have numeric codes to compute a generalizability score. Instead, I added the assumptions from Knoch and Elder's *decisions* inference. Additionally, they suggest two more pertinent assumptions under *decisions*: 1) "Learners taking up support options improve their English over the course of their studies", and 2) Learners "who fail to act on test recommendations are more likely to struggle in their academic studies" (p. 225). These are more ambitious claims and eliciting evidence for these warranted a longitudinal approach which was beyond the scope of this research; consequently, they were not incorporated. In this way, the availability of data affected this framework. In a nutshell, the conceptualization of the validity argument views diagnostic assessment as the way the test is designed, the response it yields, and the pedagogical decisions it impacts.

5.3. Data analysis

I adopted a more naturalistic inquiry and analyzed the data qualitatively on an ongoing basis, connecting it to the validity inferences. As Pellegrino et al. (2016) point out, in contrast to studies of large-scale assessments employing statistical models, in "the context of classroom assessment, the interpretation is often made less formally by the teacher and is usually based on an intuitive or qualitative model rather than a formal statistical one" (p. 64). Thus, the analysis was more interpretive and reflective rather than code-focused.

Using the assumptions framing each inference in the argument-based framework that I designed, data analysis started with a consideration of the data type and its relation to the underlying assumptions. I created a table in Microsoft Excel to map the evidence-related inferences according to specific test assumptions and pasted relevant data against each assumption. Table 2 shows an example of how data related to the first assumption of domain definition inference (i.e., "The task and evaluation criteria capture aspects of performance that are relevant to EAP course aims and goals") was categorized. I used the labels of *met*, *partially met*, and *did not meet* as delimiting categories for analysis. I purposely kept these somewhat simplistic labels to conclusively answer validity questions and to assist other teachers attempting a similar exercise. I looked for patterns in the data and made notes if the evidence was complementary or divergent. I also weighted each evidence for the intensity of its coverage of inferences. In instances where I found diversion, I made

[5] This delay was caused by a late IRB approval.

Table 1
Argument-based approach for validating diagnostic assessment of EAP writing.

Inference	Warrant	Assumption	Evidence
1. Domain definition	Observations of students' performance on the diagnostic instrument reveal knowledge, skills, processes, and strategies that align with those required for EAP writing.	1. The task and evaluation criteria capture aspects of performance that are relevant to EAP course aims and goals. 2. The characteristics of the assessment tasks offered via the test are broadly relevant to the writing tasks required in students' other university courses.	1. Diagnostic test document 2. Student's diagnostic test response 3. Course syllabus 4. Teacher's reflections 5. Student's writing samples produced for the course 6. Student's assignments from other courses
2. Evaluation	The diagnostic instrument is an accurate and useful method of broadly identifying students' strengths and weaknesses in EAP writing.	1. Test instructions, purpose, and tasks are clear to all test takers. 2. Test is pitched at appropriate difficulty level. 3. The task and evaluation criteria have the capacity to broadly identify test takers' individual strengths and weaknesses as writers.	1. Diagnostic test document 2. Student's diagnostic test response 3. Teacher's reflections 4. Student's interview responses
3. Feedback	Diagnostic feedback is fine-grained, well communicated, and well tied with future learning.	1. The descriptive feedback to students identifies their strengths and weaknesses at the subskill level and provides recommendation(s) on future actions. 2. The feedback is detailed, clear, specific, and timely. 3. The follow-up recommendation for students is appropriate and targeted at the subskill level. 4. The follow-up recommendation is closely linked to on-campus support.	1. Diagnostic test feedback 2. Teacher's reflections 3. Student's interview responses
4. Utilization	Diagnostic results on the quality of EAP writing obtained from the test are useful for students and teachers to set curricular goals and make pedagogical decisions.	1. Students can use diagnostic results to set their goals for the course and future improvement. 2. Teachers can use diagnostic feedback to sequence and plan instructional activities.	1. Student's interview responses 2. Teacher's reflections 3. Lesson plans and handouts
5. Extrapolation	Diagnostic results are relevant to students' EAP writing.	1. Test results are good reflectors of students' overall EAP writing ability and specific strengths and weaknesses in the EAP course. 2. Test results are good reflectors of students' overall EAP writing ability and specific strengths and weaknesses in other university courses.	1. Student's diagnostic test response 2. Student's writing samples produced for the course 3. Student's assignments from other courses
6. Decisions	The consequences of using the diagnostic and the decisions informed by the diagnostic test are beneficial to all stakeholders.	1. Test takers' perceptions of the test and its usefulness are positive. 2. The feedback from the test is useful and directly informs students' future learning. 3. The feedback from the test is useful for teachers to make effective instructional decisions. 4. Students act on the test recommendation (i.e., take up the proposed writing development strategies).	1. Student's interview responses 2. Teacher's reflections 3. Student's writing samples produced for the course 4. Diagnostic test feedback

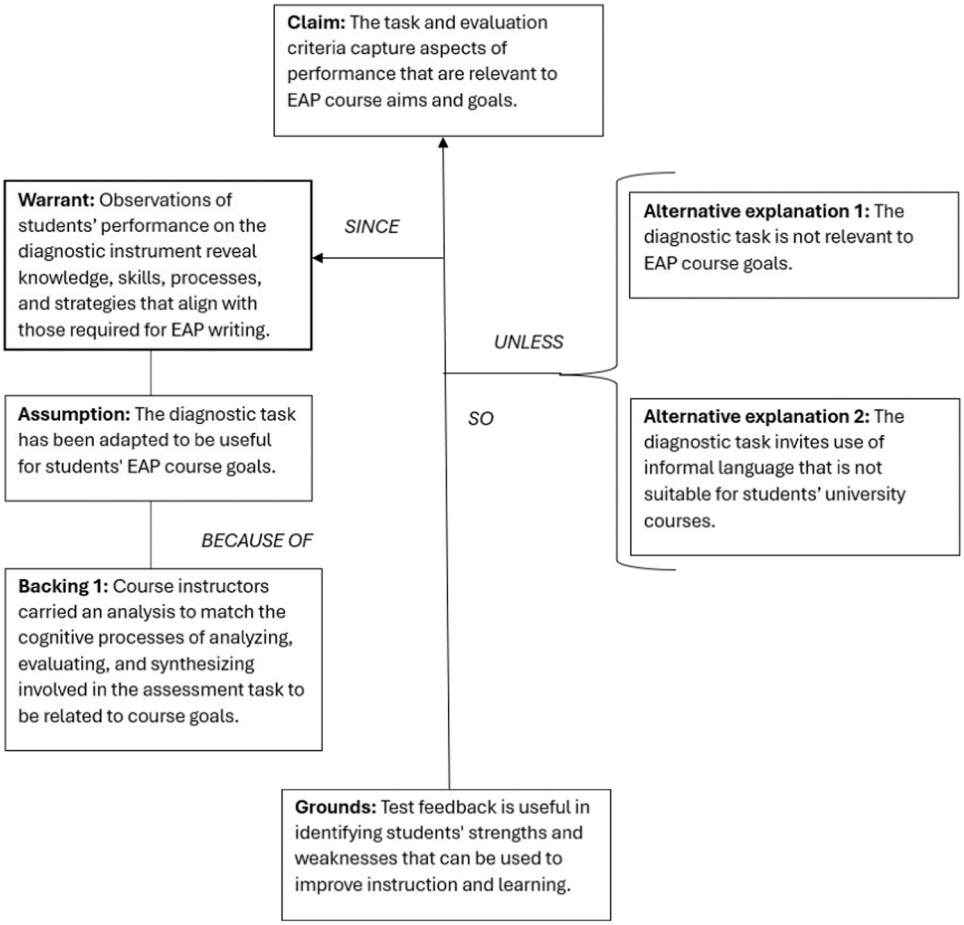


Fig. 1. Toulmin diagram of the evidentiary argument for first assumption of domain definition inference.

Table 2
Connecting evidence against the first assumption of domain definition inference.

Data Type	Relevant Excerpts	Meets Criteria
Diagnostic test document	Identify one or more points on this topic that you think are interesting or important and write a paper in which you discuss your analysis of the point(s). Support your analysis with information from the quotations, your own experience, observations, and/or background reading.	Y
Student's diagnostic test response[a]	Emily has understood the task correctly by presenting one synthesized analysis. Her response is in paragraph format. There are some expressions that she has used incorrectly.	Y
Course syllabus	By the end of the course students will be able to: analyze discipline and genre-specific academic English writing conventions and effectively apply that knowledge to graduate level writing tasks	Y
Student's writing samples produced for the course	Feedback on assignment 4: Emily chose very useful aspects to base her analysis on. Her use of examples made her analysis very clear.	Y
Teacher's reflections	The task of writing from sources is suitable. The test did predict students' performance. Emily is good. It pinpointed their [students] major areas of weaknesses and strengths.	Y
Student's assignments from other courses	Study question: Please review this paper on the following criteria: Data: Are data sufficient? Are methodologies and treatment of data appropriate? Conclusions: Are they justified by the data presented? Reasoning: Can it be followed easily? Writing: Are grammar and style appropriate? Illustrations: Are quality and quantity appropriate? Length: Is the manuscript an appropriate length? Final decision: Should the paper be accepted as is, accepted following revision, or rejected?	Y

[a] For Emily's diagnostic test response and writing produced for the EAP course, I used my impressions of her response and not the response itself, which I felt was more useful.

further investigations by re-examining the data. In cases where I could not reconcile differences, I termed those as *partially met*.

For each inference, I wrote analytic memos (Saldaña, 2009) noting what was achieved and what was not, documenting the rationale for interpretations. I drew on past research and literature in the field to come up with explanations and maintained a critical stance throughout by questioning my assumptions to gauge if my analysis was accurate. Besides, in keeping with the practice of member checking (Merriam & Tisdell, 2016), Emily read an earlier draft of this article. She did not suggest any changes to it.

6. Findings

In this section, I will present my findings for each of the six inferences.

6.1. Domain definition inference

The diagnostic instrument appeared to support the domain definition inference (Table 3). The EAP course entails analysis of disciplinary texts, and the cognitive processes captured in the test task were perceived to be aligned. After analyzing the test responses, I wrote in my journal: "The task of writing from sources is suitable. The test did predict students' performance. Emily is good. It pinpointed their [students] major areas of weaknesses and strengths."

Furthermore, analyzing Emily's assignments for her marine biology courses, I noted that properties like analytical arguments, personal opinion, and formal tone were required, especially for the study question assignment. These assignments also demanded knowledge of citing references, which matched with the test task. In this way, I tied the domain of academic writing to the task type and writing and language abilities observed in the test response, evaluated through the evaluation criteria, with both the EAP course and Emily's disciplinary courses.

Nevertheless, one issue with the test prompt was that the quotes did not carry all referencing details (e.g., publisher's name or page number) that were integral to producing an appropriate citation. Revising this element will yield more relevant responses.

6.2. Evaluation inference

The evaluation inference was partially met (Table 4). Initially, my perception was that the first assumption was completely met. In light of my observations of students' performance in class on the test day, I did not note any lack of clarity on their part. However, since Emily's previous EAP writing experience was largely limited to admission preparation, I recognized that diagnostic test tasks set expectations for international graduate students since these are among the first writing assignments they do in universities. These students often "come from educational cultures with large class sizes that only value summative standardized assessments" (Doe, 2015, p. 110). This understanding provoked critical questions: Do they understand this assessment form? Are they fully aware of its purpose?

These reflections led me to reexamine my initial assumption and look for more evidence to satisfy it fully. When I posed this question to Emily in her post-semester interview, her remark reflected that her understanding of what diagnostic assessment is bears more resemblance to proficiency testing: "I think diagnostic test is writing that help to assess my writing ability, my skill." She did not regard diagnostic assessment as understanding her strengths and weaknesses as a writer, but instead an evaluation of overall skill. Her lack of understanding indicates that I should have spent more time explaining the test purpose before administering it to students since they were not familiar with this assessment form. She also admitted that she had never attempted a diagnostic test before. To summarize, I concluded that the first assumption was partially met.

For the second assumption, I observed that the task was at appropriate difficulty level: "The quotes were on a topic general enough to be understood by all students. There is no discipline-specific focus, which is fine" (Teacher's reflections). I interpreted "difficulty" as the task presenting background knowledge or genre-based challenges and not being equally relevant to students from all disciplines. In my opinion, the subject matter of time management was general enough and I expected all students to write about it. Explaining the characteristics of an appropriate topic, Read (1990) makes the point that the contents of the question should neither challenge students' knowledge and interests, nor be too simple or predictable.

Therefore, in my view, the task was of appropriate difficulty level. The evidence from Emily's first interview, however, was to the contrary. She disclosed that she did not find the quotes "very easy" and she had to "spend some time to understand the quotation." Nevertheless, the time given for the task was deemed "appropriate." She pointed out that she had not done such a writing-from-sources task before. Emily's viewpoint is understandable because her prior writing experience was limited to her IELTS exam preparation. The two writing tasks in the IELTS exam involve writing from visual sources and an opinion-based essay (IELTS, n.d.). The properties of these tasks are distinct from the diagnostic test task, eliciting different communicative functions.

As far as the third assumption was concerned, I concluded positively. The evaluation criteria flagged individual grammar-related areas, which ranged from plural forms to parallelisms.

Table 3
Domain definition inference.

Assumption	Finding
1. The task and evaluation criteria capture aspects of performance that are relevant to EAP course aims and goals.	Met
2. The characteristics of the assessment tasks offered via the test are broadly relevant to the writing tasks required in students' other university courses.	Met

Table 4
Evaluation inference.

Assumption	Finding
1. Test instructions, purpose, and tasks are clear to all test takers.	Partially met
2. Test is pitched at appropriate difficulty level.	Partially met
3. The task and evaluation criteria have the capacity to broadly identify test takers' individual strengths and weaknesses as writers.	Met

6.3. Feedback inference

The feedback inference was refuted (Table 5) because there was no structured feedback reporting mechanism provided, making it challenging for me to convey effective feedback. Adhering to the instructions for teachers that they comment "*briefly*," I limited my feedback to a few sentences:

> Emily, you have understood the task correctly by presenting one synthesized analysis. There are some expressions that you have used incorrectly, for instance 'It is cannot deny' should be 'It cannot be denied.' In the course, initially focus on paragraphing.

This information poorly identified Emily's strengths and weaknesses at subskills levels, nor did it carry any recommendations. Though the inclusion of these subskills carried research support (Kim, 2011), I did not make effective use of the evaluation criteria, nor did I strive to check students' understanding of my feedback. I also chose to keep the post-test conference optional as suggested. Another drawback of this feedback that Emily identified was that it was restricted to task-level and not *broad* enough to apply to her general writing competency. In retrospect, this deficiency can be attributed to my own lack of understanding of diagnostic test's purpose. While whatever limited feedback was given may be categorized as clear and specific, it was not detailed and seemed inadequate.

Moreover, I was unable to communicate the feedback in a timely manner to the students due to a technical glitch on the LMS.[6] In my first interview with Emily, I discovered that she had not seen my feedback. I spent two weeks trying to resolve the problem, and finally communicated the feedback to the students in week 9.

6.4. Utilization inference

The utilization inference was met partially (Table 6). Even in the absence of clear guidelines as to how the diagnostic feedback will assist in making pedagogical decisions, I, the teacher, was able to use this information; however, the same cannot be claimed for students. Given the absence of any recommendations, it became difficult to investigate the feedback effect on learning and whether or not students followed the recommendations. In her post-semester interview, Emily was unable to identify any particular way in which she used the feedback.

Despite its limitations for student purposes, I used the feedback to make several instructional decisions. It allowed me to see the connections between course content and writing assignments. I analyzed students' performance collectively and identified two common problems: paragraphing and paraphrasing. I prioritized these two areas of need. For paragraphing I felt that more instruction time was required, so I devoted weeks 5 and 7 to it. I made these adjustments in order to make connections to students' immediate needs more tangible. I decided to first focus on composing a coherent paragraph, paying special attention to conclusion paragraphs. In addition, I commented on this subskill in my response to students' writing. For paraphrasing, I designed one lesson in week 14. When designing a syllabus, I usually keep the focus of a few sessions *to be decided* so as to utilize those to deal with students' specific needs. I capitalized on this space in the syllabus to focus on paraphrasing. Lastly, I tracked students' progress in their specific grammar areas in their future assignment feedback. I made sure I commented on these individual subskills in their writing assignments. I relied on my reflection notes throughout the process to track my instructional changes and corresponding students' progress.

6.5. Extrapolation inference

It can be concluded that the extrapolation inference was supported (Table 7). As stated previously (Section 6.1), I was able to get a sense of students' overall writing ability and specific strengths and weaknesses for designing instruction through their test responses.

In Emily's specific case, I tracked her improvement in two specific subskills: paragraphing and preposition errors. For paragraphing, she continued to display improvement. Her final assignment had information arranged neatly in sections, with well-organized paragraphs and logically sequenced sentences. For grammar, I did not notice any particular preposition errors. I did, however, notice two tense errors where she incorrectly referred to a past activity in present tense (errors circled in Fig. 2). The colored tracked changes are Emily's disciplinary-area instructors' markings.

For the second assumption, I traced the subskills identified as needing improvement in the diagnostic (i.e., paragraphing and preposition usage), in Emily's performance in her marine biology courses as evidenced by her two assignments. Regarding paragraphing, I noticed that these assignments required her to respond briefly and directly in a limited space. Apparently, the paragraphing

[6] Even completing the test had a glitch: one student could not submit his response through the LMS and had to email me his response separately.

Table 5
Feedback inference.

Assumption	Finding
1. The descriptive feedback to students identifies their strengths and weaknesses at the subskill level and provides recommendation(s) on future actions.	Not met
2. The feedback is detailed, clear, specific, and timely.	Not met
3. The follow-up recommendation for students is appropriate and targeted at the subskill level.	Not met
4. The follow-up recommendation is closely linked to on-campus support.	Not met

Table 6
Utilization inference.

Assumption	Finding
1. Students can use diagnostic results to set their goals for the course and future improvement.	Not met
2. Teachers can use diagnostic feedback to sequence and plan instructional activities.	Met

Table 7
Extrapolation inference.

Assumption	Finding
1. Test results are good reflectors of students' overall EAP writing ability and specific strengths and weaknesses in the EAP course.	Met
2. Test results are good reflectors of students' overall EAP writing ability and specific strengths and weaknesses in other university courses.	Met

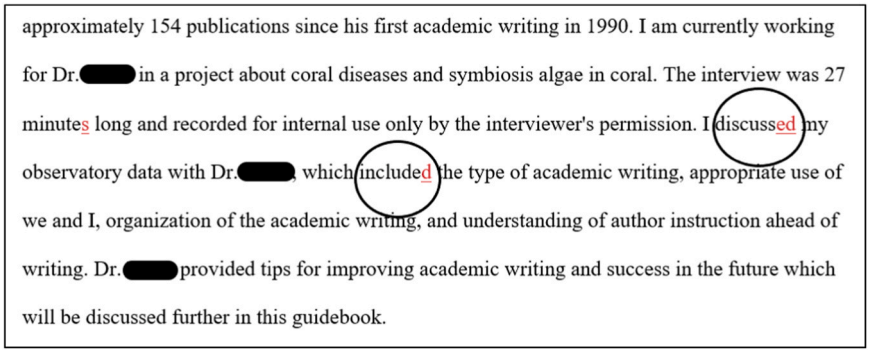
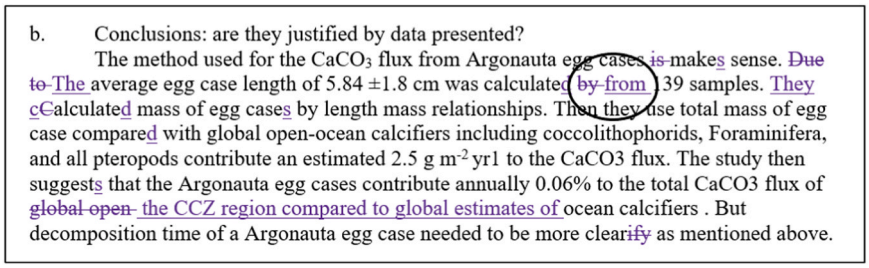

approximately 154 publications since his first academic writing in 1990. I am currently working for Dr.[redacted] in a project about coral diseases and symbiosis algae in coral. The interview was 27 minutes long and recorded for internal use only by the interviewer's permission. I discussed my observatory data with Dr.[redacted], which included the type of academic writing, appropriate use of we and I, organization of the academic writing, and understanding of author instruction ahead of writing. Dr.[redacted] provided tips for improving academic writing and success in the future which will be discussed further in this guidebook.

Fig. 2. Example from Emily's final graded assignment for the course.

standards that applied to the test in particular and the EAP course in general did not seem to apply in her disciplinary courses. Based on the limited evidence contributed by the two assignments, it can be claimed that paragraphing was perhaps not a subskill most relevant to extrapolation. Since these writing tasks were completed in the early to mid-point of the semester, I did not use them for investigating support for the feedback inference, ensuring that the evidence needed to support a decision was kept relevant to its timescale (Moss, 2013).

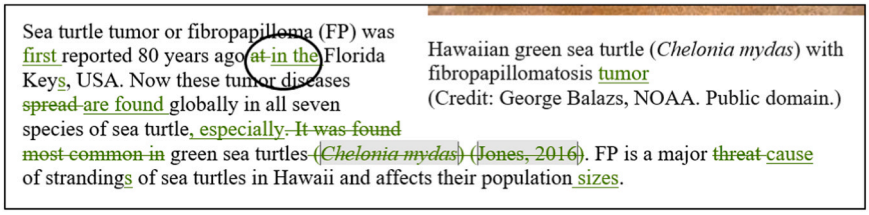
The second subskill was verb form and preposition errors. Samples from her assignments (Figs. 3 and 4) displayed that she made similar errors in her disciplinary courses as well, which provides evidence that this subskill was likely to be more relevant to her performance in other university courses.

Besides, I noticed that the two assignments were heavily edited. Whereas my written feedback approach focuses more on textual macrostructure, the gist and lines of reasoning employed in the paper, and limits grammar correction to major errors only, it appeared


b. Conclusions: are they justified by data presented?

The method used for the $CaCO_3$ flux from Argonauta egg cases ~~is~~ makes sense. ~~Due to~~ The average egg case length of 5.84 ±1.8 cm was calculated ~~by~~ from 39 samples. They c~~C~~alculated mass of egg cases by length mass relationships. Then they use total mass of egg case compared with global open-ocean calcifiers including coccolithophorids, Foraminifera, and all pteropods contribute an estimated 2.5 g m^{-2} yr1 to the CaCO3 flux. The study then suggests that the Argonauta egg cases contribute annually 0.06% to the total CaCO3 flux of ~~global open~~ the CCZ region compared to global estimates of ocean calcifiers . But decomposition time of a Argonauta egg case needed to be more clear~~ify~~ as mentioned above.

Fig. 3. Example from Emily's first assignment (Study Question).


Sea turtle tumor or fibropapilloma (FP) was first reported 80 years ago at in the Florida Keys, USA. Now these tumor diseases spread are found globally in all seven species of sea turtle, especially. It was found most common in green sea turtles (*Chelonia mydas*) (Jones, 2016). FP is a major threat cause of strandings of sea turtles in Hawaii and affects their population sizes.

Hawaiian green sea turtle (*Chelonia mydas*) with fibropapillomatosis tumor
(Credit: George Balazs, NOAA. Public domain.)

Fig. 4. Example from Emily's second assignment (Press Release).

that Emily's disciplinary-area instructors had tried to correct every error. The press release was evidently meant for publication and hence, was more high-stakes writing that perhaps required higher accuracy level, but the same could not be assumed for her study question assignment. This observation is interesting for two reasons. One, it offers an explanation why the grammar subskill appeared to be more relevant. Due to their knowledge and training on effective written feedback compared to subject/content specialists, writing instructors are likely to focus more on discourse-level aspects and are more tolerant of grammatical errors, whereas, subject/content specialists focus on grammar (Kepner, 1991). And two, this observation sheds light on how international students navigate different teacher attitudes regarding written corrective feedback. In her post-semester interview, Emily admitted that most of her feedback in her major courses tended to carry emphasis on grammatical errors, which she appreciated: "I think it's really helpful for me, and I don't mind to learn English grammar from another subject too because it's like, I think writing is not about just learning in writing class. It's like about practice." This misalignment in feedback approaches generates some insights for extrapolation inference for diagnostic writing assessment.

6.6. Decisions inference

The decisions inference was supported to a limited extent (Table 8). Reflecting on her test experience, Emily found the test fruitful in the sense that students could track their progress through the course:

> I think about diagnostic test I think it's good to use it as a beginning because so at the end student will know like how they can improve after taking the class. And I think I like the quotation thing that you can like because it's similar to the thing that like got student have to, so I think it's a good way to see what student have and can improve in the next like after you finish the class. (Emily first interview).

Analysis of Emily's writing produced for the EAP course suggested improvement, particularly in the areas of employing appropriate academic vocabulary and better organization. Information was not arranged in paragraphs in her diagnostic test response; in contrast, it was properly organized in her third assignment. Though she still tended to write paragraphs that could be broken down further and were relatively lengthy, she displayed overall improvement in this area. Her final assignment demonstrated further improvement compared to her mid-semester performance. In her post-semester interview, she agreed with this finding. Furthermore, her self-perception of her writing abilities changed. In the first interview she rated her competency as "average"; in her post-semester interview she rated it as "above average," though not feeling confident enough to label it as "advanced."

For the rest of the students, I noticed an improvement in paragraphing through their course assignments. Despite Emily's positive perception and evidence of performance improvement, the absence of follow-up activities or any specified mechanism for tracking progress marred the test's impact. In this way, this inference illuminates the ways in which I used the test and other evidence relevant to students' learning in my own context to make decisions about my practice. Hence, the study responds to a wider research call on teachers' using data in educational systems to implement positive change (Moss, 2013).

6.7. Synthesis of findings

Overall, out of six inferences, the inferences of domain definition and extrapolation were perceived to be met completely, and the evaluation inference was also perceived to be met largely. On the other hand, the utilization inference was considered to be met partially, decisions to a limited extent, and feedback was not met. It seemed to me that more or less the absence of proper feedback mechanism limited the test use, somewhat weakened its impact, and reduced the potential for learning.

Table 8
Decisions inference.

Assumption	Finding
1. Test takers' perceptions of the test and its usefulness are positive.	Met
2. The feedback from the test is useful and directly informs students' future learning.	Partially met
3. The feedback from the test is useful for teachers to make effective instructional decisions.	Met
4. Students act on the test recommendation (i.e., take up the proposed writing development strategies).	Not met

7. Discussion and proposed modification

7.1. Discussion of findings

This case study presented a more interpretive and qualitative examination of an in-house, classroom-based diagnostic assessment procedure. Unlike existing research dominated by the quantitative paradigm, this case study helped to answer how diagnostic feedback information leads to differentiated instruction and learning and why certain subskills are more valued. It elaborates on how evidence was gathered and interpreted to confirm and disconfirm the effectiveness, relevance, and appropriateness of the diagnostic assessment procedure. The rich data from the case study gave opportunities to quote Emily and to bring forward her viewpoint as a valuable data source of validity evidence. It provided me with a window on her thinking, knowledge, and understanding.

Attempting to create and apply the validity framework bolstered my understanding of diagnostic assessment. The argument-based validation approach was a useful tool because the evidence gathered and interpreted in this process allowed me to describe and document learning. During the appraisal stage, I adopted a critical stance (Kane, 2021). As I read literature and reflected on my practice, conducting the study added to my own diagnostic competence (Edelenbos & Kubanek-German, 2004; Huhta et al., 2024).

Apart from identifying shortcomings in feedback reporting, the validity investigation also identified what changes should be made to the test content (e.g., adding referencing details to the quotes), and what other language assessment issues are present (e.g., variation in written feedback approaches between EAP instructor and discipline-specific instructors). Hence, the validity argument is the centerpiece of the entire assessment process from test development to impact.

The validation investigation fleshed out the tension in the assessment process with differing teacher-student perceptions emerging in two respects: test difficulty and purpose. The finding that Emily misperceived diagnostic assessment as proficiency testing, reaffirms Doe's (2015) assertion that international graduate students often have familiarity primarily with proficiency testing. This limited prior experience shapes and colors their understanding of other test types, including diagnostic. They are less likely to have been exposed to diagnostic assessment, highlighting a need to determine ways to raise their awareness of what it is and how it functions.

Besides, the critical inquiry yielded an interesting finding with respect to the extrapolation inference, which is a difference in EAP and non-EAP instructors' approaches towards written corrective feedback. Disciplinary area instructors often disproportionately foreground surface errors relative to other content-based concerns. Therefore, they may be limited in some aspects of their feedback (Kepner, 1991), which perhaps brings to surface a useful quality of a diagnostic test and EAP writing instruction more generally—a capacity to focus on those facets of writing on which students will not get much feedback in their disciplinary courses. On the other hand, EAP teachers typically downplay the style and correctness features of writing, in favor of encouraging exploration, tentativeness, and eventual understanding. These approaches are underpinned by instructors' views of language itself. These conceptions should promote a critical social and rhetorical view of language as opposed to a prescriptivist or standard view. Hence, the validity investigation made explicit the differing values at work in the processes of test design and use.

Overall, the study draws attention to the aspect of diagnostic feedback reporting in course-embedded approaches. The findings underscore the point that a diagnostic test's purpose will be severely compromised if diagnostic feedback is not acted upon by the teacher or delivered to the students. In this regard, the instructions to teachers should be made clear.

7.2. Proposed modification to the diagnostic assessment procedure: meeting the feedback inference

In response to the shortcomings identified in the feedback inference, I put forward one revision to the diagnostic assessment procedure which consists of a sample feedback report for students (Appendix B) and guidelines for teachers. My focus on the feedback inference was driven by the recognition that it was the key inference, and the subsequent inferences of utilization and decisions logically followed from it. The fact that this inference was unmet undermined the other two inferences as well.

The proposed worked example blends key categories from Knoch (2011), written feedback locating strengths and identifying gaps, excerpts from test response (for students to know which specific areas the comments applied to), and follow up activities constituting annotated resources. Doe (2015) emphasizes that for feedback to lead to learning, students need resources and communication

Table 9
Modifications to the diagnostic assessment procedure to support the feedback inference.

Assumption	Modification
1. The feedback to students identifies their strengths and weaknesses at the subskill level and provides recommendation(s) on future actions.	Written feedback locates strengths and identifies gaps, accompanied by specific recommendations targeted at subskills level.
2. The feedback is detailed, clear, specific, and timely.	Students are given written feedback report and their original test response, followed by a one-on-one, teacher-student conference with the following agenda: 1. clarify the test purpose, 2. explain the feedback report, 3. understand the processes entailed in the production of the test response by posing reflection questions, and 4. decide a future plan for reporting progress on each area.
3. The follow-up recommendation for students is appropriate and targeted at the subskill level.	The follow-up recommendations constitute annotated resources targeted at subskills level.
4. The follow-up recommendation is closely linked to on-campus support.	Self-access resources, such as textbooks, handouts, and websites, are readily available and encourage autonomous learning.

channels. Table 9 provides a snapshot of how the proposed model compensates for the drawbacks identified earlier so that better informed and more targeted feedback could be given.

For teachers, in line with Huhta et al.'s (2024) recommendation, I suggest that this report should be followed up with a one-on-one, post-test conference with students to better communicate the feedback and come up with a plan, deciding about reporting progress on each area. This interaction is vital because merely handing out the feedback report might still leave room for misinterpretation or technical glitches. Carless (2006) described this process as "assessment dialogue" in which teachers make the assessment criteria explicit to students (p. 230). The conference could be another opportunity for the teacher to clarify the test purpose and actively engage the students in the process.

Though I could not continue this project because I did not teach the same course the following semester, I did share this feedback model with other teachers in our monthly program meeting so that they would have a roadmap of what to concentrate on when they revised assignments or updated lesson plans. The practice of sharing assessment data to drive program and instruction improvement is in line with Conference on College Composition and Communication CCCC (2022) guidelines for writing assessment. The proposed feedback model can also be particularly advantageous for institutional contexts with limited language/writing support structures. Practitioners at other institutions could recommend central support centers that they have on campus (such as, academic learning units, graduate communication or academic units, or English learning centers) to connect students with relevant resources. In the study site, since the nature of academic support students receive depends upon their home department, I chose self-access resources to encourage autonomous learning.

8. Limitations and future research

One limitation of the study is that it focuses on only one student's performance and perception as the basis of analysis. Because students in the course come from a range of disciplines, each with their unique rhetorical features, the question of representative sample of students' disciplinary writing arises. This concern of assessment in similar cross-disciplinary curricular spaces is a larger, complex issue (see Sánchez & Kenzie, 2016, for further discussion) and more research is needed in this regard.

Further research can extend the present study's analysis to evaluate the long-term benefits of the proposed modification, whether it leads to any sustained improvement in students' awareness of scholarly writing and motivation to be self-directed in their learning. The data can also be supplemented with lesson video-recordings, especially on the test day, or students' periodic reflections. Longitudinal studies can trace students' engagement with other on-campus resources.

Future research into validity of diagnostic instruments assessing EAP writing should attend to the difference in EAP and non-EAP instructors' approaches towards written corrective feedback, with respect to the extrapolation inference. This aspect is particularly relevant in light of ongoing critique of notions of standard English and models of correctness to promote critical language awareness (Shapiro, 2022).

9. Conclusion

This case study was a teacher's attempt at validating an in-house developed diagnostic assessment procedure within an EAP course. The research integrated a student's perspective of receiving, comprehending, valuing, and acting upon the diagnostic feedback. Together, these perspectives of both key stakeholders shed light on what meaning is ascribed to the diagnostic assessment procedure in promoting teaching and learning.

Instructors in EAP courses routinely assign a writing task in the first class to informally diagnose students' writing abilities. The research underscores that diagnostic assessment can play a crucial role in living up to high expectations in EAP courses and showcases how informal diagnostic procedures in these classrooms can be systematized to be more accurate. Table 1 is intended as a conceptually rich and inclusive tool for other teachers for planning and implementing validation research on existing tests and practices, while the findings provide clear guidance for the validation of particular interpretations and uses.

Unlike other studies that report on selected inferences (e.g., Doe, 2015), the present study makes a valuable contribution to argument-based validation studies by focusing on all six inferences. The research also shows that argument-based validation exercises offer potential grounds for an outreach effort between EAP teachers and disciplinary area experts to coordinate to support student needs, which is a shared responsibility. Such collaborative explorations can help to achieve consistency, curricular symmetry, and shared goals in writing instruction.

CRediT authorship contribution statement

Rabail Qayyum: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Project administration, Resources, Validation, Writing – original draft, Writing – review & editing.

Data Availability

The data that has been used is confidential.

Acknowledgements

I am deeply grateful to Dr. Daniel Isbell, Mr. Kenton Harsch, and Dr. Betsy Gilliland in their assistance with the project. I am also grateful to Emily for participating in this study. I would also like to thank the reviewers for their constructive feedback.

Appendix A

Interview Guides.

Interview 1: Mid-Semester.

Could you tell me about your educational background? Where did you complete graduate and undergraduate education? When did you come to the US?

What aspects of writing do you enjoy?

What do you want to improve upon in this course?

What are the challenges you face in your writing?

Did you face any difficulty in understanding the instructions of the test?

Did you find the timing of the test adequate in completing it?

What aspects of feedback on your diagnostic test did you find useful?

What did you understand about your writing skills after receiving feedback on the diagnostic test?

Interview 2: Post-Semester.

How relevant was the diagnostic test to the activities completed in this course?

How useful was that information? In what ways can it be made more useful?

What in your opinion is the purpose of a diagnostic test?

Did you learn anything about yourself as a writer after getting the feedback?

Did you take any steps after getting the feedback?

Would you like there to be a word count for the response?

Did you use the writing center? Why? Why not?

Appendix B

Sample diagnostic feedback report for Emily.

Category	Feature	Writer's strengths and areas to improve	Sample text	Follow-up recommendations
Content	Task completion	The response fulfils task requirements by presenting one synthesized analysis using two quotations. However, there is no clear thesis statement.	It is cannot deny that time management is the key to success in work and life.	The following handout outlines how to structure analytical essays: https://writingcenter.unc.edu/tips-and-tools/essay-exams/
	Use of source material	The writer correctly incorporates the ideas from two quotations with no major errors of referencing.	It's true as Guitton said, "It is important to prepare oneself and one's environment for peak performance".	This handout carries explanation and activities on how to summarize, paraphrase or quote another source: https://writing.wisc.edu/handbook/assignments/quotingsources/
Reader/writer interaction	Style and stance	Although the writer uses appropriate tone and register throughout the essay, the writer repeatedly uses 'we', which at times includes the reader and at times not. It is unclear if 'we' refers to students specifically or people generally.	To be good in time management, we have to access what is the proper environment for work.	This book presents concrete strategies for improving prose: Williams, J., & Bizup, J. (2016). *Style: Lessons in clarity and grace* (12th ed.). The University of Chicago Press.
Accuracy	Vocabulary	Overall, the writer uses adequate vocabulary sufficient for the task with minor errors in places.	To be good in time management, we have to access what is the proper environment for work.	This resource includes explanation of a wide variety of common usage errors: https://brians.wsu.edu/common-errors/
	Grammar	The writer writes accurately on most occasions, but makes few errors in areas of verb tense, missing pronouns, and preposition.	According to, we have a different preference, for example, some people more productive in the early morning time because it is quieter. In my opinion, I do agree that sometime finish easy work first because it will give me a sense of productive and motivation for next tasks.	This self-help book carries activities to improve grammar: Wallwork, A. (2013). *English for academic research: Grammar, usage and style*. Springer. This website carries interactive exercises to explore various grammar topics: http://sentencesyntax.com/

(*continued on next page*)

(*continued*)

Category	Feature	Writer's strengths and areas to improve	Sample text	Follow-up recommendations
			But in the same time I don't prefer to accumulate big work and procrastinate it until the deadline come.	
Fluency	Text length	The response carries 324 words (class average was 483) which can be improved further.	What is key success in time management ?	This handout describes a range of brainstorming strategies: https://writingcenter.unc.edu/tips-and-tools/brainstorming/
Organization	Cohesion and coherence	The paragraphs in the middle are clearly organized, but the introduction and the conclusion carry single sentences and need further development.	In conclusion, understanding the nature of working environment of ourselves and priorities tasks is the key to the success of time management.	This handout carries signposting activities with explanation: https://www.monash.edu/learnhq/write-like-a-pro/improve-your-writing/write-clearly/signpost-to-guide-your-readers
Mechanics	Spelling and punctuation	The response is generally free from major errors in spelling, formatting, and punctuation.	It's true as Glutton said, "It is important to prepare oneself and one's environment for peak performance".	Google documents can be used to compose texts. Errors appear underlined in blue and corrections are also suggested.

Teacher Post-Conference Notes:

References

Alderson, J. C. (2005). Diagnosing foreign language proficiency: The interface between learning and assessment. *Continuum*.

Alderson, J. C., Brunfaut, T., & Harding, L. (2015a). Towards a theory of diagnosis in second and foreign language assessment: Insights from professional practice across diverse fields. *Applied Linguistics, 36*(2), 236–260. https://doi.org/10.1093/applin/amt046

Alderson, J. C., Haapakangas, E. L., Huhta, A., Nieminen, L., & Ullakonoja, R. (2015b). *The diagnosis of reading in a second or foreign language*. Routledge.

Bhatia, V. K. (1997). Applied genre analysis and ESP. In T. Miller (Ed.), *Functional approaches to written text: Classroom applications* (pp. 134–149). United States Information Agency.

Carless, D. (2006). Differing perceptions in the feedback process. *Studies in Higher Education, 31*(2), 219–233. https://doi.org/10.1080/03075070600572132.

Chapelle, C. A., Cotos, E., & Lee, J. (2015). Validity arguments for diagnostic assessment using automated writing evaluation. *Language Testing, 32*(3), 385–405. https://doi.org/10.1177/0265532214565386

Ching, K. L. (2014). The instructor-led peer conference: Teachers as participants in peer response. In S. Corbett, M. LaFrance, & T. E. Decker (Eds.), *Peer pressure, peer power: Theory and practice in peer review and response for the writing classroom* (pp. 15–28). Fountainhead Press.

Conference on College Composition and Communication [CCCC]. (2022). Writing assessment: A position statement. Retrieved from ⟨https://cccc.ncte.org/cccc/resources/positions/writingassessment⟩.

Creswell, J.W. (2013). *Qualitative inquiry and research design* (3rd ed.). Sage.

Crusan, D. (2010). Assess thyself lest others assess thee. In T. Silva, & P. K. Matsuda (Eds.), *Practicing theory in second language writing* (pp. 245–262). Parlor Press.

Doe, C. (2015). Student interpretations of diagnostic feedback. *Language Assessment Quarterly, 12*, 110–135. https://doi.org/10.1080/15434303.2014.1002925

Dolgova, N., & Siczek, M. (2019). Assessment from the ground up: Developing and validating a usage-based diagnostic assessment procedure in a graduate EAP context. *Journal of English for Academic Purposes, 41*. https://doi.org/10.1016/j.jeap.2019.100771

Dursun, A., & Li, Z. (2021). A systematic review of argument-based validation studies in the field of language testing (2000–2018). In C. A. Chapelle, & E. Voss (Eds.), *Validity argument in language testing: Case studies of validation research* (pp. 45–70). Cambridge University Press. https://doi.org/10.1017/9781108669849.005.

Edelenbos, P., & Kubanek-German, A. (2004). Teacher assessment: The concept of 'diagnostic competence. *Language Testing, 21*(3), 259–283. https://doi.org/10.1191/0265532204lt284oa

Fox, J. D. (2009). Moderating top-down policy impact and supporting EAP curricular renewal: Exploring the potential of diagnostic assessment. *Journal of English for Academic Purposes, 8*(1), 26–42. https://doi.org/10.1016/j.jeap.2008.12.004

Fox, J., Haggerty, J., & Artemeva, N. (2016). Mitigating risk: The impact of a diagnostic assessment procedure on the first-year experience in engineering. In J. Read (Ed.), *Postadmission language assessment of university students* (pp. 43–65). Springer.

Freeman, D. (1998). *Doing teacher research: From inquiry to understanding*. Heinle & Heinle.

Huhta, A. (2008). Diagnostic and formative assessment. In B. Spolsky, & F. M. Hult (Eds.), *The handbook of educational linguistics* (pp. 469–482). Blackwell.

Huhta, A., Harsch, C., Leontjev, D., & Nieminen, L. (2024). *The diagnosis of writing in a second or foreign language*. Routledge.

IELTS. (n.d.). Academic writing: What is the IELTS writing test?. Retrieved April 30, 2022, from ⟨https://www.ielts.org/en-us/how-to-use-ielts-results/four-skills/academic-writing/format#tab-2⟩.

Jang, E. E., & Sinclair, J. (2021). Diagnostic assessment in language classrooms. In G. Fulcher, & L. Harding (Eds.), *The Routledge handbook of language testing* (pp. 187–205). Routledge.

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

Kane, M. T. (2021). Articulating a validity argument. In G. Fulcher, & L. Harding (Eds.), *The Routledge handbook of language testing* (pp. 32–47). Routledge.

Kepner, C. G. (1991). An experiment in the relationship of types of written feedback to the development of second-language writing skills. *The Modern Language Journal, 75*(3), 305–313.

Kim, Y. H. (2011). Diagnosing EAP writing ability using the Reduced Reparameterized Unified model. *Language Testing, 28*(4), 509–541. https://doi.org/10.1177/0265532211400860ltj.sagepub.com

Knoch, U. (2011). Rating scales for diagnostic assessment of writing: What should they look like and where should the criteria come from? *Assessing Writing, 16*(2), 81–96. https://doi.org/10.1016/j.asw.2011.02.003

Knoch, U., & Elder, C. (2016). Post-entry English language assessments at university: How diagnostic are they? In V. Aryadoust, & J. Fox (Eds.), *Trends in language assessment research and practice: The view from the Middle East and the Pacific Rim* (pp. 210–230). Cambridge Scholars Publishing.

Koltovskaia, S. (2020). Student engagement with automated written corrective feedback (AWCF) provided by Grammarly: A multiple case study. *Assessing Writing, 44*, 1–12. https://doi.org/10.1016/j.asw.2020.100450

Mangelsdorf, K., & Ruecker, T. (2018). Peer reviews and graduate writers: Engagements with language and disciplinary differences while responding to writing. *Journal of Response to Writing, 4*(1), 4–33. ⟨https://scholarsarchive.byu.edu/journalrw/vol4/iss1/2⟩.

Merriam, S. B., & Tisdell, E. J. (2016). *Qualitative research: A guide to design and implementation.* John Wiley & Sons.

Moss, P. A. (2013). Validity in action: Lessons from studies of data use. *Journal of Educational Measurement, 50*(1), 91–98.

Oliveri, M. E., Lawless, R., & Mislevy, R. J. (2019). Using evidence-centered design to support the development of culturally and linguistically sensitive collaborative problem-solving assessments. *International Journal of Testing, 19*, 270–300. https://doi.org/10.1080/15305058.2018.1543308

Pellegrino, J. W., DiBello, L. V., & Goldman, S. R. (2016). A framework for conceptualizing and evaluating the validity of instructionally relevant assessments. *Educational Psychologist, 51*(1), 59–81. https://doi.org/10.1080/00461520.2016.1145550

Randall, J., Poe, M., Slomp, D., & Oliveri, M. E. (2024). Our validity looks like justice. Does yours? *Language Testing, 41*(1), 203–219. https://doi.org/10.1177/02655322231202947

Read, J. (1990). Providing relevant content in an EAP writing test. *English for Specific Purposes, 9*, 109–121.

Saldaña, J. (2009). *The coding manual for qualitative researchers*. Sage Publications Ltd.

Sánchez, F., & Kenzie, D. (2016). Of evolutions and mutations: Assessment as tactics for action in WAC partnerships. *The WAC Journal, 27*, 119–141.

Shapiro, S. (2022). *Cultivating critical language awareness in the writing classroom* (first ed.). Routledge. https://doi.org/10.4324/9781003171751

Swales, J. M., & Feak, C. B. (2012). *Academic writing for graduate students: Essential skills and tasks* (3rd ed..,). University of Michigan Press.

Toulmin, S. E. (2003). *The uses of argument (Updated ed.)*. Cambridge University Press.

Urmston, A., Raquel, M., & Tsang, C. (2013). Diagnostic testing of Hong Kong tertiary students' English language proficiency: The development and validation of DELTA. *Hong Kong Journal of Applied Linguistics, 14*(2), 60–82.

Xie, Q. (2017). Diagnosing university students' academic writing in English: Is cognitive diagnostic modelling the way forward? *Educational Psychology, 37*(1), 26–47. https://doi.org/10.1080/01443410.2016.1202900

Rabail Qayyum is a PhD candidate in Second Language Studies at *University of Hawai'i at Mānoa*. She spent over a decade teaching academic writing to university students in Karachi, Pakistan. She is an editorial board member of *The Peer Review*, a journal promoting the work of emerging writing center scholars.