CRS-Que: A User-centric Evaluation Framework for Conversational Recommender Systems

YUCHENG JIN, LI CHEN, WANLING CAI, and XIANGLIN ZHAO, Hong Kong Baptist University, China

An increasing number of recommendation systems try to enhance the overall user experience by incorporating conversational interaction. However, evaluating conversational recommender systems (CRSs) from the user's perspective remains elusive. The GUI-based system evaluation criteria may be inadequate for their conversational counterparts. This article presents our proposed unifying framework, ***CRS-Que***, to evaluate the user experience of CRSs. This new evaluation framework is developed based on *ResQue*, a popular user-centric evaluation framework for recommender systems. Additionally, it includes user experience metrics of conversation (e.g., understanding, response quality, humanness) under two dimensions of *ResQue* (i.e., Perceived Qualities and User Beliefs). Following the psychometric modeling method, we validate our framework by evaluating two conversational recommender systems in different scenarios: *music exploration* and *mobile phone purchase*. The results of the two studies support the validity and reliability of the constructs in our framework and reveal how conversation constructs and recommendation constructs interact and influence the overall user experience of the CRS. We believe this framework could help researchers conduct standardized user-centric research for conversational recommender systems and provide practitioners with insights into designing and evaluating a CRS from users' perspectives.

CCS Concepts: • **Information systems** → **Recommender systems;** • **Computing methodologies** → *Discourse, dialogue and pragmatics;* • **Human-centered computing** → **User studies**; **Heuristic evaluations**;

Additional Key Words and Phrases: Recommender systems, conversational systems, user experience, music recommenders, questionnaire, user-centric evaluation

ACM Reference format:
Yucheng Jin, Li Chen, Wanling Cai, and Xianglin Zhao. 2024. *CRS-Que*: A User-centric Evaluation Framework for Conversational Recommender Systems. *ACM Trans. Recomm. Syst.* 2, 1, Article 2 (March 2024), 34 pages.
https://doi.org/10.1145/3631534

This work was supported by the Hong Kong RGC GRF project (RGC/HKBU12201620), the Hong Kong Baptist University IG-FNRA project (RC-FNRA-IG/21-22/SCI/01), and the Hong Kong Baptist University Start-up Grant (RC-STARTUP/21-22/23).

Authors' address: Y. Jin, L. Chen, W. Cai, and X. Zhao, Hong Kong Baptist University, Hong Kong, China; e-mails: yuchengjin@hkbu.edu.hk, {lichen, cswlcai, csxlzhao}@comp.hkbu.edu.hk.

2770-6699/2024/03-ART2 $15.00
https://doi.org/10.1145/3631534

1 INTRODUCTION

Conversational recommender systems (CRSs) enable users to interact with recommendations using natural human language. Traditionally, users click a rating button to tell the system if they like/dislike the recommended item. However, while using a CRS, users may say, for example, *"I like the melody of this song,"* to express their preferences in more detail. In essence, the CRS is a task-oriented chatbot that aims to generate quality recommendations by offering a more natural and interactive way to elicit user preferences, provide an explanation for the recommendation, and/or ask for user feedback on the recommendation [48]. The CRS's prominent feature is the natural interaction, which could facilitate user critiquing [51, 80], user exploration [9, 80], and explanations for recommendations [42, 81, 90] in recommender systems.

The implementation of CRS involves multiple technical modules, such as the recommendation module, the natural language processing module, and the dialogue management module [30], which poses a new challenge in holistically evaluating such a complex system. For example, the success of CRS may depend on whether the system correctly understands user expressions, recognizes user intents, responds to user intents, and makes proper recommendations. Previous studies on traditional recommender systems have shown the limitations of only considering objective metrics in evaluating the system and have proposed several user-centric evaluation frameworks to measure user perceptions of recommendations such as user satisfaction, user trust, and intention to use [31, 69, 78]. Likewise, the user-centric evaluation of a CRS should be equally or even more important, since a CRS intends to improve the overall **user experience (UX)** of recommendations through more natural human-computer interaction.

The current user-centric evaluation frameworks for recommender systems primarily focus on recommendations but overlook conversations, which is insufficient to assess the actual quality of a CRS from users' perspectives. However, recent user evaluations on CRSs [9, 51, 90] have usually adopted some questions from the general user-centric evaluation frameworks for recommender systems [68, 95] and some questions from the questionnaires for evaluating conversational agents [35, 57, 130]. This kind of customization and combination of existing evaluation questions is based primarily on the specific needs of researchers, which, however, might need more standardization to compare different studies for evaluating the CRS [47]. Therefore, we intend to develop a **consolidated and unifying evaluation framework** based on *ResQue* [95]. Specifically, we extend *ResQue* by incorporating six critical conversation constructs (i.e., CUI Understanding, CUI Adaptability, CUI Response quality, CUI Attentiveness, CUI Rapport, and CUI Humanness). The development of this framework follows psychometric research methodology [86]. We validate this new framework and the hypothesized paths by evaluating two conversational recommender systems in different scenarios: *music exploration* and *mobile phone purchase*. The results of our study show that this framework has good validity and reliability in assessing the UX of CRSs. Furthermore, this framework also reveals how conversation constructs are fitted to two dimensions, i.e., *Perceived Qualities* and *User Beliefs*, of *ResQue*, and how they interact with recommendation constructs and influence the constructs of two other dimensions, *User Attitudes* and *Behavioral Intentions*. Therefore, we believe that *CRS-Que* enables practitioners to comprehensively evaluate a CRS by considering user perceptions of both recommendation and conversation.

The contributions of our work are four-fold:

(1) We developed a consolidated and unifying user-centric evaluation framework for CRSs, allowing practitioners to evaluate conversational recommender systems from users' perspectives. This new framework reveals how the conversation constructs correlate with the recommendation constructs and how they influence the overall UX of CRSs.

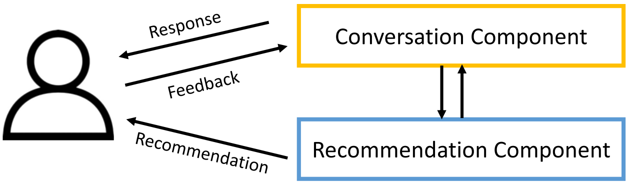


Fig. 1. Conversational Recommender Systems.

(2) We followed psychometric research methodology to validate the framework by conducting two user studies with different experimental conditions (e.g., scenarios, manipulated system design factors, platforms).
(3) The evaluation results re-validated one of the most influential user-centric evaluation frameworks, *ResQue*, when the recommendations are delivered via conversation. We also revealed how this new interaction method has changed the original *ResQue* framework.
(4) Our framework provides a standardized user-centric research and evaluation approach for conversational recommender systems.

We structure the article as follows: We first review the related work. After that, we explain the development process of the new evaluation model ***CRS-Que***, including the constructs and the hypothesized relations. We then present two user studies to validate our framework, including each study's experimental setup, results, and discussion. Finally, we discuss the validation and use of the framework.

2 RELATED WORK

In this section, we first review conversational recommender systems that support natural language interaction, followed by the existing user-centric evaluation frameworks for recommender systems and the UX metrics of conversational agents.

2.1 Conversational Recommender Systems

Traditional recommender systems only support a one-shot interaction, i.e., presenting one or a list of recommended items based on users' past behavior [99]. In contrast, CRSs allow users to find recommendations of their interests with multi-turn interactions [48]. Furthermore, the CRS can interactively elicit users' current preferences from their feedback and build a more comprehensive user model to make better recommendations [21]. According to a recent survey on CRSs [48], some earlier CRSs work with **graphical user interface (GUI)** widgets rather than dialogue-based interaction, such as critiquing-based systems [17, 77] where users can give feedback on recommendations by picking some pre-defined or auto-generated critiques. However, the recent advance of natural language technology has led to an increased interest in building a CRS based on natural language (called dialogue-based CRSs; see Figure 1) that enables natural interaction between users and the system [8, 60]. This work mainly focuses on the latter type of CRSs that support natural language interaction.

However, most existing studies evaluated their proposed methods using offline experiments, which usually simulated user behavior, for example, answering preference-related questions or giving feedback on recommendations, based on their historical data [111]. With simulated data, they separately measured the recommendation performance by adopting accuracy metrics (e.g., Average Precision, RMSE, and Recall) [21] and/or assessed the system's responses using linguistic metrics like BLEU score [88]. However, such a simulated evaluation ignores that users may develop new preferences after exploring recommendations during the conversation. Thus,

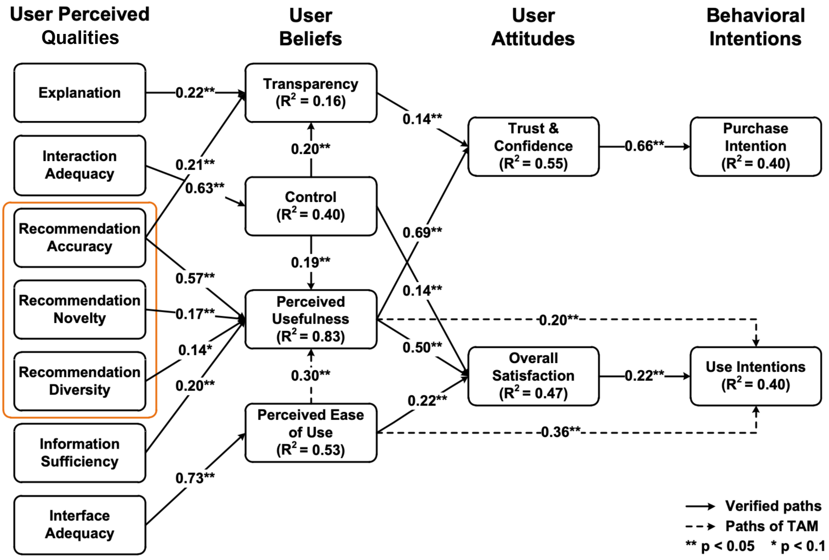


Fig. 2. A structural equation model of *ResQue* [95].

it may not inform us about the evaluation results in real-world situations. In contrast, empirical studies carried out with user-centric evaluation can better measure the system's effectiveness in realistic scenarios. It typically requires participants to use the system to complete a specific task (e.g., finding music for a party) and then assesses their perceived quality of the system [9, 51]. A recent paper has discussed the importance of user-centric evaluation for dialogue-based CRS [47]; however, to the best of our knowledge, there is no user-centric evaluation framework that is specific to dialogue-based CRS. We aim to develop a consolidated and unifying evaluation framework to fill this gap. As text-based interaction is the major modality of existing CRSs, we have focused on validating the evaluation framework with the text-based CRSs in this work.

2.2 User-centric Evaluation of Recommender Systems

Given the limitations of evaluation methods based on objective metrics, as mentioned before, several studies proposed different user-centric evaluation frameworks for recommender systems.

2.2.1 ResQue. ResQue is a unifying evaluation framework [95] that was developed based on two well-known usability evaluation models, i.e., **Technology Acceptance Model (TAM)** [25] and **Software Usability Measurement Inventory (SUMI)** [62]. It measures the user experience of recommender systems from four dimensions: Perceived Qualities, User Beliefs, User Attitudes, and Behavioral Intentions [95]. Figure 2 shows a structural equation model of *ResQue*. Each dimension contains multiple constructs that measure different aspects of the dimension. For example, the dimension of Perceived Qualities contains Explanation, Interaction Adequacy, Recommendation Accuracy, and so on. The arrows between the constructs indicate casual relationships. For example, the construct of explanation can positively influence transparency, in turn leading to higher trust and intention to purchase. The full model contains 15 constructs and 32 questions, allowing practitioners and researchers to evaluate the success of their own developed recommender systems from users' perspectives.

2.2.2 Explaining User Experience of Recommender Systems. Knijnenburg et al. [68] proposed a framework to explain users' behavior through a set of constructs organized in a structure relating the objective system aspects, subjective system aspects (i.e., the perceived qualities of the system), experience constructs (i.e., how users experience the system), personal characteristics (e.g., demographics, domain knowledge, initial trust), and situational characteristics (e.g., privacy concern, familiarity, choice goal). The authors validated the framework with multiple field trials and experimental settings by manipulating various objective system aspects (e.g., recommendation algorithm, interaction, presentation). This framework provides a more holistic view of explaining the user experience of recommender systems by incorporating personal and situational characteristics into the model.

These two most influential user-centric evaluation frameworks for recommender systems have been extensively adopted to evaluate various types of recommender systems, including social recommendations [28, 65, 121], media recommendations [56, 71, 102], and product recommendations [18, 54]. In addition, some metrics were proposed to measure specific aspects of recommender systems, such as explanation [119], trust [94], inspectability [65], and user control [65]. In our work, we have developed ***CRS-Que*** mainly based on the four dimensions of *ResQue* [95] and adopted some questions of the framework of Knijnenburg et al. [68].

2.3 UX Metrics of Conversational Agents

Similar to the evaluation of recommendations, objective measures are insufficient to gauge the effectiveness and user experience of a CRS. Since a CRS is a task-oriented conversational agent, the evaluation should consider the success of dialogue: whether the agent helps users find the recommended items of their interests.

From a technical point of view, the evaluation metrics of **conversational agents (CA)** have identified several key components, such as the performance of **natural language understanding (NLU)** component [6] and **natural language generation (NLG)** component [24, 33]. Given that CA usability can significantly influence the demonstration and user perception of CA functionality [123], we especially review the metrics measuring the conversational experience based on both objective and subjective measures.

PARADISE is a popular evaluation framework for CA [125], a general performance model of system usability for spoken dialogue agents, including a subjective user satisfaction metric and three objective metrics about dialogue efficiency, dialogue quality, and task success. In another work, according to the quality attributes of chatbot development and implementation, Radziwill and Benton [97] suggested quantifying the quality of CA from four aspects: performance, humanity, affect, and accessibility and proposed an **analytic hierarchy process (AHP)** for quality metrics selection. Furthermore, evaluating embodied conversational agents introduces additional metrics that reflect the quality of communication, such as likeability, entertainment, engagement, helpfulness, and naturalness [101]. Toward commercial conversational agents, Kuligowska [70] proposed some sophisticated metrics, such as the visual look of the chatbot, conversational abilities, language skills, and context sensitiveness. More specifically, the measure of a CA's response quality could be assessed by content informativeness and interaction fluency [50].

In recent years, the rise of task-oriented chatbots has called for a new methodology to assess the impact of the agent's interaction strategies on the quality of experience, mainly considering **Quality of Service (QoS)** and **Quality of Experience (QoE)** [35]. In addition, some metrics for assessing communication and social skills have been proposed to evaluate social CA. For example, an evaluation model characterizes the interaction into dimensions of rapport [116], such as positivity, attentiveness, and coordination, based on the theories of negotiation and communication [138]. PEACE model [113] identifies four essential qualities of a chatbot, including politeness,

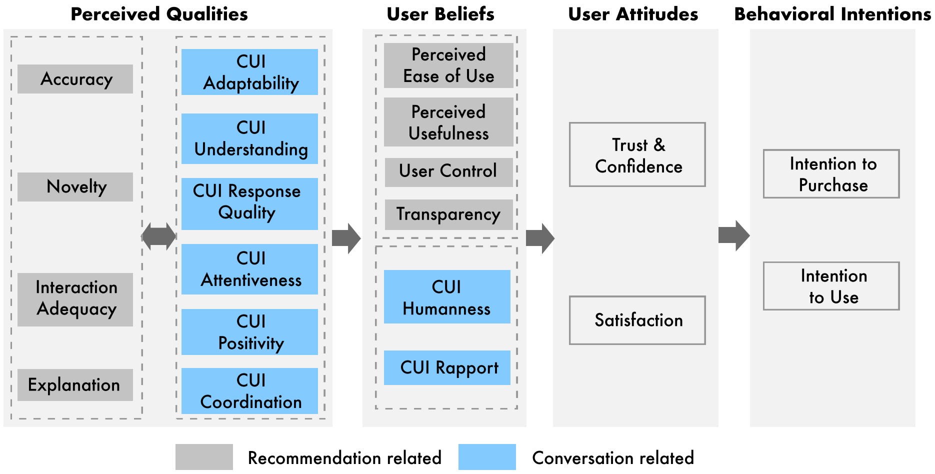


Fig. 3. General evaluation framework with hypothesized relationships (CRS-Que).

entertainment, attentive curiosity, and empathy, that can influence users' intention to use open-domain chatbots.

It can be seen from the above-described evaluation frameworks that some metrics are common (e.g., task ease, performance, and satisfaction), while some focus more on communications, such as language skill, coordination, and rapport [70, 138], and some include constructs about user beliefs and behavioral intentions, such as future use [124], affect [97], trust [101], and ease of use [35]. Besides, some metrics specialize in particular types of CA, for example, evaluating visual look and knowledge base for commercial CAs [70], and assessing feelings about negotiation and information disclosure for social negotiation CAs [138].

3 FRAMEWORK DEVELOPMENT

Figure 3 shows our general evaluation framework for conversational recommender systems with hypothesized relationships. We develop this evaluation framework, called ***CRS-Que***, based on the widely used user-centric evaluation framework *ResQue* consisting of four dimensions [95]: Perceived System Qualities, User Beliefs, User Attitudes, and Behavioral Intentions (see Figure 2). Each dimension contains several constructs carefully derived from prior work related to the user experience of recommender systems. Instead of generating a linear model, *ResQue* organizes the question items into four dimensions to clearly describe how Users' Perceived Qualities of the recommender system influence their Beliefs, Attitudes, and Behavioral Intentions. In our framework, *CRS-Que*, all constructs, except conversation-related constructs (blue boxes), are from *ResQue* (see Figure 2).

We collected the metrics for conversation-related constructs mainly used for assessing conversation quality. We distinguish conversation metrics from recommendation metrics by adding a prefix of **CUI (Conversational User Interface)**. Specifically, based on the rapport theory in conversation [116], we adopted CUI Adaptability, CUI Attentiveness, and CUI Rapport to measure user perception of conversations. Moreover, CUI Understanding and CUI Response Quality are the metrics for gauging user-perceived quality of NLU and NLG [125], which are two prominent functional modules in a CA. Besides, CUI Humanness, a kind of user belief, measures the extent to which a CA behaves like a human, which has been studied in the conversational recommender system for electronic products [34].

We conducted a study (i.e., Study 1) to identify several key qualities of conversational recommender systems by investigating which constructs of Perceived Qualities and User Beliefs could influence the user's intention to use the CRS [52]. This study helps us refine the constructs of CRS-Que after merging closely correlated constructs and modifying some redundant and confusing questions. The following sections explain what each construct is supposed to measure and review the relevant studies that have inspired our model development. The validated questions for measuring each construct are shown in Tables 2 and 4.

3.1 Perceived Qualities

As the first dimension of ***CRS-Que***, Perceived Qualities mainly measure how users perceive the significant characteristics of the system, including qualities of recommendations (e.g., accuracy, novelty, and interaction adequacy) and qualities of conversation (e.g., attentiveness, understanding, and response quality). Note that for this dimension, we omit three constructs (i.e., diversity, interface adequacy, information sufficiency) of the original *ResQue* model due to the unique characteristics of CRS. Specifically, we omit *diversity,* because it measures a set of recommended items rather than a single item usually delivered by a CRS during each recommendation round [48]. Although some CRSs may use special UI widgets (e.g., list view, carousel) to display multiple items simultaneously, most CRSs often show a single item in conversation due to the limited display size on mobile devices. In addition, we exclude the construct of *interface adequacy,* as it mainly focuses on the design elements of graphical user interfaces, such as control buttons and layout. However, a CRS usually has a standard, natural-language-based user interface. For *information sufficiency*, it can be an important indicator of system informativeness for decision-making. But, since questions of CUI Response Quality have assessed it, we also exclude it.

3.1.1 Accuracy. Perceived accuracy measures how users feel the recommendation matches their interests and preferences. It can compensate for the limitation of objective accuracy [23] to indicate how good the recommendation could be from the user's perspective.

3.1.2 Novelty. Novelty is one of the most discussed beyond-accuracy metrics for recommender systems, which gauges the extent to which the recommendation is new or unknown to users. Novelty is particularly important to a recommender system that aims to support user exploration and discovery of new items. Novelty is sometimes discussed with "serendipity"; however, Herlocker [41] argued that the recommendation of high serendipity should be new and surprising. Despite the nuances of the two words, we do not distinguish them in our user study to avoid user confusion. Novelty is usually positively correlated with some constructs such as diversity and coverage [58].

3.1.3 Interaction Adequacy. This construct mainly measures the system's ability to elicit and refine user preferences through user interaction; some recommender systems may implicitly adapt to user preferences based on their interaction behaviors. Because a CRS tends to improve user experience through dialogue-based conversation [81], preference elicitation is its integral process [92]. Similar to the typical interaction strategies of critiquing-based recommender systems [16], a CRS allows users to give feedback by rating items or specifying the attributes of their preferred items.

3.1.4 Explanation. This construct measures the system's ability to explain its recommendations. Explainable recommender systems tend to improve the trustworthiness and transparency of the system [117]. Several works [32, 118] have proposed different approaches to designing and evaluating explanations of recommender systems. It was shown that explanations could positively influence various aspects of recommender systems [22, 118], such as user acceptance and trust.

3.1.5 CUI Positivity. Positivity is the first component of rapport theory [116], corresponding to the perceived mutual friendliness and caring in communication. For example, positivity may determine the tone and vocabulary of a conversation [116].

3.1.6 CUI Attentiveness. Attentiveness is the second component of "rapport." It measures if a system establishes a focused and cohesive interaction by expressing mutual attention and involving each other. CUI attentiveness closely relates to the other two components, Positivity and Coordination [116].

3.1.7 CUI Coordination. Coordination is the third component that examines whether communication is synchronous and harmonious [116]. Coordination is more critical to rapport than the other two components in the late communication phase, which could arouse communicators' empathy in conversation.

3.1.8 CUI Adaptability. Adaptability measures a system's ability to adapt to users' behavior and preferences during the conversation. Adaptability is usually associated with personalization, e.g., whether a system can personalize its replies by adapting to the user's emotions or historical behavior [61]. Other adaptive agents can learn users' vocabularies to engage with community members [105] and adjust the conversation length according to the context [126]. In our framework, this construct particularly assesses if the system adapts to the user's preferences for items.

3.1.9 CUI Understanding. Understanding is the key performance indicator of conversational agents, which measures an agent's ability to understand users' intents. The performance of the NL) module of a CA is usually measured by the correct rate and confidence scores of intent classification and entities extraction [1]. In our work, we aim to measure user-perceived understanding of a CRS.

3.1.10 CUI Response Quality. Response Quality refers to content quality (informativeness) and the pace of interaction (fluency), which have been frequently adopted to assess response quality in chatbots [50, 59, 74, 137]. It has been found that informativeness and fluency often positively influence users' perceived humanness of the conversation agent [104]. However, from the referred literature, we cannot find concrete questions for measuring CUI Response Quality, so in our framework, the corresponding questions were composed according to the definitions of informativeness and fluency.

3.2 User Beliefs

The constructs of User Beliefs in our framework measure a higher level of user perception of a system, which the constructs of Perceived Qualities could influence. This dimension could reflect the effectiveness of a CRS in supporting users to perform specific tasks, such as decision-making and exploring diverse items to develop new interests. In addition to the constructs of *ResQue* (e.g., Perceived Ease of Use, Perceived Usefulness, User Control, Transparency) [95], this dimension also contains two constructs of conversation: CUI rapport and CUI Humanness. We determined these two conversation constructs, because rapport and humanness are two higher levels of user perception and are closely associated with users' perceived quality of the conversation [13, 103, 107].

3.2.1 Perceived Ease of Use. Perceived ease of use can be measured physically and mentally. The physical measures include the completion time of a task and the learning curve of using a new system. Regarding mental measures, many user studies of recommender systems employ the NASA-TLS evaluation framework [38] to assess users' cognitive load. Both physical efforts and cognitive load will influence the perceived ease of use of a CRS. Similar to *ResQue* [95], we use subjective questions to measure this construct.

3.2.2 *Perceived Usefulness.* Perceived usefulness measures the system's competence in supporting users in performing tasks [94]. Users may judge the usefulness of a recommender by the support they get from recommendations when performing a task. It was found that the perceived usefulness influences the users' willingness to share their data for improving recommendations [79]. In our model, perceived usefulness mainly measures the extent to which the system supports decision-making.

3.2.3 *User Control.* User control measures the controllability users perceive while interacting with the recommender. Previous studies show the positive effects of controllability on multiple user experience factors such as recommendation accuracy [53] and overall user satisfaction [43]. To address the challenges of designing personalized user control mechanisms for recommender systems [49], some researchers suggest different user control mechanisms that are tailored to some personal characteristics such as domain knowledge, trusting propensity, and persistence [55, 66].

3.2.4 *Transparency.* The transparency of a system enables users to understand the inner logic of the recommendation process. Moreover, transparency closely relates to user control and explanation, which tends to positively influence users' perceived accuracy [108], intention to buy [114], and overall satisfaction [54]. Although transparency is supposed to influence user trust positively, Kizilcec [63] argues that designers should find a proper degree of interface transparency for building trust, as too much transparency may impair user trust.

3.2.5 *CUI Rapport.* It is an overall measure of rapport that users perceive while communicating with the conversational agent. According to the rapport theory [116], it contains three components: Positivity, Attentiveness, and Coordination. The three components closely correlate, and each emphasizes the important trait at different stages of communication. For example, CUI Positivity is particularly important at the early stage of communication, while CUI Attentiveness is more critical when the conversation has started for a while. Several studies investigate approaches to help agents develop and maintain a communication rapport with users. For instance, Novick and Gris [84] suggest increasing the amplitude of nonverbal behaviors to establish a rapport, and Riek et al. [100] enable human-robot rapport via real-time head gesture mimicry.

3.2.6 *CUI Humanness.* Humanness is also an overall quality measure of conversation, as it gauges the extent to which an agent behaves like a human. Many studies show various design factors that may influence user perception of humanness, such as anthropomorphic visual cues [34], the presence of typos and capitalized words in the responses [132], typeface [12], and conversational skills [103]. However, a study suggests avoiding small talk and maintaining a formal tone to reduce humanness in a service-oriented context [112].

3.3 User Attitudes

User Attitudes assess users' overall feelings towards a conversational recommender system. Compared with the constructs of User Beliefs, the constructs of Attitudes are less likely to be influenced by the short-term experience of using the system. The typical constructs of Attitudes include user trust, confidence, and satisfaction.

3.3.1 *Trust & Confidence.* Trust significantly influences the overall success of a recommender system. The trust can be influenced by recommendations, conversations, or both for a CRS. Incorporating the concept of trust into a collaborative filtering framework tends to increase the predictive accuracy of recommendations [75, 87]. Kunkel et al. [71] suggest that recommenders should provide richer explanations to increase a system's trustworthiness. Pu and Chen [94] explore the potential of building users' trust with explanation interfaces for recommender systems. Besides, personal characteristics (e.g., personality [10, 71], cultural differences [4, 15]) and situational characteristics (e.g., system reputation [15], task type [127], and system familiarity [26]) may also influence user trust. Although Przegalinska et al. [93] propose a new methodology to measure chatbot performance based on user trust, the trust in conversations does not have a unified definition and measurement yet [27].

Confidence indicates whether the system can convince users of recommended items. In other words, it measures the user's confidence in accepting the recommendation. Hoxmeier et al. [46] investigate the effects of gender and technical experience on user confidence in electronic communication. For decision support systems, the level of presenting uncertainty information can influence user confidence in decision-making [2].

3.3.2 Satisfaction. This construct is an overall measure of users' attitudes and opinions toward a conversational recommender system in our framework. It allows subjects to provide general feedback to the whole system. Several studies show increased user satisfaction by integrating user personality traits [82] and domain knowledge [67] into the process of generating recommendations. Besides, a large-scale user study shows a positive effect of recommendation serendipity on user satisfaction [19].

3.4 Behavioral Intentions

Behavioral Intentions toward a system are related to user loyalty, which measures the likelihood that users will use the system in the future, accept/purchase resulting recommendations, and recommend the system to their acquaintances [95]. Therefore, we mainly consider Intention to Use and Intention to Purchase in this dimension. Users' behavioral intentions tend to be influenced by performance expectancy, effort expectancy, social influence, and trust [128]. By definition, performance expectancy is similar to Perceived Usefulness, and effort expectancy is similar to Perceived Ease of Use, both measured by our framework. We did not consider social influence, because our framework focuses on personal perceptions rather than the influence of others' opinions on using the recommender systems. Moreover, user trust and satisfaction positively influence users' intention to use [106].

4 EMPIRICAL VALIDATION

4.1 Validation Approach

Figure 3 shows a full version of the evaluation framework that contains all constructs related to the user experience of a CRS. This framework provides a holistic view for evaluating CRSs from users' perspectives. This framework is not restricted to a specific type of CRS or an application domain. To assess the validity and reliability of the evaluation framework, we conducted two user studies to evaluate two different CRSs by using this framework. According to the research questions of each study, we selected the most relevant constructs for each of the four dimensions in the evaluation model. We employed **Confirmatory factor analysis (CFA)**, a multivariate statistical technique, to verify the factor structure and framework [36]. Following the psychometric methods [86], we evaluated the framework from various aspects, such as internal reliability and convergent validity. In the following, we will introduce the details of our validation approach, including measurements, system manipulation, study design, and analysis method.

4.1.1 Measurements. All constructs of this user-centric evaluation framework ***CRS-Que*** are based on subjective measures. We employed questionnaires to measure constructs of Perceived

Qualities of recommendations and conversations (e.g., Accuracy, CUI Response Quality), User Beliefs (e.g., Perceived Ease of Use, CUI Rapport), User Attitudes (e.g., Trust), and Behavioral Intentions (e.g., Intention to Use). We developed all evaluation constructs based on existing UX metrics for recommenders and conversational systems. To ensure we can keep enough question items for each construct after dropping some items that do not contribute to the measurement of any construct, we kept at least three questions per construct to "*provide minimum coverage of the construct's theoretical domain*" [36]. Therefore, we composed some questions for some constructs with less than three question items. All the self-composed questions are validated following the requirements of CFA and are marked by the symbol "*" in Tables 2 and 4. All the question items are rated on a 7-point Likert scale, ranging from "Strongly Disagree" to "Strongly Agree." In addition, to filter out the poor responses, we put some attention-check questions in the questionnaire to identify the inattentive responses, for example, "*Please respond to this question with '2'.*"

4.1.2 System Manipulation. Typical evaluations of recommender systems (including user-centric evaluations) involve A/B comparisons between different versions of the same system. The manipulations allow us to investigate how the design factors could influence users' responses to the measurement of various constructs and provide insight into the ability of these constructs of our framework to gauge the user experience of a CRS. For example, would users of a human-like chatbot score higher on the CUI Humanness scale than users of a machine-like chatbot? Therefore, we manipulated several prominent design factors of CRSs in our studies, including critiquing initiative (user-initiated vs. system-suggested) [9], explanation display (true vs. false) [32], and humanization level (low vs. high) [34].

To fairly investigate the effect of manipulation, we keep everything constant between compared versions of the system except the manipulated design factors. In addition, we often choose a version as the baseline condition (e.g., without explanation), being compared with other conditions.

4.1.3 Study Design. All our studies aim to evaluate different versions of a system. As the evaluation of the system depends on user perception of recommended content (e.g., music, mobile phones), we decided to choose a between-subjects study design to avoid carryover effects and the high effort of answering a long questionnaire repeatedly in a within-subjects study.

We recruited subjects from Prolific,[1] a popularly used platform for academic surveys. To ensure the quality of the study, we pre-screened users in Prolific using the following criteria: (1) participants should be fluent in English; (2) the number of the participant's previous submissions should be more than 100; (3) approval rate should be greater than 95%. The **Research Ethics Committee (REC)** of our university approved this study.

The procedure of the study contains the following steps:

(1) Participants must sign a consent form to accept **General Data Protection Regulation (GDPR)** before signing into our system.
(2) Participants are asked to read a brief introduction about using the experimental CRS and fill out a pre-study questionnaire.
(3) Participants are asked to try the system.
(4) Participants are asked to perform a task using the system. For example, create a music playlist (Study 1) or add mobile phones to the shopping cart (Study 2).
(5) After finishing the task, we ask users to fill out a post-study questionnaire according to ***CRS-Que***.

[1] https://www.prolific.co/

4.1.4 Analysis Method. We first applied CFA to establish internal reliability, convergent validity, and discriminant validity. A CFA model consists of latent variables, which cannot be measured directly and should be measured by at least three indicators [7]. Convergent validity ensures that a group of questions (indicators) measure the same latent factor. In contrast, discriminant validity ensures that the two latent factors' indicators measure different things. To keep the discriminant validity of constructs, we can merge the two highly correlated constructs (greater than 0.85), or the **average variance extracted (AVE)** of the constructs should be higher than the correlation value of the construct with other constructs [20]. Furthermore, we employed **Structural Equation Modeling (SEM)** to investigate the relationship between constructs within one dimension (e.g., Explainability and CUI Attentiveness, under Perceived Qualities) or constructs belonging to different dimensions (e.g., CUI Adaptability and Perceived Usefulness, under Perceived Qualities and User Beliefs, respectively). We validate the general evaluation framework (Figure 3) in two studies, resulting in two models (Figures 5 and 7) that illustrate relationships among the four dimensions in different experimental settings. Compared with the traditional multivariate analysis methods, SEM has four significant advantages: (1) estimating variables that cannot be directly measured (latent variables) via observed variables, (2) taking measurement error into account in the model, (3) validating multiple hypotheses simultaneously as a whole, (4) testing a model regarding its fit of the data [131].

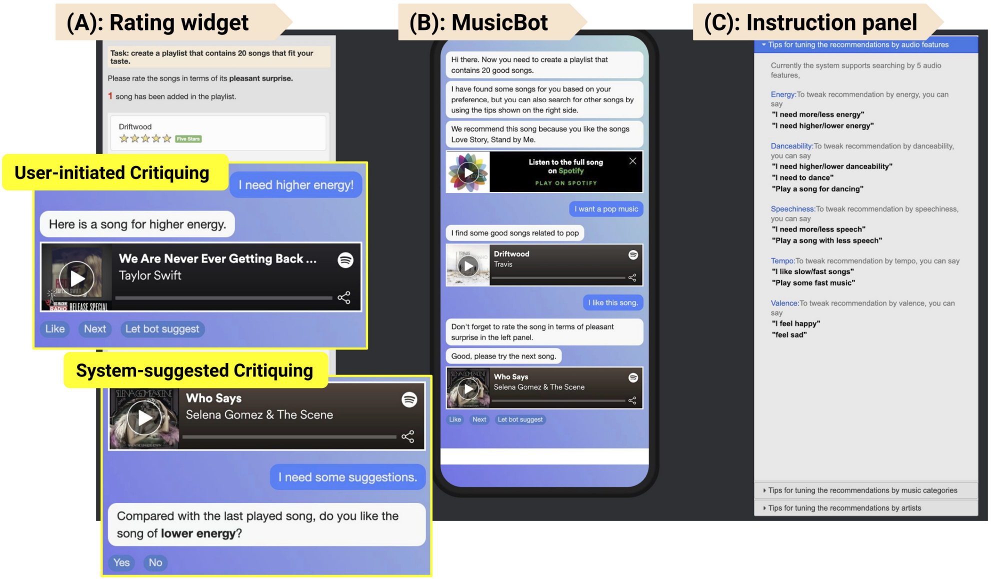
4.1.5 Visual Presentation of SEM. The visual presentation of SEM results consists of boxes and arrows referring to constructs and significant relationships, respectively. Each single-headed arrow is associated with two numbers. The first number is the parameter β representing the regression coefficient, and the second number in the parentheses represents the standard error of the regression coefficient. To be more specific, the regression coefficient β indicates the amount of changes in a dependent variable (y) attributed to a unit change in an independent variable (x). Since the β coefficient is not standardized, the large coefficient values (> 1) do not necessarily mean multicollinearity. The direction (the arrow), strength (β), and significance (p value) of the path between two constructs indicate how they are related. For example, for a path (Explanation → Transparency) in Figure 2, you may conclude that Explanation has a positive and significant effect on Transparency, suggesting that explaining recommendations could lead to a better understanding of recommendation logic. It is worth mentioning that interpreting an SEM is not solely based on individual paths but also involves understanding the overall theoretical model. In addition, the double-headed arrows represent the correlations between two variables, and the numbers on the edge represent the estimate and standard error of covariance.

We use different notations to denote the statistical significance level.[2] In addition, we use different colors to signify system design factors (orange color), the constructs of recommendations (gray color), the constructs of conversations (blue color), and the constructs of user attitudes and behavioral intentions (white color).

4.2 Study 1: *MusicBot* for Music Exploration

MusicBot is a critiquing-based recommender system for exploring music recommendations [9]. In the context of recommender systems, critiques refer to the users' feedback made on the recommendations [17]. For example, *"I want a cheaper computer"* can be a critique of a recommended computer. This study mainly compares two critiquing approaches (i.e., user-initiated and system-suggested). Two dialogue examples with yellow labels (see Figure 4) illustrate how the user can make **user-initiated critiquing (UC)** and **system-suggested critiquing (SC)**, respectively.

[2] The significance level: *** $p < .001$, ** $p < .01$, * $p < .05$.



Task: create a playlist that contains 20 songs that fit your taste.

Please rate the songs in terms of its **pleasant surprise.**

1 song has been added in the playlist.

Driftwood

I need higher energy!

Here is a song for higher energy.

We Are Never Ever Getting Back ...
Taylor Swift

Who Says
Selena Gomez & The Scene

I need some suggestions.

Compared with the last played song, do you like the song of **lower energy**?

Hi there. Now you need to create a playlist that contains 20 good songs.

I have found some songs for you based on your preference, but you can also search for other songs by using the tips shown on the right side.

We recommend this song because you like the songs Love Story, Stand by Me.

I want a pop music

I find some good songs related to pop

Driftwood
Travis

I like this song.

Don't forget to rate the song in terms of pleasant surprise in the left panel.

Good, please try the next song.

Who Says
Selena Gomez & The Scene

Tips for tuning the recommendations by audio features

Currently the system supports searching by 5 audio features,

Energy:To tweak recommendation by energy, you can say
"I need more/less energy"
"I need higher/lower energy"

Danceability:To tweak recommendation by danceability, you can say
"I need higher/lower danceability"
"I need to dance"
"Play a song for dancing"

Speechiness:To tweak recommendation by speechiness, you can say
"I need more/less speech"
"Play a song with less speech"

Tempo:To tweak recommendation by tempo, you can say
"I like slow/fast songs"
"Play some fast music"

Valence:To tweak recommendation by valence, you can say
"I feel happy"
"feel sad"

Tips for tuning the recommendations by music categories

Tips for tuning the recommendations by artists

Fig. 4. The user interface of *MusicBot*.

User-initiated critiquing allows users to critique the recommended song by themselves. For example, they can critique a recommended song using audio features, *"I need a song with higher energy."* In contrast, with *system-suggested critiquing*, users get the agent's suggestions for exploring music recommendations, for example, *"Compared to the last played song, do you like the song of lower tempo?"* Users can decide whether to accept the suggested critique or not.

Evaluating MusicBot with *CRS-Que* involves two research questions.

RQ1: How does the *critiquing initiative* (user-initiated vs. system-suggested) influence users' perceived qualities of recommendations and conversations?

RQ2: How do the changes in Perceived Qualities influence the constructs of other dimensions (i.e., User Beliefs, User Attitudes, and Behavioral Intentions)?

4.2.1 Setup. The music CRS is implemented as a desktop web application. The web application consists of three parts: a rating widget (Figure 4(A)), *MusicBot* (Figure 4(B)), and an instruction panel (Figure 4(C)). A dialogue window of *MusicBot* shows the dialogue between the user and the system, where the cards show the recommended songs and a row of buttons is for users to give feedback on the recommendation (i.e., "Like," "Next," "Let bot suggest"). The users can click the "Next" button to skip the currently recommended song or the "Like" button to add the song to the playlist. The "Let bot suggest" button can trigger a system-suggested critiquing on the recommended song (see details below). The system-suggested critiquing is implemented based on **Multi-Attribute Utility Theory (MAUT)** [136] and the diversity calculation based on Shannon's entropy [135]. The implementation details can be found in our prior work [9]. We use off-the-shelf technologies to implement the recommendation component and conversation component. The recommendation is powered by Spotify recommendation service,[3] and the natural language understanding is enabled by Dialogflow ES (standard) API.[4]

[3] https://developer.spotify.com/console/get-recommendations/
[4] https://cloud.google.com/dialogflow/es/docs

Table 1. Demographics of the Participants in Study 1

	Item	Frequency	Percentage (%)
Age	18–24	36	16.67%
	25–34	74	34.26%
	35–44	54	25.00%
	45–54	27	12.50%
	55–64	18	8.33%
	≥ 65	7	3.24%
Gender	Male	117	54.17%
	Female	96	44.44%
	Other	3	1.39%
Nationality	UK	135	62.50%
	Canada	15	6.94%
	USA	15	6.94%
	Germany	6	2.78%
	Netherlands	5	2.31%
	Others	40	18.52%

We conducted a between-subjects study to investigate the effects of the critiquing initiative on the user experience constructs of the CRS. The **task** of the user study is to use the *MusicBot* to discover new and diverse songs and create a playlist that contains 20 pieces of music that fit the user's music taste.

4.2.2 Participants. The experiment took 20 minutes, on average, and we compensated each participant £2.4. A total of 265 users participated in our study. We removed 38 participants' responses for extremely long duration in the study and 54 participants who failed to pass the attention check questions.[5] We finally kept the data of 173 participants, which meet the minimum sample size according to a CFA/SEM rule of thumb that 5:1 is the recommended ratio of subjects to observable variables (N:q) [3]. The number of observable variables is the total number of questions contained in all constructs. Table 1 presents the demographics of those participants.

4.2.3 Validity and Reliability of Evaluation Model. We performed a CFA to establish convergent and discriminant validity. We iteratively adjust the model based on the factor loadings and correlation coefficient between the two factors. For example, removing an indicator until the **average variance extracted (AVE)** of a factor is less than 0.4 [7], or merging two factors if they strongly correlate. A latent variable should contain at least three indicators [36]. Specifically, we dropped some constructs containing only a single item (e.g., Accuracy, Explainability, CUI Attentiveness, and CUI Engagingness), which cannot assess measurement error and check the validity of scales applied in a new context. To validate these dropped constructs, we modified the questions of these constructs and tested them in Study 2. In addition, we merged some strongly correlated constructs (e.g., CUI Positivity & CUI Rapport, CUI Adaptability & CUI Coordination, and Trust & Confidence) to keep the discriminant validity of the constructs.

In addition, we chose Cronbach's alpha and correlated item-total correlations to measure the construct's internal reliability for considered latent variables. The scores of all constructs are above

[5] To ensure the quality of user responses, we set attention checking questions (for example, "*Please indicate which number is an odd number?*"). Besides, we checked if users' responses have dubious patterns, for example, "AAAA," "ABAB," or contain conflicts to similar or reversing questions.

Table 2. Reliability for Latent Factors (Constructs) Validated in Study 1

		Internal Reliability		**Convergent Validity**	
Construct	**Items**	Cronbach alpha (0.5)	Item-total correlation (0.4)	Factor loading (R^2) (0.4)	Variance extracted (AVE) (0.4)
Perceived Qualities					
1. Novelty [76, 95]	4		0.922		0.757
The music chatbot helps me discover new songs.		0.728		0.593	
The music chatbot provides me with surprising recommendations that helped me discover new music that I wouldn't have found elsewhere.		0.896		0.902	
The music chatbot provides me with recommendations that I had not considered in the first place but turned out to be a positive and surprising discovery.		0.816		0.726	
The music chatbot provides me with recommendations that were a pleasant surprise to me because I would not have discovered them somewhere else.		0.850		0.816	
2. Interaction Adequacy [95]	3		0.784		0.560
I find it easy to inform the music chatbot if I dislike/like the recommended song.		0.592		0.549	
The music chatbot allows me to tell what I like/dislike.		0.571		0.455	
I find it easy to tell the system what I like/dislike.		0.722		0.717	
3. CUI Adaptability [116, 129]	3		0.805		0.584
I felt I was in sync with the music chatbot.		0.628		0.605	
The music chatbot adapts continuously to my preferences.		0.642		0.545	
I always have the feeling that this music chatbot learns my preferences.		0.692		0.596	
4. CUI Response Quality [137]	3		0.722		0.473
The music chatbot's responses are readable and fluent.		0.581		0.464	
Most of the chatbot's responses make sense.		0.560		0.475	
The pace of interaction with the music chatbot is appropriate.		0.503		0.479	
User Beliefs					
1. Perceived Usefulness [95]	3		0.816		0.593
The music chatbot helps me find the ideal item.		0.694		0.555	
Using the music chatbot to find what I like is easy.		0.661		0.570	
The music chatbot gives me good suggestions.		0.653		0.659	
2. CUI Rapport [116]	5		0.893		0.629
The music chatbot is warm and caring.		0.750		0.653	
The music chatbot cares about me.		0.803		0.761	
I like and feel warm toward the music chatbot.		0.764		0.715	
I feel that I have no connection with the music chatbot.		0.628		0.431	
The music chatbot and I establish rapport.		0.764		0.627	
User Attitudes					
1. Trust & ***Confidence*** [95]	3		0.801		0.607
This music chatbot can be trusted.		0.528		0.400	
I am convinced of the items recommended to me.		0.731		0.758	
I am confident I will like the items recommended to me.		0.698		0.669	
Behavioral Intentions					
1. Intention to Use [95]	3		0.922		0.798
I will use this music chatbot again.		0.843		0.824	
I will use this music chatbot frequently.		0.872		0.861	
I will tell my friends about this music chatbot.		0.812		0.720	

The symbol "*" indicates the self-composed questions.

the moderate level of 0.5 [44].[6] The scores of item-total correlations are above the cut-off value (0.4) for all constructs [91].

After several iterations, we obtained values as indicated in Table 2. They meet the cut-off values of all validity and reliability indicators. By running these validity tests, we refine each construct's questions and increase the validity of our evaluation model's constructs. After proving the model's reliability and validity, we validated the relationships between constructs using the SEM [64]. Table 2 shows eight validated constructs under four dimensions: Novelty, Interaction Adequacy, CUI Adaptability, CUI Response Quality, Perceived Usefulness, CUI Rapport, Trust & Confidence, and Intention to Use. Each construct contains at least three question items. There are 27 valid question items in the validated questionnaire of Study 1.

[6] Excellent reliability (>0.90), high reliability (0.70–0.90), moderate reliability (0.50–0.70), and low reliability (<0.50).

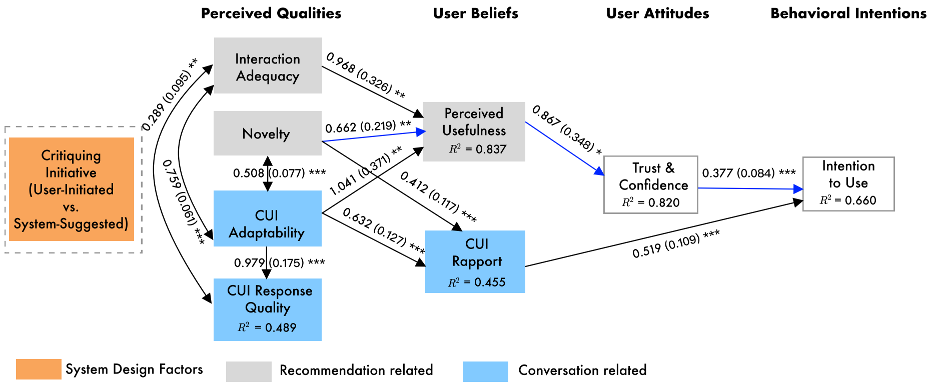


Fig. 5. The Structural Equation Modeling (SEM) results of Study 1. Significance: *** $p < .001$, ** $p < .01$, * $p < .05$, • $p < .10$. R^2 is the proportion of variance explained by the model. Factors are scaled to have a standard deviation of 1.

4.2.4 Structural Model. We employed SEM to build a path model for validating the relationships between constructs. Figure 5 shows the resulting model. Overall, the model has an acceptable fit indicated by the following indices: $\tilde{\chi}^2 = 555.300$ (d.f. = 311), $p < 0.001$, TLI = 0.926, CFI = 0.934, RMSEA = 0.062, 90% CI [0.052, 0.072], which meet the recommended standards of these fit indices [45]. $\tilde{\chi}^2$ is an absolute fit index for the model, but it could be affected by multiple factors, such as sample size, model size, and the distribution of variables. **TLI (Tucker-Lewis Index)** is a relative fit index whose cut-off value is 0.9. **CFI (Comparative Fit Index)** and **RMSEA (Root Mean Square Error)** are non-centrality-based indices for model fit.[7] Besides, R squared values for all the constructs are larger than 0.40, which indicates that the model can examine the significance of the paths associated with these constructs.

According to the path between four dimensions, i.e., Perceived Qualities → User Beliefs → User Attitudes → Behavioral Intentions, we categorize and associate constructs as shown in Figure 5. We use blue color to indicate paths that have been verified in the original *ResQue* model, and the remaining paths are related to the constructs of conversation that are new in our framework *CRS-Que* (except the path Interaction Adequacy → Perceived Usefulness).

The results do not show a significant effect of the manipulated design factor (critiquing initiative) on any measured construct. Therefore, the system design factor box is isolated from the model. The paths between the first two dimensions show that Novelty and CUI Adaptability positively influence Perceived Usefulness and CUI Rapport, and Interaction Adequacy also positively influences Perceived Usefulness. The significant path, Perceived Usefulness → Trust → Intention to Use, indicates that increasing Perceived Usefulness could lead to higher user trust in the system and intention to use the system, which has been validated by *ResQue* [95]. Moreover, the significant path, CUI Rapport → Intention to Use, shows a positive effect of CUI Rapport on Intention to Use. More notably, several short paths, Interaction Adequacy ↔ CUI Adaptability,[8] Interaction

[7] CFI: excellent (>0.99), close (0.95–0.99), fair (0.92–0.95), poor (<0.90); RMSEA: excellent (<0.01), close (0.01–0.05), fair (0.05–0.08), poor (>0.1).

[8] The high covariance between two latent variables could be interpreted as a strong correlation between two factors. According to the definitions of the two factors, they gauge different aspects of the CRS: "Interaction Adequacy" is mainly from users' perspective to see whether they feel easy to tell their preferences or not, while "CUI Adaptability" is about the conversational interface's general ability to adapt to the user's request. Therefore, we keep these two factors in the model instead of merging them into one factor.

Adequacy ↔ CUI Response Quality, and Novelty ↔ CUI Adaptability demonstrate positive correlations between conversation constructs and recommendation constructs of Perceived Qualities.

4.2.5 Discussion of the Results. Several previous studies have compared the user-initiated and system-suggested critiquing systems. For example, one study found that user-initiated critiquing systems outperform system-suggested critiquing systems regarding decision accuracy, decision confidence, and cognitive effort [14]. However, these experimental systems support critiquing features by traditional UI widgets (e.g., buttons and dropdown menus), not natural language interaction.

In Study 1, the comparison of the two critiquing techniques does not yield a significant difference in users' perceived qualities, echoing the findings of our prior studies that compared user-initiated critiquing with system-suggested critiquing in a conversational recommender system [9, 51]. Moreover, Study 1 focuses on a scenario (music recommendations) requiring low user involvement in decision support and critiquing for music exploration, which may influence user perception of the two critiquing initiatives.

Nevertheless, the resulting model still validates the hypothesized relationships between the four dimensions of our framework. Specifically, users think a music CRS is useful if it supports rich user interaction and exploration of new music. Several studies proposed different approaches to explore music for serendipity and novelty, such as exploring music preferences [115] and music genres [73]. Furthermore, for the aspects of conversation, the more *MusicBot* can adapt to user preferences, the higher response quality and rapport the user can perceive. In addition, we validate the correlations between the perceived qualities of recommendations (i.e., Interaction Adequacy and Novelty) and the perceived qualities of conversation (i.e., CUI Adaptability and CUI Response Quality), which imply the importance of associating conversation constructs with recommendation constructs when evaluating a CRS.

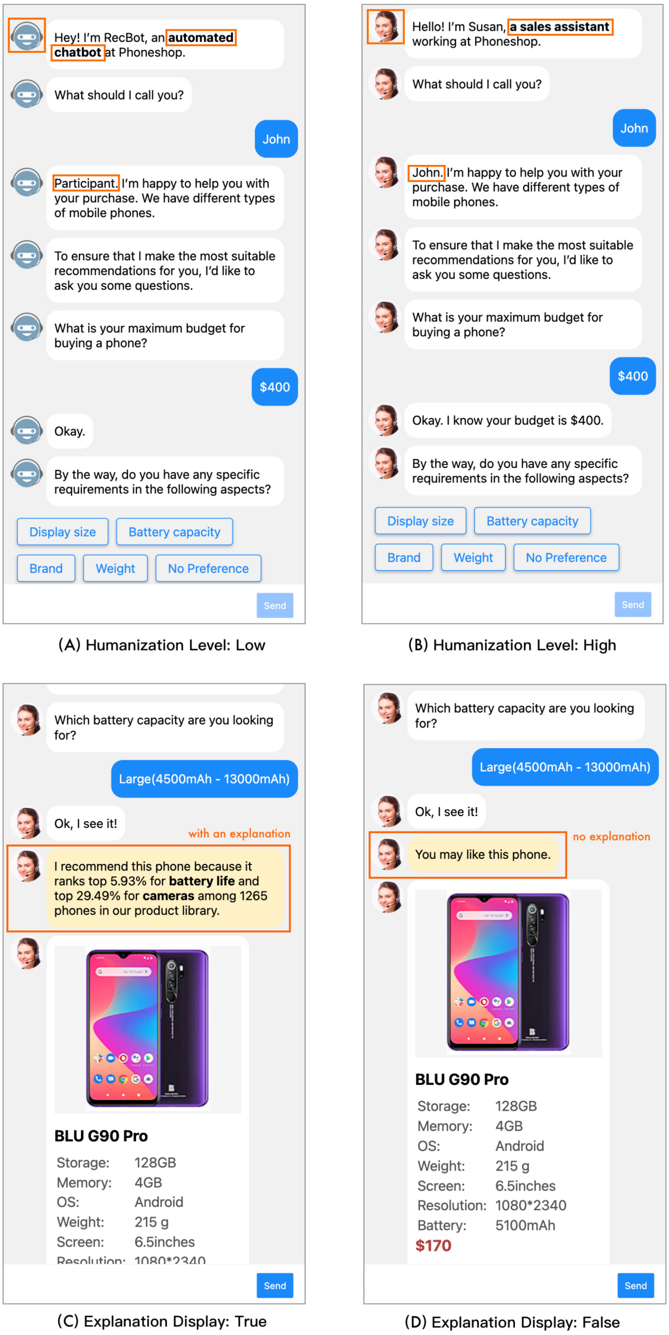
4.3 Study 2: *PhoneBot* for Purchase Decision-making

In this study, we evaluated another CRS, *PhoneBot*, which helps users purchase mobile phones. Compared with *MusicBot* in Study 1, *PhoneBot* allows us to validate our framework in a different scenario (mobile phone recommendations) that requires high user involvement in decision support. Moreover, *PhoneBot* is a mobile application that requires participants to chat with the bot on their mobile devices, which helps us validate the evaluation framework on different platforms. Same as *MusicBot*, we implement the recommendation component based on MAUT [136] and the mobile phones database of GSMArena.com. Moreover, we use DialogFlow ES (standard) to implement the conversation component by defining intents for critiquing mobile phone recommendations on various attributes, such as price, screen resolution, and battery life.

Previous studies have investigated how humanization level and recommendation explanations influence user trust on traditional recommender systems [71, 110, 120, 134]. This study aims to investigate how the two design factors, i.e., *humanization level* and *explanation display* (see Figure 6), influence user trust in CRS. Specifically, we will address the following research questions:

RQ1: How does the *humanization level* of the system influence user trust in the CRS?
RQ2: How do the *recommendation explanations* influence user trust in the CRS?
RQ3: What are the interaction effects of *humanization level* and *recommendation explanations*?

4.3.1 Setup. First, *PhoneBot* elicits user preferences by asking several questions about the user's budget and more specific requirements for other phone attributes (e.g., display size, battery capacity, and brand). After that, the bot presents a recommended phone in a dialog turn.



Hey! I'm RecBot, an **automated chatbot** at Phoneshop.

What should I call you?

John

Participant. I'm happy to help you with your purchase. We have different types of mobile phones.

To ensure that I make the most suitable recommendations for you, I'd like to ask you some questions.

What is your maximum budget for buying a phone?

$400

Okay.

By the way, do you have any specific requirements in the following aspects?

(A) Humanization Level: Low

Hello! I'm Susan, **a sales assistant** working at Phoneshop.

What should I call you?

John

John. I'm happy to help you with your purchase. We have different types of mobile phones.

To ensure that I make the most suitable recommendations for you, I'd like to ask you some questions.

What is your maximum budget for buying a phone?

$400

Okay. I know your budget is $400.

By the way, do you have any specific requirements in the following aspects?

(B) Humanization Level: High

Which battery capacity are you looking for?

Large(4500mAh - 13000mAh)

Ok, I see it!

with an explanation

I recommend this phone because it ranks top 5.93% for **battery life** and top 29.49% for **cameras** among 1265 phones in our product library.

BLU G90 Pro

Storage: 128GB
Memory: 4GB
OS: Android
Weight: 215 g
Screen: 6.5inches
Resolution: 1080*2340

(C) Explanation Display: True

Which battery capacity are you looking for?

Large(4500mAh - 13000mAh)

Ok, I see it!

no explanation

You may like this phone.

BLU G90 Pro

Storage: 128GB
Memory: 4GB
OS: Android
Weight: 215 g
Screen: 6.5inches
Resolution: 1080*2340
Battery: 5100mAh
$170

(D) Explanation Display: False

Fig. 6. The user interfaces of *PhoneBot*.

Table 3. Demographics of the Participants in Study 2

	Item	**Frequency**	**Percentage (%)**
Age	19–25	80	46.24%
	26–30	35	20.23%
	31–35	19	10.98%
	41–50	13	7.51%
	36–40	13	7.51%
	51–60	9	5.20%
	>60	4	2.31%
Gender	Male	90	52.02%
	Female	80	46.24%
	Other	3	1.73%
Nationality	UK	41	23.70%
	USA	38	21.97%
	Portugal	18	10.40%
	Poland	15	8.67%
	Italy	13	7.51%
	Others	48	27.73%

We manipulated the humanization level of *PhoneBot* by adopting multiple humanization features for chatbots [98], including a human avatar, human identity in self-introduction, addressing users by their names, and adaptive response speed (Figure 6(B)). For explanation display, the bot can explain the current recommended item by ranking it in the recommendation pool by some attributes the user cares about (Figure 6(C)), which is compared with its counterpart that does not show any explanations (Figure 6(D)).

We conducted a 2 × 2 between-subjects study to investigate how these two system design factors (humanization level and explanation display) influence purchase decision-making and other user experience aspects according to ***CRS-Que***. The **task** of this study is to help a fictional character pick three mobile phones based on her budget, battery, and display size requirements.

4.3.2 Participants. We recruited 256 participants from the Prolific platform. The average completion time of the experiment is around 12 minutes. We paid £1.5 for each participant. In addition, we applied the same criteria used in Study 1 to filter out some low-quality responses. Specifically, we removed 6 responses that failed to pass attention-check questions, 29 that contained a straight line of answers or a similar pattern, and 5 that took an extremely long duration in the study. Finally, we kept 216 valid responses, an acceptable sample size for running SEM according to the recommended ratio of subjects to observable variables (5:1) [3]. We present the demographics of valid participants in Table 3.

4.3.3 Validity and Reliability of Evaluation Model. Same as the procedure of adjusting the model in Study 1, we performed a CFA to exclude some question items that have a low factor loading (< 0.4) or a strong correlation with other latent factors indicated by modification indices [133]. In the end, our model demonstrates convergent validity, discriminant validity, and internal reliability. Moreover, Table 4 presents the scores of various validity and reliability indices (e.g., Cronbach's alpha, factor loading) and 11 validated constructs, including Accuracy, Explainability, CUI Attentiveness, CUI Understanding, Transparency, Perceived Ease of Use, User Control, CUI Humanness, Trust & Confidence, Satisfaction, and Intention to Purchase. The final questionnaire contains 37 items; at least three questions measure each latent variable (construct).

Table 4. Reliability for Latent Factors (Constructs) Validated in Study 2

Construct	Items	Internal Reliability		Convergent Validity	
		Cronbach alpha (0.5)	Item-total correlation (0.4)	Factor loading (R^2) (0.4)	Variance extracted (AVE) (0.4)
Perceived Qualities					
1. ***Accuracy*** [68, 95]	3		0.805		0.600
The recommended phones were well-chosen.		0.717		0.680	
The recommended phones were relevant.		0.663		0.631	
The recommended phones were interesting.*		0.606		0.482	
2. ***Explainability*** [95]	3		0.916		0.800
The chatbot explained why the phones were recommended to me.		0.893		0.937	
The chatbot explained the logic of recommending phones.*		0.750		0.607	
The chatbot told me the reason why I received the recommended phones.*		0.854		0.847	
3. ***CUI Attentiveness*** [116, 138]	3		0.812		0.598
The chatbot tried to know more about my needs.		0.631		0.514	
The chatbot paid attention to what I was saying.*		0.708		0.700	
The chatbot was respectful to me and considered my needs.*		0.662		0.592	
4. ***CUI Understanding*** [5]	3		0.930		0.822
The chatbot understood what I said.		0.852		0.797	
I found that the chatbot understood what I wanted.		0.899		0.904	
I felt that the chatbot understood my intentions.		0.823		0.767	
User Beliefs					
1. ***Transparency*** [40, 95]	3		0.821		0.614
I understood why the phones were recommended to me.		0.645		0.551	
I understood how the system determined the quality of the phones.		0.680		0.556	
I understood how well the recommendations matched my preferences.		0.711		0.720	
2. ***Perceived Ease of Use*** [95]	4		0.944		0.808
I could easily use the chatbot to find the phones of my interests.*		0.865		0.799	
Using the chatbot to find what I like was easy.		0.871		0.809	
Finding a phone to buy with the help of the chatbot was easy.		0.844		0.763	
It was easy to find what I liked by using the chatbot.*		0.881		0.860	
3. ***User Control*** [95]	3		0.913		0.785
I felt in control of modifying my taste using this chatbot.		0.857		0.861	
I could control the recommendations the chatbot made for me.*		0.761		0.645	
I felt in control of adjusting recommendations based on my preference.*		0.859		0.855	
4. ***CUI Humanness*** [107]	3		0.914		0.787
The chatbot behaved like a human.		0.881		0.903	
I felt like conversing with a real human when interacting with this chatbot.		0.770		0.663	
This chatbot system has human properties.		0.841		0.823	
User Attitudes					
1. ***Trust*** & ***Confidence*** [29, 95]	6		0.955		0.781
The recommendations provided by the chatbot can be trusted.*		0.758		0.666	
I can rely on the chatbot when I need to buy a mobile phone.*		0.821		0.771	
I feel I could count on the chatbot to help me purchase the mobile phone I need.		0.838		0.842	
I was convinced of the phones recommended to me.		0.848		0.806	
I was confident I would like the phones recommended to me.		0.821		0.780	
I had confidence in accepting the phones recommended to me.		0.865		0.815	
2. ***Satisfaction***	3		0.932		0.825
I was satisfied with the recommendations made by the chatbot.*		0.869		0.851	
The recommendations made by the chatbot were satisfying.*		0.833		0.748	
These recommendations made by the chatbot made me satisfied.*		0.879		0.865	
Behavioral Intentions					
1. ***Intention to Purchase*** [37]	3		0.937		0.831
Given a chance, I predict that I would consider buying the phones recommended by the chatbot in the near future.		0.873		0.859	
I will likely buy the phones recommended by the chatbot in the near future.		0.880		0.847	
Given the opportunity, I intend to buy the phones recommended by the chatbot.		0.855		0.788	

The symbol "$*$" indicates the self-composed questions.

4.3.4 Structural Model. Based on the validated 11 constructs, we built a path model to analyze the causal relationships between different constructs using SEM. Overall, our SEM model shows a good fit indicated by the following indices: $\tilde{\chi}^2$ = 1,295.438 (d.f. = 685), $p < 0.001$, TLI = 0.947, CFI = 0.951, RMSEA = 0.049, 90% CI [0.049, 0.060], which meet the recommended standards of these fit indices [45].

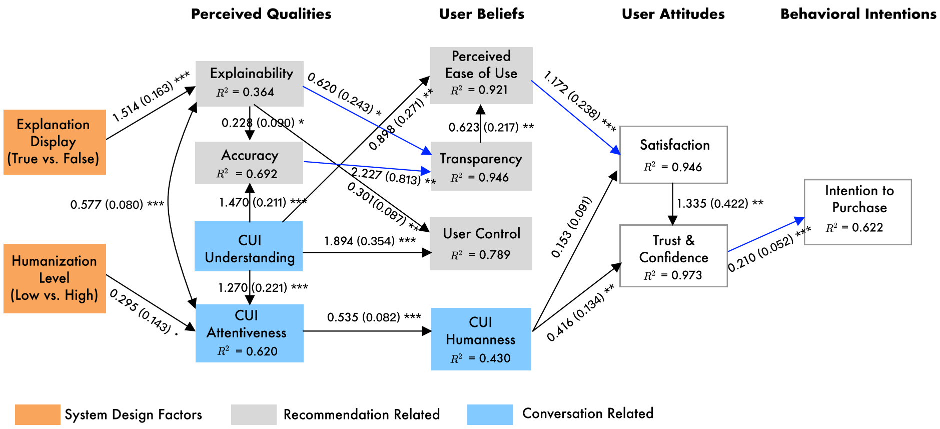


Fig. 7. The structural equation modeling (SEM) results of Study 2. Significance: *** $p < .001$, ** $p < .01$, * $p < .05$, • $p < .10$. R^2 is the proportion of variance explained by the model. Factors are scaled to have a standard deviation of 1.

The resulting model (see Figure 7) shows that our manipulated system design factors influence Perceived Qualities. Specifically, the *explanation* condition has a direct positive effect on Explainability ($p < 0.001$). Furthermore, the path Explainability → (mediators) → Intention to Purchase indicates that Explainability positively influences Transparency, which in turn leads to higher Trust & Confidence and Intention to Purchase.

Despite a marginal significance ($p = 0.06$), the *humanization* level of the system tends to influence CUI Attentiveness positively. The path CUI Attentiveness → (mediators) → Intention to Purchase shows an indirect positive effect of the humanization level on CUI Humanness, Satisfaction, Trust & Confidence, and Intention to Purchase.

In addition, Transparency and User Control are positively influenced by the constructs of Perceived Qualities (i.e., Accuracy, Explainability, and CUI Understanding); however, they do not influence any constructs of User Attitudes and Behavioral Intentions. Similar to the model of Study 1, this model also shows relationships between conversation constructs and recommendation constructs of Perceived Qualities as indicated by the paths Explainability ↔ CUI Attentiveness, and CUI Understanding → Accuracy.

4.3.5 Discussion of the Results. The model of Study 2 confirms the hypothesized relationships between the four dimensions of our framework. Moreover, we manipulated two design factors, i.e., humanization level and explanation to recommendations, which have been well studied by several existing studies [32, 34, 98, 119]. The effects of the two design factors on the constructs indicate that our framework can capture the variability of users' responses on the various measurement scales. Since we only validated a part of the constructs in Study 1, the additional validated constructs in Study 2 (see Table 4) can complement the validation of constructs in our evaluation framework.

Most of our identified positive effects of explanations align with the findings of studies on explanations in traditional recommender systems. For example, explaining recommendations could benefit Transparency [85, 122], User Control [65], Perceived Ease of Use [83], Satisfaction [122], Trust & Confidence [83, 85], and Intention to Purchase [83]. However, the significant correlation between Explainability and CUI Attentiveness suggests that explaining recommendations could also positively influence user perceptions of conversations, for example, the attentiveness and humanness of the agent.

However, the high humanization level tends to increase user-perceived attentiveness of the CRS, which was not reported in the previous studies on the humanization of chatbots [34, 98]. We speculate that one feature of high humanization level, i.e., addressing participants by their names (see Figure 6 (B)), may make participants feel more respect and attention from *PhoneBot*. Moreover, users tend to be satisfied with the recommendation and trust the CRS when they perceive a high humanness of the CRS, echoing the positive effects of humanness on user satisfaction and trust reported in the studies on a survey chatbot [98] and a news chatbot [107].

5 DISCUSSION

5.1 Framework Validation

This article aims to provide a consolidated and unifying framework for the user-centric evaluation of conversational recommender systems rather than investigating the effects of specific design factors on a CRS's **user experience (UX)**. User-centric evaluation has gained extensive attention in the community of recommender systems. Recently, researchers have discussed the importance of subjective evaluation metrics (perception-oriented) in addition to objective metrics (computation-oriented), such as algorithmic accuracy, understanding rate, and dialogue turns [47]. The previous studies show that the user's perception of conversations strongly influences the overall user experience of a CRS [52, 72, 89]. Therefore, we have developed a unifying user-centric evaluation framework called ***CRS-Que*** for conversational recommender systems. Compared with the original *ResQue* model that primarily focuses on traditional recommender systems [95], our framework seamlessly integrates several important user experience constructs of conversations into the *ResQue* model, allowing researchers and practitioners to evaluate a CRS more comprehensively. Specifically, by reviewing the existing UX metrics of conversational agents, we identified eight constructs (e.g., CUI Adaptability, CUI Response Quality, and CUI Understanding) that are closely related to the quality of conversations based on the theory of rapport [116] and humanlikeness [34]. Then, by performing CFA, we merged several conversation constructs and integrated them into *ResQue*. Ultimately, ***CRS-Que*** model accommodates adaptability, understanding, attentiveness, response quality, rapport, and humanness. To validate our proposed evaluation framework, we conducted two user studies to evaluate two conversational recommender systems (i.e., *MusicBot* and *PhoneBot*). The two studies target different recommendation domains (i.e., low user involvement and high user involvement) and devices (i.e., personal computers and mobile phones), which help us confirm the robustness and generalizability of our framework.

5.2 Effects on Recommendation and Conversation Constructs

Most paths between recommendation constructs shown in our models have been validated in *ResQue* [95] (marked with blue; see Figures 5 and 7). Additionally, our models identify some new paths between recommendation constructs, for example, the positive effect of Interaction Adequacy on Perceived Usefulness (in Study 1) and the positive effect of Explainability on User Control (in Study 2). These new effects may be attributed to the influence of incorporating conversation constructs into our *CRS-Que* framework. For conversation constructs, the model of Study 1 validates a previously verified path between conversation constructs, i.e., the positive effects of CUI Adaptability on CUI Rapport [116]. Moreover, the model of Study 2 shows a relationship between Humanization Level and Intention to Purchase mediated by CUI Attentiveness, CUI Humanness, and Trust & Confidence. This relationship suggests that increasing the humanization level by applying various social features to a CRS could increase user trust, which is particularly important to the recommendations for decision-making with high user involvement (e.g., high-cost product recommendations) [96]. More importantly, the positive correlations between the constructs of conversations and recommendations explicitly show the added value of ***CRS-Que*** in explaining the user experience of CRS. For example, we find that the novelty of recommendations positively influences the rapport of conversation (in Study 1); and improving the understanding of a CRS could increase perceived ease of use and user control (in Study 2).

In most cases, both recommendation and conversation constructs influence the constructs of Behavioral Intention through the constructs of User Attitudes. For example, CUI Humanness positively influences Intention to Purchase via Trust & Confidence in Study 2. However, in Study 1, we find that the rapport of conversation could directly influence the user's intention to use a CRS, which implies that increasing the friendliness of a CRS (e.g., caring and warm expressions) could immediately stimulate users' behavioral intentions. Compared with a traditional recommender system, a CRS provides a more natural and free way for users to control the system, which may attract satisfied users to use the CRS repeatedly in the future [39].

5.3 The Use of *CRS-Que*

We validate the evaluation framework through two user studies, indicating the validity of all contained constructs and causal relationships among some constructs. Although the identified relationships among these constructs may depend on the system settings (e.g., algorithm configuration, interaction design, and application domains), the causal relationships among the four dimensions (i.e., Perceived Qualities, User Beliefs, User Attitudes, and Behavioral Intentions) should always be held. Therefore, we suggest evaluators select which constructs are adopted in their questionnaires based on their objectives and needs and ensure that the selected constructs span four dimensions of *CRS-Que* for making an integrative model to evaluate a CRS. To generalize the description in the evaluation framework, we use more general terms, for example, "items" in the questionnaire provided in appendices. However, users may use more specific terms to describe objects to be evaluated. For example, recommended books and recommended movies.

According to the evaluation framework of recommender systems proposed by Knijnenburg et al. [68], the user experience of recommender systems also depends on personal characteristics (e.g., demographics, domain knowledge) and situational characteristics (e.g., privacy concerns, choice goals). Therefore, it is possible to incorporate personal and situational characteristics into *CRS-Que* when evaluating a CRS with different types of users or in different contexts [10, 11].

In addition, we provide the original questionnaires we validated through the two studies as supplementary material (Appendix A), allowing researchers to customize the evaluation model according to their interests and needs. We also mark question items dropped in our final questionnaires to advise researchers to be cautious about adopting them in their studies. For evaluators aiming to conduct a quick study to evaluate a CRS, they may use one question for each of the selected constructs. We provided a short version of the questionnaire, which is detailed in Table 5 in Appendix B. We developed the short version based on the loadings in factor analysis and the similarity of model structure [109].

5.4 Limitations

We validated the evaluation framework by assessing two conversational recommender systems in two different application domains. However, the development and validation of this evaluation framework contain several limitations. *First*, the design of the study may limit the generalizability of our framework. For instance, the dialogue design in Study 1 primarily focuses on critiquing-based interaction [17], representing a specific user feedback acquisition in conversational recommender systems. *Second*, we evaluated two different systems in two different application domains, which may make it challenging to examine the exact impact of the domain on this evaluation framework. Given these two limitations, future evaluation efforts for conversational recommender systems might involve a more diverse range of dialogue designs for recommendation scenarios, as well as identifying the domain independence of the evaluation framework (e.g., testing the same CRS in different domains). *Third*, although most framework constructs are not restricted to text-based conversational recommender systems, these constructs have only been validated with text-based conversations in the two studies. To assess its validity in a voice-based system, researchers may need to conduct additional studies with some constructs related to the quality of voice-based interaction. *Last*, we validate the framework with two systems with some technical limitations. For example, our predefined intents may not cover all user intents to explore music (in Study 1), and the conversation skills cannot be comparable with those of an agent powered by **large language models (LLMs)**. In the future, we plan to evaluate more advanced conversational agents (e.g., based on ChatGPT) for recommendation scenarios. To maintain this evaluation framework, we will track how practitioners use *CRS-Que* to evaluate different CRSs.

6 CONCLUSION

We propose a unifying user-centric evaluation framework for the **conversational recommender system (CRS)** based on *ResQue* [95]. We review the subjective metrics of measuring user experience of conversational systems and seamlessly integrate them into the four dimensions of *ResQue*: Perceived System Qualities, User Beliefs, User Attitudes, and Behavioral Intentions. We conducted two online user studies to identify the validity and reliability of the evaluation constructs and adjusted the evaluation models based on factor analysis results. Eventually, we kept 64 question items under 18 constructs in our final framework. Moreover, we have identified several influencing paths that show how conversation constructs and recommendation constructs influence each other.

Our framework, ***CRS-Que***, allows practitioners to systematically evaluate a CRS by looking at the UX factors of both recommendations and conversation. The questionnaire can assess the user experience of various conversational recommender systems for different application domains. In the end, we discuss the use of ***CRS-Que*** in practice for meeting various needs of evaluating a CRS from users' perspectives.

APPENDICES

A THE ORIGINAL QUESTIONNAIRE

The following is the original questionnaire used in our user studies containing all constructs. Users may customize the questionnaire according to their needs and then follow our presented method to validate the validity and reliability of their own models. We use "*" to mark the dropped question items that did not contribute to the validated constructs in the final models.

Perceived Qualities

Accuracy

- The recommended items were well-chosen.
- The recommended items were relevant.
- The recommended items were interesting.
- The items recommended to me matched my interests. *

Novelty

- The chatbot helped me discover new items.
- The chatbot provided me with surprising recommendations that helped me discover new items that I wouldn't have found elsewhere.
- The chatbot provided me with recommendations that I had not considered in the first place but turned out to be a positive and surprising discovery.

- The chatbot provided me with recommendations that were a pleasant surprise to me because I would not have discovered them somewhere else.
- The items recommended to me are novel. *

Interaction Adequacy

- I found it easy to inform the chatbot if I dislike/like the recommended item.
- The chatbot allows me to tell what I like/dislike.
- I found it easy to tell the system what I like/dislike.
- The chatbot allows me to modify my taste profile. *
- I found it easy to modify my taste profile in the chatbot. *

Explainability

- The chatbot explained why the items were recommended to me.
- The chatbot explained the logic of recommending items.
- The chatbot told me the reason why I received the recommended items.
- The chatbot helped me understand why the items were recommended to me. *

CUI Adaptability

- The chatbot adapted continuously to my preferences.
- I always had the feeling that this chatbot learns my preferences.
- I felt I was in sync with the chatbot.

CUI Understanding

- The chatbot understood what I said.
- I found that the chatbot understood what I wanted.
- I felt that the chatbot understood my intentions.

CUI Response Quality

- The chatbot's responses are readable and fluent.
- Most of the chatbot's responses make sense.
- The pace of interaction with the chatbot is appropriate.
- The chatbot's responses are informative. *
- The chatbot responded to my query/request quickly. *
- I found the chatbot easy to understand in this conversation. *

CUI Attentiveness

- The chatbot tried to know more about my needs.
- The chatbot was interested in what I was saying. *
- The chatbot was respectful to me and considered my needs.
- The chatbot paid attention to what I was saying.

User Beliefs

Perceived Ease of Use

- I could easily use the chatbot to find the items of my interests.
- Using the chatbot to find what I like was easy.
- Finding an item to buy with the help of the chatbot was easy.
- It was easy to find what I liked by using the chatbot.

Perceived Usefulness

- The chatbot helped me find the ideal item.
- Using the chatbot to find what I like is easy.
- The chatbot gave me good suggestions.

User Control

- I felt in control of modifying my taste using this chatbot.
- I could control the recommendations the chatbot made for me.
- I felt in control of adjusting recommendations based on my preference.
- I felt in control of telling the chatbot what I want. *

Transparency

- I understood why the items were recommended to me.
- I understood how the system determined the quality of the items.
- I understood how well the recommendations matched my preferences.
- I understand the underlying logic of the recommendation service. *

CUI Humanness

- The chatbot behaved like a human.
- I felt like conversing with a real human when interacting with this chatbot.
- This chatbot system has human properties.

CUI Rapport

- The chatbot was warm and caring.
- The chatbot cared about me.
- I liked and felt warm toward the chatbot.
- I felt that I had no connection with the chatbot.
- The chatbot and I established rapport.
- I felt rapport between this chatbot and myself. *
- The chatbot was friendly to me. *

User Attitudes

Trust & Confidence

- The recommendations provided by the chatbot can be trusted.
- I can rely on the chatbot when I need to buy an item.
- I feel I could count on the chatbot to help me purchase the item I need.
- I was convinced of the items recommended to me.
- I was confident I would like the items recommended to me.
- I had confidence in accepting the items recommended to me.

Satisfaction

- I was satisfied with the recommendations made by the chatbot.
- The recommendations made by the chatbot were satisfying.
- These recommendations made by the chatbot made me satisfied.

Behavioral Intentions

Intention to Use

- I will use this chatbot again.
- I will use this chatbot frequently.
- I will tell my friends about this chatbot.

Intention to Purchase

- Given a chance, I predict that I would consider buying the items recommended by the chatbot in the near future.
- I will likely buy the items recommended by the chatbot in the near future.
- Given the opportunity, I intend to buy the items recommended by the chatbot.

B THE SHORT VERSION

Table 5. The Short Version of *CRS-Que*

Perceived Qualities	
Accuracy	The recommended items were well-chosen.
Novelty	The chatbot provided me with surprising recommendations that helped me discover new items that I wouldn't have found elsewhere.
Interaction Adequacy	I found it easy to tell the system what I like/dislike.
Explainability	The chatbot explained why the items were recommended to me.
CUI Adaptability	I felt I was in sync with the chatbot.
CUI Understanding	I found that the chatbot understood what I wanted.
CUI Response Quality	Most of the chatbot's responses make sense.
CUI Attentiveness	The chatbot paid attention to what I was saying.
User Beliefs	
Perceived Ease of Use	It was easy to find what I liked by using the chatbot.
Perceived Usefulness	The chatbot gave me good suggestions.
User Control	I felt in control of modifying my taste using this chatbot.
Transparency	I understood how well the recommendations matched my preferences.
CUI Humanness	The chatbot behaved like a human.
CUI Rapport	The chatbot cared about me.
User Attitudes	
Trust & Confidence	I feel I could count on the chatbot to help me choose/purchase the items I need.
Satisfaction	These recommendations made by the chatbot made me satisfied.
Behavioral Intentions	
Intention to Use	I will use this chatbot frequently.
Intention to Purchase	Given a chance, I predict that I would consider buying the items recommended by the chatbot in the near future.

REFERENCES

[1] Ahmad Abdellatif, Khaled Badran, Diego Elias Costa, and Emad Shihab. 2021. A Comparison of Natural Language Understanding Platforms for Chatbots in Software Engineering. *IEEE Transactions on Software Engineering* 48, 8 (May 2021), 3087–3102. https://doi.org/10.1109/tse.2021.3078384

[2] Syed Z. Arshad, Jianlong Zhou, Constant Bridon, Fang Chen, and Yang Wang. 2015. Investigating user confidence for uncertainty presentation in predictive decision making. In *Proceedings of the Annual Meeting of the Australian Special Interest Group for Computer Human Interaction*. 352–360. DOI:https://doi.org/10.1145/2838739.2838753

[3] Peter M. Bentler. 1990. Comparative fit indexes in structural models.*Psychol. Bull.* 107, 2 (1990), 238. DOI:https://doi.org/10.1037/0033-2909.107.2.238

[4] Shlomo Berkovsky, Ronnie Taib, and Dan Conway. 2017. How to recommend? User trust factors in movie recommender systems. In *Proceedings of the 22nd International Conference on Intelligent User Interfaces (IUI'17)*. Association for Computing Machinery, New York, NY, 287–300. DOI: https://doi.org/10.1145/3025171.3025209

[5] Simone Borsci, Alessio Malizia, Martin Schmettow, Frank Van Der Velde, Gunay Tariverdiyeva, Divyaa Balaji, and Alan Chamberlain. 2022. The chatbot usability scale: The design and pilot of a usability scale for interaction with AI-based conversational agents. *Person. Ubiq. Comput.* 26, 1 (2022), 95–119. DOI: https://doi.org/10.1007/s00779-021-01582-9

[6] Daniel Braun, Adrian Hernandez Mendez, Florian Matthes, and Manfred Langen. 2017. Evaluating natural language understanding services for conversational question answering systems. In *Proceedings of the 18th Annual SIGdial Meeting on Discourse and Dialogue*. 174–185. DOI: https://doi.org/10.18653/v1/w17-5522

[7] Timothy A. Brown. 2015. *Confirmatory Factor Analysis for Applied Research: 3. Introduction to CFA*. Guilford Publications.

[8] Wanling Cai and Li Chen. 2020. Predicting user intents and satisfaction with dialogue-based conversational recommendations. In *Proceedings of the 28th ACM Conference on User Modeling, Adaptation and Personalization (UMAP'20)*. Association for Computing Machinery, New York, NY, 33–42. DOI: https://doi.org/10.1145/3340631.3394856

[9] Wanling Cai, Yucheng Jin, and Li Chen. 2021. Critiquing for music exploration in conversational recommender systems. In *Proceedings of the International Conference on Intelligent User Interfaces (IUI'21)*. Association for Computing Machinery, New York, NY, 480–490. DOI: https://doi.org/10.1145/3397481.3450657

[10] Wanling Cai, Yucheng Jin, and Li Chen. 2022. Impacts of personal characteristics on user trust in conversational recommender systems. In *Proceedings of the CHI Conference on Human Factors in Computing Systems (CHI'22)*. Association for Computing Machinery, New York, NY. DOI: https://doi.org/10.1145/3491102.3517471

[11] Wanling Cai, Yucheng Jin, and Li Chen. 2022. Task-Oriented User Evaluation on Critiquing-Based Recommendation Chatbots. *IEEE Transactions on Human-Machine Systems* 52, 3 (Jan 2022), 354–366. https://doi.org/10.1109/thms.2021.3131674

[12] Heloisa Candello, Claudio Pinhanez, and Flavio Figueiredo. 2017. Typefaces and the perception of humanness in natural language chatbots. In *Proceedings of the CHI Conference on Human Factors in Computing Systems (CHI'17)*. Association for Computing Machinery, New York, NY, 3476–3487. DOI: https://doi.org/10.1145/3025453.3025919

[13] Ana Paula Chaves and Marco Aurelio Gerosa. 2020. How should my chatbot interact? A survey on social characteristics in human–chatbot interaction design. *Int. J. Hum.–Comput. Interact.* 37, 8 (Nov. 2020), 729–758. DOI: https://doi.org/10.1080/10447318.2020.1841438

[14] Li Chen and Pearl Pu. 2006. Evaluating critiquing-based recommender agents. In *Proceedings of the 21st National Conference on Artificial Intelligence (AAAI'06)*. AAAI Press, 157–162. DOI: https://doi.org/10.5555/1597538.1597564

[15] Li Chen and Pearl Pu. 2008. A cross-cultural user evaluation of product recommender interfaces. In *Proceedings of the ACM Conference on Recommender Systems (RecSys'08)*. Association for Computing Machinery, New York, NY, 75–82. DOI: https://doi.org/10.1145/1454008.1454022

[16] Li Chen and Pearl Pu. 2009. Interaction design guidelines on critiquing-based recommender systems. *User Model. User-adapt Interact.* 19, 3 (Oct. 2009), 167–206. DOI: https://doi.org/10.1007/s11257-008-9057-x

[17] Li Chen and Pearl Pu. 2012. Critiquing-based recommenders: Survey and emerging trends. *User Model. User-adapt. Interact.* 22, 1–2 (Apr. 2012), 125–150. DOI: https://doi.org/10.1007/s11257-011-9108-6

[18] Li Chen and Feng Wang. 2017. Explaining recommendations based on feature sentiments in product reviews. In *Proceedings of the International Conference on Intelligent User Interfaces (IUI'17)*. Association for Computing Machinery, New York, NY, 17–28. DOI: https://doi.org/10.1145/3025171.3025173

[19] Li Chen, Yonghua Yang, Ningxia Wang, Keping Yang, and Quan Yuan. 2019. How serendipity improves user satisfaction with recommendations? A large-scale user evaluation. In *Proceedings of the World Wide Web Conference (WWW'19)*. Association for Computing Machinery, New York, NY, 240–250. DOI: https://doi.org/10.1145/3308558.3313469

[20] Gordon W. Cheung and Chang Wang. 2017. Current approaches for assessing convergent and discriminant validity with SEM: Issues and solutions. In *Academy of Management Proceedings*, Vol. 2017. Academy of Management Briarcliff Manor, NY.

[21] Konstantina Christakopoulou, Filip Radlinski, and Katja Hofmann. 2016. Towards conversational recommender systems. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'16)*. ACM, New York, NY, 815–824. DOI: https://doi.org/10.1145/2939672.2939746

[22] Henriette Cramer, Vanessa Evers, Satyan Ramlal, Maarten Van Someren, Lloyd Rutledge, Natalia Stash, Lora Aroyo, and Bob Wielinga. 2008. The effects of transparency on trust in and acceptance of a content-based art recommender. *User Model. User-adapt. Interact.* 18, 5 (2008), 455–496. DOI: https://doi.org/10.1007/s11257-008-9051-3

[23] Paolo Cremonesi, Franca Garzotto, Sara Negro, Alessandro Vittorio Papadopoulos, and Roberto Turrin. 2011. Looking for "good" recommendations: A comparative evaluation of recommender systems. In *Proceedings of the IFIP Conference on Human-Computer Interaction*. Springer, 152–168. DOI: https://doi.org/10.1007/978-3-642-23765-2_11

[24] Robert Dale and Chris Mellish. 1998. Towards evaluation in natural language generation. In *Proceedings of 1st International Conference on Language Resources and Evaluation.*

[25] Fred D. Davis. 1989. Perceived usefulness, perceived ease of use, and user acceptance of information technology. *MIS Quarterly* 13, 3 (1989), 319–340. https://doi.org/10.2307/249008

[26] Ewart J. De Visser, Samuel S. Monfort, Ryan McKendrick, Melissa A. B. Smith, Patrick E. McKnight, Frank Krueger, and Raja Parasuraman. 2016. Almost human: Anthropomorphism increases trust resilience in cognitive agents.*J. Experim. Psychol.: Appl.* 22, 3 (2016), 331–349. DOI:https://doi.org/10.1037/xap0000092

[27] Justin Edwards and Elaheh Sanoubari. 2019. A need for trust in conversational interface research. In *Proceedings of the 1st International Conference on Conversational User Interfaces (CUI'19).* Association for Computing Machinery. DOI:https://doi.org/10.1145/3342775.3342809

[28] Soude Fazeli, Hendrik Drachsler, Marlies Bitter-Rijpkema, Francis Brouns, Wim van der Vegt Brouns, and Peter B. Sloep. 2017. User-centric evaluation of recommender systems in social learning platforms: Accuracy is just the tip of the iceberg. *IEEE Trans. Learn. Technol.* 11, 3 (July 2017), 294–306. DOI:https://doi.org/10.1109/tlt.2017.2732349

[29] Mark A. Fuller, Mark A. Serva, and John "Skip" Benamati. 2007. Seeing is believing: The transitory influence of reputation information on e-commerce trust and decision making. *Decis. Sci.* 38, 4 (2007), 675–699. DOI:https://doi.org/10.1111/j.1540-5915.2007.00174.x

[30] Chongming Gao, Wenqiang Lei, Xiangnan He, Maarten de Rijke, and Tat-Seng Chua. 2021. Advances and challenges in conversational recommender systems: A survey. *AI Open* 2 (2021), 100–126. DOI:https://doi.org/10.1016/j.aiopen.2021.06.002

[31] Mouzhi Ge, Carla Delgado-Battenfeld, and Dietmar Jannach. 2010. Beyond accuracy: Evaluating recommender systems by coverage and serendipity. In *Proceedings of the Conference on Recommender Systems (RecSys'10).* Association for Computing Machinery, New York, NY, 257–260. DOI:https://doi.org/10.1145/1864708.1864761

[32] Fatih Gedikli, Dietmar Jannach, and Mouzhi Ge. 2014. How should I explain? A comparison of different explanation types for recommender systems. *Int. J. Hum.-Comput. Stud.* 72, 4 (2014), 367–382. DOI:https://doi.org/10.1016/j.ijhcs.2013.12.007

[33] Asma Ghandeharioun, Judy Hanwen Shen, Natasha Jaques, Craig Ferguson, Noah Jones, Agata Lapedriza, and Rosalind Picard. 2019. *Approximating Interactive Human Evaluation with Self-Play for Open-Domain Dialog Systems.* Curran Associates Inc., Red Hook, NY, USA.

[34] Eun Go and S. Shyam Sundar. 2019. Humanizing chatbots: The effects of visual, identity and conversational cues on humanness perceptions. *Comput. Hum. Behav.* 97 (2019), 304–316. DOI:https://doi.org/10.1016/j.chb.2019.01.020

[35] Marco Guerini, Sara Falcone, and Bernardo Magnini. 2019. A methodology for evaluating interaction strategies of task-oriented conversational agents. In *Proceedings of the EMNLP Workshop SCAI: The 2nd International Workshop on Search-Oriented Conversational AI.* DOI:https://doi.org/10.18653/v1/w18-5704

[36] Joseph F. Hair Jr, William C. Black, Barry J. Babin, and Rolph E. Anderson. 2009. *Multivariate Data Analysis* (7th ed.). Prentice Hall, Upper Saddle River, NJ.

[37] Nick Hajli. 2015. Social commerce constructs and consumer's intention to buy. *Int. J. Inf. Manag.* 35, 2 (2015), 183–191. DOI:https://doi.org/10.1016/j.ijinfomgt.2014.12.005

[38] Sandra G. Hart and Lowell E. Staveland. 1988. Development of NASA-TLX (task load index): Results of empirical and theoretical research. In *Human Mental Workload*, Peter A. Hancock and Najmedin Meshkati (Eds.). *Advances in Psychology*, Vol. 52. North-Holland, 139–183. DOI:https://doi.org/10.1016/S0166-4115(08)62386-9

[39] Marcel Heerink, Ben Kröse, Bob Wielinga, and Vanessa Evers. 2008. Enjoyment intention to use and actual use of a conversational robot by elderly people. In *Proceedings of the 3rd ACM/IEEE International Conference on Human–Robot Interaction.* 113–119. DOI:https://doi.org/10.1145/1349822.1349838

[40] Marco Hellmann, Diana Carolina Hernandez Bocanegra, and Jürgen Ziegler, 2022. Development of an Instrument for Measuring Users' Perception of Transparency in Recommender Systems. In *Workshops at the International Conference on Intelligent User Interfaces (IUI) 2022: Proceedings of the IUI 2022 Workshops: APEx-UI, HAI-GEN, HEALTHI, HUMANIZE, TExSS, SOCIALIZE*, A. Smith-Renner and O. Amir (Eds.). Vol. 3124, RWTH Aachen, 156–165. https://doi.org/10.17185/duepublico/75905

[41] Jonathan L. Herlocker, Joseph A. Konstan, Loren G. Terveen, and John T. Riedl. 2004. Evaluating collaborative filtering recommender systems. *ACM Trans. Inf. Syst.* 22, 1 (2004), 5–53. DOI:https://doi.org/10.1145/963770.963772

[42] Diana C. Hernandez-Bocanegra and Jürgen Ziegler. 2021. Conversational review-based explanations for recommender systems: Exploring users' query behavior. In *Proceedings of the 3rd Conference on Conversational User Interfaces (CUI'21).* Association for Computing Machinery, New York, NY. DOI:https://doi.org/10.1145/3469595.3469596

[43] Yoshinori Hijikata, Yuki Kai, and Shogo Nishida. 2014. A study of user intervention and user satisfaction in recommender systems. *J. Inf. Process.* 22, 4 (2014), 669–678. DOI:https://doi.org/10.2197/ipsjjip.22.669

[44] Perry Roy Hinton, Charlotte Brownlow, and Isabella McMurray. 2004. *SPSS Explained.* Psychology Press.

[45] Daire Hooper, Joseph Coughlan, and Michael R. Mullen. 2008. Structural equation modelling: Guidelines for determining model fit. *Electron. J. Busin. Res. Meth.* 6, 1 (2008), 53–60.

[46] John A. Hoxmeier, Winter Nie, and G. Thomas Purvis. 2000. The impact of gender and experience on user confidence in electronic mail. *J. Organiz. End User Comput.* 12, 4 (2000), 11–20. DOI: https://doi.org/10.4018/joeuc.2000100102

[47] Dietmar Jannach. 2022. Evaluating conversational recommender systems. *Artif. Intell. Rev.* 56, 3 (July 2022), 2365–2400. DOI: https://doi.org/10.1007/s10462-022-10229-x

[48] Dietmar Jannach, Ahtsham Manzoor, Wanling Cai, and Li Chen. 2021. A survey on conversational recommender systems. *ACM Comput. Surv.* 54, 5, Article 105 (May 2021), 36 pages. DOI: https://doi.org/10.1145/3453154

[49] Dietmar Jannach, Sidra Naveed, and Michael Jugovac. 2017. User control in recommender systems: Overview and interaction challenges. *Lect. Notes Busin. Inf. Process.* 278 (2017), 21–33. DOI: https://doi.org/10.1007/978-3-319-53676-7_2

[50] Jiepu Jiang and Naman Ahuja. 2020. Response quality in human-chatbot collaborative systems. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'20).* Association for Computing Machinery, New York, NY, 1545–1548. DOI: https://doi.org/10.1145/3397271.3401234

[51] Yucheng Jin, Wanling Cai, Li Chen, Nyi Nyi Htun, and Katrien Verbert. 2019. MusicBot: Evaluating critiquing-based music recommenders with conversational interaction. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management (CIKM'19).* Association for Computing Machinery, New York, NY, 951–960. DOI: https://doi.org/10.1145/3357384.3357923

[52] Yucheng Jin, Li Chen, Wanling Cai, and Pearl Pu. 2021. Key qualities of conversational recommender systems: From users' perspective. In *Proceedings of the 9th International Conference on Human-Agent Interaction (HAI'21).* Association for Computing Machinery, New York, NY, 93–102. DOI: https://doi.org/10.1145/3472307.3484164

[53] Yucheng Jin, Nyi Nyi Htun, Nava Tintarev, and Katrien Verbert. 2019. ContextPlay: Evaluating user control for context-aware music recommendation. In *Proceedings of the User Modeling, Adaptation and Personalization Conference (UMAP'19).* Association for Computing Machinery, New York, NY, 294–302. DOI: https://doi.org/10.1145/3320435.3320445

[54] Yucheng Jin, Karsten Seipp, Erik Duval, and Katrien Verbert. 2016. Go with the flow: Effects of transparency and user control on targeted advertising using flow charts. In *Proceedings of the International Working Conference on Advanced Visual Interfaces (AVI'16).* Association for Computing Machinery, New York, NY, 68–75. DOI: https://doi.org/10.1145/2909132.2909269

[55] Yucheng Jin, Nava Tintarev, Nyi Nyi Htun, and Katrien Verbert. 2020. Effects of personal characteristics in control-oriented user interfaces for music recommender systems. *User Model. User-adapt. Interact.* 30, 2 (2020), 199–249. DOI: https://doi.org/10.1007/s11257-019-09247-2

[56] Yucheng Jin, Nava Tintarev, and Katrien Verbert. 2018. Effects of personal characteristics on music recommender systems with different levels of controllability. In *Proceedings of the 12th ACM Conference on Recommender Systems (RecSys'18).* Association for Computing Machinery, New York, NY, 13–21. DOI: https://doi.org/10.1145/3240323.3240358

[57] Mohammed Kaleem, Omar Alobadi, James O'Shea, and Keeley Crockett. 2016. Framework for the formulation of metrics for conversational agent evaluation. In *Proceedings of the Workshop on Collecting and Generating Resources for Chatbots and Conversational Agents-Development and Evaluation Workshop Programme (RE-WOCHAT'16).*

[58] Marius Kaminskas and Derek Bridge. 2016. Diversity, serendipity, novelty, and coverage: A survey and empirical analysis of beyond-accuracy objectives in recommender systems. *ACM Trans. Interact. Intell. Sys.* 7, 1, Article 2 (Dec. 2016), 42 pages. DOI: https://doi.org/10.1145/2926720

[59] Dongyeop Kang, Anusha Balakrishnan, Pararth Shah, Paul Crook, Y-Lan Boureau, and Jason Weston. 2019. Recommendation as a communication game: Self-supervised bot-play for goal-oriented dialogue. *arXiv preprint arXiv:1909.03922* (2019).

[60] Jie Kang, Kyle Condiff, Shuo Chang, Joseph A. Konstan, Loren Terveen, and F. Maxwell Harper. 2017. Understanding how people use natural language to ask for recommendations. In *Proceedings of the 11th ACM Conference on Recommender Systems (RecSys'17).* ACM, New York, NY, 229–237. DOI: https://doi.org/10.1145/3109859.3109873

[61] Pratik Kataria, Kiran Rode, Akshay Jain, Prachi Dwivedi, and Sukhada Bhingarkar. 2018. User adaptive chatbot for mitigating depression. *Int. J. Pure Appl. Math.* 118, 16 (2018), 349–361.

[62] Jurek Kirakowski and Mary Corbett. 1993. SUMI: The software usability measurement inventory. *Brit. J. Educ. Technol.* 24, 3 (1993), 210–212.

[63] René F. Kizilcec. 2016. How much information? Effects of transparency on trust in an algorithmic interface. In *Proceedings of the CHI Conference on Human Factors in Computing Systems.* 2390–2395.

[64] Rex B. Kline and Darcy A. Santor. 1999. Principles & practice of structural equation modelling. *Canad. Psychol.* 40, 4 (1999), 381.

[65] Bart P. Knijnenburg, Svetlin Bostandjiev, John O'Donovan, and Alfred Kobsa. 2012. Inspectability and control in social recommenders. In *Proceedings of the 6th ACM Conference on Recommender Systems.* 43–50.

[66] Bart P. Knijnenburg, Niels J. M. Reijmer, and Martijn C. Willemsen. 2011. Each to his own: How different users call for different interaction methods in recommender systems. In *Proceedings of the 5th ACM Conference on Recommender Systems*. 141–148.

[67] Bart P. Knijnenburg and Martijn C. Willemsen. 2009. Understanding the effect of adaptive preference elicitation methods on user satisfaction of a recommender system. In *Proceedings of the 3rd ACM Conference on Recommender Systems*. 381–384.

[68] Bart P. Knijnenburg, Martijn C. Willemsen, Zeno Gantner, Hakan Soncu, and Chris Newell. 2012. Explaining the user experience of recommender systems. *User Model. User-adapt. Interact.* 22, 4 (2012), 441–504.

[69] Joseph A. Konstan and John Riedl. 2012. Recommender systems: From algorithms to user experience. *User Model. User-adapt. Interact.* 22, 1 (2012), 101–123.

[70] Karolina Kuligowska. 2015. Commercial chatbot: Performance evaluation, usability metrics and quality standards of embodied conversational agents. *Profess. Cent. Busin. Res.* 2 (2015).

[71] Johannes Kunkel, Tim Donkers, Lisa Michael, Catalin-Mihai Barbu, and Jürgen Ziegler. 2019. Let me explain: Impact of personal and impersonal explanations on trust in recommender systems. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. 1–12.

[72] SeoYoung Lee and Junho Choi. 2017. Enhancing user experience with conversational agent for movie recommendation: Effects of self-disclosure and reciprocity. *Int. J. Hum.-Comput. Stud.* 103 (2017), 95–105.

[73] Yu Liang and Martijn C. Willemsen. 2022. Promoting music exploration through personalized nudging in a genre exploration recommender. *International Journal of Human–Computer Interaction* 39, 7 (2022), 1–24. https://doi.org/10.1080/10447318.2022.2108060

[74] Lizi Liao, Ryuichi Takanobu, Yunshan Ma, Xun Yang, Minlie Huang, and Tat-Seng Chua. 2019. Deep conversational recommender in travel. *ArXiv* abs/1907.00710 (2019).

[75] Paolo Massa and Paolo Avesani. 2007. Trust-aware recommender systems. In *Proceedings of the ACM Conference on Recommender Systems*. 17–24.

[76] Christian Matt, Alexander Benlian, Thomas Hess, and Christian Weiß 2014. Escaping from the Filter Bubble? The effects of novelty and serendipity on users' evaluations of online recommendations. In *ICIS 2014 Proceedings*. (2014). https://doi.org/10.7892/boris.105398

[77] Lorraine Mcginty and Barry Smyth. 2006. Adaptive selection: An analysis of critiquing and preference-based feedback in conversational recommender systems. *Int. J. Electron. Commerce* 11, 2 (2006), 35–57.

[78] Sean M. McNee, John Riedl, and Joseph A. Konstan. 2006. Being accurate is not enough: How accuracy metrics have hurt recommender systems. In *Extended Abstracts on Human Factors in Computing Systems (Montréal, Québec, Canada) (CHI EA'06)*. Association for Computing Machinery, New York, NY, 1097–1101. https://doi.org/10.1145/1125451.1125659

[79] Daniel Mican, Dan-Andrei Sitar-Tăut, and Ovidiu-Ioan Moisescu. 2020. Perceived usefulness: A silver bullet to assure user data availability for online recommendation systems. *Decis. Supp. Syst.* 139 (2020), 113420.

[80] Fedelucio Narducci, Pierpaolo Basile, Marco de Gemmis, Pasquale Lops, and Giovanni Semeraro. 2020. An investigation on the user interaction modes of conversational recommender systems for the music domain. *User Model. User-adapt. Interact.* 30, 2 (2020), 251–284. DOI: https://doi.org/10.1007/s11257-019-09250-7

[81] Fedelucio Narducci, Marco de Gemmis, Pasquale Lops, and Giovanni Semeraro. 2018. Improving the user experience with a conversational recommender system. In *Proceedings of the International Conference of the Italian Association for Artificial Intelligence*. Springer, 528–538.

[82] Tien T. Nguyen, F. Maxwell Harper, Loren Terveen, and Joseph A. Konstan. 2018. User personality and user satisfaction with recommender systems. *Inf. Syst. Front.* 20, 6 (2018), 1173–1189.

[83] Mehrbakhsh Nilashi, Dietmar Jannach, Othman bin Ibrahim, Mohammad Dalvi Esfahani, and Hossein Ahmadi. 2016. Recommendation quality, transparency, and website quality for trust-building in recommendation agents. *Electron. Commerce Res. Applic.* 19 (2016), 70–84.

[84] David Novick and Iván Gris. 2014. Building rapport between human and ECA: A pilot study. In *Proceedings of the International Conference on Human-Computer Interaction*. Springer, 472–480.

[85] Ingrid Nunes and Dietmar Jannach. 2017. A systematic review and taxonomy of explanations in decision support and recommender systems. *User Model. User-adapt. Interact.* 27, 3 (2017), 393–444.

[86] Jum C. Nunnally. 1994. *Psychometric Theory 3E*. Tata McGraw-Hill Education.

[87] John O'Donovan and Barry Smyth. 2005. Trust in recommender systems. In *Proceedings of the 10th International Conference on Intelligent User Interfaces*. 167–174.

[88] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: A method for automatic evaluation of machine translation. In *Proceedings of the Conference of the Association for Computational Linguistics (ACL'02)*. 311–318.

[89] Florian Pecune, Lucile Callebert, and Stacy Marsella. 2020. A socially-aware conversational recommender system for personalized recipe recommendations. In *Proceedings of the 8th International Conference on Human-Agent Interaction.* 78–86.

[90] Florian Pecune, Shruti Murali, Vivian Tsai, Yoichi Matsuyama, and Justine Cassell. 2019. A model of social explanations for a conversational movie recommendation system. In *Proceedings of the 7th International Conference on Human-Agent Interaction.* 135–143.

[91] Robert A. Peterson. 1994. A meta-analysis of Cronbach's coefficient alpha. *J. Consum. Res.* 21, 2 (1994), 381–391.

[92] Bilih Priyogi. 2019. Preference elicitation strategy for conversational recommender system. In *Proceedings of the 12th ACM International Conference on Web Search and Data Mining.* 824–825.

[93] Aleksandra Przegalinska, Leon Ciechanowski, Anna Stroz, Peter Gloor, and Grzegorz Mazurek. 2019. In bot we trust: A new methodology of chatbot performance measures. *Busin. Horiz.* 62, 6 (2019), 785–797.

[94] Pearl Pu and Li Chen. 2006. Trust building with explanation interfaces. In *Proceedings of the 11th International Conference on Intelligent User Interfaces.* 93–100.

[95] Pearl Pu, Li Chen, and Rong Hu. 2011. A user-centric evaluation framework for recommender systems. In *Proceedings of the 5th ACM Conference on Recommender Systems.* Association for Computing Machinery, New York, NY, 157–164. DOI: https://doi.org/10.1145/2043932.2043962

[96] Lingyun Qiu and Izak Benbasat. 2009. Evaluating anthropomorphic product recommendation agents: A social relationship perspective to designing information systems. *J. Manag. Inf. Syst.* 25, 4 (2009), 145–182.

[97] Nicole M. Radziwill and Morgan C. Benton. 2017. Evaluating quality of chatbots and intelligent conversational agents. *arXiv preprint arXiv:1704.04579* (2017).

[98] Jungwook Rhim, Minji Kwak, Yeaeun Gong, and Gahgene Gweon. 2022. Application of humanization to survey chatbots: Change in chatbot perception, interaction experience, and survey data quality. *Comput. Hum. Behav.* 126 (2022), 107034.

[99] Francesco Ricci, Lior Rokach, Bracha Shapira, and Paul B. Kantor. 2015. *Recommender Systems Handbook* (2nd ed.). Springer-Verlag.

[100] Laurel D. Riek, Philip C. Paul, and Peter Robinson. 2010. When my robot smiles at me: Enabling human-robot rapport via real-time head gesture mimicry. *J. Multimodal User Interf.* 3, 1 (2010), 99–108.

[101] Zsófia Ruttkay, Claire Dormann, and Han Noot. 2004. Embodied conversational agents on a common ground. In *From Brows to Trust.* Springer, 27–66.

[102] Alan Said, Ben Fields, Brijnesh J. Jain, and Sahin Albayrak. 2013. User-centric evaluation of a k-furthest neighbor collaborative filtering recommender algorithm. In *Proceedings of the Conference on Computer Supported Cooperative Work.* 1399–1408.

[103] Ryan M. Schuetzler, G. Mark Grimes, and Justin Scott Giboney. 2020. The impact of chatbot conversational skill on engagement and perceived humanness. *J. Manag. Inf. Syst.* 37, 3 (2020), 875–900.

[104] Abigail See, Stephen Roller, Douwe Kiela, and Jason Weston. 2019. What makes a good conversation? How controllable attributes affect human judgments. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT'19).*

[105] Joseph Seering, Michal Luria, Connie Ye, Geoff Kaufman, and Jessica Hammer. 2020. It takes a village: Integrating an adaptive chatbot into an online gaming community. In *Proceedings of the CHI Conference on Human Factors in Computing Systems.* 1–13.

[106] Donghee Shin. 2020. How do users interact with algorithm recommender systems? The interaction of users, algorithms, and performance. *Comput. Hum. Behav.* 109 (2020), 106344.

[107] Donghee Shin. 2022. The perception of humanness in conversational journalism: An algorithmic information-processing perspective. *New Media Societ.* 24, 12 (2022), 2680–2704.

[108] Itamar Simonson. 2005. Determinants of customers' responses to customized offers: Conceptual framework and research propositions. *J. Market.* 69, 1 (2005), 32–45.

[109] Gregory T. Smith, Denis M. McCarthy, and Kristen G. Anderson. 2000. On the sins of short-form development. *Psychol. Assess.* 12, 1 (2000), 102.

[110] Stephen Wonchul Song and Mincheol Shin. 2022. Uncanny valley effects on chatbot trust, purchase intention, and adoption intention in the context of e-commerce: The moderating role of avatar familiarity. *International Journal of Human–Computer Interaction* 0, 0 (2022), 1–16. https://doi.org/10.1080/10447318.2022.2121038

[111] Yueming Sun and Yi Zhang. 2018. Conversational recommender system. In *Proceedings of the 41st International ACM SIGIR Conference on Research & Development in Information Retrieval (SIGIR'18).* 235–244. Retrieved from http://doi.acm.org/10.1145/3209978.3210002

[112] Nina Svenningsson and Montathar Faraon. 2019. Artificial intelligence in conversational agents: A study of factors related to perceived humanness in chatbots. In *Proceedings of the 2nd Artificial Intelligence and Cloud Computing Conference.* 151–161.

[113] Ekaterina Svikhuushina and Pearl Pu. 2021. Key qualities of conversational chatbots—The PEACE model. In *Proceedings of the 26th International Conference on Intelligent User Interfaces.*

[114] Kirsten Swearingen and Rashmi Sinha. 2002. Interaction design for recommender systems. In *Designing Interactive Systems*, Vol. 6. Citeseer, 312–334.

[115] Maria Taramigkou, Efthimios Bothos, Konstantinos Christidis, Dimitris Apostolou, and Gregoris Mentzas. 2013. Escape the bubble: Guided exploration of music preferences for serendipity and novelty. In *Proceedings of the 7th ACM Conference on Recommender Systems.* 335–338.

[116] Linda Tickle-Degnen and Robert Rosenthal. 1990. The nature of rapport and its nonverbal correlates. *Psychol. Inquir.* 1, 4 (1990), 285–293.

[117] Nava Tintarev. 2007. Explanations of recommendations. In *Proceedings of the ACM Conference on Recommender Systems.* 203–206.

[118] Nava Tintarev and Judith Masthoff. 2011. Designing and evaluating explanations for recommender systems. In *Recommender Systems Handbook.* Springer, 479–510.

[119] Nava Tintarev and Judith Masthoff. 2012. Evaluating the effectiveness of explanations for recommender systems. *User Model. User-adapt. Interact.* 22, 4–5 (2012), 399–439.

[120] Thi Ngoc Trang Tran, Alexander Felfernig, and Nava Tintarev. 2021. Humanized recommender systems: State-of-the-art and research issues. *ACM Trans. Interact. Intell. Sys.* 11, 2 (2021), 1–41.

[121] Chun-Hua Tsai and Peter Brusilovsky. 2018. Beyond the ranked list: User-driven exploration and diversification of social recommendation. In *Proceedings of the 23rd International Conference on Intelligent User Interfaces.* 239–250.

[122] Chun-Hua Tsai and Peter Brusilovsky. 2021. The effects of controllability and explainability in a social recommender system. *User Model. User-adapt. Interact.* 31, 3 (2021), 591–627.

[123] Markku Turunen, Jaakko Hakulinen, and Anssi Kainulainen. 2006. Evaluation of a spoken dialogue system with usability tests and long-term pilot studies: Similarities and differences. In *Proceedings of the 9th International Conference on Spoken Language Processing.*

[124] Marilyn Walker, Candace Kamm, and Diane Litman. 2000. Towards developing general models of usability with PARADISE. *Nat. Lang. Eng.* 6, 3–4 (2000), 363–377.

[125] Marilyn Walker, Diane J. Litman, Candace A. Kamm, and Alicia Abella. 1997. PARADISE: A framework for evaluating spoken dialogue agents. In *Proceedings of the 35th Annual Meeting of the Association of Computational Linguistics.* 271–280.

[126] Disen Wang and Hui Fang. 2020. Length adaptive regularization for retrieval-based chatbot models. In *Proceedings of the ACM SIGIR on International Conference on Theory of Information Retrieval.* 113–120.

[127] Lin Wang, Pei-Luen Patrick Rau, Vanessa Evers, Benjamin Krisper Robinson, and Pamela Hinds. 2010. When in Rome: The role of culture & context in adherence to robot recommendations. In *Proceedings of the 5th ACM/IEEE International Conference on Human-Robot Interaction (HRI'10).* IEEE, 359–366.

[128] Yen-Yao Wang, Andy Luse, Anthony M. Townsend, and Brian E. Mennecke. 2015. Understanding the moderating roles of types of recommender systems and products on customer behavioral intention to use recommender systems. *Inf. Syst. e-Busin. Manag.* 13, 4 (2015), 769–799.

[129] Pontus Wärnestål. 2005. User evaluation of a conversational recommender system. In *Proceedings of the 4th Workshop on Knowledge and Reasoning in Practical Dialogue Systems.* Citeseer.

[130] Benjamin Weiss, Ina Wechsung, Christine Kühnel, and Sebastian Möller. 2015. Evaluating embodied conversational agents in multimodal interfaces. *Comput. Cognit. Sci.* 1, 1 (2015), 1–21.

[131] Karin Schermelleh-Engel, Helfried Moosbrugger, and Hans Müller. 2003. Evaluating the Fitof Structural Equation Models: Tests of Significance and Descriptive Goodness-of-Fit Measures. *Methods of Psychological Research* 8, 2 (2003), 23–74. https://doi.org/10.23668/psycharchives.12784

[132] David Westerman, Aaron C. Cross, and Peter G. Lindmark. 2019. I believe in a thing called bot: Perceptions of the humanness of "Chatbots."*Commun. Stud.* 70, 3 (2019), 295–312.

[133] Tiffany A. Whittaker. 2012. Using the modification index and standardized expected parameter change for model modification. *J. Experim. Educ.* 80, 1 (2012), 26–44.

[134] Anders Hauge Wien and Alessandro M. Peluso. 2021. Influence of human versus AI recommenders: The roles of product type and cognitive processes. *J. Busin. Res.* 137 (2021), 13–27.

[135] Wen Wu, Li Chen, and Yu Zhao. 2018. Personalizing recommendation diversity based on user personality. *User Model. User-adapt. Interact.* 28, 3 (2018), 237–276.

[136] Jiyong Zhang and Pearl Pu. 2006. A comparative study of compound critique generation in conversational recommender systems. In *Proceedings of the 4th International Conference on Adaptive Hypermedia and Adaptive Web-based Systems (AH'06).* Springer, 234–243.

[137] Saizheng Zhang, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, and Jason Weston. 2018. Personalizing dialogue agents: I have a dog, do you have pets too? In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*. 2204–2213.

[138] Ran Zhao, Oscar J. Romero, and Alex Rudnicky. 2018. SOGO: A social intelligent negotiation dialogue system. In *Proceedings of the 18th International Conference on Intelligent Virtual Agents*. 239–246.

Received 19 December 2022; revised 27 September 2023; accepted 30 September 2023