On the Experimental Process in Evaluations of Brazilian Conversational Agents in Education

Leo Natan Paschoal AND Tayana Uchôa Conte AND and Simone do Rocio Senger de Souza

***Resumo*—Conversational agents represent applications that can interact in a natural language with humans. Due to the ability to process the human language, conversational agents have been explored in diverse areas, like e-commerce, health, entertainment, and education. Particularly, in the education field, they have been utilized as mechanisms to support the teaching of a second language, to recommend educational resources, to solve students' doubts, as a learning partner, among others. In Brazil, there is a particular interest in the educational technology community in studying these software systems. Although there is a strong interest in developing these systems, the real impact of these systems when used by students is little studied. One way to better understand the impact of technology on education based on evidence is through experimental studies. Therefore, this article presents an overview of empirical research in the context of conversational agents in education. This study presents a mapping of the Brazilian community, identifying how researchers are systematically evaluating the conversational agents produced in their research work. Among the results, we have identified the absence of experimental methodologies and taxonomies that support the planning, conduct, and reporting of experimental studies. Despite this, it was possible to recognize a set of variables that are commonly used in studies. In addition, to support the systematization of studies whose intention is to experimentally evaluate the pedagogical conversational agents, we have developed an initial set of guidelines to guide experiments in this context.**

***Index Terms*—Chatbot, chatterbot, systematic mapping.**

I. Introdução

Em razão das tarefas e operações em benefício ao usuário que os agentes conversacionais podem realizar, ao longo dos anos, um interesse particular vem surgindo em comunidades que estudam práticas e ambientes educacionais apoiados por tecnologias. Em especial, o interesse está concentrado na capacidade dos agentes conversacionais interagirem em língua natural com os usuários, o que possibilita o seu uso em diferentes níveis educacionais e áreas do conhecimento [1], [2]. Além disso, podem ser estabelecidos com diferentes propósitos, tendo as mais variadas funções, tais como: responder dúvidas dos alunos sobre o conteúdo [3], ajudar no desenvolvimento de competências linguísticas [4], acompanhar o conhecimento adquirido por alunos ao longo de uma disciplina [5], recomendar materiais didáticos e, quando inspirados em modelos de computação afetiva, demonstrar afeto pelo aluno [6].

Os agentes conversacionais são estabelecidos para contextos de uso específicos. Por exemplo, há agentes conversacionais sendo concebidos com conhecimento sobre orientação a objetos e a linguagem de programação "Java" para ajudar alunos que estão começando a aprender programação [7], agentes que conhecem regras gramaticais do inglês com a intenção de auxiliar no aprendizado de uma segunda língua [8], agentes para treinar estudantes de saúde a desenvolver habilidades interpessoais como a empatia [9], dentre outros. Muito se discute sobre o estabelecimento dos agentes conversacionais [10]. A comunidade educacional que tem interesse sobre esse tipo de sistema de software vem produzindo trabalhos nessa temática sobre utilizá-los em práticas de ensino presencial, no ensino a distância e em modelos pedagógicos que mesclam momentos presenciais e a distância [1]. Embora haja uma diversidade de estudos, os quais vêm sendo quantificados por estudos secundários recentes [10], pouco se sabe sobre a qualidade dos mesmos e o impacto real deles para os contextos aos quais eles são estabelecidos [11].

A qualidade de agentes conversacionais é um tema que tem promovido interesse na comunidade que dedica esforços para contribuir com a área. Estudos recentes estão definido métricas e abordagens adequadas para subsidiar a avaliação desse tipo de software [12], [13], [14]. Em particular, quando são estabelecidos para apoiar a temática que permeia o contexto educacional, há também pesquisas sendo realizadas, uma vez que estudos são feitos por pesquisadores de diferentes áreas, com diferentes *backgrounds*, e não se conhece procedimentos uniformes e sistemáticos dedicados a demonstração da eficácia dos mesmos enquanto mecanismo de apoio aos processos de ensino-aprendizagem [15]. De acordo com Winkler e Sollner [16], reconhecer a eficácia do agente conversacional como mecanismo de apoio ao ensino pode não ser trivial e deve depender de vários fatores, que possivelmente incluem a qualidade e o design do agente conversacional, confiança do aluno em relação à tecnologia, dentre outros.

Embora pareça existir uma complexidade acerca da avaliação sistemática dos agentes conversacionais definidos com propósitos educacionais, há de se entender que as pesquisas nessa temática estão abordando alguma evidência sobre a eficácia desse mecanismo de apoio ao ensino, uma vez que somente perspectivas e suposições não são suficientes para estimular a produção desses mecanismos. Além disso, há de se considerar que a área de agentes conversacionais tem pesquisa no contexto de agentes não funcionais, por meio do método Wizard-of-Oz[1] [18], que tem elementos característicos da pesquisa experimental. A pesquisa experimental, por sua vez, envolve observação de variáveis de interesse e dos efeitos

Manuscript received August, 2020; revised January, 2021.

[1]O método Wizard-of-Oz é usado para prototipar sistemas custosos de serem implementados, como os agentes inteligentes. Nesse método, um humano simula as funcionalidades do sistema inteligente e interage com usuários sem que eles tenham conhecimento sobre estarem interagindo com um humano ao invés de um sistema [17].

dessas variáveis [19]. Portanto, apesar de existir um déficit no que se refere à avaliação sistemática no contexto de agentes conversacionais educacionais, pode-se imaginar que estudos experimentais têm sido realizados, mesmo que de forma prematura. Assim, é preciso entender como esses estudos estão sendo planejados, conduzidos e reportados. Somente com esse entendimento será possível ajudar a comunidade de pesquisa a definir procedimentos que auxiliam na averiguação do potencial dos agentes conversacionais em práticas e contextos educacionais, dado que os estudos experimentais podem contribuir com a construção do conhecimento confiável e reduzir incertezas sobre teorias e ferramentas [20].

Com esse entendimento inicial, este artigo tem a intenção de apresentar uma pesquisa cujo objetivo principal é oferecer uma visão geral sobre a avaliação experimental da aplicação de agentes conversacionais no contexto da educação (*i.e.*, agentes conversacionais pedagógicos). Ainda, com base nessa visão geral, busca oferecer orientações para planejamento de avaliações experimentais sobre agentes conversacionais pedagógicos. Em particular, da-se ênfase a visão do campo no Brasil, uma vez que o país tem demonstrado uma quantidade significativa de produções na área sobre a temática [21]. É importante esclarecer que as pesquisas brasileiras foram selecionadas porque a comunidade está sensibilizada com a temática mas tem concentrado suas pesquisas em veículos de divulgação nacional, o que tem prejudicado a visibilidade desses estudos.

Para conduzir a pesquisa, optou-se por analisar trabalhos que originaram teses de doutorado e dissertações de mestrado. A escolha se baseia em alguma justificativas: (i) esse formato de estudo pode ser considerado como a única fonte de informação que contém o resultado de pesquisa realizada em um período de quatro a cinco anos de intenso trabalho [22]; (ii) esses trabalhos como objetos de estudos são considerados oportunos, porque refletem as tendências da pesquisa universitária na área, além de abordarem trabalhos tidos como originais, com problemáticas relevantes para a área [22]; (iii) os artigos acadêmicos publicados em veículos de divulgação científica do Brasil limitam-se a demonstrar uma funcionalidade ou a discutir um tópico específico, estando, geralmente, associados ou inseridos em contextos de pesquisas maiores, como em projetos de mestrado e doutorado [23].

Para apresentar a pesquisa, este artigo está organizado da seguinte forma. Durante a Seção II, são apresentados os procedimentos adotados no mapeamento dos estudos. Os resultados do mapeamento sistemático são apresentados na Seção III. Na Seção IV são discutidas as ameaças à validade. Na Seção V são descritas as limitações deste estudo. Na Seção VI diretrizes que apoiam o planejamento de avaliações experimentais são definidas. Por fim, na Seção VII são apresentadas as considerações finais e levantadas perspectivas de trabalhos futuros.

II. Materiais e Métodos

O mapeamento sistemático foi conduzido no período que compreende os meses de outubro de 2019 e janeiro de 2020, considerando os procedimentos para estudos secundários sistemáticos definidos por Petersen et al. [24] e Kitchenham and Charters [25]. Nesse sentido, o mapeamento englobou um processo que pode ser resumido em três fases: planejamento, condução e relatório. Cada uma dessas fases será apresentada nas seções a seguir.

A. Planejamento

O planejamento de um mapeamento sistemático envolve: (i) definição de objetivo e questões de pesquisa; (ii) estratégia de busca; (iii) critérios de seleção; (iv) procedimentos para seleção dos estudos; (v) método para extração e síntese.

1) Objetivos de pesquisa & questões de pesquisa: o objetivo da pesquisa foi definido como oferecer uma visão geral da avaliação experimental no contexto de agentes conversacionais pedagógicos. Diante disso, uma meta que auxilia a atingir esse objetivo foi estabelecida: mapear os princípios básicos do processo experimental que são adotados nos estudos selecionados. Tendo como base essa meta, a abordagem GQM (Goal–Question– Metric) [26] foi utilizada para apoiar a geração de questões que possam representar essa meta de modo quantitativo e especificar maneiras de mensurar os resultados e responder essas questões. A Tabela I apresenta as questões que foram derivadas a partir da meta do estudo e as métricas para cada questão.

2) Estratégia de busca: com a intenção de localizar estudos sobre a temática, publicados no Brasil, recorreu-se ao repositório "Biblioteca Digital Brasileira de Teses e Dissertações (BDTD)[2]". Esse repositório é um sistema que cataloga todos as teses e dissertações defendidas em programas de pós-graduação do Brasil.

Para identificar esses documentos, foi necessário definir termos de busca (palavras-chave) associados ao objetivo principal deste estudo. Inspirando-se em trabalhos anteriores (*e.g.*, [9], [15], [16]), uma *string* de busca foi formulada:

```
(''conversational agent'' OR ''chatbot'' OR
''chat bot'' OR ''chatterbot'' OR ''dialogue
    system'' OR ''pedagogical agent'')
```

Salienta-se que a busca foi feita no idioma português e inglês, utilizando o plural de cada termo de busca. Durante a busca inicial, optou-se por não utilizar termos como `''educação''` e palavras derivadas para que apenas os documentos envolvendo o assunto de agentes conversacionais fossem considerados em um primeiro momento.

3) Critérios de seleção: para apoiar a seleção de estudos relevantes que possibilitem responder as questões de pesquisa, critérios de seleção foram definidos. Além de apoiar a inclusão de estudos relevantes, eles ajudam a excluir os estudos que não são adequados ao escopo deste trabalho. Os critérios de inclusão (CI) e exclusão (CE) foram:

- CI_1: O estudo apresenta o estabelecido, aplicação, utilização ou avaliação de um agente conversacional como mecanismo de apoio ao ensino.
- CE_1: O estudo não atende ao critério de inclusão.
- CE_2: O texto completo do estudo não está disponível.

[2]Mais informações disponíveis em: <http://bdtd.ibict.br/>.

Tabela I
QUESTÕES DE PESQUISA E MÉTRICA

Questão (Q)	Métrica (M)
Q1 - Os estudos mencionaram processos que foram utilizados para apoiar a definição das avaliações experimentais?	M1.1: Número de estudos que menciona o uso de um processo ou metodologia experimental. M1.2: Número de estudos que não menciona o uso de um processo ou metodologia experimental.
Q2 - Os estudos mencionaram os objetivos das avaliações experimentais?	M2.1: Número de estudos que menciona o objetivo da avaliação experimental. M2.2: Número de estudos que não menciona o objetivo da avaliação experimental.
Q3 -Quais são os contextos das avaliações experimentais?	M3.1: Característica da turma de alunos que é utilizada na avaliação experimental (*i.e.*, nível de ensino, assunto trabalhado no estudo e que o agente conversacional possui conhecimento).
Q4 - Os estudos apresentam hipóteses para serem testadas nas avaliações experimentais?	M4.1: Número de estudos que menciona hipóteses para serem testadas ao longo da avaliação. M4.2: Número de estudos que não menciona hipóteses para serem testadas ao longo da avaliação.
Q5 - Os estudos mencionam as variáveis investigadas durante a avaliação experimental?	M5.1: Variáveis mencionadas nos estudos. M5.2: Tipo de variável mencionada. M5.3: Número de ocorrências de cada variável.
Q6 - Os estudos definem as métricas que são utilizadas nas avaliações experimentais?	M6.1: Métricas mencionadas nos estudos. M6.2: Número de ocorrências de cada métrica.
Q7 - Quais são as configurações das avaliações experimentais?	M7.1: Configurações do design experimental. M7.2: Número de ocorrência dessas configurações.
Q8 - Os estudos apresentam as análises de dados feitas durante as avaliações experimentais?	M8.1: Número de estudos que menciona a análise de dados. M8.2: Número de estudos que não menciona a análise de dados. M8.3: Tipo de análise de dados. M8.4: Número de ocorrência de cada tipo de análise.
Q9 - Os estudos apresentam as ameaças à validade das avaliações experimentais?	M9.1: Número de estudos que descreve as ameaças à validade. M9.2: Número de estudos que não descreve as ameaças à validade.
Q10 - Os estudos apresentam um *lab package* para replicação da avaliação experimental?	M10.1: Número de estudos que menciona *lab package*. M10.2: Número de estudos que não menciona *lab package*.

4) Processo de seleção: a seleção dos estudos do mapeamento sistemático envolve um conjunto de atividades [25]. Assim, em um primeiro momento, a *string* de busca é aplicada no repositório BDTD. Na sequência, os metadados dos documentos retornados pelo mecanismo de busca são coletados. A partir dos metadados, é necessário identificar e remover documentos duplicados. Após a remoção dos documentos duplicados, é feita a leitura dos títulos, resumos e palavras-chave dos estudos que restaram, identificando aqueles que atendem ao CI, descartando aqueles que servem aos CE. Por último, é feita a leitura completa dos estudos que restaram, considerando novamente os critérios de seleção.

5) Extração de dados: um formulário de extração de dados foi definido para apoiar a extração de dados. Esse formulário foi inspirado nos instrumentos definidos por Dybå e Dingsøyr [27] e Melo et al. [28]. Para abordar essas terminologias, as definições de Wohlin et al. [19] foram consideradas.

B. Condução

Seguindo o planejamento estabelecido, 177 estudos foram localizados na BDTD. Os metadados desses estudos foram coletados por meio da ferramenta Zotero[3]. Esses metadados foram importados na ferramenta Parsifal[4], que auxiliou identificar os estudos duplicados. Ao total, foram detectados 98 réplicas. Diante disso, foi feita a leitura dos títulos, resumo e palavras-chave dos documentos que restaram. Utilizando os critérios de seleção, 84 estudos foram removidos e 13 incluídos. Na sequência, a leitura completa desses estudos foi realizada e nenhum estudo foi excluído.

C. Finalização

Ao final do mapeamento, foram encontrados 13 estudos (cinco dissertações e oito teses). A lista completa dos estudos selecionados é apresentada na Tabela II. Os resultados são discutidos na Seção III.

Tabela II
LISTA FINAL DE ESTUDOS SELECIONADOS

Título do artigo	Referência
A avaliação do uso de chatterbots no ensino através de uma ferramenta de autoria	Castanho [29]
Avaliação de Faqbots através da ferramenta Autochatter	Campos [30]
Mídia e aprendizagem: um estudo comparativo entre hipertexto e chatterbot	Domingues [31]
Doroty: um chatterbot para treinamento de profissionais atuantes no gerenciamento de redes de computadores	Leonhardt [32]
i-collaboration: Um modelo de colaboração inteligente personalizada para ambientes de EAD	Oliveira [33]
A inserção de um agente conversacional animado em um ambiente virtual de aprendizagem a partir da teoria da carga cognitiva	Santos [34]
A produção de sentidos na conversação com chatterbots	Leite [35]
Aprimoramento das habilidades cognitivas de resolução de problemas com o apoio de um agente conversacional	Aguiar [36]
Desenvolvimento de chatterbots educacionais: um estudo de caso voltado ao ensino de algoritmos	Lemos [37]
Estudo de implementação de um robô de conversação em curso de língua estrangeira em ambiente virtual: um caso de estabilização do Sistema Adaptativo Complexo	Lima [38]
A mediação de um agente pedagógico na aprendizagem colaborativa de inglês como língua estrangeira	Pinho [39]
Explorando autodeterminação, utilizando novas tecnologias para ensejar autocuidado em obesos	Sgobbi [40]
Contribuições ao ensino de teste de software com o modelo flipped classroom e um agente conversacional	Paschoal [41]

III. RESULTADOS E DISCUSSÕES

Os dados coletados foram organizados e analisados, considerando as contribuições descritas nos estudos. Para apresentá-los, inicialmente os estudos serão caracterizados. Na seção de resultados, é dada ênfase na apresentação e discussão dos resultados para as questões de pesquisa.

[3]Mais informações disponíveis em: <https://www.zotero.org/>.
[4]Mais informações disponíveis em: <https://parsif.al/>.

A. Caracterizando os estudos

A origem dos estudos primários é apresentada na Tabela III. Foi possível identificar que seis universidades possuem pesquisas realizadas na temática. A instituição de ensino superior (IES) com a maior quantidade de estudos é a Universidade Federal do Rio Grande do Sul (UFRGS). Em um estudo secundário anterior [21], a UFRGS já havia sido reconhecida como a instituição de ensino brasileira com a maior produção no assunto. Nesse sentido, a descoberta vai ao encontro do que é descrito pelo trabalho de outros pesquisadores. De um modo geral, a maioria dos estudos é originário da região Sul do Brasil. Em razão disso, os grupos de pesquisa com maior produção estão concentrados nessas regiões. Com destaque para a professora Dra. Liane M. R. Tarouco[5] que foi responsável por liderar mais de 25% dessas pesquisas.

Tabela III
CARACTERIZAÇÃO INICIAL DOS ESTUDOS

	Respostas	Qt.E	%
IES de origem	Universidade Federal do Rio Grande do Sul	5	38,46
	Universidade Federal de Santa Catarina	3	23,08
	Universidade Federal de Pernambuco	2	15,38
	Universidade Federal de Minas Gerais	1	7,69
	Universidade Federal do Rio Grande do Norte	1	7,69
	Universidade de São Paulo	1	7,69
Supervisores	Liane M. R. Tarouco	4	30,77
	Raul S. Wazlawick	3	23,08
	André M. C. Campos	1	7,69
	Eliseo B. Reategui	1	7,69
	Luciano Meira	1	7,69
	Patricia C. A. R. Tedesco	1	7,69
	Ricardo A. Souza	1	7,69
	Simone R. S. Souza	1	7,69

Também investigou-se as áreas de conhecimento que têm pesquisa no contexto dos agentes conversacionais para fins educacionais. Por meio da análise dos estudos, contatou-se que a maioria dos estudos provêm de programas de pós-graduação que dialogam com Computação – seis estudos produzidos na Ciência da Computação e quatro na Informática na Educação. Conforme esperado, em uma menor proporção, foram detectados estudos realizados em outras áreas de conhecimento. Essa descoberta está alinhada com o posicionamento de Hobert [15], que já havia apontado que pesquisadores com formações diferentes da Computação têm investido esforços para estudar o uso desse tipo de sistema em seus domínios.

Acredita-se que existe uma tendência no surgimento de estudos sobre a temática em diferentes domínios. Com base nesse entendimento, o interesse pelo tema em relação às áreas de conhecimento foi analisado. Diante dos resultados apresentados na Figura 1, notou-se que os primeiros estudos foram publicados em 2002, na Computação. Um estudo anterior que mapeou estudos indexados na Scopus [42], indicou que as primeiras publicações sobre a temática que são indexadas nessa base também são de 2002. Por conta disso, acredita-se que o interesse de pesquisadores brasileiros em relação ao tema começou quando os primeiros estudos na área tiveram início. Ainda, nota-se que a quantidade de estudos se mantém estável. Na década passada foram produzidos sete estudos enquanto que na década atual existem seis estudos.

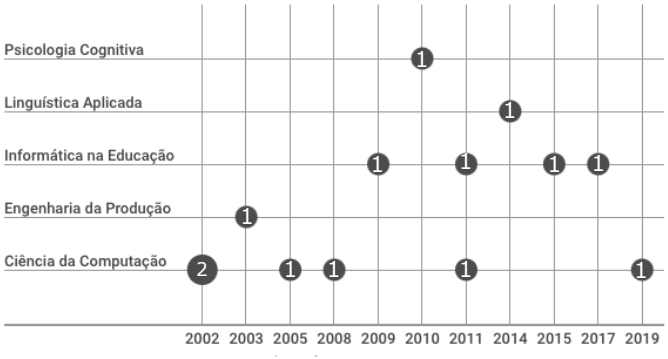

Figura 1. Relação entre o ano e a área de publicação

B. Práticas experimentais observadas

Considerando o objetivo principal deste artigo, visando responder as questões de pesquisa, a Tabela IV foi elaborada.

Tabela IV
RESULTADOS PARA CADA QUESTÃO DE PESQUISA

Q	Resposta	Qt.E	%
Q1	Não	13	100,00
Q2	Não	7	53,85
	Sim	6	46,15
Q3	Figura 2		
Q4	Não	10	76,92
	Sim	3	23,08
Q5 $*$	Não	9	69,23
	Sim	4	30,77
Q5 $**$	Sim	12	92,31
	Não	1	7,69
Q6	Sim	11	84,62
	Não	2	15,38
Q7	Não	13	100,00
Q8	Sim	11	84,62
	Não	2	15,38
Q8$\star$	Estatística descritiva	7	53,85
	Estatística inferencial	4	30,77
Q9	Não	13	100
Q10	Não	12	92,31
	Sim	1	7,69

Legenda da tabela:
$*$ Variável independente. $**$ Variável dependente. $\star$ Tipo de análise estatística.

Na primeira questão (Q1), foi analisado se os estudos citaram alguma metodologia para o planejamento e condução do experimento (*e.g.*, o processo de experimentação de Wohlin et al. [19]). Esse tipo de metodologia apoia a elaboração do estudo experimental e guia o pesquisador na tomada de decisões. Observou-se que nenhum estudo mencionou o uso de alguma metodologia ou processo experimental. Esse resultado já era esperado, pois os estudos anteriores já mencionavam que não há sistematização nesse tipo de estudo que vem sendo realizado pela nossa comunidade [15]. Acredita-se que a comunidade ainda não tem um entendimento concreto sobre experimentos, não conhecendo processos e taxonomias.

Quando se vai conduzir um estudo experimental, é primordial que durante o planejamento seja definido o objetivo

[5]Mais informações disponíveis em <http://lattes.cnpq.br/0878410768350416>.

da avaliação. Nesse sentido, foi verificado se os estudos mencionam o objetivo para o estudo experimental (em Q2). Constatou-se que mais de 50% dos estudos não definiram um objetivo específico para o experimento. Salienta-se ainda que em dois estudos os autores repetem no objetivo do experimento o objetivo principal das suas teses. Nesse sentido, é possível que não sejam objetivos experimentais. É importante esclarecer que se considerou como objetivo do experimento aquilo que os autores definiram como o objetivo da avaliação que eles realizaram. Durante a análise, poderia ter sido usado o abordagem de Basili [43] para tentar compreender se há um objetivo claro em cada estudo. No entanto, em razão de não ter sido observado um processo experimental durante a análise da Q1, optou-se por considerar apenas o que os autores assumem como objetivo.

Outro aspecto importante é o contexto em que o experimento é realizado. Por exemplo, se o agente conversacional é definido para apoiar alunos que estão aprendendo teoria geral da administração e ele for avaliado com aluno que não estão estudando esse assunto, é provável que um contexto inadequado tenha sido selecionado. Diante disso, a procedência dos sujeitos do experimento foi analisada (Q3). Para tanto relacionou-se a área de origem dos participantes do estudo com o nível escolar dos mesmos (Figura 2). Tendo como exemplo a Ciência da Computação, notou-se que em três estudos, os agentes foram avaliados por alunos de graduação de Computação. Em uma avaliação, os participantes não estavam vinculados a cursos de educação formal, portanto, podem não ter formação em computação. Também contatou-se que dois estudos não mencionaram o perfil de aluno que utilizou o agente conversacional. Não mencionar o contexto ao qual o experimento é executado pode dificultar o entendimento sobre a viabilidade do agente quanto mecanismo de apoio àquele assunto. Finalmente, foi identificado um agente conversacional dedicado ao ensino de língua estrangeira, avaliado por pessoas com diferentes níveis de formação e, portanto, o contexto selecionado e usado para o estudo foi classificado como "outro".

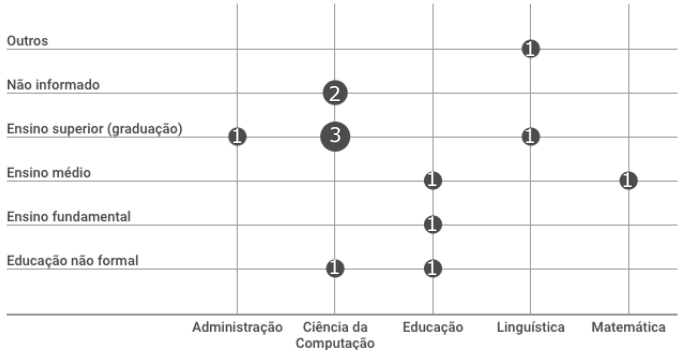

Figura 2. Contextos das avaliações

Os experimentos surgem a partir de hipótese do pesquisador. Diante disso, é de se imaginar que os estudos que exploram os agentes conversacionais como mecanismo de apoio ao ensino descrevam as hipóteses do seu estudo. Por exemplo, se um agente é desenvolvido para resolver dúvida dos alunos sobre uma dada temática, pode-se criar uma hipótese que ele resolve dúvidas de maneira mais eficaz que um FAQ, por exemplo, sob determinadas condições. Nesse sentido, no mapeamento, foi feita uma verificação se os estudos mencionam hipótese de pesquisa para as avaliações (Q4). Constatou-se que menos de 25% dos estudos mencionam hipóteses. Esse resultado pode ser considerado inquietante, pois indicaria que as pesquisas da comunidade estão sendo construídas sem comprovação sobre o impacto gerado pela tecnologia de agentes conversacionais para o contexto educacional.

Os efeitos promovidos por um agente conversacional devem ser mensurados por intermédio de variáveis. Nesse sentido, é necessário analisar os efeitos promovidos por um tratamento em uma determinada variável ao final do experimento [19]. Tendo como base o entendimento, após a investigação sobre a presença de hipóteses, em Q5 as variáveis abordadas ao longo das avaliações experimentais foram analisadas. Vale ressaltar que se buscou reconhecer as variáveis independentes[6] e as variáveis dependentes[7] mencionadas pelos autores.

Nenhum estudo mencionou com clareza as variáveis independentes. Apesar disso, é possível assumir que quatro estudos consideraram variáveis independentes, são elas: (i) mídias educacionais, (ii) ferramentas de comunicação, (iii) exercícios de aprendizagem, (iv) ambiente virtual de aprendizagem. Cada uma dessas variáveis possuem tratamentos específicos. As mídias envolvem agente conversacional, mídia impressa e um sistema hipertexto. Ferramentas de comunicação englobam interações com ferramenta síncrona (agente conversacional e ferramenta assíncrona (e-mails). No que se refere aos exercícios de aprendizagem, eles são propostos para serem solucionados com suporte de um agente conversacional e sem suporte do agente conversacional. Em relação ao ambiente virtual de aprendizagem, assume-se duas versões do ambiente virtual, em uma versão o ambiente é disponibilizado em conjunto com um agente conversacional e na outra versão do ambiente não há um agente.

Os estudos também não descrevem as variáveis dependentes. Em virtude disso, fez-se uma análise dedutiva inspirando-se em Roos [44]. Por meio dessa análise, foram reconhecidas variáveis que podem ser consideradas como dependentes em 92,31% dos estudos. Essas variáveis são apresentadas na Tabela V. Ainda que os pesquisadores não tenham explicitado essas variáveis como dependentes, há de se considerar que a preocupação dos pesquisadores tem sido bastante diversificada. O que mais se observou foi a identificação da percepção que os alunos tiveram sobre o agente conversacional.

Considerando que uma variável dependente precisa ser medida, em Q6 investigou-se as métricas citadas nos estudos. De maneira similar às variáveis, os autores/pesquisadores não costumam definir quais são as métricas utilizadas. Por conta disso, buscou-se inferi-las. Não foi possível perceber as métricas de dois estudos. Nos demais, a maioria das métricas é a quantidade de resposta emitida por alunos em questões de formulários definidos pelos próprios pesquisadores. Isso

[6]Chama-se de variáveis independentes aquelas que produzem algum efeito sobre as variáveis dependentes [19].

[7]Essas variáveis representam o efeito que causado pelas variáveis independentes [19].

Tabela V
VARIÁVEIS ABORDADAS NOS ESTUDOS

Variáveis dependentes	**Descrição**	**Qt.E**
Percepção dos estudantes	Caracteriza-se por representar as opiniões dos estudantes sobre o agente conversacional	6
Potencialização do aprendizado	Caracteriza-se por representar o desempenho obtido pelos estudantes em atividades a partir do uso do agente conversacional	4
Corretude	Caracteriza-se por representar se a mensagem gerada pelo agente conversacional está correta	2
Motivação	Caracteriza-se por representar as mudanças nos aspectos motivacionais dos estudantes a partir da interação com o agente conversacional	2
Percepção dos professores	Caracteriza-se por representar as opiniões dos professores sobre o agente conversacional	2
Usabilidade	Caracteriza-se por representar a capacidade do agente conversacional ser usado por usuários específicos para alcançar objetivos específicos em um dado contexto	2
Demanda mental	Caracteriza-se por representar o esforço que o aluno teve para realizar atividades com apoio do agente conversacional.	1
Ergonômia	Caracteriza-se por representar a conformidade da interface do agente conversacional com recomendações ergonômicas	1
Frustração	Caracteriza-se por representar o grau de insegurança e descontentamento do aluno durante a realização de uma atividade	1
Pressão de tempo	Caracteriza-se por representar a percepção do aluno sobre o tempo demandado para realizar um atividade	1
Representação de ilusão de vida	Caracteriza-se por representar os traços que transmitem 'ilusão de vida' para um humano	1
Satisfação	Caracteriza-se por representar o grau de satisfação do aluno com o agente conversacional	1

significa que a reutilização de instrumentos já desenvolvidos, principalmente aqueles que tem um processo de validação reconhecido em determinadas temáticas (*e.g.*, usabilidade e motivação), não tem sido adotada.

Após a investida no mapeamento das métricas, tentou-se entender as configurações dos experimentos (Q7). Observou-se que nenhum estudo descreve o design experimental e, por conta disso, não foi possível observar as configurações. Nesse sentido, houve uma tentativa de assimilar os fatores investigados e seus respectivos tratamentos. No entanto, algumas dificuldades surgiram, porque os estudos misturavam conceitos. Um estudo, por exemplo, descreveu que a qualidade do vocabulário do agente conversacional é um fator. Em outro estudo, o recurso instrucional utilizado pelo aluno é considerado o fator. Como não foi possível obter um entendimento correto sobre fatores, nem sobre tratamentos, não foi possível inferir as configurações dos experimentos.

Os testes estatísticos usados para analisar os resultados dos estudos também foram observados (Q8). Contatou-se que dois estudos não mencionam quaisquer testes. Eles apenas relatam sobre o uso dos agentes conversacionais e as percepções dos alunos não são quantificadas. Dentre os 11 estudos que mencionam algum teste estatístico, somente quatro utilizaram algum teste de estatística inferencial. Os demais utilizaram medidas de dispersão e posição para descrever os seus resultados.

Como todo experimento está sujeito a ameaças à validade, em Q9 buscou-se mapear se os estudos descrevem algum tipo de ameaça. Apesar da tentativa, não foi possível observar quaisquer citações à ameaças à validade. Isso oferece respaldo a possíveis discussões quanto à validade dos estudos, uma vez que não se sabe como os elementos que são capazes de prejudicar a qualidade dos estudos foram controlados e fatores que afetam a execução do estudo foram mitigados. Por exemplo, como a maioria dos estudos utiliza formulários próprios, seria necessário que os autores/pesquisadores explicassem como eles garantiram que aquele formulário coleta as informações adequadas. Vale salientar que as ameaças à validade precisam ser descritas e os autores precisam esclarecer como lidar com tais ameaças. A apresentação e discussão das ameaças à validade visa ampliar a confiança de outros pesquisadores nos resultados que foram encontrados [45].

Por fim, na última questão (Q10), buscou-se verificar se os autores prepararam um ambiente que permitisse a replicação dos seus estudos e respetivas avaliações. Isso implica em disponibilização do agente conversacional, formulários, dados coletados e análises realizadas (*i.e.*, um *lab package* [19]). Foi possível observar que somente um estudo disponibilizou o material utilizado durante o estudo. Portanto, é possível concluir que os estudos têm dificultado a replicação e reúso de materiais produzidos. A falta da disponibilização dos materiais usados no experimento afeta a credibilidade das avaliações realizadas e pode gerar discussões como a probabilidade das experiências publicadas relatarem resultados errôneos [46].

IV. AMEAÇAS À VALIDADE

O estudo foi projetado cuidadosamente, mas, por ser um estudo baseado em evidências, ele está sujeito a ameaças que possam invalidá-lo. Nesse sentido, visando mitigar possíveis ameaças, algumas ações foram tomadas:

Ausência de estudos relevantes: foi usada uma biblioteca digital específica para recuperar as teses e dissertações. Embora essa biblioteca tenha o compromisso de mapear todos os resultados produzidos por programas de pós-graduação do Brasil, é possível que algumas teses e dissertações não tenham sido catalogadas pela BDTD. Para mitigar essa ameaça, foi usada a técnica *snowballing* [19]. Para o contexto deste estudo foi utilizado o *backward snowballing*, assumindo como conjunto inicial *snowballing* os estudos 13 estudos obtidos por meio da busca feita no mapeamento sistemático. A partir disso, os estudos que foram citados pelo conjunto inicial foram analisados. Como não foi possível identificar dissertações ou teses associadas com a temática e citadas nas referências do conjunto inicial, foi considerada somente uma interação e assumiu-se que o *snowballing* não retornou outros (novos) estudos. Apesar desse esforço, nenhum outro estudo relevante foi identificado.

Confiabilidade da seleção: a seleção dos estudos foi baseada na recuperação dos mesmos por meio de uma *string* de busca e na adequação dos estudos aos critérios de inclusão e exclusão. Como a ausência de alguns termos pode diminuir a precisão do mapeamento, optou-se por selecionar um número significativo de palavras-chave como candidatas. Além disso,

a *string* de busca não foi especializada para o domínio da educação, abordando assim o termo genérico e excluindo estudos não relevantes de forma manual, com auxílio de critérios de inclusão e exclusão. Para complementar, em razão do uso do *snowballing*, a cadeia de pesquisa foi evoluída, atenuando ainda mais essa ameaça.

Viés na extração de dados: o modo como os dados são extraídos dos estudos selecionados pode comprometer o mapeamento. Nesse sentido, em um primeiro momento um formulário de extração de dados foi, seguindo as recomendações de Petersen et al. [24] e Kitchenhan and Charters [25]. Adicionalmente, em virtude da necessidade de extrair informações associadas a experimentação, o formulário foi inspirado nos estudos de Dybå e Dingsøyr [27] e Melo et al.[28]. Nesse sentido, os campos do formulário continham opções de resposta de padrões da experimentação, mais especificamente de classificações descritas em Wohlin et al. [19]. Essas classificações são conhecidas pelos pesquisadores deste estudo e dois deles são especialistas em experimentação. Por conta disso, acreditamos que o viés da extração tenha sido mitigado.

Viés dos pesquisadores: o autor principal deste estudo pode ter introduzido seu viés durante o planejamento e execução do mapeamento. Para mitigar essa ameaça, dois pesquisadores mais experientes analisaram o protocolo e se envolveram na revisão do mapeamento (incluindo desde o protocolo até a síntese e escrita dos resultados). Ainda assim, é possível que o viés tenha sido introduzido porque um dos trabalhos selecionados é de autoria de um dos pesquisadores que conduziu este mapeamento. Nesse sentido, é possível que isso tenha influenciado na seleção dos estudos para refletir nossas preferencias pessoais. Para mitigar essa ameaça, foram usados critérios de seleção objetivos.

Replicabilidade do estudo: a replicação de um mapeamento sistemático por outros pesquisadores pode levar a diferente conjunto de estudos, dado o viés que os pesquisadores incluem ao conduzir a seleção. Para mitigar essa ameaça, todas as medidas e ações tomadas durante o mapeamento foram documentadas e reportadas neste estudo. Além disso, todos os formulários e dados coletados foram incluídas em um repositório aberto o repositório Zenodo[8], no seguinte endereço: <http://doi.org/10.5281/zenodo.4625860>.

V. LIMITAÇÕES

Apesar da seção anterior descrever algumas ações que visam ampliar a generalização dos achados, não se pode deixar de considerar que este estudo possui algumas limitações. Acredita-se que a principal esteja relacionada ao mapeamento contemplar somente trabalhos realizados por pesquisadores brasileiros. Portanto, os resultados não podem ser representativos em uma escala mundial. Apesar dos trabalhos anteriores já oferecerem alguns argumentos que há falta de sistematização nos procedimentos usados na avaliação, considerando o âmbito internacional, este estudo analisou produções brasileiras. Outra limitação é a fonte de informação utilizada para extrair os dados. Como foram consideradas teses e dissertações, é possível que trabalhos relevantes tenham sido desconsiderados, por serem feitos por pesquisadores não vinculados à programas de pós-graduação.

[8]Mais informações disponíveis em: <https://zenodo.org/>.

VI. DIRETRIZES PARA AVALIAÇÃO EXPERIMENTAL DE AGENTES CONVERSACIONAIS

Esta seção apresenta diretrizes para o planejamento de avaliações experimentais no contexto dos agentes conversacionais pedagógicos. Serão fornecidas orientações para seguir o processo experimental, de modo contemplar aspectos como o objetivo da avaliação, a configuração do experimento, seleção de participantes, variáveis para determinar o efeito de uma variável sobre a outra. Esses direcionamentos provêm dos resultados obtidos por meio do mapeamento sistemático, assim como, de experiências que os pesquisadores possuem na temática.

As subseções que incorporam esta seção abordam as etapas do processo de planejamento de uma avaliação experimental. Espera-se que os direcionamentos auxiliem pesquisadores a elaborarem avaliações experimentais melhor planejadas, de modo que possam ser replicadas e comparadas com outros estudos.

A. Definição do objetivo

A primeira etapa do planejamento de um estudo/avaliação experimental é a definição do seu objetivo. Nesse sentido, é preciso deixar claro qual a meta da avaliação, de modo que seja possível verificar ao final da avaliação se a meta foi alcançada. Nessa perspectiva, surge o *template* para definição de escopo de Basili e Rombach [47], que pode ajudar os pesquisadores a definir o objetivo do estudo. Esse *template* organiza o objetivo em cinco parte, de modo a contemplar o objeto que está sob estudo, o propósito do estudo, o aspecto que será contemplado durante o estudo, a perspectiva de quem interpretará os resultados do estudo, e o contexto ao qual o estudo será conduzido [19].

A Tabela VI apresenta um exemplo de uso do *template* de Basili e Rombach [47] para o contexto de avaliações de agentes conversacionais pedagógicos. Os projetistas das avaliações experimentais podem utilizar o exemplo para definir o objetivo do seu experimento. Um exemplo de uma instância da tabela é apresentado a seguir: Analisar o agente conversacional TOB–STT[9] [48], com o propósito de avaliar a qualidade das respostas fornecidas, no que diz respeito à corretude, do ponto de vista dos pesquisadores, no contexto de estudantes de engenharia de software, estudando o conteúdo de teste de software.

B. Formulação de hipóteses

Ao longo de uma avaliação experimental são coletadas evidências. Essas evidências podem atingir alguma expectativa do pesquisador. Essas expectativas precisam ser formalizadas no planejamento da avaliação, no molde de hipóteses. Em geral, esse tipo de avaliação costuma apresentar dois tipos de hipótese: (i) hipótese nula e (ii) hipótese alternativa [19].

[9]TOB–STT é um agente conversacional para apoiar alunos que estão com dúvidas ao aprender teste de software.

Tabela VI
TEMPLATE PARA DEFINIÇÃO DE OBJETIVO

	Descrição	Valores de referência
Objeto de estudo	Aspecto que descreve as entidades que serão investigadas no decorrer da avaliação experimental	Agente conversacional pedagógico
Propósito	Aspecto que descreve a intenção do estudo	Avaliar, analisar, comparar, compreender, mensurar
Foco na qualidade	Aspecto que descreve o que está sendo avaliado	Potencialização do aprendizado, corretude, satisfação, etc (Veja a Tabela V)
Perspectiva	Aspecto que descreve o ponto de vista que os resultados serão interpretados	Pesquisador, alunos
Contexto	Aspecto que descreve o ambiente em que a avaliação experimental será realizada	Disciplina, perfil dos estudantes, ambiente de aprendizagem

A hipótese nula deve retratar em formato de afirmação que não há relação ou diferença entre variáveis observadas em um estudo. Por exemplo, um agente conversacional ajuda os alunos a resolver suas dúvidas com a mesma eficácia que um FAQ (*Frequently Asked Questions*) [49]. Formalmente, essa hipótese pode ser descrita como:

$$\text{H0:} \quad \text{Eficácia}_{\text{Agente Conversacional}} = \text{Eficácia}_{\text{FAQ}}$$

A hipótese alternativa difere-se da hipótese nula porque aborda que existe alguma correlação ou diferença entre variáveis que estão sendo observadas. Por exemplo, um agente conversacional ajuda os alunos a resolver suas dúvidas com mais eficácia que um FAQ. Formalmente, essa hipótese pode ser descrita como:

$$\text{Ha:} \quad \text{Eficácia}_{\text{Agente Conversacional}} > \text{Eficácia}_{\text{FAQ}}$$

Ainda, é possível que uma avaliação experimental contemple mais de uma hipótese alternativa. Considerando o exemplo usado anteriormente, poderiam existir outras duas hipóteses: (i) um agente conversacional ajuda os alunos a resolver suas dúvidas com menos eficácia que um FAQ; ou, (ii) um agente conversacional ajuda os alunos a resolver suas dúvidas com uma eficácia diferente de um FAQ. Formalmente, essas hipóteses seriam descritas da seguinte forma:

$$\text{Ha2:} \quad \text{Eficácia}_{\text{Agente Conversacional}} < \text{Eficácia}_{\text{FAQ}}$$
$$\text{Ha3:} \quad \text{Eficácia}_{\text{Agente Conversacional}} \neq \text{Eficácia}_{\text{FAQ}}$$

C. Seleção de variáveis

As variáveis representam as causas e os efeitos observados ao longo de uma avaliação experimental. As variáveis independentes representam as causas e as variáveis dependentes representam os efeitos. Uma descrição sobre cada uma delas é apresentada a seguir:

- **Variáveis independentes:** variáveis independentes podem representar os fatores e tratamentos considerados na avaliação experimental. Esses fatores e tratamentos são usados pelo pesquisador para compreender os resultados de uma avaliação e o impacto dessas variáveis nas variáveis dependentes. Para o contexto de avaliações experimentais relacionadas aos agentes conversacionais pedagógicos, pode-se tomar como variáveis independentes: (i) agentes conversacionais desenvolvidos com diferentes técnicas de projeto (*e.g.*, agente X desenvolvido com a técnica de casamento de padrões e agente Y desenvolvido com a técnica baseada em intenções); (ii) diferentes mecanismos de apoio para oferecer suporte educacional (*e.g.*, agente conversacional, FAQ, sistema hipermídia, sistema tutor inteligente); (iii) ambientes de aprendizagem (*e.g.*, agente disponível em um AVA, agente disponível em um objeto de aprendizagem, agente disponível em um aplicativo); dentre outras.
- **Variáveis dependentes:** as variáveis dependentes são responsáveis por reportarem os efeitos das variáveis independentes. Um exemplo é a satisfação do aluno. A Tabela V apresenta um conjunto de variáveis com base nos estudos identificados neste mapeamento sistemático.

D. Seleção de participantes

Os participantes são responsáveis por se envolver na execução do experimento. No contexto das avaliações experimentais sobre agentes conversacionais pedagógicos, os participantes podem ser alunos, professores e até mesmo tutores, sujeitos que irão interagir com os objetos de estudo. Esses sujeitos, idealmente, precisam ser selecionados de forma aleatória de uma população. Por exemplo, se o objeto de estudo for um agente conversacional para ensinar a língua inglesa, deveriam ser considerados participantes aleatórios de toda a população de indivíduos que estuda a língua inglesa. Entretanto, muitas vezes é inviável obter tal amostragem. Assim, os pesquisadores podem utilizar amostragem por conveniência, ou seja, alunos de um curso de língua inglesa de uma determinada escola de idiomas.

E. Escolha do tipo de configuração do experimento

O tipo de configuração da avaliação experimental depende da quantidade de fatores e tratamentos considerados na definição das variáveis independentes. Nesse sentido, é possível projetar um experimento com diferentes configurações. De acordo com Wohlin et al. [19] as configurações mais típicas em estudos que abordam avaliações experimentais são: um fator com dois tratamentos; um fator com mais de dois tratamentos; dois fatores com dois tratamentos; mais de dois fatores, cada um com dois tratamentos.

Ao se considerar os exemplos de variáveis independentes descritos na Seção VI-C, foram descritos diferentes fatores e tratamentos. A Tabela VII agrupa os fatores e tratamentos exemplificados e indica a configuração de cada um deles. Nota-se que nenhum dos exemplos possui mais de um fator, mas há casos em que o pesquisador precisa considerar mais de um fator. Considerando esses exemplos, em experimentos do tipo "um fator com dois tratamentos", busca-se comparar os dois tratamentos, um com o outro. Por outro lado, em experimentos "um fator, com mais de dois tratamentos", objetiva-se comparar um tratamento com os outros.

Vale salientar que o tipo de configuração do experimento delimita os tipos de testes de inferência estatística que podem ser abordados. A seguir são apresentadas especificações de

Tabela VII
EXEMPLOS DE CONFIGURAÇÕES DE AVALIAÇÕES EXPERIMENTAIS

Fatores	Tratamentos	Configuração
Técnicas de projeto	Casamento de padrões Intenções	Um fator, dois tratamentos
Mecanismos de apoio ao ensino	Agente conversacional FAQ Sistema hipermídia Sistema tutor inteligente	Um fator, quatro tratamentos
Ambientes de aprendizagem	Agente disponível em um AVA Agente disponível em um objeto de aprendizagem Agente disponível em um aplicativo	Um fator, três tratamentos

teste estatísticos para as configurações apresentadas na Tabela VII. Outras sugestões de testes podem ser conferidas em Morettin e Bussab [50] ou em Allua e Thompson [51], estudos que abordam testes para avaliações experimentais que possuem mais de um fator sob estudo.

- **Um fator, dois tratamentos:** para avaliações com configuração do tipo "um fator com dois tratamentos", considerando que cada tratamento foi executado com sujeitos independentes (*i.e.*, cada participante se envolveu em um único tratamento), deve-se observar a normalidade dos dados (para isso, pode-se usar o teste de Shapiro-Wilk) e se a distribuição for normal pode-se usar o teste t de Student. Caso os dados não sigam uma distribuição normal, pode-se usar o teste de Mann-Whitney.
- **Um fator, mais de dois tratamentos:** para avaliações com configuração "um fator com três tratamentos" ou "um fator com quatro tratamentos", considerando que cada tratamento foi executado com sujeitos independentes (*i.e.*, cada participante se envolveu em somente um tratamento), deve-se observar a normalidade dos dados (sugere-se o teste de Kolmogorov - Smirnov) e se a distribuição for normal pode-se usar ANOVA unifatorial. Caso os dados não sigam uma distribuição normal, pode-se usar o teste de Kruskal-Wallis (KW).

F. Instrumentação

Durante o planejamento da avaliação experimental é preciso reconhecer e elaborar os instrumentos que serão utilizados nas sessões experimentais. A instrumentação consistirá em diferentes artefatos, tais como:

- **Termo de consentimento livre e esclarecido:** o termo deve conter orientações com as informações que buscam esclarecer aos participantes do experimento sobre o objetivo da avaliação e em relação ao uso dos dados que serão coletados no decorrer da avaliação.
- **Objetos do experimento:** diferentes objetos podem ser utilizados na avaliação, tais como: ambientes virtuais, mecanismos de apoio ao ensino, versões de agentes conversacionais, dentre outros. Os objetos dependem dos tratamentos e/ou fatores que serão utilizados na avaliação experimental.
- **Mecanismos para coleta de dados:** a coleta de dados pode ser feita por formulários desenvolvidos pelos pesquisadores ou recuperados na literatura, ferramentas que armazenam *logs* das interações entre os participantes com o (s) agente (s) conversacional (is), capturas de telas, gravações, dentre outros.

G. Avaliação da validade

Durante o planejamento da avaliação experimental, os pesquisadores precisam identificar possíveis ameaças que podem prejudicar a validade do estudo. Ao identificá-las, ainda, é preciso estabelecer estratégias para mitigá-las. Por exemplo, se o experimento será realizado em sala de aula, uma ameaça à validade pode ser o barulho dos corredores feitos por sujeitos externos aos experimentos. Uma forma de mitigar essa ameaça é realizar o experimento em um horário em que não há aulas próximas a sala em que o experimento será realizado. Outra ameaça pode ter relação com os formulários usados para coletadas de dados. Se os pesquisadores produzirem seus próprios formulários, é possível que os instrumentos não consigam satisfazer o objetivo da avaliação. Uma forma de mitigar essa ameaça é utilizar instrumentos padronizados e consolidados na literatura.

VII. CONCLUSÕES

Neste artigo foi apresentado um estudo que mapeou a comunidade brasileira, reportando como os pesquisadores brasileiros estão avaliando sistematicamente os agentes conversacionais pedagógicos. Foram utilizados procedimentos sistemáticos definidos por Petersen et al. [24] e Kitchenham e Charters [25] para identificar e selecionar estudos relevantes. Com base nisso, constatou-se que faltam procedimentos sistemáticos nos estudos. A partir dos resultados, percebe-se que o entendimento por parte da comunidade sobre o processo experimental é deficiente. Acredita-se que os pesquisadores não conhecem procedimentos e taxonomias existentes (*e.g.*, [19], [52]) e o vocabulário usado na experimentação em outras áreas não é conhecido pelos membros dessa comunidade. As análises feitas evidenciam que intervenções precisam ser realizadas Neste sentido, foram propostos diretrizes para os projetos de avaliações experimentais. Como trabalhos futuros, pretende-se desenvolver um template de apoio para guiar o planejamento de estudos experimentais na área, incluindo a definição de um vocabulário para a área e recomendação de variáveis e medidas. Ainda, como pesquisa futura, pretende-se replicar esse estudo, considerando o cenário internacional.

AGRADECIMENTOS

Os autores gostariam de agradecer a CAPES - Código de Financiamento 001, a FAPESP (Processo 2018/26636-2) e ao CNPq (Processo 311494/2017-0 e Processo 312922/2018-3) pelo apoio financeiro.

REFERÊNCIAS

[1] A. Kerry, R. Ellis, and S. Bull, "Conversational agents in e-learning," in *Applications and Innovations in Intelligent Systems XVI*, T. "Allen, R. Ellis, and M. Petridis, Eds., Springer. London: Springer London, 2009, pp. 169–182.

[2] S. Tamayo-Moreno and D. Pérez-Marín, "Adapting the design and the use methodology of a pedagogical conversational agent of secondary education to childhood education," in *International Symposium on Computers in Education*, 2016, pp. 1–6.

[3] A. C. Graesser, Z. Cai, B. Morgan, and L. Wang, "Assessment with computer agents that engage in conversational dialogues and trialogues with learners," *Computers in Human Behavior*, vol. 76, pp. 607–616, 2017.

[4] E. Ayedoun, Y. Hayashi, and K. Seta, "Adding communicative and affective strategies to an embodied conversational agent to enhance second language learners' willingness to communicate," *International Journal of Artificial Intelligence in Education*, vol. 29, pp. 29–57, 2019.

[5] L. N. Paschoal, M. M. Oliveira, and P. M. M. Chicon, "A chatterbot sensitive to student's context to help on software engineering education," in *Latin American Computer Conference*, 2018, pp. 839–848.

[6] L. Lin, R. K. Atkinson, R. M. Christopherson, S. S. Joseph, and C. J. Harrison, "Animated agents and learning: Does the type of verbal feedback they provide matter?" *Computers & Education*, vol. 67, pp. 239–249, 2013.

[7] M. Coronado, C. A. Iglesias, Á. Carrera, and A. Mardomingo, "A cognitive assistant for learning java featuring social dialogue," *International Journal of Human-Computer Studies*, vol. 117, pp. 55–67, 2018.

[8] C. Troussas, A. Krouska, and M. Virvou, "Integrating an adjusted conversational agent into a mobile-assisted language learning application," in *International Conference on Tools with Artificial Intelligence*, 2017, pp. 1153–1157.

[9] J. L. Z. Montenegro, C. A. Costa, and R. R. Righi, "Survey of conversational agents in health," *Expert Systems with Applications*, vol. 129, pp. 56–67, 2019.

[10] L. N. Paschoal, A. L. Krassmann, F. B. Nunes, M. M. de Oliveira, M. Bercht, E. F. Barbosa, and S. d. R. S. de Souza, "A systematic identification of pedagogical conversational agents," in *Annual Frontiers in Education*, 2020, pp. 1–9.

[11] L. N. Paschoal, "A framework to support the experimental evaluation process of the pedagogical conversational systems," in *Ibero-American Conference on Software Engineering*, 2020, pp. 1–8.

[12] B. AbuShawar and E. Atwell, "Usefulness, localizability, humanness, and language-benefit: additional evaluation criteria for natural language dialogue systems," *International Journal of Speech Technology*, vol. 19, no. 2, pp. 373–383, 2016.

[13] N. M. Radziwill and M. C. Benton, "Evaluating quality of chatbots and intelligent conversational agents," *Software Quality Professional*, vol. 19, no. 3, pp. 25–36, 2017.

[14] E. Ruane, T. Faure, R. Smith, D. Bean, J. Carson-Berndsen, and A. Ventresque, "Botest: A framework to test the quality of conversational agents using divergent input examples," in *International Conference on Intelligent User Interfaces Companion*, 2018, pp. 1–2.

[15] S. Hobert, "How are you, chatbot? evaluating chatbots in educational settings," *Lecture Notes in Informatics*, vol. 17, pp. 259–270, 2019.

[16] R. Winkler and M. Söllner, "Unleashing the potential of chatbots in education: A state-of-the-art analysis," in *Academy of Management Annual Meeting*, 2018, pp. 1–40.

[17] J. Wilson and D. Rosenberg, "Rapid prototyping for user interface design," in *Handbook of Human-Computer Interaction*, M. HELANDER, Ed. Amsterdam: North-Holland, 1988, pp. 859 – 875.

[18] J. Ahn, P. Watson, M. Chang, S. Sundararajan, T. Ma, N. Mukhi, and S. Prabhu, "Wizard's apprentice: Cognitive suggestion support for wizard-of-oz question answering," in *International Conference on Artificial Intelligence in Education*, 2017, pp. 630–635.

[19] C. Wohlin, P. Runeson, M. Hst, M. C. Ohlsson, B. Regnell, and A. Wessln, *Experimentation in Software Engineering*, 1st ed. Heidelberg, Germany: Springer-Verlag Berlin Heidelberg, 2012.

[20] R. Conradi, V. R. Basili, J. Carver, F. Shull, and G. H. Travassos, "A pragmatic documents standard for an experience library: Roles, documen, contents and structure," University of Maryland, College Park, Maryland, Tech. Rep. 1, 2001.

[21] N. L. Kuyven, C. A. Antunes, V. J. B. Vanzin, J. L. T. Silva, A. L. Krassmann, and L. M. R. Tarouco, "Chatbots na educação: uma revisão sistemática da literatura," *RENOTE - Revista Novas Tecnologias na Educação*, vol. 16, no. 1, pp. 123–132, 2018.

[22] R. Kuri, "Use of doctoral thesis as a source of information: A study of researchers of karnataka university dharwad," *International Journal of Information Dissemination and Technology*, vol. 7, no. 1, pp. 14–18, 2017.

[23] A. Krassmann, A. Falcade, R. Jardim, R. Medina, and M. Bercht, "Um panorama de teses e dissertações brasileiras sobre mundos virtuais 3d na educação," in *Brazilian Symposium on Computers in Education*, 2017, pp. 71–81.

[24] K. Petersen, R. Feldt, S. Mujtaba, and M. Mattsson, "Systematic mapping studies in software engineering," in *International Conference on Evaluation and Assessment in Software Engineering*, 2008, pp. 68–77.

[25] B. Kitchenham and S. Charters, "Guidelines for performing systematic literature reviews in software engineering," Keele University and University of Durham, Keele, UK, Tech. Rep. 2.3, 2007.

[26] V. R. Basili, G. Caldiera, and H. D. Rombach, "The goal question metric approach," in *Encyclopedia of Software Engineering*, 1st ed., J. J. Marciniak, Ed. Hoboken, New Jersey: John Wiley & Sons, 1994, vol. 2, pp. 528–232.

[27] T. Dybå and T. Dingsøyr, "Empirical studies of agile software development: A systematic review," *Information and Software Technology*, vol. 50, no. 9, pp. 833 – 859, 2008.

[28] S. M. Melo, J. C. Carver, P. S. L. Souza, and S. R. Souza, "Empirical research on concurrent software testing: A systematic mapping study," *Information and Software Technology*, vol. 105, pp. 226 – 251, 2019.

[29] C. L. O. Castanho, "A avaliação do uso de chatterbots no ensino através de uma ferramenta de autoria," Master's thesis, Federal University of Santa Catarina, Florianópolis, State of Santa Catarina, 2002.

[30] C. C. M. P. Campos, "AvaliaÇÃo de faqbots atravÉs da ferramenta autochatter," Master's thesis, Federal University of Santa Catarina, Florianópolis, State of Santa Catarina, 2002.

[31] M. J. C. S. Domingues, "Mídia e aprendizagem: um estudo comparativo entre hipertexto e chatterbot," Ph.D. dissertation, Federal University of Santa Catarina, Florianópolis, State of Santa Catarina, 2003.

[32] M. D. Leonhardt, "Doroty," Master's thesis, Federal University of Rio Grande do Sul, Porto Alegre, State of Rio Grande do Sul, 2005.

[33] E. A. Oliveira, "i-collaboration (in portuguese)," Master's thesis, Federal University of Pernambuco, Recife, State of Pernambuco, 2008.

[34] L. M. A. Santos, "A inserção de um agente conversacional animado em um ambiente virtual de aprendizagem a partir da teoria da carga cognitiva," Ph.D. dissertation, Federal University of Rio Grande do Sul, Porto Alegre, State of Rio Grande do Sul, 2009.

[35] I. D. C. Leite, "A produção de sentidos na conversação com chatterbots," Ph.D. dissertation, Federal University of Pernambuco, Recife, State of Pernambuco, 2010.

[36] E. V. B. Aguiar, "Aprimoramento das habilidades cognitivas de resolução de problemas com o apoio de um agente conversacional," Ph.D. dissertation, Federal University of Rio Grande do Sul, Porto Alegre, State of Rio Grande do Sul, 2011.

[37] E. C. Lemos, "Desenvolvimento de chatterbots educacionais: um estudo de caso voltado ao ensino de algoritmos," Master's thesis, Federal University of Rio Grande do Norte, Natal, State of Rio Grande do Norte, 2011.

[38] L. A. Lima, "Estudo de implementação de um robô de conversação em curso de língua estrangeira em ambiente virtual: um caso de estabilização do sistema adaptativo complexo," Ph.D. dissertation, Federal University of Minas Gerais, Belo Horizonte, State of Minas Gerais, 2014.

[39] I. C. Pinho, "A mediação de um agente pedagógico na aprendizagem colaborativa de inglês como língua estrangeira," Ph.D. dissertation, Federal University of Rio Grande do Sul, Porto Alegre, State of Rio Grande do Sul, 2015.

[40] F. S. Sgobbi, "Explorando autodeterminação, utilizando novas tecnologias para ensejar autocuidado em obesos," Ph.D. dissertation, Federal University of Rio Grande do Sul, Porto Alegre, State of Rio Grande do Sul, 2017.

[41] L. N. Paschoal, "Contribuições ao ensino de teste de software com o modelo flipped classroom e um agente conversacional," Master's thesis, University of São Paulo, São Carlos, State of São Paulo, 2019.

[42] A. A. Bernardini, A. A. Sônego, and E. Pozzebon, "Chatbots: An analysis of the state of art of literature," in *Workshop on Advanced Virtual Environments and Education*, 2018, pp. 1–6.

[43] V. R. Basili, "The experimental paradigm in software engineering," in *International Workshop on Experimental Software Engineering Issues: Critical Assessment and Future Directions*, 1992, pp. 3—12.

[44] S. Roos, "Chatbots in education," Master's thesis, Uppsala University, Disciplinary Domain of Humanities and Social Sciences, Faculty of Social Sciences, Department of Informatics and Media., Uppsala, Sweden, 2018.

[45] B. B. N. França and G. H. Travassos, "Experimentation with dynamic simulation models in software engineering: planning and reporting guidelines," *Empirical Software Engineering*, vol. 21, no. 3, pp. 1302–1345, 2016.

[46] M. Shepperd, N. Ajienka, and S. Counsell, "The role and value of replication in empirical software engineering results," *Information and Software Technology*, vol. 99, pp. 120 – 132, 2018.

[47] V. R. Basili and H. D. Rombach, "The tame project: towards improvement-oriented software environments," *IEEE Transactions on Software Engineering*, vol. 14, no. 6, pp. 758–773, 1988.
[48] L. N. Paschoal, L. F. Turci, T. U. Conte, and S. R. S. Souza, "Towards a conversational agent to support the software testing education," in *Brazilian Symposium on Software Engineering*, 2019, p. 57–66.
[49] B. R. Ranoliya, N. Raghuwanshi, and S. Singh, "Chatbot for university related faqs," in *International Conference on Advances in Computing, Communications and Informatics*, 2017, pp. 1525–1530.
[50] P. A. MORETTIN and W. O. Bussab, *Estatística básica*, 9th ed. São Paulo: Editora Saraiva, 2017.
[51] S. Allua and C. B. Thompson, "Inferential statistics," *Air Medical Journal*, vol. 28, no. 4, pp. 168 – 171, 2009.
[52] F. Shull, J. Singer, and D. I. K. Sjøberg, *Guide to Advanced Empirical Software Engineering*, 1st ed. Berlin, Germany: Springer-Verlag London, 2008.

Leo Natan Paschoal Atualmente é aluno de doutorado em Ciências de Computação e Matemática Computacional pela Universidade de São Paulo. Mestre em Ciências Ciências de Computação e Matemática Computacional pela Universidade de São Paulo. Dentre os seus interesses de pesquisa, tem investido na temática de agentes conversacionais como mecanismos de apoio ao ensino. E-mail: paschoalln@usp.br.

Tayana Uchôa Conte Atualmente é professora da Universidade Federal do Amazonas. Doutora em Engenharia de Sistemas e Computação pela Universidade Federal do Rio de Janeiro. Tem experiência na área de Ciência da Computação, com ênfase em Engenharia de Software e IHC, atuando principalmente nos seguintes temas: Engenharia de Aplicações Web, Avaliação de Usabilidade, Engenharia de Software Experimental, Usabilidade de Aplicações Web e Qualidade de Software. E-mail: tayana@icomp.ufam.edu.br

Simone do Rocio Senger de Souza Atualmente é professora da Universidade de São Paulo. Doutora em Física Computacional pela Universidade de São Paulo. Suas áreas de interesse em pesquisa são teste de software, ensino de teste de software, teste de programas concorrentes, experimentação em engenharia de software e qualidade de software. E-mail: srocio@icmc.usp.br.