Brazilian Journal of Music Therapy
BRjmt Revista Brasileira de Musicoterapia

BRJMT, N42, p 05 -25

DOI 10.51914/brjmt.42.2026.486
ISSN 2764-5541
https://musicoterapia.revistademusicoterapia.mus.br/index.php/rbmt
OPEN ACCESS

# BACKTRANSLATIONLLM: UMA ARQUITETURA AGÊNTICA PARA TRADUÇÃO E VALIDAÇÃO DE CONTEÚDO DE INSTRUMENTOS PSICOMÉTRICOS

*BackTranslationLLM: An Agentic Architecture for Translation and Validation of Psychometric Instruments Content*

*BackTranslationLLM: Una Arquitectura Agéntica de IA para la Traducción y Validación del Contenido de Instrumentos Psicométricos*

*Frederico Gonçalves Pedrosa*[1]
ORCID:0000-0002-0682-0734

**Resumo** - A adaptação transcultural de instrumentos psicométricos é um processo metodologicamente complexo, lento e custoso. Este estudo apresenta e avalia o BackTranslationLLM, uma arquitetura de Inteligência Artificial (IA) multiagente projetada para automatizar a tradução e validação de conteúdo de testes. Guiado por princípios de engenharia de IA, o sistema utiliza uma pipeline sequencial com agentes especializados para tradução, retrotradução e refinamento, seguida por um comitê diversificado de juízes de IA para avaliação. O sistema foi aplicado na adaptação do *Cuestionario de Impacto de la Sesión de Musicoterapia* do espanhol para o português brasileiro. Os resultados foram comparados com uma validação paralela conduzida por cinco juízes humanos. A Análise de Grafo Exploratória confirmou a preservação da estrutura semântica do instrumento. Tanto o comitê composto por juízes de IA quanto por humanos atribuiu Coeficientes de Validade de Conteúdo (CVC) adequados. Contudo, a ausência de correlação entre os CVCs item a item (e.g., Clareza, $\rho = -0,37$) revelou padrões de julgamento distintos: a IA atuou como uma auditora de conformidade metodológica, enquanto os humanos se destacaram na avaliação de nuances culturais. Conclui-se que o BackTranslationLLM é uma ferramenta eficaz que complementa, em vez de substituir, a *expertise* humana, otimizando o processo de validação.
**Palavras chave**: psicometria, musicoterapia, tradução, inteligência artificial

**Abstract** - Cross-cultural adaptation of psychometric instruments is a methodologically complex, time-consuming, and resource-intensive process. This study introduces and evaluates BackTranslationLLM, a multi-agent Artificial Intelligence (AI) architecture designed to automate the translation and content validation of psychological tests. Guided by principles of AI agent engineering, the system employs a sequential pipeline with specialized agents for translation, back-translation, and refinement, followed by a diversified committee of AI judges for evaluation. The framework was applied to adapt the *Cuestionario de Impacto de la Sesión de Musicoterapia* from Spanish to Brazilian Portuguese. Results were benchmarked against a parallel validation conducted by five human expert judges. Exploratory Graph Analysis

[1]Doutor em Música (UFMG, 2023), Mestre em Música (UFPR, 2017), Bacharel em Musicoterapia (FAP, 2011). http://lattes.cnpq.br/9227138663195042, e- mail fredericopedrosa@ufmg.br

*Recebido em 14 de Janeiro 2026, aceito em 12 de Maio de 2026*

confirmed that the automated translation preserved the instrument's semantic structure. Both the AI and human committees assigned adequate Content Validity Coefficients (CVCs) to the adapted instrument. However, a lack of correlation between the item-level CVCs (e.g., Clarity, $\rho = -0.37$) revealed divergent judgment heuristics: the AI acted as an auditor of methodological compliance, whereas human experts excelled at evaluating pragmatic and cultural nuances. We conclude that BackTranslationLLM is an effective framework that complements, rather than replaces, human expertise by optimizing the validation process.
**Keywords:** psychometrics, music therapy, translation, artificial intelligence

**Resumen -** La adaptación transcultural de instrumentos psicométricos es un proceso metodológicamente complejo, lento y costoso. Este estudio presenta y evalúa BackTranslationLLM, una arquitectura de Inteligencia Artificial (IA) multiagente diseñada para automatizar la traducción y validación de contenido de tests psicológicos. Guiado por principios de ingeniería de IA, el sistema utiliza un pipeline secuencial con agentes especializados para traducción, retrotraducción y refinamiento, seguido por un comité diversificado de jueces de IA para la evaluación. El sistema fue aplicado en la adaptación del Cuestionario de Impacto de la Sesión de Musicoterapia del español al portugués brasileño. Los resultados fueron comparados con una validación paralela realizada por cinco jueces humanos. El Análisis Exploratorio de Grafos confirmó la preservación de la estructura semántica del instrumento. Tanto el comité de IA como el de humanos asignaron Coeficientes de Validez de Contenido (CVC) adecuados. Sin embargo, la ausencia de correlación entre los CVCs a nivel de ítem (ej., Claridad, $\rho = -0,37$) reveló patrones de juicio distintos: la IA actuó como una rigurosa auditora de conformidad metodológica, mientras que los humanos destacaron en la evaluación de matices culturales. Se concluye que BackTranslationLLM es una herramienta eficaz que complementa, en lugar de sustituir, la pericia humana, optimizando el proceso de validación.
**Palabra clave**: psicometría, musicoterapia, traducción, inteligencia artificial

## 1. Introdução

A mensuração de construtos psicológicos por meio de testes com evidências de validade e fidedignidade é um pilar fundamental da psicometria, permitindo a quantificação de traços latentes e a avaliação objetiva de intervenções (Pasquali, 2009; Borsa & Seize, 2017). Em áreas aplicadas como a Musicoterapia, a utilização de testes psicométricos assume uma importância crítica, pois oferece um meio sistemático para avaliar a eficácia das sessões terapêuticas, monitorar o progresso dos pacientes e gerar evidências robustas que sustentem a prática clínica (Cripps et al., 2016; Zmitrowicz & Moura, 2018; Pedrosa, 2023). A escassez de instrumentos específicos para este campo impulsiona tanto a criação de novas medidas quanto a adaptação de instrumentos desenvolvidos em outras culturas, um processo conhecido como adaptação transcultural (Boateng et al., 2018).



O processo de adaptação transcultural de um instrumento psicométrico é uma tarefa complexa e multifacetada, que vai muito além de uma simples tradução literal. Para garantir a equivalência entre a versão original e a adaptada exige uma série de procedimentos metodológicos rigorosos (Cassepp-Borges, 2010; Silveira et al., 2018). O desafio central reside em assegurar a equivalência semântica, idiomática e conceitual dos itens, de modo que eles meçam o mesmo construto latente em ambas as culturas (Kyriazos & Stalikas, 2018). Tradicionalmente, este processo envolve múltiplas etapas, como a tradução, a retrotradução (*back-translation*), e a validação de conteúdo por um comitê de juízes especialistas (Pasquali, 2010; Hernandez-Nieto, 2002). Esta etapa, embora essencial, é intensiva em recursos, demandando tempo e a disponibilidade de múltiplos especialistas, o que pode tornar o processo lento e custoso (Nakano & Siqueira, 2012).

Diante desses desafios, avanços recentes em Inteligência Artificial (IA), particularmente em Modelos de Linguagem Grandes (*Large Language Models* - LLMs), oferecem uma oportunidade inédita para otimizar e automatizar etapas do desenvolvimento de instrumentos. Estudos recentes demonstram o potencial dos LLMs para gerar e refinar itens psicométricos com alta qualidade, reduzindo significativamente a dependência da intervenção humana (Russell-Lasalandra et al., 2026; Laverghetta Jr. et al., 2024; Laverghetta Jr. & Licato, 2023). A capacidade de construir sistemas de IA mais complexos, conhecidos como agentes, abre uma nova fronteira. É possível projetar arquiteturas multiagentes em que cada agente possua uma

função especializada e colabore dentro de um *workflow*, que é uma sequência de passos automatizados e pré-definidos, orquestrado para executar tarefas complexas (Bhagwat, 2024). Esta abordagem é guiada por *prompts*, ou seja, instruções detalhadas em linguagem natural, que permitem às ferramentas de linguagem decompor um problema, como o levantamento de validade de conteúdo[2], em subtarefas lógicas, atribuindo a cada agente responsabilidades específicas, como traduzir ou refinar itens com base em regras metodológicas predefinidas.

Inspirado por esses avanços, este estudo propõe e avalia uma nova ferramenta, denominada BackTranslationLLM, um sistema multiagente projetado para automatizar a *pipeline* – uma sequência de etapas computacionais automatizadas – de tradução, retrotradução e validação de conteúdo de instrumentos psicométricos. Desta forma, o objetivo principal deste trabalho é apresentar a arquitetura do BackTranslationLLM e avaliar sua eficácia na adaptação transcultural do *Cuestionario de Impacto de la Sesión de Musicoterapia* (CISMA; Vercher et al., 2023) do espanhol para o português brasileiro. Para isso, os resultados gerados pela validação de conteúdo do comitê de IA serão comparados com os resultados de uma validação paralela conduzida por um comitê de juízes humanos especialistas, utilizando o Coeficiente de Validade de Conteúdo (CVC; Hernandez-Nieto, 2002) como métrica principal. Adicionalmente, a equivalência da estrutura semântica entre os itens originais e traduzidos será investigada por meio da EGA (Golino & Epskamp, 2017), reforçando o levantamento de validade do instrumento.

A escolha do CISMA (Vercher et al., 2023) como instrumento-alvo para a validação desta arquitetura justifica-se por sua relevância clínica e complexidade psicométrica. No cenário brasileiro, a Musicoterapia carece de instrumentos validados que permitam o monitoramento sistemático e imediato do impacto das sessões, tornando a adaptação do CISMA uma contribuição prática direta à área. Além disso, por se tratar de um teste que avalia estados subjetivos e relacionais, o CISMA oferece um cenário de teste rigoroso para a arquitetura agêntica ao demandar que o sistema faça a tradução gramatical e demonstre sensibilidade a nuances conceituais e pragmáticas que são críticas em instrumentos de autorrelato. Dessa forma, o CISMA funciona como um caso metodológico que permite avaliar se a colaboração

[2] Neste texto usou-se preferencialmente a expressão levantamento de evidência de validade, em consonância com as normas do *Standards for Educational and Psychological Testing* (AERA *et al*., 2014). No entanto, oportunamente também se usará "validação" como sinônimo do mesmo processo.

entre múltiplos agentes de IA é capaz de preservar a densidade semântica necessária para a prática clínica e a pesquisa em musicoterapia.

## 2. Metodologia

Este estudo foi estruturado em duas etapas principais. A primeira compreende o desenvolvimento da ferramenta computacional BackTranslationLLM, um sistema multiagente baseado nos princípios da engenharia de IA. A segunda etapa descreve os procedimentos de levantamento de evidência de validade do instrumento CISMA, utilizando tanto a ferramenta de IA desenvolvida quanto a avaliação por juízes humanos e técnicas de psicometria de redes. A pesquisa faz parte do projeto "Medindo a influência das relações com a música no desenvolvimento humano: uma abordagem psicométrica" e foi aprovado pelo Comitê de Ética da UFMG (CAAE: 74309523.6.0000.5149).



É de suma importância ressaltar que o BackTranslationLLM foi projetado especificamente para o levantamento de evidências de validade de conteúdo no âmbito da adaptação transcultural de instrumentos. Isto significa que a arquitetura não é destinada à análise qualitativa de dados clínicos ou transcrições de sessões terapêuticas, tarefas que exigem outros parâmetros éticos e metodológicos.

### 2.1. Desenvolvimento do Sistema BackTranslationLLM

A construção do BackTranslationLLM foi conceitualmente guiada pelos princípios de desenvolvimento de agentes de IA (Bhagwat, 2024). O sistema integra diferentes tecnologias para executar suas tarefas. Para a tradução inicial, foi utilizada a API[3] da DeepL (DeepL SE, 2024), uma plataforma de tradução automática neural. As tarefas subsequentes de raciocínio, verificação e julgamento foram executadas por uma família diversificada de LLMs desenvolvidos pela Google[4], incluindo os modelos Gemini Pro e Flash (Google, 2024a) e

[3] API (*Application Programming Interface)* é uma interface de comunicação que permite que diferentes softwares interajam de forma segura e padronizada. Neste estudo, o uso de chaves de API nos permitiu enviar requisições programáticas para os serviços da DeepL e da Google e receber os resultados de traduções e avaliações diretamente em nosso ambiente de codificação, automatizando o processo.

[4] Foram usados modelos disponibilizados pela Google dado que esta dispõe de um limiar generoso de utilização gratuita, o que é suficiente para a tarefa aqui realizada.

Gemma (Google, 2024b). A escolha por múltiplos modelos visa introduzir diferentes "perspectivas" computacionais no processo de avaliação, simulando a diversidade de um comitê de especialistas.

A arquitetura do sistema foi dividida em duas fases automatizadas: (1) uma *pipeline* de tradução e refinamento, e (2) um comitê de juízes de IA para levantamento de validade de conteúdo. A Tabela 1 sintetiza os princípios que nortearam o design do sistema (Bhagwat, 2024) e como cada um foi implementado, garantindo um processo robusto e metodologicamente transparente.

**Tabela 1.**

*Princípios para implementação na arquitetura do BackTranslationLLM*

| **Princípio** | **A Prática Essencial** | **Implementação no BackTranslationLLM** |
|---|---|---|
| 1. Decomposição de Tarefas | Evitar um único *prompt* monolítico. Quebrar problemas complexos em uma sequência de tarefas menores e mais simples. Cada tarefa deve ter uma responsabilidade clara, tornando o sistema mais robusto e depurável. | O problema complexo de tradução e validação foi decomposto em duas fases claras: uma *pipeline* de tradução sequencial e um Comitê de Juízes paralelo. |
| 2. Design do Agente: *Prompts* | Criar instruções claras e detalhadas. Fornecer contexto, regras explícitas, definir a "persona" do agente e especificar o formato de saída desejado para garantir consistência e qualidade. | Utiliza um contexto rico e detalhado como um "*prompt* de sistema" para guiar todos os agentes. Cada agente tem um *prompt* específico que define seu papel e formato de saída. |
| 3. Design do Agente: Modelos | Selecionar o modelo certo para cada tarefa. Usar modelos mais rápidos e econômicos para operações simples, como a retrotradução e modelos mais poderosos para tarefas que exigem raciocínio complexo ou adesão estrita a regras, como julgamento. | Seleciona modelos de forma estratégica: DeepL para tradução, Gemini Flash para tarefas simples e rápidas, Gemini Pro para raciocínio complexo, e uma diversidade de modelos da Google para o comitê de juízes. |
| 4. Design do Agente: Ferramentas | Capacitar os agentes a interagir com sistemas externos. Dar-lhes acesso a APIs, bancos de dados ou outras fontes de informação para superar as limitações do modelo e executar ações especializadas. | O Agente Tradutor utiliza uma ferramenta externa especializada, a API do DeepL, para a tarefa inicial, garantindo uma base de alta qualidade. |
| 5. Orquestração de *Workflow* | Definir a lógica de como os agentes colaboram. Utilizar padrões como encadeamento, ramificação e paralelismo ou fusão para construir fluxos de trabalho previsíveis. | Fase 1: Implementa um *workflow* de encadeamento com ramificação condicional (se aprovado, termina; se reprovado, chama o refinador). Fase 2: Implementa um *workflow* de paralelismo e fusão, em que vários juízes avaliam, e os resultados são agregados. |

| | | |
|---|---|---|
| 6. Saída Estruturada | Forçar o LLM a responder em um formato de dados previsível, como JSON. Isso torna a saída do agente confiável, fácil de ser processada por código e permite a integração robusta entre as diferentes etapas do sistema. | O Agente Juiz da Fase 1 e todos os Juízes do Comitê da Fase 2 são instruídos a retornar respostas em formato JSON, tornando a saída previsível e programaticamente utilizável. |
| 7. Memória e Estado | Implementar mecanismos para os agentes reterem contexto. Isso pode ser memória de curto prazo como passar o estado entre etapas de um *workflow*, ou memória de longo prazo como armazenar conhecimento entre execuções para aprender. | Implementa memória de curto prazo ao passar o "raciocínio" sobre o erro do Agente Juiz para o Agente Refinador, permitindo uma ação contextualizada. |
| 8. Observação e Rastreabilidade | Criar "rastros" detalhados para cada execução. Capturar os *inputs*, *outputs* e metadados de cada agente e etapa, permitindo que desenvolvedores depurem falhas e entendam o "raciocínio" do sistema. | Gera relatórios finais detalhados em HTML e JSON, que servem como um "rastro" completo de cada item processado, permitindo a inspeção de cada etapa. |
| 9. Avaliação | Ir além de testes binários. Construir avaliações que medem a qualidade da saída do agente com métricas quantificáveis. Utilizar os resultados, especialmente da revisão humana, para criar um circuito de *feedback* e melhorar o sistema. | A Fase 2, com o Comitê de IA, funciona como um sistema de avaliação automatizada. Os resultados desta fase foram então submetidos a uma avaliação de padrão-ouro, comparando-os diretamente com as avaliações de um comitê de juízes humanos para analisar a concordância e identificar as forças complementares de cada abordagem. |



Cabe ressaltar que, embora a presente arquitetura tenha sido elaborada com a família de modelos Gemini 1.5, a BackTranslationLLM foi estruturada para permitir a integração de modelos de linguagem mais recentes, quando for o caso. A lógica da pipeline de agentes é adaptável, possibilitando que versões atualizadas dos LLMs sejam incorporadas conforme disponibilizadas via API, ainda que tais transições possam exigir ajustes pontuais no refinamento dos *prompts* para manter a consistência dos resultados frente às diferentes heurísticas de cada nova geração de modelos.

### 2.1.1. Fase 1: Sequência de Tradução e Refinamento

Esta fase consistiu em uma sequência com quatro agentes especializados, projetados para produzir uma tradução inicial de alta fidelidade metodológica (Cassepp-Borges, 2010). O Agente Tradutor utilizou a API da plataforma DeepL; o Agente Verificador empregou o modelo

de IA Gemini 1.5 Flash para realizar a retrotradução (*back-translation*) do item traduzido de volta para o idioma original (espanhol, no caso estudado). A escolha deste modelo teve como objetivo a otimização da velocidade e do custo para uma tarefa de verificação semântica. O Agente Juiz foi alimentado pelo modelo Gemini 1.5 Pro que atuou como auditor de qualidade ao comparar o item original, o retrotraduzido e a tradução inicial com base em um conjunto de regras metodológicas explicitadas anteriormente no Manual de Estilo (Quadro 1), fornecido em *prompt*. Sua tarefa foi retornar uma avaliação em formato JSON, indicando se o item foi aprovado e a justificativa para a decisão. Por fim, o Agente Refinador, também utilizando o Gemini 1.5 Pro, foi ativado condicionalmente apenas se o Agente Juiz reprovasse um item. Este fluxo está representado na figura 1.

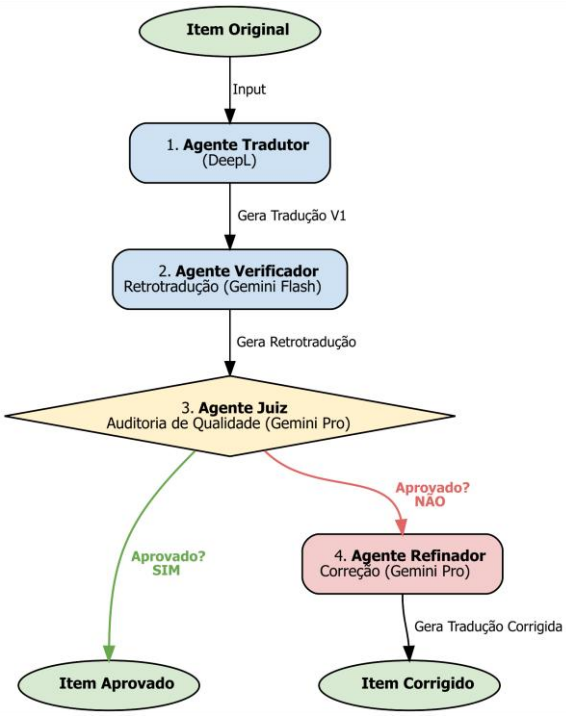
**Quadro 1**

*Manual de Estilo Fornecido aos Agentes de IA para a tradução do CISMA*

| | |
|---|---|
| Equivalência Semântica | A tradução deve manter o significado exato do item original em espanhol. Deve-se preservar a temporalidade imediata dos itens (foco no “aqui e agora”), evitando interpretações que generalizem o estado do paciente. |
| Estrutura do Item (CRÍTICO) | As perguntas devem ser interrogativas diretas (e.g. 'Eu sinto... ?', 'A musicoterapia me ajudou...?'). |
| Perspectiva Narrativa (CRÍTICO) | Este é um instrumento de autorrelato. Os itens devem ser formulados na perspectiva da primeira pessoa ('Eu', 'meu', 'me'). A tradução NÃO PODE mudar para a segunda pessoa ('você', 'seu'). Esta regra é crítica. |
| Clareza e Simplicidade | A linguagem deve ser simples, direta e sem ambiguidades. Cada item deve fazer apenas UMA pergunta. As traduções devem ser concisas e sem jargões técnicos. |
| Tom | O tom deve ser neutro e profissional, adequado para um instrumento formal de autorrelato. |
| Padrão de Avaliação | A avaliação deve ser extremamente rigorosa. Desvios menores, especialmente das regras CRÍTICAS (#2 e #3), são considerados falhas significativas. |

**Figura 1**

*Fluxo de Agentes para Tradução e Retrotradução*




A solução final de itens foi comparada com a solução inicial transformando os itens em *embeddings*, ou seja, representações vetoriais numéricas que traduzem o significado semântico do texto para um espaço multidimensional, permitindo que o computador calcule a proximidade conceitual entre as frases, e estimando a dimensionalidade pela EGA (Golino & Epskamp, 2017) e a estabilidade de cada item pela sua versão com *bootstrap*[5] (bootEGA; Christensen & Golino, 2021). Os embeddings foram gerados pelo modelo gpt 3,5 turbo.

### 2.1.2. Fase 2: Comitê de Juízes de IA

Após a tradução, refinamento e verificação da estrutura latente dos itens, um painel diversificado de LLMs atua como um comitê de juízes especialistas para realizar a validação de conteúdo. Para simular um painel multidisciplinar e reduzir o viés de um único modelo, foram empregados cinco agentes com diferentes "personas" e modelos de IA, conforme

[5] Uma técnica de reamostragem computacional que gera milhares de variações dos dados para testar se a estrutura identificada é robusta e se os itens permanecem consistentemente agrupados nas mesmas dimensões.

detalhado na Tabela 2. Cada juiz foi instruído a avaliar os itens traduzidos em três critérios: Clareza da Linguagem, Pertinência Prática e Relevância Teórica; atribuindo uma nota em uma escala Likert de 5 pontos e fornecendo uma breve justificativa para sua avaliação.

**Tabela 2**

*Personas e modelos de linguagem para cada Juiz*

| Persona do Juiz | Modelo | Justificativa da Escolha |
|---|---|---|
| Juiz 1 - Psicometrista | Gemini 1.5 *flash* | Um modelo rápido e moderno, ideal para a persona do psicometrista que precisa de avaliações objetivas e eficientes, focando na clareza e estrutura sem o custo do modelo Pro. |
| Juiz 2 - Linguista | Gemini 1.5 pro | Atribuímos o modelo mais robusto, entre os escolhidos, à análise linguística, que exige uma compreensão profunda de nuances gramaticais, fluidez e naturalidade da linguagem, tarefas em que o Pro se destaca. |
| Juiz 3 Musicoterapia com doutorado | Gemma-2-9b-it | Usar um modelo da família Gemma introduz uma arquitetura de treinamento diferente. Atribuído à persona teórica, ele pode fornecer uma "opinião" com vieses distintos dos modelos Gemini, enriquecendo a avaliação do construto. |
| Juiz 4 - Tradutor Cultural | Gemini Pro Preview | Utilizar uma versão "*preview*" ou experimental do Gemini Pro simula um especialista que está na vanguarda, potencialmente com acesso a dados mais recentes. É uma escolha para a persona que avalia vieses e regionalismos. |
| Juiz 5 - Musicoterapeuta Clínico | Gemma-7b-it | Um modelo Gemma menor simula um especialista com foco na praticidade e aplicação direta. Sua avaliação pode ser mais pragmática, e sua arquitetura diferente complementa o painel, garantindo máxima diversidade de perspectivas. |

## 2.2. Procedimentos de Validação Humana do Instrumento CISMA

### 2.2.1. Participantes

A validação por especialistas foi conduzida com um comitê de cinco juízes humanos (N = 5), selecionados intencionalmente para compor um painel multidisciplinar capaz de avaliar o instrumento sob diversas óticas. O painel foi composto por três mulheres e dois homens, com titulação variando de graduação a doutorado e atuação profissional em diferentes regiões do Brasil. A composição do comitê foi projetada para abranger as seguintes áreas de expertise:

- Expertise Psicométrica**:** representada por um psicólogo com doutorado e vasta experiência em construção e validação de instrumentos, garantindo o rigor metodológico da avaliação.

- Expertise de Domínio**:** aprofundada por três musicoterapeutas com perfis complementares: uma musicoterapeuta e linguista, trazendo a dupla perspectiva da adequação terapêutica e da precisão linguística; uma musicoterapeuta clínica com experiência em contexto hospitalar, focada na aplicabilidade prática dos itens; e uma musicoterapêuta doutoranda, que possui a vantagem única de ter residido na Argentina e já ter experiência prévia na aplicação da versão original do CISMA.
- Expertise Cultural e Idiomática: garantindo a equivalência transcultural, o painel contou com um músico profissional argentino residente no Brasil, que atuou como um "tradutor cultural", avaliando as nuances idiomáticas entre o espanhol e o português brasileiro.



A composição heterogênea do comitê foi fundamental para garantir uma avaliação robusta, que abrangesse não apenas a validade de conteúdo teórica, mas também a pertinência prática, a clareza linguística e a adequação cultural dos itens traduzidos pela pipeline BackTranslationLLM.

### 2.2.2. Instrumentos e Materiais

O instrumento-alvo foi o *Cuestionario de Impacto de la Sesión de Musicoterapia* (CISMA), composto por 14 itens (Vercher et al., 2023). Os materiais incluíram a versão original em espanhol, a versão traduzida gerada pela IA, e um formulário de avaliação para os juízes humanos, que espelhava os critérios de Clareza, Pertinência e Relevância e a avaliação gerada pelo comitê de IA, somados à caracterização sociodemográfica (e.g. idade, formação).

### 2.2.3. Procedimentos de Análise de Dados

As análises foram realizadas em quatro etapas. Primeiramente, para avaliar a equivalência da estrutura latente, foi conduzida uma Análise de Estrutura Semântica seguindo a metodologia proposta por Russell-Lasalandra et al. (2026). Os 14 itens originais (espanhol) e os 14 itens traduzidos (português) foram convertidos em vetores numéricos (*embeddings*) utilizando o modelo text-embedding-3-small da OpenAI via Python. Em seguida, foi realizada uma Análise de Grafo Exploratória (EGA) para ambos os conjuntos de *embeddings* utilizando

o software R v. 4.5.0 (R Core Team, 2025)[6], a fim de comparar visualmente a estrutura dimensional e a organização semântica dos itens antes e depois da tradução. Para estimar a concordância entre as duas versões utilizou-se o índice de Informação Mútua Normalizada (*Normalized Mutual Information* – NMI) comparando as duas redes. A literatura informa que NMIs < 0,40 indicam concordância fraca, NMIs entre 0,41 e 0,80 concordância de moderada a forte, e NMIs > 0,80, concordâncias quase perfeitas (Danon et al., 2005; Vinh et al., 2010).

Em um segundo momento, tanto as avaliações do comitê de IA quanto de humanos foram utilizadas para estimar o CVC para cada item e para o total do instrumento, de forma separada para os três critérios de Clareza, Pertinência e Relevância. O cálculo seguiu a metodologia de Hernandez-Nieto (2002), que considera a média das notas dos juízes, o valor máximo da escala e corrige o resultado para vieses de concordância ao acaso. Ainda que seja indicado por Hernandez-Nieto o ponto de corte de 0,80, a literatura nacional indica valores acima de 0,70 como aceitáveis (Silveira et al., 2018; Nakano & Siqueira, 2012). Desta forma, estratificamos os resultados em três categorias: inadequado (CVC < 0,70), limítrofe (CVC entre 0,70 e 0,79) e adequado (CVC ≥ 0,80), permitindo uma análise mais granular dos itens.

Em um terceiro momento, para investigar a concordância entre os dois comitês, foi calculada a Correlação de Spearman entre os CVCs de cada item obtidos pelos juízes humanos e pela IA. Esta análise foi complementada por visualizações de dados, como gráficos de dispersão e de linhas, para discutir as semelhanças e diferenças nos padrões de avaliação.

Por fim, foi realizada uma análise qualitativa para aprofundar a compreensão das divergências encontradas. Para esta etapa, foram selecionados os itens com as maiores discrepâncias de CVC entre os dois grupos. As justificativas textuais geradas pelos juízes de IA para esses itens foram sistematicamente comparadas com o padrão de avaliação dos juízes humanos, buscando identificar os diferentes julgamentos empregados por cada comitê.

## 3. Resultados

Nesta seção, são apresentados os resultados das três etapas de análise conduzidas para avaliar a eficácia do sistema BackTranslationLLM e a qualidade da adaptação transcultural do instrumento CISMA.

[6] Essa abordagem computacional está disponível em https://github.com/FredPedrosa/backtranslationllm.

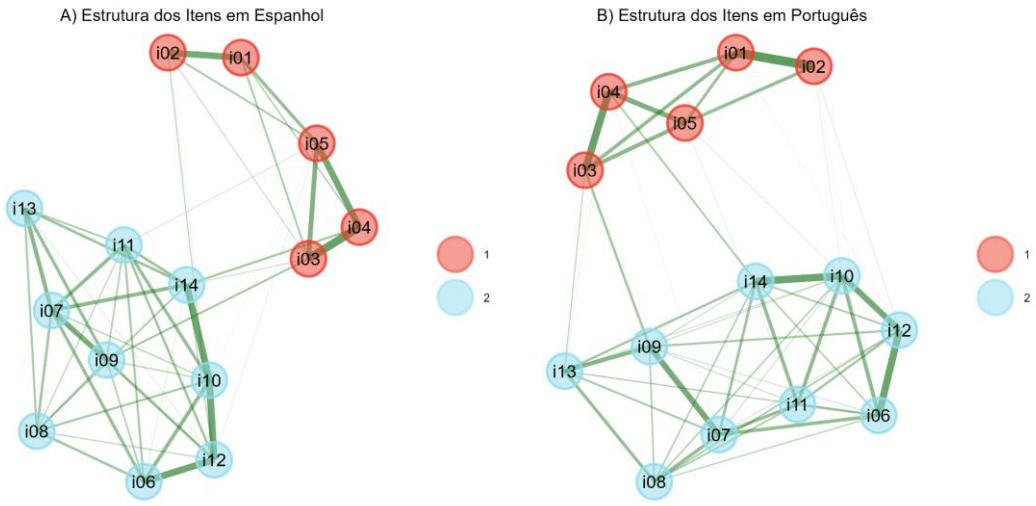
### 3.1. Equivalência da Estrutura Semântica

A comparação da estrutura latente dos itens originais em espanhol com a dos itens traduzidos para o português pelo BackTranslationLLM, realizada pela EGA dos *embeddings* de ambos os conjuntos de itens, demonstrou notável similaridade estrutural, com um NMI = 1 ($p < 0,01$). A análise visual dos grafos (Figura 2) revela que ambos os conjuntos de itens se organizaram em duas comunidades distintas, indicadas pelas cores vermelho e azul. Os itens que se agruparam na comunidade 1 no instrumento original permaneceram, em sua totalidade, agrupados na mesma comunidade após a tradução, o mesmo ocorrendo para a comunidade 2. A estabilidade dimensional, avaliada pelo processo de *bootEGA*, também se mostrou robusta em ambas as versões, com todos os itens apresentando estabilidade de replicação de 100% em suas respectivas comunidades.



**Figura 2**

*Comparação da Estrutura de Rede entre os Itens Originais e os Traduzidos*




### 3.2. Validade de Conteúdo

As avaliações de ambos os comitês (humanos e IA) foram utilizadas para calcular o CVC para os critérios de Clareza, Pertinência e Relevância. Os CVCs totais, que representam a validade geral do instrumento segundo cada grupo, foram consistentemente altos. Para o

comitê de juízes humanos, os CVCs totais foram de 0,87 para Clareza, 0,88 para Pertinência e 0,88 para Relevância. Para o comitê de IA da abordagem BackTranslationLLM, os valores foram de 0,86 para Clareza, 0,95 para Pertinência e 0,95 para Relevância.

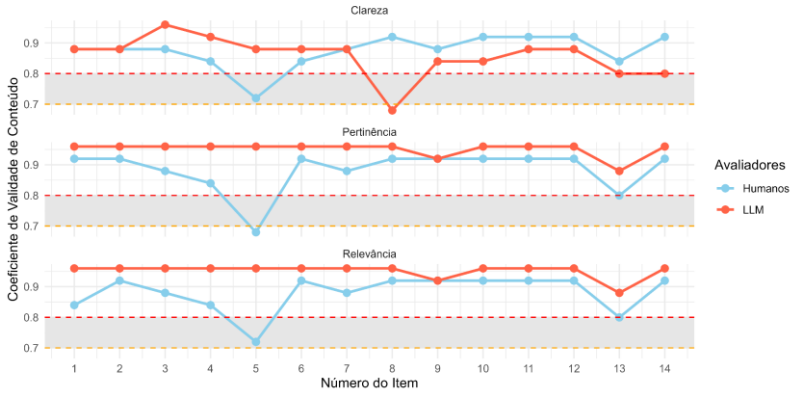
Considerando o ponto de corte de 0,80 (Hernandez-Nieto, 2002), ambos os comitês avaliaram o instrumento como possuidor de uma excelente validade de conteúdo em todos os critérios. No entanto, apenas adotando um critério mais flexível, com ponto de corte de 0,70, todos os itens individuais também foram considerados adequados na avaliação geral de ambos os grupos (figura 3).

### 3.3. Análise de Concordância entre os Comitês Humanos em relação ao composto por Juízes de IA

A análise da concordância item a item revelou padrões de avaliação distintos. A Figura 3 apresenta a comparação dos CVCs para cada um dos 14 itens, nos três critérios.

**Figura 3**

*Comparação do CVC por Item entre o Comitê de Juízes Humanos e o Comitê de IA*




*Nota*. Os resultados do CVC são estratificados em três níveis: CVC adequado (≥ 0,80), indicado acima da linha vermelha tracejada; CVC limítrofe (entre 0,70 e 0,80), destacado pela área cinzenta; e CVC inadequado (< 0,70), abaixo da linha laranja tracejada.

Visualmente, o gráfico de linhas evidencia que, embora as médias gerais sejam similares, os padrões de avaliação para itens específicos divergem. Por exemplo, para o critério de Clareza, o Item 5 foi avaliado com um CVC limítrofe pelos humanos (CVC = 0,72), enquanto o comitê de IA o avaliou positivamente (CVC = 0,88). Inversamente, o Item 8 recebeu

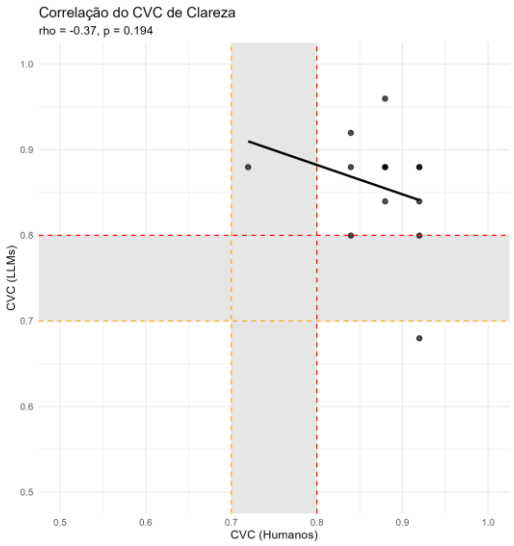
a menor avaliação do comitê de IA (CVC = 0,68), mas foi bem avaliado pelos humanos (CVC = 0,84).

Para quantificar essa relação, foi calculada a Correlação de Spearman entre os CVCs dos dois grupos. Os resultados indicaram uma correlação fraca e não significativa para todos os critérios: Clareza (ρ = -0,37, p = 0,194), Pertinência (ρ = 0,21, p = 0,470) e Relevância (ρ = 0,17, p = 0,551). A correlação negativa para o critério de Clareza, ilustrada no gráfico de dispersão da Figura 4, é particularmente notável, sugerindo uma tendência inversa no julgamento de quais itens são mais ou menos claros entre os dois grupos.



**Figura 4**

*Gráfico de Dispersão da Correlação de Spearman entre os Coeficientes de Validade de Conteúdo de Clareza*




*Nota*. Os resultados do CVC são estratificados em três níveis: CVC adequado (≥ 0,80), indicado acima da linha vermelha tracejada; CVC limítrofe (entre 0,70 e 0,80), destacado pela área cinzenta; e CVC inadequado (< 0,70), abaixo da linha laranja tracejada.

Em suma, os resultados indicam que, enquanto ambos os sistemas de avaliação chegam a conclusões semelhantes na macroestrutura, em relação à validade de conteúdo do instrumento. Por outro lado, ao nível do item as avaliações não são correlacionadas.

### 3.4. Análise qualitativa dos relatórios

Observou-se um padrão consistente em que as avaliações da IA estavam diretamente atreladas às regras metodológicas definidas no *prompt*. Por exemplo, para o Item 8 ("A sessão de musicoterapia me ajudou a lembrar ou a ativar minha memória?"), que recebeu o CVC de Clareza mais baixo do comitê (0,68), as justificativas apontaram explicitamente para violações de regras do Manual (Quadro 1). O Juiz 4 afirmou que o item violava a regra de clareza ao apresentar uma pergunta dupla (lembrar e ativar), enquanto o Juiz 1 o classificou como uma "frase declarativa, não uma pergunta interrogativa", em desacordo com a regra de número 2.

Inversamente, para o Item 5 ("Eu me sinto acompanhado?"), que obteve um baixo CVC dos humanos (0,72) mas um alto CVC da IA (0,88), as justificativas do comitê de IA foram uniformemente positivas, citando o cumprimento das regras. O Juiz 1, por exemplo, notou que o item adere perfeitamente a todas as regras, incluindo a Estrutura do Item e a Perspectiva Narrativa.

## 4. Discussão

### 4.1. Divergência nos Padrões de Julgamento

A análise qualitativa revela que o comitê de IA atua primariamente como um auditor de conformidade metodológica, avaliando os itens com base em uma aplicação literal das regras fornecidas no *prompt*. Isso ficou evidente no Item 8 quando o comitê de IA o penalizou o CVC e os juízes virtuais apontaram explicitamente a violação da regra contra perguntas duplas. Em contraste, os juízes humanos avaliaram o item positivamente, demonstrando uma flexibilidade interpretativa que prioriza o significado pragmático sobre a rigidez estrutural.

O contraponto a essa rigidez lógica é encontrado no Item 5 que foi severamente penalizado pelos especialistas humanos, mas recebeu uma boa avaliação do comitê de IA. As justificativas da IA para este item foram unânimes em sua aprovação, com comentários como "adere perfeitamente a todas as regras, incluindo a Estrutura do Item (Regra 2) e a Perspectiva Narrativa (Regra 3)". O *workflow* agêntico, ao verificar que todas as regras formais foram cumpridas, não encontrou falhas. A baixa avaliação humana, portanto, não decorre de um erro metodológico, mas de um problema de nuance semântica e cultural. O termo "acompanhado", embora tecnicamente correto, pode não ressoar adequadamente no contexto terapêutico

brasileiro, sendo percebido como frio, ambíguo ou menos apropriado que alternativas mais calorosas. Este achado sublinha a função insubstituível dos juízes humanos como validadores de adequação pragmática e cultural.

### 4.2. Implicações para a Prática Psicométrica



A divergência nos padrões de julgamento representa uma oportunidade para um novo paradigma no desenvolvimento de instrumentos. O BackTranslationLLM demonstrou ser uma possível ferramenta de interesse para uma primeira camada de validação automatizada, funcionando como um assistente em psicometria. Seu relatório detalhado e as justificativas baseadas em regras podem ser usados proativamente para refinar os itens antes de serem submetidos à avaliação humana.

Por exemplo, com base na análise do comitê de IA sobre o Item 8, um pesquisador poderia tomar uma decisão informada: manter o item, ciente de sua ambiguidade estrutural, ou reformulá-lo em duas perguntas separadas para maior clareza, como "A sessão de musicoterapia me ajudou a lembrar de coisas?" e "A sessão de musicoterapia me ajudou a ativar minha memória?". Da mesma forma, a aprovação unânime do Item 5 pela IA, combinada com a baixa avaliação humana, sinaliza que o problema não é de regra, mas de escolha lexical, direcionando o pesquisador a explorar alternativas para o termo "acompanhado".

Este processo otimizaria o trabalho dos especialistas humanos, que recebem um conjunto de itens já depurado de falhas metodológicas, permitindo que eles concentrem naquilo que a IA não conseguiu fazer: o julgamento fino de nuances culturais, idiomáticas e de fluidez. A integração de sistemas agênticos como o BackTranslationLLM representa, portanto, um avanço significativo, prometendo um fluxo de trabalho mais eficiente e colaborativo entre a precisão da máquina e a sabedoria humana.

### 4.3. Limitações e Estudos Futuros

Apesar dos resultados promissores, este estudo possui limitações que devem ser consideradas e que abrem caminhos para futuras investigações. A avaliação do instrumento CISMA se ateve à etapa de validação de conteúdo. Conforme preconiza a teoria psicométrica

(Pasquali, 2009), os próximos passos essenciais envolvem o levantamento de validação empírica do instrumento, aplicando-o a uma amostra representativa da população-alvo para conduzir análises de sua estrutura fatorial, consistência interna bem como evidências de validade baseadas na relação com variáveis externas.

Por fim, o próprio sistema BackTranslationLLM, embora eficaz, representa uma primeira iteração. A arquitetura atual não implementa um *loop* de *feedback* automatizado, em que os resultados da avaliação são usados para refinar autonomamente os *prompts* ou o comportamento dos agentes. A melhoria do sistema ainda depende da intervenção do pesquisador, um ponto que será discutido nas considerações finais. Futuros desenvolvimentos podem explorar a implementação de memória de longo prazo e mecanismos de aprendizado para tornar o sistema agêntico progressivamente mais alinhado ao julgamento de especialistas.

## 5. Considerações Finais

Este estudo se propôs a explorar uma nova fronteira na intersecção entre a Inteligência Artificial agêntica e a psicometria, desenvolvendo e avaliando uma ferramenta para automatizar a complexa tarefa de adaptação transcultural de instrumentos. A pesquisa demonstrou que uma arquitetura multiagente, guiada por princípios de engenharia de IA (Bhagwat, 2024), pode replicar com alta fidelidade o processo de tradução e retrotradução bem como realizar um processo inicial de validação de conteúdo, reduzindo consideravelmente o tempo do processo canônico e produzindo um instrumento com sólida equivalência semântica estrutural, com evidências de validade levantadas pela EGA dos embeddings, e alta validade de conteúdo geral, com alto CVC.

O principal achado, no entanto, foi que a IA não substitui os especialistas humanos, dado que opera com uma "lógica" fundamentalmente diferente. Ao atuar como uma rigorosa auditora de conformidade metodológica, a IA complementa, em vez de replicar, a capacidade humana de avaliar nuances pragmáticas e culturais. Este estudo, portanto, oferece um modelo de colaboração, no qual a máquina otimiza a precisão técnica e o ser humano aprofunda a relevância contextual.

É importante reconhecer que a ferramenta aqui apresentada, embora funcional, ainda não satisfaz todos os princípios avançados da engenharia de agentes. O conceito de avaliação iterativa, em que os resultados de uma rodada de validação são usados para refinar

automaticamente o agente para a próxima, e o de memória de longo prazo não foram implementados. O BackTranslationLLM atual opera como um sistema de passo único, em que o refinamento ainda é uma tarefa delegada ao pesquisador, que interpreta os relatórios gerados. A criação de um sistema capaz de se auto aperfeiçoar com base no feedback de especialistas, permanece como o próximo grande desafio e a evolução deste trabalho.

Mesmo com essas limitações o BackTranslationLLM não propõe a substituição da perícia humana, mas uma colaboração: enquanto a ferramenta assegura o rigor da conformidade metodológica, o clínico e o pesquisador ganham maior espaço para aprofundar a sensibilidade cultural e ética, elevando a qualidade das evidências que sustentam o cuidado ao paciente.



É fundamental ressaltar que a utilização da arquitetura agêntica na adaptação de instrumentos não deve ser confundida com traduções automatizadas simples ou sem supervisão. O sucesso do processo descrito neste estudo depende da manutenção de um crivo sistemático e do rigor metodológico da pipeline proposta, que inclui obrigatoriamente a validação final por especialistas humanos. O uso isolado de modelos de linguagem para tradução de instrumentos, sem o suporte de uma arquitetura de múltiplos agentes e sem a validação por comitês de juízes, compromete as evidências de validade e a segurança ética da aplicação clínica.

## 6. Referências

AERA - American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing*. American Educational Research Association.

Bhagwat, S. (2024). Principles of Building AI Agents (2nd ed.). Mastra.ai.

Boateng, G. O., Neilands, T. B., Frongillo, E. A., Melgar-Quiñonez, H. R., & Young, S. L. (2018). Best practices for developing and validating scales for health, social, and behavioral research: A primer. *Frontiers in Public Health*, 6, 149.

Borsa, L. C.; Seize, M.. (2017). Construção e adaptação de instrumentos psicológicos: dois caminhos possíveis. In: Damásio, B. F.; Borsa, J. C.. *Manual de desenvolvimento de instrumentos psicológicos*. São Paulo: Vetor.

Cassepp-Borges, V., Balbinotti, M. A. A., & Teodoro, M. L. M. (2010). Tradução e validação de conteúdo: uma proposta para a adaptação de instrumentos. In L. Pasquali (Org.), Instrumentação psicológica: Fundamentos e práticas (pp. 506-520). Artmed.

Cripps, C.; Tsiris, G.; Spiro, N. (2016). *Outcome measures in music therapy: A resource developed by the Nordoff Robbins research team*. 1. ed. London: Nordoff Robbins. https://eresearch.qmu.ac.uk/handle/20.500.12289/4429

Danon, L., Díaz-Guilera, A., Duch, J., & Arenas, A. (2005). Comparing community structure identification. *Journal of Statistical Mechanics: Theory and Experiment*, 2005(09), P09008–P09008. https://doi.org/10.1088/1742-5468/2005/09/P09008.

DeepL SE. (2024). *DeepL Translator*. https://www.deepl.com/

Google. (2024a). *Gemini: A family of highly capable multimodal models*. https://deepmind.google/technologies/gemini/

Google. (2024b). *Gemma: Open models based on Gemini research and technology*. https://ai.google/discover/gemma/

Hernandez-Nieto, R. A. (2002). Contributions to statistical analysis. Universidad de Los Andes.

Kyriazos, T. A., & Stalikas, A. (2018). Applied psychometrics: The steps of scale development and standardization process. *Psychology*, 9(11), 2531-2560.

Laverghetta Jr., A., & Licato, J. (2023). Generating better items for cognitive assessments using large language models. In E. Kochmar, J. Burstein, A. Horbach, R. Laarmann-Quante, N. Madnani, A. Tack, V. Yaneva, Z. Yuan, & T. Zesch (Eds.), Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023) (pp. 414–428). Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.bea-1.34

Laverghetta Jr., A., Luchini, S., Linell, A., Reiter-Palmon, R., & Beaty, R. (2024). *The creative psychometric item generator: A framework for item generation and validation using large language models*. arXiv. https://arxiv.org/abs/2409.00202

Nakano, T. C., & Siqueira, L. G. G. (2012). Validade de conteúdo da Gifted Rating Scale (versão escolar) para a população brasileira. Avaliação Psicológica, 11(1), 123-140.

Pasquali, L. (2009). Psicometria: Teoria dos testes na psicologia e na educação. Vozes.

Pasquali, L. (2010). Instrumentação psicológica: Fundamentos e práticas. Artmed.

Pedrosa, F. G. (2023). Escala de Avaliação dos Efeitos da Musicoterapia em Grupo na Dependência Química (MTDQ) [Tese, Universidade Federal de Minas Gerais]. https://doi.org/10.13140/RG.2.2.21769.04962

R Core Team. (2025). R: A language and environment for statistical computing. R Foundation for Statistical Computing. https://www.R-project.org/

Russell-Lasalandra, L.L., Christensen, A.P. & Golino, H. (2026) Generative psychometrics via AI-GENIE: Automatic item generation and validation with network-integrated evaluation. *Behavior Research Methods* 58, 217. https://doi.org/10.3758/s13428-026-03082-

Silveira, M. B., et al. (2018). Construção e validade de conteúdo de um instrumento para avaliação de quedas em idosos. *Einstein* (São Paulo), 16(2).

Vinh, N. X., Epps, J., & Bailey, J. (2010). Information theoretic measures for clusterings comparison. *Journal of Machine Learning Research*. 11(95):2837−2854.

Vercher, I. B., Soler, A. A., & Ferrari, K. D. (2023). Cuestionario CISMA - Cuestionario del Impacto de las Sesiones de Musicoterapia en Pacientes Adultos. *Brazilian Journal of Music Therapy*, (33). https://doi.org/10.51914/brjmt.33.2022.385

Zmitrowicz, M., & Moura, C. de F. (2018). Musicoterapia e reabilitação: uma revisão de escopo. Cadernos Brasileiros de Terapia Ocupacional, 26(4), 891-903.