pratica. Revista Multimédia de Investigação em Inovação Pedagógica e Práticas de e-LEARNING

**PRATICA**, VOL. 9 (1)
ABRIL 2026

# AUTOSCOPIA NA AVALIAÇÃO FORMATIVA DA ORALIDADE EM ESPANHOL NO ENSINO SUPERIOR EM PORTUGAL

# VIDEO SELF-OBSERVATION FOR FORMATIVE ASSESSMENT IN SPANISH SPEAKING IN HIGHER EDUCATION IN PORTUGAL

# AUTOSCOPIA EN LA EVALUACIÓN FORMATIVA DE LA EXPRESIÓN ORAL EN ESPAÑOL EN EDUCACIÓN SUPERIOR EN PORTUGAL

Isabel Cabo[1] [https://orcid.org/0000-0002-6482-4653]

Mário Cruz[2] [https://orcid.org/0000-0001-8894-8821]

Paula Querido[3] [https://orcid.org/0009-0006-8473-4045]

Mailing Rodil Méndez[4] [https://orcid.org/0009-0006-2188-0280]

Raquel Cardoso[5] [https://orcid.org/0000-0001-5989-959X]

Berenice Peñaloza Santos[6] [https://orcid.org/0000-0003-4008-8986]

*[1]Instituto Politécnico de Coimbra & Centro de Estudos em Educação e Inovação (CI&DEI), Portugal, icabo@iscac.pt*

*[2]Centro de Investigação e Inovação em Educação, Escola Superior de Educação, Instituto Politécnico do Porto, Portugal, mariocruz@ese.ipp.pt*

*[3]Universidade de Vigo, Espanha, pique.quepi@gmail.com*

*[4]Instituto Superior Miguel Torga, Portugal, mailingrodilmen@gmail.com*

*[5]CEOS.PP Coimbra, Instituto Politécnico de Coimbra, Coimbra, Portugal, e Instituto Politécnico de Coimbra, Portugal, rcardoso@iscac.pt*

*[6]Universidad Nacional Autónoma de México, UNAM, Mexico, berenicepenaloza28@gmail.com*

## Resumo

Este artigo descreve e operacionaliza um procedimento de autoscopia (gravação e visionamento do próprio desempenho) integrado na avaliação formativa da oralidade em Espanhol Língua Estrangeira (ELE) no ensino superior em Portugal. A intervenção decorreu em contexto real, associada a uma tarefa final de simulação ("ir de compras"), com apoio digital para trabalho extra-aula. O dispositivo combinou rubrica com quatro descritores (eficácia comunicativa, uso da língua, nível discursivo e correção), autoavaliação individual após visionamento extra-aula, avaliação entre pares e avaliação docente, seguida de confronto de avaliações e definição de metas. Os resultados quantitativos (N=30) indicam desempenho global positivo na tarefa (M_docente=14.75, DP=1.65, escala 0–20), mas revelam um gap avaliativo sistemático: a autoavaliação foi, em média, superior à avaliação docente (M_auto=16.96, DP=1.70; Docente-Auto: M=-2.21, DP=2.07; t(29)=-5.85, p=0.000002; d_z=-1.07). Por descritor, o maior desfasamento ocorreu em Eficácia Comunicativa (Docente-Auto = -3.33), seguindo-se Uso da Língua (-2.67), Nível Discursivo (-1.67) e Correção (-1.17). A evidência qualitativa da ficha de reflexão sugere que o vídeo apoiou o diagnóstico de fragilidades e a formulação de metas, persistindo, contudo, a referência a nervosismo associado à filmagem. Discute-se como implementar a autoscopia com baixo custo, mitigando efeitos de ameaça, e como usar o vídeo para promover literacia de avaliação e autorregulação na oralidade em ELE.

**Palavras-chave:** autoscopia; avaliação formativa; autorregulação; oralidade; ELE.

## Abstract

This paper describes a replicable procedure for video self-observation (autoscopy) integrated into formative assessment of L2 Spanish speaking in higher education in Portugal. The intervention took place in an authentic classroom context around a final role-play simulation ("shopping"), supported by digital tools for out-of-class work. The assessment design combined an analytic rubric with four descriptors (communicative effectiveness, language use, discourse level and accuracy), individual self-assessment after out-of-class video viewing, peer assessment, and teacher assessment, followed by comparison of ratings and goal setting. Quantitative results (N=30) show positive overall performance (teacher M=14.75, SD=1.65, 0–20 scale) but a systematic rating gap: self-ratings were higher than teacher ratings (self M=16.96, SD=1.70; Teacher–Self: M=−2.21, SD=2.07; t(29)=−5.85, p=0.000002; d_z=−1.07). Across descriptors, the largest gap occurred in Communicative Effectiveness (Teacher–Self = −3.33), followed by Language Use (−2.67), Discourse Level (−1.67) and Accuracy (−1.17). Student reflections suggest that video viewing supported diagnosis of weaknesses and the formulation of concrete improvement goals, while also highlighting heightened anxiety related to being recorded. Practical recommendations are provided for low-cost implementation and for reducing perceived threat while strengthening assessment literacy and self-regulation in L2 speaking.

**Keywords:** video self-observation; formative assessment; self-regulation; speaking; L2 Spanish.

## Resumen

Este artículo describe y operacionaliza un procedimiento de autoscopia (grabación y visionado del propio desempeño) integrado en la evaluación formativa de la expresión oral en Español como Lengua Extranjera (ELE) en educación superior en Portugal. La intervención se desarrolló en un contexto real de aula, a partir de una simulación final ("ir de compras"), con apoyo digital para el trabajo extraaula. El dispositivo combinó una rúbrica analítica con cuatro descriptores (eficacia comunicativa, uso de la lengua, nivel discursivo y corrección), autoevaluación individual tras el visionado fuera del aula, evaluación entre pares y evaluación docente, con comparación de valoraciones y fijación de metas. Los resultados cuantitativos (N=30) muestran un desempeño global positivo (M_docente=14.75, DT=1.65, escala 0–20) pero un gap evaluativo sistemático: la autoevaluación fue, en promedio, superior a la evaluación docente (M_auto=16.96, DT=1.70; Docente–Auto: M=−2.21, DT=2.07; t(29)=−5.85, p=0.000002; d_z=−1.07). Por descriptor, el mayor desfase se observó en Eficacia Comunicativa (Docente–Auto = −3.33), seguido de Uso de la Lengua (−2.67), Nivel Discursivo (−1.67) y Corrección (−1.17). Los testimonios indican que el vídeo apoyó el diagnóstico de dificultades y la formulación de metas de mejora, aunque se mantuvo la referencia a nerviosismo asociado a la filmación. Se presentan recomendaciones para una implementación de bajo coste y para reducir la sensación de amenaza, reforzando la alfabetización evaluativa y la autorregulación.

**Palabras-clave:** autoscopia; evaluación formativa; autorregulación; expresión oral; ELE.

## 1 INTRODUÇÃO

Apesar da relevância da oralidade nos currículos de línguas estrangeiras, a sua avaliação permanece um ponto crítico das práticas pedagógicas, porque o desempenho oral é efémero, co-construído e dependente do contexto, tornando a observação e a justificação do juízo avaliativo particularmente exigentes. Frisén (2024) indica que, ao analisar processos de classificação em provas de produção oral, os avaliadores recorrem a documentos de pontuação para estabilizar decisões e explicitar o que conta como qualidade, confirmando que critérios e descritores operacionais são condição de consistência e transparência.

Em contextos de ensino superior, incluindo em Portugal, a necessidade de avaliação defensável convive com a exigência de que a avaliação seja formativa e produza informação acionável para melhoria. Goh e Aryadoust (2025) sublinham que a integração de tecnologia na avaliação de compreensão oral e de produção oral só cria valor quando está subordinada a um desenho pedagógico claro, contemplando tarefas autênticas, critérios compreensíveis e oportunidades reais de revisão.

Nesse quadro, a autoavaliação é relevante, mas só se for concebida como prática criterial baseada em evidência. Li e Zhang (2021) aludem a uma associação global moderada entre autoavaliação e desempenho linguístico, sugerindo

utilidade em usos formativos quando ancorada em descritores claros. Em produção oral, Winke (2023) indica que autoavaliações estruturadas podem relacionar-se, substancialmente, com uma referência externa, reforçando o potencial do método para monitorização e *feedback*.

A literacia de avaliação do estudante é determinante para que a autoavaliação se traduza em autorregulação e não em perceções vagas. Weng et al. (2024) demonstram relações entre literacia de avaliação, motivação e ansiedade, implicando que a compreensão de critérios e o uso de evidência devem integrar o desenho didático, sobretudo em oralidade, frequentemente associada a exposição social.

A autoscopia, entendida como gravação do desempenho e visionamento orientado, responde a este problema ao transformar a produção oral em evidência revisível, permitindo observar o próprio desempenho com base em critérios e formular metas. Bernard et al. (2022) evidenciam que *feedback* mediado por vídeo, articulado com avaliação por pares, pode ativar processos de autorregulação e co-regulação em tarefas orais.

Este artigo descreve uma investigação aplicada em ELE no ensino superior em Portugal e incorpora um procedimento replicável de autoscopia integrado na avaliação formativa da oralidade, combinando rubricas, autoavaliação após visionamento extra-aula, avaliação por pares e confronto com a avaliação docente. O objetivo é consolidar o enquadramento académico recente e disponibilizar um guia passo-a-passo para docentes, incluindo cuidados de implementação e limites reportados pelos estudantes.

## 2 ENQUADRAMENTO TEÓRICO

O enquadramento teórico organiza-se em quatro subtemas articulados com o dispositivo de autoscopia: (2.1) clarifica-se a avaliação da oralidade em línguas estrangeiras, enfatizando o constructo e o papel de rubricas/descritores na consistência e transparência do julgamento; (2.2) discute-se a autoavaliação e a literacia de avaliação do estudante, bem como a evidência recente sobre validade para fins formativos; (2.3) analisa-se a autoscopia e o *feedback* multimodal como mecanismos para tornar o desempenho oral evidência revisível e potenciar autorregulação; e (2.4) discutem-se dimensões afetivas e éticas (ansiedade, exposição, privacidade) que condicionam a implementação em contextos reais, incluindo no ensino superior em Portugal.

### 2.1 Avaliação da oralidade em línguas estrangeiras: constructo, rubricas e consistência

Em avaliação de línguas, a oralidade é um constructo multifacetado que integra dimensões linguísticas (por exemplo, precisão e variedade), discursivas (por exemplo, coesão e organização) e interacionais/pragmáticas (por exemplo, adequação e gestão do turno), o que torna inevitável algum grau de julgamento profissional. Goh e Aryadoust (2025) assinalam que, mesmo quando tecnologias recentes (incluindo soluções baseadas em Inteligência Artificial) são usadas para apoiar a avaliação, a questão central permanece na definição do constructo e na evidência que a representa em tarefas concretas; sem clarificação do que se pretende avaliar e de como se observa, a tecnologia tende a amplificar ruído em vez de qualidade.

Quando o julgamento humano é inevitável, rubricas e documentos de pontuação tornam-se instrumentos de estabilização e de transparência. Frisén (2024) revela que avaliadores, ao classificarem a produção oral em provas padronizadas, recorrem a documentos de *scoring* (pontuação ou avaliação do desempenho de um aluno) para orientar a atenção e justificar decisões; em termos pedagógicos, o que implica a necessidade de rubricas com descritores observáveis em sala de aula, reduzindo arbitrariedade e permitindo que estudantes compreendam o que conta como desempenho de qualidade.

No âmbito da avaliação da produção oral em contextos educativos, os problemas de praticabilidade e a consistência tornam-se mais visíveis em tarefas autênticas (*role-plays*, simulações), precisamente porque estas incluem variabilidade comunicativa. Koizumi (2022) discute desafios da avaliação da produção oral em contexto de sala de aula, salientando tensões entre autenticidade, fiabilidade e alinhamento curricular; a implicação para práticas no ensino superior em Portugal é a necessidade de articular tarefas com objetivos e de usar rubricas que preservem autenticidade, mas tornem o julgamento comunicável e comparável.

Para além da consistência entre avaliadores, rubricas podem apoiar a consistência intra-avaliador, reduzindo efeitos situacionais e comparações implícitas entre desempenhos. Frisén (2024) descreve como o documento de *scoring*

funciona como foco atencional durante o julgamento, ajudando a converter impressões holísticas em evidência ancorada em critérios; em sala de aula, o que se traduz em rubricas curtas, com descritores claros, usadas antes e após a tarefa para alinhar expectativas e tornar o *feedback* verificável.

Em síntese, as rubricas não representam apenas instrumentos de atribuição de nota; pelo contrário, constituem artefactos de mediação pedagógica que organizam *feedback* e permitem aprendizagem baseada em critérios. Goh e Aryadoust (2025) defendem que, quando a avaliação é desenhada como parte do ciclo de aprendizagem, as rubricas podem funcionar como 'mapas de qualidade' que orientam a preparação, a execução e a revisão, sobretudo quando articuladas com evidência revisível (por exemplo, gravações) e oportunidades de reexecução.

## 2.2 Autoavaliação, literacia de avaliação e validade para fins formativos

A autoavaliação tem sido debatida na investigação em avaliação de línguas, devido a preocupações com enviesamento e precisão, indicando que a sua utilidade depende do uso e do desenho. Li e Zhang (2021) identificam uma correlação global moderada entre autoavaliação e desempenho linguístico, e consideram que moderadores como tipo de tarefa, domínio e características do instrumento, influenciam a magnitude da relação, sugerindo que a autoavaliação é mais defensável quando criterial, contextualizada e usada para finalidades formativas.

Em produção oral, as autoavaliações estruturadas podem ter validade prática em usos de baixo impacto. A este propósito, Winke (2023) refere que uma autoavaliação adaptativa de produção oral em aprendentes de Espanhol se relaciona de forma substancial com uma medida externa, dando como exemplo a *Oral Proficiency Interview Computer* (exame digital que avalia a proficiência oral em línguas – OPIc), sugerindo que o método pode apoiar monitorização e *feedback,* desde que se assumam as limitações inerentes ao autorrelato e se privilegiem interpretações pedagógicas.

Do ponto de vista conceptual, a autoavaliação é mais produtiva quando integrada em processos de autorregulação, ao definir objetivos, monitorizar a execução, refletir sobre evidência e ajustar estratégias. Butler (2024) argumenta que a autoavaliação pode fortalecer a autorregulação ao promover definição de metas, monitorização e reflexão, sobretudo quando o estudante dispõe de critérios claros e de oportunidades para transformar o juízo em ação (por exemplo, reescrever, repetir, regravar).

Contudo, a capacidade do estudante para usar critérios e *feedback* não é automática, requerem literacia de avaliação. Weng et al. (2024) opinam que as dimensões de literacia de avaliação em estudantes se associam a motivação e ansiedade, indicando que compreender critérios, interpretar resultados e gerir emoções são competências interdependentes. Além disso, em contextos portugueses de aprendizagem de línguas, os referidos autores sugerem que práticas avaliativas devem incluir momentos explícitos de 'ensino de critérios' e de discussão de qualidade, em vez de assumir que estudantes interpretam rubricas do mesmo modo que docentes.

Quando a autoavaliação é tratada como atividade formativa, e não como substituto da nota do professor, tende a apoiar a aprendizagem, sobretudo quando acompanhada por treino e com critérios. No âmbito da autoavaliação e do desempenho académico, Yan (2023) concluiu que intervenções de autoavaliação se associam a melhorias de desempenho; aplicado à oralidade, defende que o valor central está no processo reflexivo e na monitorização, mais do que em "acertar" na nota.

## 2.3 Autoscopia e *feedback* multimodal: vídeo como evidência revisível para aprender

A autoscopia (gravar e rever o próprio desempenho) pode ser entendida como estratégia de 'objetivação' da produção oral, que transforma uma atuação efémera numa evidência revisível, possibilitando observação repetida, comparação com critérios, e discussão com pares e docentes. Bernard e Kermarrec (2022) analisaram uma tarefa oral em que a avaliação por pares é acompanhada por *feedback* em vídeo e identificam processos de autorregulação, co-regulação e regulação socialmente partilhada, argumentando que o vídeo pode ativar reflexão metacognitiva e negociação de significado sobre qualidade.

A relevância do vídeo para avaliação formativa reside, também, na forma como reduz dependência de memória e de impressões globais, aumentando a rastreabilidade do *feedback*. Bernard e Kermarrec (2022) indicam que o suporte audiovisual permite ao estudante revisitar episódios específicos do desempenho para justificar juízos e planear

melhorias; em termos de desenho aplicado, o que favorece ciclos de melhoria (gravar → rever com rubrica → receber/give *feedback* → reexecutar), particularmente adequados a tarefas de simulação e *role-play* em ELE.

Além disso, o *feedback literacy*, embora estudado com regularidade em escrita, fornece princípios transferíveis para a produção oral, nomeadamente no que respeita a preparar estudantes para compreender critérios, avaliar evidência e agir sobre *feedback*. Zhang e Mao (2023) destacam que o desenvolvimento de literacia de *feedback* é favorecido por abordagens sistemáticas que combinam atividades preparatórias, múltiplas fontes de *feedback* e oportunidades de revisão; por analogia prudente, em oralidade, a autoscopia pode cumprir função semelhante ao tornar o desempenho observável e ao criar condições para *feedback* multiorigem e ação regulatória.

Em ambientes com componente digital, ainda que a tecnologia possa ser simultaneamente indutora de ansiedade e instrumento de suporte, depende do desenho e das normas de participação. Bárkányi e Brash (2025) afirmam que o uso de câmara pode atuar como gatilho de ansiedade em contextos *online*, mas que ferramentas tecnológicas também podem apoiar estratégias de mitigação; por analogia cuidadosa, a autoscopia em contexto presencial deve prever preparação, controlo de acesso e opção de ensaio, para que o vídeo seja percecionado como apoio à aprendizagem e não como mecanismo de vigilância.

## 2.4 Dimensão afetiva e ética de implementação: ansiedade, exposição e salvaguardas

A oralidade é frequentemente descrita como a competência mais ansiógena em aprendizagem de línguas, e o recurso a gravação pode aumentar a autoconsciência e o medo de avaliação. Bárkányi e Brash (2025) debatem a ansiedade na produção oral na língua estrangeira e a sua relação com bem-estar e aprendizagem em ambientes digitais, salientando que intervenções eficazes devem considerar fatores de saúde mental e criar condições de segurança psicológica; esta orientação é diretamente relevante quando se introduz vídeo em tarefas avaliativas.

A ansiedade pode ser amplificada por exigências de *performance* pública e por incerteza sobre critérios, sobretudo quando a mediação tecnológica aumenta a autoconsciência. Neste sentido, Bárkányi e Brash (2025) sublinham que dispositivos como a câmara podem funcionar como gatilhos de ansiedade em atividades síncronas, mas que estratégias de preparação e normas de participação podem amenizar o efeito; aplicado à autoscopia em contexto presencial, o que implica iniciar com gravações de baixo risco, clarificar finalidade e acesso ao vídeo, bem como normalizar o erro como parte do processo.

A gestão ética do vídeo envolve, ainda, decisões sobre privacidade, consentimento e armazenamento, que condicionam aceitabilidade e equidade. Goh e Aryadoust (2025) alertam que a integração tecnológica em avaliação deve ser acompanhada por salvaguardas e por decisões transparentes sobre dados e usos; em contexto institucional português, o que se traduz em consentimento informado, anonimização sempre que possível e regras explícitas de não partilha, garantindo que o vídeo serve a aprendizagem e não a exposição.

Assim, a autoscopia é mais eficaz quando integrada numa cultura de avaliação para aprender, com critérios claros, *feedback* acionável, oportunidades de reexecução e ambiente psicologicamente seguro. Butler (2024) sublinha que a autoavaliação e práticas de reflexão funcionam melhor quando os estudantes percebem utilidade e justiça do processo; aplicado ao vídeo, o que reforça que o dispositivo deve ser apresentado como ferramenta de melhoria, com progressão gradual e *feedback* organizado sob a forma de rubricas, reduzindo o peso de comparação social.

## 3 METODOLOGIA

A seção metodológica está organizada em quatro subsecções. Primeiro, descreve-se o desenho e o contexto do estudo de caso (3.1). De seguida, apresenta-se a tarefa oral e a integração de ferramentas digitais que apoiam o ciclo de preparação, execução e revisão (3.2). Posteriormente, descrevem-se os instrumentos utilizados nesta investigação (rubricas, autoavaliação, avaliação por pares e folha de reflexão) (3.3). Por fim, apresenta-se o procedimento de implementação passo a passo, desde o registo da atividade até à comparação das avaliações e a definição dos objetivos de melhoria (3.4), o que garante a replicabilidade do estudo e a coerência entre os objetivos, as evidências recolhidas e as práticas de avaliação adotadas.

## 3.1 Desenho e contexto

Adotou-se um desenho de estudo de caso com intervenção pedagógica em contexto natural de sala de aula, adequado quando se pretende compreender, em profundidade, como um dispositivo didático-avaliativo funciona em condições reais e com múltiplas variáveis contextuais que não podem ser controladas experimentalmente. Greenhalgh et al. (2025) caracterizam o estudo de caso como relato detalhado e contextualizado de um fenómeno delimitado no mundo real, frequentemente apoiado por dados qualitativos e, quando pertinente, por indicadores quantitativos articulados com objetivos da intervenção.

A opção por estudo de caso é coerente com a natureza aplicada da investigação em aprendizagem de línguas, na medida em que permite descrever o desenho, a implementação e as condições de viabilidade de uma prática inovadora (autoscopia), sem descontextualizar a tarefa comunicativa. Koizumi (2022) reflete que, na avaliação da produção oral, em sala de aula, a praticabilidade e a coerência pedagógica são constrangimentos centrais; por isso, estudos em contexto real são particularmente úteis para produzir conhecimento transferível sobre formatos de tarefa, critérios e rotinas de avaliação.

O contexto corresponde a uma unidade curricular de Espanhol (nível de iniciação) no ensino superior em Portugal, aplicado em turma com 36 estudantes e participação efetiva de 30 no dia da tarefa final. A delimitação do caso inclui uma turma, tarefa de simulação, período de implementação e instrumentos de avaliação, permitindo uma descrição suficientemente detalhada para replicação, conforme recomendações metodológicas para estudos de caso orientados à implementação (Greenhalgh et al., 2025).

## 3.2 Tarefa e integração digital

A atividade central concentrou-se numa tarefa comunicativa de simulação ("ir de compras..."), realizada em grupos, desenhada para estimular a comunicação funcional, a gestão de turnos e a adequação pragmática, e gravada em vídeo para possibilitar a autoscopia. Goh e Aryadoust (2025) alegam que o valor pedagógico de tecnologias na avaliação da produção oral depende da coerência entre a tarefa, o constructo e a evidência; neste estudo, a gravação serve a recolha de evidência revisível, e não a automatização da avaliação.

A integração digital ocorreu, sobretudo, como suporte ao trabalho extra-aula (por exemplo, através de correio eletrónico e da plataforma *Moodle*), permitindo acesso a materiais e continuidade do ciclo de aprendizagem fora do tempo presencial, incluindo a disponibilização do registo para visionamento. Goh e Aryadoust (2025) sublinham que esta extensão temporal da tarefa (preparar, executar e rever) é particularmente relevante para competências orais, porque cria oportunidades de reflexão e reexecução que raramente são possíveis em uma única aula.

A escolha da gravação e o visionamento extra-aula é, também, consistente com abordagens que usam estímulos em vídeo para apoiar recordação e reflexão, reduzindo dependência de memória e de impressões globais. Zhai et al. (2024) destacam que estímulos em vídeo são amplamente usados para aceder a processos de pensamento e promover metacognição, reforçando a pertinência de tornar o desempenho oral observável e discutível.

## 3.3 Instrumentos

Os instrumentos privilegiaram rubricas/grelhas com critérios e descritores para avaliação da oralidade, organizados em dimensões como eficácia comunicativa, nível discursivo, uso da língua e correção, de modo a tornar explícitos os parâmetros de qualidade a observar durante e após a tarefa (Tabela 1). Frisén (2024) considera que documentos de *scoring* orientam a atenção do avaliador para critérios relevantes e asseguram decisões mais justificáveis; em contexto de sala de aula, rubricas cumprem função análoga ao estabilizar o julgamento e ao viabilizar *feedback* criterial.

**Tabela 1**

*Rubrica analítica (N1–N5) para avaliação da produção oral: descritores por nível*

| Descritor | N1 (10) | N2 (20) | N3 (30) | N4 (40) | N5 (50) |
|---|---|---|---|---|---|
| Eficácia Comunicativa | Não cumpre a tarefa; mensagem não se entende no contexto. | Cumpre parcialmente; falhas frequentes na adequação; mensagem limitada. | Cumpre o essencial; adequação global aceitável; falhas pontuais. | Cumpre bem; adequação consistente; mensagem clara na maior parte. | Cumpre plenamente; adequação muito consistente; intenção comunicativa clara. |
| Uso da Língua | Erros muito frequentes; léxico/gramática impedem comunicação; muita interferência. | Erros frequentes; controlo limitado; comunicação por vezes afetada. | Controlo suficiente de estruturas frequentes; léxico funcional; comunicação mantida. | Bom controlo; variedade adequada; erros ocasionais. | Controlo muito bom; variedade e precisão elevadas; erros raros. |
| Nível Discursivo | Discurso desorganizado; sem coesão; difícil de seguir. | Alguma organização; coesão fraca; saltos de sentido. | Organização básica; conectores simples; compreensível. | Boa organização; conectores adequados; poucas incoerências. | Organização excelente; coesão variada e eficaz; discurso muito claro. |
| Correção | Pronúncia/entoação dificultam seriamente a compreensão; muitos desvios. | Compreensão por vezes difícil; vários desvios segmentais/prosódicos. | Globalmente inteligível; desvios não impedem compreensão. | Inteligibilidade boa; poucos desvios; prosódia adequada. | Inteligibilidade excelente; prosódia muito adequada; desvios mínimos. |

*Nota*. A escala N1–N5 corresponde a 10–50 pontos e foi convertida para 0–20 (0–4) por transformação linear nos cálculos.

Para captar a dimensão formativa, o dispositivo incluiu autoavaliação e avaliação entre pares, com base na rubrica, assim como confronto com a avaliação docente. Li e Zhang (2021) afirmam que a autoavaliação se relaciona, de forma moderada, com desempenho linguístico, e tende a ser mais defensável quando estruturada e ancorada em descritores; o que justifica a adoção de rubricas explícitas e o uso da autoavaliação como instrumento de autorregulação, e não como substituto direto da classificação docente.

Complementarmente, foi aplicada uma ficha final de reflexão sobre a experiência de autoscopia (utilidade, dificuldades e metas), permitindo recolher evidência qualitativa sobre perceções e processos autorregulatórios. Bernard e Kermarrec (2022) defendem que tarefas orais com avaliação por pares e *feedback* mediado por vídeo podem ativar processos de autorregulação, co-regulação e regulação socialmente partilhada, o que depreende a relevância de instrumentos que, além de captar resultados, também compreende processos e condições de segurança psicológica.

## 3.4 Procedimento

O procedimento organizou-se em quatro momentos: preparação com explicitação de critérios e treino da tarefa; gravação da simulação; disponibilização do vídeo para visionamento extra-aula; e sessão seguinte centrada em autoavaliação, avaliação por pares, confronto com a avaliação docente e definição de metas. Bernard e Kermarrec (2022) descrevem que a combinação de formulário de avaliação e vídeo pode estruturar comportamentos regulatórios em tarefas orais, sendo crucial que a avaliação culmine em decisões concretas de melhoria.

A disponibilização extra-aula do vídeo teve como finalidade criar tempo para observação repetida e para aplicação criterial da rubrica, reduzindo decisões precipitadas e favorecendo a justificação do juízo. Zhai et al. (2024) indicam que o uso de estímulos em vídeo em processos de reflexão tende a apoiar metacognição, o que, aplicado à autoscopia, se traduz em diagnóstico mais fino de aspetos como fluência, pronúncia e recursos não verbais.

Por fim, o confronto de avaliações foi concebido como atividade de literacia de avaliação, promovendo discussão sobre critérios e perceções de justiça. Weng e Liu (2024) sugerem que a literacia de avaliação dos estudantes se relaciona com motivação e ansiedade; deste modo, o procedimento procurou tornar os critérios transparentes e enquadrar o vídeo como ferramenta de aprendizagem, minimizando o potencial de ameaça associado à exposição e ao registo audiovisual.

## 3.5 Ética e proteção de dados

A participação envolveu estudantes do ensino superior maiores de idade (≥18 anos) e foi voluntária, mediante consentimento informado, incluindo autorização para a gravação. Foi garantido o direito de não participar e/ou de não autorizar a gravação, sem qualquer penalização académica. Os vídeos foram utilizados, exclusivamente, para fins pedagógicos e de investigação, no âmbito do estudo, sem divulgação pública. O acesso aos registos foi operacionalizado em ambiente institucional, através da plataforma *Moodle* e da *Google Drive*, em pastas com permissões restritas, configuradas para acesso exclusivo à docente e ao(s) estudante(s) do respetivo grupo. Para efeitos de análise e relatório, os dados foram anonimizados e apresentados de forma agregada; os excertos qualitativos incluídos foram desidentificados. A gravação não constituiu um elemento autónomo adicional na avaliação sumativa, mantendo-se no enquadramento da tarefa já prevista na unidade curricular. Concluído o ciclo de avaliação/reflexão e o tratamento dos dados, os registos foram removidos das pastas restritas, garantindo a sua eliminação dos repositórios utilizados.

## 4 PROCEDIMENTO REPLICÁVEL PARA DOCENTES

A seguir, apresenta-se um procedimento de baixo custo para integrar autoscopia na avaliação formativa da oralidade. Pode ser aplicado em ELE ou noutras línguas no ensino superior.

Custo temporal típico (referência prática): em sala de aula, reservar ~10–15 min para relembrar critérios e organizar a gravação, e ~25–35 min para apresentações/gravação (por exemplo, 5 grupos × 5–7 min). Extra-aula, o docente tende a despender ~30–45 min na transferência/nomeação e disponibilização dos ficheiros e ~60–90 min para aplicar a rubrica de forma criterial (2–3 min por estudante, dependendo do nível de detalhe). O visionamento pelos estudantes pode ser orientado para ~15–25 min (2 passagens do vídeo + preenchimento da rubrica).

1. Definir resultados de aprendizagem e critérios: explicitar o que conta como desempenho de qualidade (por exemplo, adequação, inteligibilidade, fluência, coerência).

2. Preparar a rubrica: usar 3-5 níveis por descritor, com descritores observáveis; partilhar a rubrica antes da tarefa e discutir exemplos.

3. Reduzir ameaça antes da gravação: realizar 1-2 micro-ensaios sem nota; permitir uma gravação-teste curta; acordar regras de confidencialidade e respeito.

4. Gravar a tarefa: garantir áudio audível e enquadramento simples; privilegiar situações de comunicação autêntica (*role-play*, simulação).

5. Disponibilizar o vídeo para visionamento extra-aula: indicar o que observar; sugerir 2 visionamentos (um global, outro com rubrica).

6. Autoavaliação individual: preencher a rubrica e registar 2 pontos fortes, 2 aspetos a melhorar e uma prioridade para o próximo desempenho.

7. Avaliação entre pares: pares avaliam com a mesma rubrica e justificam com evidência observável (momentos do vídeo).

8. Confronto e calibração: discutir divergências (auto *vs* pares *vs* docente), com foco em critérios, não em pessoas; clarificar descritores ambíguos.

9. Fecho com metas e plano: cada estudante define metas específicas e uma ação concreta (por exemplo, treino de expressões idiomáticas ou frases comuns, repetição orientada, gravações curtas semanais).

10. Acompanhar com mini-ciclos: repetir o procedimento em tarefas curtas ao longo do semestre para estabilizar critérios e reduzir ansiedade.

# 5 RESULTADOS

## 5.1 Desempenho global e por grupo (avaliação docente)

Com base nas rubricas preenchidas pelos alunos (autoavaliação após visualização), na avaliação pelos pares e na avaliação do professor (N = 30), observou-se um desempenho geralmente favorável na tarefa final. Após a transformação da escala N1–N5 (10–50 pontos) para uma escala de 0–20, a nota média global atribuída pelo professor foi de 14,75 (DP = 1,65). A autoavaliação apresentou uma média de 16,96 (DP = 1,70), sendo que a avaliação pelos pares alcançou uma média de 13,75 (DP = 2,13). Estes dados são apresentados nas Tabela 2 e Tabela 3.

Quanto ao nível de cada grupo (Tabela 2), foram detetadas algumas variações relevantes entre os três métodos de avaliação. Alguns grupos, como o Grupo 5, apresentaram pontuações relativamente altas, tanto na autoavaliação quanto na avaliação por pares, enquanto o Grupo 2 apresentou discrepâncias mais acentuadas, particularmente na avaliação por pares, em que as pontuações foram significativamente menores. Essas diferenças sugerem a possível influência de fatores grupais, como coesão, percepção do desempenho do grupo ou familiaridade com os critérios de avaliação.

Neste contexto, observa-se um *gap* negativo, entendido como a diferença (avaliação docente – autoavaliação) < 0, o que evidencia que, em média, a autoavaliação dos estudantes é superior à classificação atribuída pela docente, traduzindo uma sobreavaliação relativa face ao referencial avaliativo docente.

**Tabela 2**

*Médias por grupo (0–20) para Auto, Pares e Docente na tarefa final*

| Grupo | N | M Docente | M Auto | M Pares |
|---|---|---|---|---|
| 1 | 6 | 12.50 | 15.00 | 13.75 |
| 2 | 6 | 16.25 | 17.08 | 11.25 |
| 3 | 6 | 15.00 | 16.25 | 13.75 |
| 4 | 6 | 16.25 | 17.50 | 12.50 |
| 5 | 6 | 13.75 | 18.96 | 17.50 |

*Nota*. *Gap* negativo indica autoavaliação superior à avaliação docente.

Assim, uma análise do descritor (Tabela 3) confirma uma tendência geral, o que significa que a autoavaliação produz as pontuações mais elevadas em todos os critérios: eficácia comunicativa (M = 17,50), uso da linguagem (M = 17,67), nível do discurso (M = 16,33) e precisão (M = 16,33). As avaliações da professora, por outro lado, apresentam pontuações mais moderadas, enquanto a avaliação pelos pares se situa consistentemente abaixo de ambas.

**Tabela 3**

*Médias por descritor (0–20) e gap (Docente - Auto)*

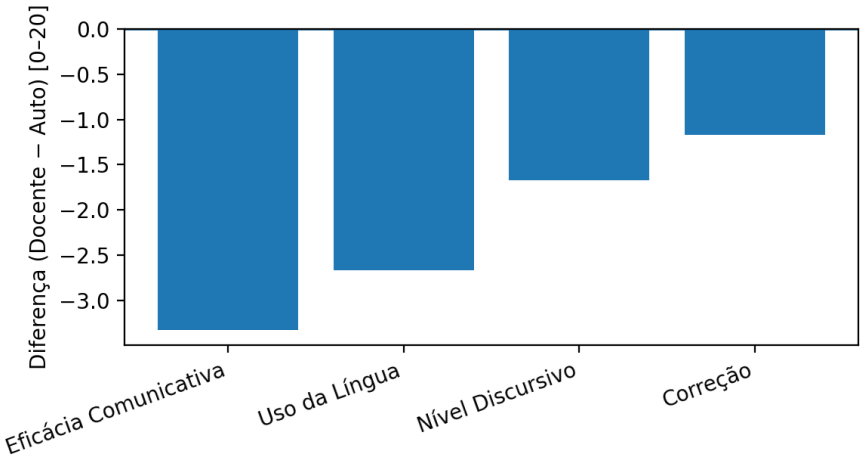
| Descritor | M Auto | M Docente | M Pares | Gap Docente-Auto |
|---|---|---|---|---|
| Eficácia Comunicativa | 17.50 | 14.17 | 14.00 | -3.33 |
| Uso da Língua | 17.67 | 15.00 | 14.17 | -2.67 |
| Nível Discursivo | 16.33 | 14.67 | 13.50 | -1.67 |
| Correção | 16.33 | 15.17 | 13.33 | -1.17 |

## 5.2 *Gap* avaliativo por descritor (Auto *vs* Docente; Pares como referência adicional)

A análise comparativa por descritor revela diferenças sistemáticas entre a autoavaliação e a avaliação do professor, confirmando uma tendência consistente – observada, anteriormente, nos resultados gerais –. É importante destacar que, para todos os descritores, a média da autoavaliação é superior à da avaliação da docente, resultando em valores negativos para a diferença (Professor – Autoavaliação). Observa-se, ainda, uma tendência geral de maior tolerância dos alunos à autoavaliação, mesmo quando a visualização de vídeos é incorporada como ferramenta de apoio à reflexão sobre o desempenho (Figura 1).

**Figura 1.**

*Diferença média entre autoavaliação e avaliação docente por descritor (0–20)*




*Fonte*: Elaboração própria (2026)

Em particular, as diferenças mais marcantes ocorrem no descritor de eficácia comunicativa (-3,33) e no descritor de uso da linguagem (-2,67), o que indica que os alunos tendem a superestimar os aspetos concernentes à comunicação global e à competência linguística em maior grau.

No entanto, os descritores para o nível de discurso (-1,67) e correção (-1,17) apresentam discrepâncias menores, o que pode indicar maior clareza ou especificidade nos critérios associados a estes aspetos. Neste sentido, é possível que os elementos mais observáveis e normativos, como a correção linguística, facilitem uma avaliação mais coerente com os padrões de ensino, reduzindo, assim, o fosso entre as duas avaliações.

De modo geral, os resultados revelam que, embora o uso de recursos como vídeos promova a reflexão sobre o próprio desempenho, não é suficiente por si só para garantir uma autoavaliação precisa. Essa constatação está em consonância com estudos que destacam a necessidade de apoio pedagógico explícito para desenvolver a capacidade avaliativa dos alunos e aprimorar a precisão de seus julgamentos (Panadero, 2017; Yan, 2023).

### 5.3 Diferença global Auto–Docente (teste pareado)

No total (média dos quatro descritores), verificou-se diferença estatisticamente significativa entre avaliação docente e autoavaliação (Docente − Auto: M=−2.21, DP=2.07; t(29)=−5.85, p=0.000002). O tamanho do efeito para diferenças emparelhadas foi elevado (d_z=−1.07), indicando divergência consistente entre os dois julgamentos.

### 5.4 Evidência qualitativa: diagnóstico e metas após autoscopia

As respostas abertas na ficha de reflexão foram tratadas como indicadores descritivos, com leitura integral e codificação temática leve (diagnóstico linguístico, autorregulação/metas e dimensão afetiva). De um modo geral, os estudantes referiram que o vídeo tinha facilitado a identificação de falhas e a definição de ações concretas para o desempenho seguinte; em paralelo, surgiram referências a nervosismo e a aspetos não verbais (tiques, postura), reforçando a necessidade de dessensibilização e de regras claras de confidencialidade.

E1: "Ter a perceção de como agimos perante uma câmara... ver as falhas e poder corrigi-las."

E2: "O uso da autoscopia permitiu ver os erros... deu para perceber o que devo melhorar."

E3: "Nunca tive a experiência de estar diante de uma câmara... foi uma experiência divertida."

E4: "Usar muitas expressões faciais e tiques nervosos."

## 6 DISCUSSÃO

Os resultados confirmam a viabilidade de integrar a autoscopia num ciclo de avaliação formativa da oralidade em ELE no ensino superior, mas também expõem tensões previsíveis entre transparência criterial, autojuízo e condições afetivas de participação. Em primeiro lugar, o desempenho global positivo na tarefa sugere que a simulação comunicativa, apoiada por visionamento extra-aula, gerou evidência suficiente para atribuir *feedback* e definição de metas. Esta interpretação é coerente com a perspetiva de que tecnologias na avaliação da produção oral só acrescentam valor quando subordinadas a um desenho pedagógico explícito (tarefa autêntica, critérios claros e oportunidades de revisão), e não quando funcionam como "camada" instrumental desligada do ensino (Goh & Aryadoust, 2025). Neste estudo, o vídeo não é tratado como automatização, mas como evidência revisível para fundamentar o julgamento e a regulação.

O resultado central é o padrão sistemático auto > docente, com diferença global estatisticamente significativa e efeito elevado, e com *gaps* negativos em todos os descritores. Este padrão tem duas interpretações que não se excluem. Por um lado, indica que a mera disponibilização de evidência (vídeo) não garante, por si, a calibração do autojuízo; a autoavaliação pode permanecer influenciada por crenças de competência, normas de grupo e apropriação incompleta dos critérios, mesmo quando a tarefa é revista (Li & Zhang, 2021; Butler, 2024). Por outro lado, a existência de autoavaliações estruturadas e a possibilidade de confronto com a avaliação docente constituem, precisamente, o mecanismo formativo pretendido: tornar visível o desfasamento, discutir o porquê e converter discrepâncias em metas observáveis. Esta interpretação aproxima-se de trabalhos que defendem a autoavaliação

como prática de autorregulação, mais do que como "acerto" na nota (Butler, 2024), e é consistente com evidência de que o valor da autoavaliação aumenta quando é criterial e quando alimenta ações subsequentes (Li & Zhang, 2021; Yan, 2023).

A análise por descritor revela que o maior desfasamento se concentrou na Eficácia Comunicativa e no Uso da Língua, enquanto que o Nível Discursivo e a Correção apresentaram *gaps* menores. Este padrão é aceitável à luz da literatura sobre julgamento em produção oral, uma vez que, nas dimensões mais globais e integradas (por exemplo, "cumprimento da tarefa" e "uso funcional da língua"), tendem a exigir maior competência criterial e comparação com padrões de desempenho, enquanto aspetos mais salientes (por exemplo, inteligibilidade) podem ser mais facilmente auto-observáveis, mesmo por aprendentes iniciantes. A evidência sobre processos de classificação constata que avaliadores recorrem a documentos de *scoring* para orientar a atenção e estabilizar decisões (Frisén, 2024); por analogia, estudantes podem necessitar de treino explícito para "ver" nos próprios vídeos o que a rubrica exige, sobretudo em dimensões menos tangíveis. O que reforça que a rubrica, por si só, é necessária, mas não suficiente, já que precisa de ser ensinada, exemplificada e usada em ciclos repetidos de aplicação e recalibração.

A presença de avaliação por pares como referência adicional ajuda a enquadrar o *gap*, visto que, quando as avaliações entre pares não acompanham a leniência da autoavaliação, o que sugere que a discrepância não é só a "diferença de padrão" entre docente e estudante, mas, possivelmente, um efeito de autorrelato (proteção da autoimagem, critérios ainda pouco internalizados, ou leitura benevolente do próprio esforço). A literacia de avaliação do estudante associa-se à motivação e à ansiedade, e que diferenças na apropriação de critérios influenciam o modo como o *feedback* e as rubricas são interpretados (Weng & Liu, 2024). Assim, o confronto sistemático auto–pares–docente pode ser defendido como estratégia didática de literacia de avaliação, tornando o julgamento um objeto de aprendizagem e não só um resultado.

A componente qualitativa (reflexão final) é consistente com um efeito pedagógico desejável, uma vez que os relatos de diagnóstico e a formulação de metas após visionamento sugerem que o vídeo apoiou a transição de *feedback* genérico para *feedback* utilizável, isto é, para decisões de melhoria associadas a evidência concreta. Esta interpretação confirma que *feedback* mediado por vídeo, combinado com avaliação por pares, pode ativar processos de autorregulação e co-regulação em tarefas orais, justamente porque permite revisitar episódios específicos do desempenho e justificar juízos (Bernard & Kermarrec, 2022). Mesmo quando a calibração do autojuízo não é imediata, a autoscopia pode cumprir a função formativa essencial, transformando a oralidade, efémera e difícil de "reter", em objeto observável e discutível.

Este estudo também evidencia um custo pedagógico que não deve ser minimizado ao referir-se à ansiedade associada à gravação. No âmbito da ansiedade, ao falar em contextos digitais e mediados por câmara, alerta-se que o dispositivo pode aumentar autoconsciência e ameaça percebida, o que exige medidas de segurança psicológica e desenho gradual (Bárkányi & Brash, 2025). Neste sentido, a primeira gravação tende a funcionar como tarefa de adaptação, e a melhoria da oralidade depende – frequentemente – de repetição do ciclo, normalização progressiva da câmara e regras claras de confidencialidade. Consequentemente, a autoscopia deve ser apresentada e aplicada como ferramenta de aprendizagem, com preparação, treino de critérios, ensaios de baixo risco e controlo rigoroso de acesso e partilha, para evitar que o custo afetivo anule o ganho formativo.

Em termos de contributo, este estudo é, especialmente, prático e demonstra que a autoscopia pode ser integrada num ciclo de avaliação formativa sem infraestrutura complexa, recorrendo a um único dispositivo de gravação e a um canal institucional de disponibilização. Esta característica aumenta a transferibilidade para contextos portugueses de ELE, em que limitações de tempo e recursos condicionam a adoção de práticas inovadoras. Ao mesmo tempo, a evidência quantitativa do *gap* (auto > docente) fornece um argumento pedagógico robusto para o uso do procedimento, na medida em que, além de melhorar a rastreabilidade do *feedback*, também expõe a necessidade de trabalhar literacia de avaliação, permitindo discutir critérios com base em evidência comum e promover autorregulação (Butler, 2024; Weng & Liu, 2024).

Por fim, em termos de limitações, a generalização dos resultados deve ser prudente e orientada para a transferibilidade do procedimento, visto que se trata de um estudo de caso. Embora as grelhas permitam análises descritivas e comparação entre autoavaliação, pares e docente, a interpretação do *gap* avaliativo deve considerar que autoavaliações podem refletir simultaneamente crenças sobre competência, normas de grupo e diferentes níveis de apropriação dos critérios (Li & Zhang, 2021; Weng & Liu, 2024). Recomenda-se replicação com múltiplos ciclos de tarefa ao longo do semestre e, quando possível, verificação de consistência interavaliadora na avaliação docente, de

modo a reforçar a robustez das inferências e distinguir efeitos de calibração criterial de efeitos de contexto (Frisén, 2024; Koizumi, 2022). Para concluir, o estudo apresenta a autoscopia como um dispositivo promissor de avaliação para as aprendizagens, desde que implementado como prática criterial, iterativa e psicologicamente segura, e não como mero registo de desempenho (Goh & Aryadoust, 2025; Bárkányi & Brash, 2025).

## 7 IMPLICAÇÕES PRÁTICAS E CHECKLIST DE IMPLEMENTAÇÃO

*Checklist* mínimo para docentes:

- Rubrica partilhada e discutida antes da tarefa.
- Ensaio sem nota e gravação-teste curta.
- Regras explícitas de confidencialidade (quem vê o vídeo e com que finalidade).
- Autoavaliação sempre posterior ao visionamento (não só de memória).
- Confronto de avaliações com foco em evidência observável.
- Fecho com metas e uma ação concreta por meta.
- Repetição do ciclo em tarefas curtas para estabilizar critérios e reduzir ameaça.

## CONCLUSÕES E IMPLICAÇÃO NA PRÁTICA PEDAGÓGICA

Este estudo evidencia que a autoscopia (gravação e visionamento orientado do próprio desempenho) é exequível, com baixa exigência tecnológica, como componente de um ciclo de avaliação formativa da oralidade em Espanhol Língua Estrangeira (ELE) no ensino superior. Ao converter um desempenho tipicamente efémero em evidência revisível, o vídeo permitiu ancorar o *feedback* em episódios observáveis e aumentar a rastreabilidade das decisões avaliativas, articulando preparação da tarefa, execução, revisão com rubrica e definição de metas.

O resultado quantitativo mais consistente é o *gap* avaliativo sistemático auto > docente. Este padrão não deve ser lido como "erro" do estudante, mas como indicador de literacia de avaliação ainda em desenvolvimento, já que, mesmo com acesso ao vídeo, o autojuízo pode permanecer permeável a crenças de competência, normas de grupo e apropriação incompleta dos descritores. Pedagogicamente, o valor do procedimento reside, precisamente, em tornar o desfasamento visível e discutível, permitindo calibrar critérios com base numa evidência comum e orientar a autorregulação (metas, estratégias e monitorização).

A análise qualitativa aponta, contudo, um custo afetivo relevante, relativo ao nervosismo e à autoconsciência perante a câmara, o que implica que a autoscopia deve ser introduzida de forma progressiva e apoiada por uma planificação explícita da oralidade (objetivos, guião de tarefa, etapas de preparação e critérios), de maneira a reduzir constrangimento e evitar que o registo audiovisual seja percecionado como vigilância. A primeira gravação deve funcionar como experiência de adaptação, com ensaios de baixo risco e regras claras de confidencialidade e de controlo de acesso.

Para a prática pedagógica, propõem-se cinco orientações operacionais: explicitar e treinar a rubrica antes da tarefa com exemplos curtos; assegurar que o visionamento gera ação — autoavaliação e avaliação por pares culminam em metas específicas e numa pequena reexecução/ensaio orientado; usar o confronto de avaliações como atividade de literacia de avaliação (justificar com evidência do vídeo), e não como disputa de notas; proteger segurança psicológica e dados (consentimento informado, limites de partilha, armazenamento institucional); repetir mini-ciclos, ao longo do semestre, para estabilizar critérios, reduzir ansiedade e consolidar ganhos na oralidade.

Por último, a autoscopia configura-se como um dispositivo didático-avaliativo promissor, com o intuito de fortalecer a avaliação para aprender na oralidade em ELE, já que aumenta a qualidade do *feedback*, apoia o diagnóstico e metas, e cria condições para desenvolver literacia de avaliação. A transferibilidade do procedimento deve ser entendida de forma prudente (estudo de caso), mas o conjunto de passos e salvaguardas apresentado constitui um caminho realista para docentes que pretendam melhorar a avaliação da oralidade sem depender de infraestruturas avançadas.

## AGRADECIMENTOS

Este trabalho é financiado por fundos nacionais através da FCT – Fundação para a Ciência e a Tecnologia, I.P., no âmbito do projeto UID/05198/2025 – Centro de Investigação e Inovação em Educação (inED), com o identificador DOI https://doi.org/10.54499/UID/05198/2025 e https://doi.org/10.54499/UID/PRR2/05198/2025.

## REFERÊNCIAS

Bárkányi, Z., & Brash, B. (2025). Foreign language speaking anxiety, mental health, and online learning: Overcoming barriers in the digital age. Language Teaching, 1–19 https://doi.org/10.1017/S0261444825101080

Black, P., & Wiliam, D. (1998). Assessment and classroom learning. *Assessment in Education: Principles, Policy & Practice, 5*(1), 7–74. https://doi.org/10.1080/0969595980050102

Butler, Y. G. (2024). Self-assessment in second language learning. *Language Teaching, 57*(1), 42–56. https://doi.org/10.1017/S0261444822000489

Clayton Bernard, R., & Kermarrec, G. (2022). Peer assessment and video *feedback* for fostering self, co, and shared regulation of learning in a higher education language classroom. *Frontiers in Education, 7*, 732094. https://doi.org/10.3389/feduc.2022.732094

Conselho da Europa. (2001). *Quadro europeu comum de referência para as línguas: Aprendizagem, ensino, avaliação*. Conselho da Europa.

Frisén, L. B. (2024). *Eyes on the ball: Teachers' rating processes when assessing a national L2 speaking test through the lens of their scoring document* [Doctoral dissertation, The University of Melbourne]. https://arts.unimelb.edu.au/__data/assets/pdf_file/0008/5109065/1f6cfbdc25329ef6ad0ef886ed487ab9d3cc761f.pdf

Goh, C. C. M., & Aryadoust, V. (2025). Developing and assessing second language listening and speaking: Does AI make it better? *Annual Review of Applied Linguistics, 45*, 179–199. https://doi.org/10.1017/S0267190525100111

Greenhalgh, T. (2025). Case studies: A guide for researchers, educators, and implementers. *BMJ Medicine, 4*(1), e001623. https://doi.org/10.1136/bmjmed-2025-001623

Koizumi, R. (2022). L2 speaking assessment in secondary school classrooms in Japan. *Language Assessment Quarterly, 19*(2), 142–161. https://doi.org/10.1080/15434303.2021.2023542

Li, M., & Zhang, X. (2021). A meta-analysis of self-assessment and language performance in language testing and assessment. *Language Testing, 38*(2), 189–218. https://doi.org/10.1177/0265532220932481

Nicol, D. J., & Macfarlane-Dick, D. (2006). Formative assessment and self-regulated learning: A model and seven principles of good *feedback* practice. *Studies in Higher Education, 31*(2), 199–218. https://doi.org/10.1080/03075070600572090

Panadero, E. (2017). A review of self-regulated learning: Six models and four directions for research. *Frontiers in Psychology, 8*, 422. https://doi.org/10.3389/fpsyg.2017.00422

Sadler, D. R. (1989). Formative assessment and the design of instructional systems. *Instructional Science, 18*, 119–144. https://doi.org/10.1007/BF00117714

Weng, F., & Liu, X. (2024). Exploring second language students' language assessment literacy: Impact on test anxiety and motivation. *Frontiers in Psychology, 15*, 1289126. https://doi.org/10.3389/fpsyg.2024.1289126

Winke, P., Zhang, X., & Pierce, S. J. (2023). A closer look at a marginalized test method: Self-assessment as a measure of speaking proficiency. *Studies in Second Language Acquisition, 45*(2), 416–441. https://doi.org/10.1017/S0272263122000079

Yan, Z. (2023). The effect of self-assessment on academic performance and the role of explicitness: A meta-analysis. *Assessment & Evaluation in Higher Education*. Advance online publication. https://doi.org/10.1080/02602938.2021.2012644

Zhai, X., Chu, X., Wang, M., Tsai, C.-C., Liang, J.-C., & Spector, J. M. (2024). A systematic review of stimulated recall (SR) in educational research from 2012 to 2022. *Humanities and Social Sciences Communications, 11*, 489. https://doi.org/10.1057/s41599-024-02987-6

Zhang, T., & Mao, Z. (2023). Exploring the development of student *feedback* literacy in the second language writing classroom. *Assessing Writing, 55*, 100697. https://doi.org/10.1016/j.asw.2023.100697

Zimmerman, B. J. (2002). Becoming a self-regulated learner: An overview. Theory Into Practice, 41(2), 64–70. https://doi.org/10.1207/s15430421tip4102_2