Original Paper

Eliciting Reader Preferences for Automated Systematic Gray Literature Reviews: Best-Worst Scaling Study

Zen Yang Ang[1,2], PharmB; Shakirah MdSharif[1]; Wern Han Lim[3]; Shaun Wen Huey Lee[2]

[1]Institute for Health Systems Research, Shah Alam, Selangor, Malaysia
[2]School of Pharmacy, Monash University Malaysia, Subang Jaya, Selangor, Malaysia
[3]School of Information Technology, Monash University Malaysia, Subang Jaya, Selangor, Malaysia

Corresponding Author:

Zen Yang Ang, PharmB
Institute for Health Systems Research
National Institutes of Health, 1, Jalan Setia Murni U13/52, Setia Alam
Shah Alam, Selangor 40170
Malaysia
Phone: 60 134889045
Email: angzenyang@gmail.com

Abstract

Background: Automation can be a useful efficiency tool for saving time in systematic reviews, but a trust gap and a steep learning curve persist concerning its accuracy and ease of use. It remains unknown whether these adoption barriers apply to the more complex process of systematic gray literature reviews (SGLRs), especially given the rapid evolution of large language models.

Objective: This study seeks to identify the preferred features of SGLRs generated using an automation tool (automated SGLRs).

Methods: Participants aged 18 years or older with experience using evidence from systematic reviews were recruited. A cross-sectional online Best-Worst Scaling Questionnaire was distributed through the authors' contacts, professional networks, and social media platforms. Descriptive statistics and multinomial logistic regression were used to estimate the baseline model for preferred features of automated SGLRs, while mixed logit analyses were used to explore subgroup heterogeneity.

Results: A total of 168 respondents, comprising researchers, academicians, health care professionals, health policymakers, and postgraduates, completed the study. Multinomial logistic regression found that the top 3 preferred features for automated SGLRs were (1) high precision and sensitivity comparable to manual methods (best-worst [BW] score 269, and M-logit estimate B=2.26, odds ratio [OR] 9.60, 95% CI 7.72-11.93; P<.001), (2) acknowledgment and detailed explanation of the AI algorithms (BW score: 218, B=2.06, OR 7.85, 95% CI 6.35-9.7; P<.001), and (3) involvement of more than 2 independent human reviewers in the process (BW score: 118, B=1.54, OR 4.68, 95% CI 3.84-5.72; P<.001), when compared to the reference feature, which was the acceptance of automated SGLR output by peer colleagues. Subgroup analysis using a mixed logit model suggested that respondents aged 18 to 29 years exhibited the highest relative utility weights for detailed explanations (B=0.60, OR 1.81, 95% CI 1.29-2.55) and acceptance by everyone (B=0.35, OR 1.41, 95% CI 1.05-1.90) compared to other age groups. In contrast, academicians (B=1.36, OR 3.91, 95% CI 1.51-10.17) might prefer the detailed explanation feature of automated SGLRs more than those in other professions.

Conclusions: The study concludes that automation tool developers should prioritize the accuracy of automated SGLRs, ensure algorithmic transparency, and integrate human-in-the-loop processes, as respondents most strongly preferred these features, which may support user trust.

JMIR Form Res 2026;10:e97813; doi: 10.2196/97813

Keywords: gray literature; systematic review; best-worst scaling; desirable features; machine learning

Introduction

Health policy decision-making operates in a dynamic environment where policymakers require timely, contextually relevant evidence to inform decisions about programs, interventions, and resource allocations [1]. Unlike clinical medicine, where evidence is often derived from controlled studies published in peer-reviewed journals, health system decision-making and policy development are often based on evidence from gray literature (GL), such as real-world implementation data, program evaluations, or government reports, most of which are rarely documented in traditional academic publications [2,3].

As such, many health policy and system research and associated systematic reviews (SRs) often rely on GL to supplement findings, as it provides richer, contextually relevant information. This requires reviewers to focus their efforts on the GL portion of the SR or sometimes even perform an SR exclusively using GL (ie, systematic gray literature review [SGLR]) when medical literature lacks the necessary evidence for stakeholder decision-making [2,3]. However, SGLRs typically need to be completed within a short time frame due to urgent policy decisions. Unlike SRs synthesized from indexed peer-reviewed articles, the unstructured and unindexed nature of GL further impedes the speed of SGLR production by adding approximately 30% to the time required to conduct reviews using peer-reviewed papers alone [3,4]. The advancement of machine learning and AI has led systematic reviewers to seek ways to shorten the time to produce SRs and SGLRs. While research on machine learning–assisted SRs gained traction in the 2010s, research on automation in SGLRs remains rare [5]. To the best of our knowledge, there is no automation tool specifically designed for SGLRs. Consequently, research on perceptions of SGLR automation remains limited.

The rapid advancement of large language models (LLMs) has fundamentally transformed the landscape of SR automation, introducing sophisticated natural language processing capabilities for screening, data extraction, and synthesis tasks [6]. Understanding reader perspectives on AI-generated SGLRs is crucial, as these end users ultimately determine whether evidence presented in journal papers can inform health policy or clinical decision-making in an era when AI increasingly supports evidence synthesis.

A survey by Scott et al [7] found that 80% of respondents agreed that automation saves time; however, about 44% remained neutral on whether automation tools could increase SR accuracy, indicating a trust gap regarding the quality of automation tools. Moreover, the study found that the primary obstacle to adoption was a lack of knowledge about automation tools rather than a lack of tools (51%), with most reviewers self-taught (72%), suggesting that a steep learning curve and poor dissemination further hinder adoption. Nonetheless, these findings were based on conventional SRs and, therefore, their applicability to SGLRs remains unknown. Finally, the velocity of LLM development and model optimization [8-11] reflects a changing landscape in evidence synthesis, including users' perceptions of it. Therefore, it is essential that the perceived barriers and preferred attributes and features regarding SGLRs generated with assistance from automation tools are ranked. Importantly, we wish to understand and quantify the balance and trade-offs among features of SGLRs generated partially or entirely with automation tools (automated SGLRs).

To understand the factors influencing the acceptance and use of technology, we used the unified theory of acceptance and use of technology (UTAUT) model. Specifically, we sought to determine which design features, that is, accuracy, algorithmic transparency, levels of human oversight, and institutional endorsement, are most valued by evidence consumers when deciding whether to trust and use an automated SGLR. We measured the trade-offs among preferred features of automated SGLRs from the perspectives of different readers through a best-worst scaling (BWS) experiment. These findings will guide automation tool developers on which attributes to prioritize when developing such tools in the future.

Methods

Selection of a Survey Tool

Several survey methods are commonly used to assess participants' opinions or preferences: Likert scales, discrete choice experiments (DCEs), and BWS experiments. A Likert-scale survey might not be suitable for identifying the most desired features of these automation tools, as participants tend to rate all features as desired due to social desirability bias [12,13]. This prevents differentiation between the most and least desired features for tool development. On the other hand, instead of choosing the preferred feature for each profile, a DCE requires participants to choose the preferred profile among 2 or more profiles comprising several features, making it unsuitable for measuring the preferred feature.

To overcome these limitations, we used profile (case 2) BWS, which requires respondents to choose the most and least preferred features in a hypothetical task (ie, an automated SGLR profile) [14,15]. Features of this automated SGLR are described by several attributes at different levels across different tasks (Table A2 in Multimedia Appendix 1). These choices were used to produce an estimate of relative attribute-level utility weights, providing explicit trade-off data on preferences for automated SGLRs to guide the development of SGLR automation tools [15,16]. The paired model in BWS assumes that respondents select the best and worst features based on the 2 attribute levels that have the greatest personal utility/preference difference [15].

Establishing the Demographic Questions, Attributes, and Levels of BWS

All 4 domains of the UTAUT framework were included in the questionnaire, namely performance expectancy, effort expectancy, social influence, and facilitating conditions [17]. The initial attributes and levels, as well as the questions

related to participants' demographics and their perceptions toward SGLR automation tools, were adapted from a literature review [7,18-23]. For example, the accuracy of synthesized reviews was adopted as an indicator of automated SGLRs' performance to apply the performance expectancy domain within the UTAUT. To reduce the cognitive burden on respondents, the precision and sensitivity attributes were combined into unified performance levels (90%-100%, 70%, and 50%) across all stages of an automated SGLR (from the initial literature search and screening to data extraction and synthesis), rather than being presented as separate attributes.

Following this, the refined attributes and levels underwent face validation, the results of which are reported in a prior publication [24]. Briefly, 6 experts provided content validity ratings for simplicity, relevance, and clarity and commented on each attribute and level. A scale-level content validity index, based on the average method score of 95.5% [24], fell within an acceptable range [25]. This resulted in 4 attributes (each with 3 mutually exclusive levels) as shown in Table 1.

Table 1. The final set of attributes and their levels used for the best-worst scaling (BWS) choice task[a].

Attribute (UTAUT[b] domain) and level	The resultant attribute's level after the content validity and usability test
Transparency of AI involvement (FC[c]) and explainability of the outputs (EE[d])	
First	AI involvement is acknowledged, but no further explanation is provided.
Second	AI involvement is acknowledged, with a brief explanation of the steps used to produce the SGLR[e].
Third	AI involvement is acknowledged, with a detailed explanation of the AI algorithms, assumptions, and steps used to produce the SGLR.
Accuracy of synthesized reviews (PE[f])	
First	SGLR generated by automation has high (90%-100%) precision and sensitivity compared to the manual method.
Second	SGLR generated by automation has moderate (70%) precision and sensitivity compared to the manual method.
Third	SGLR generated by automation has low (50%) precision and sensitivity compared to the manual method.
Acceptance by the research community (SI[g])	
First	Everyone (including the institutional top management, policymakers, and colleagues) accepts the outputs.
Second	Only the institutional top management and policymakers accept the outputs.
Third	Only accept the outputs.
Human involvement (PE)	
First	No human reviewer is involved (fully automated).
Second	One human reviewer and one robot reviewer are involved (semiautomated).
Third	At least 2 independent human reviewers are involved (semiautomated).

[a]Each attribute within each attribute was arranged during the orthogonal array in Multimedia Appendix 1.
[b]UTAUT: unified theory of acceptance and use of technology.
[c]FC: facilitating condition.
[d]EE: effort expectancy.
[e]SGLR: systematic gray literature review.
[f]PE: performance expectancy.
[g]SI: social influence.

Experimental Design

Adults aged 18 years and older who had read or used evidence from SRs (or other review types) in a health care setting were eligible for this study. As the study primarily aimed to investigate sample-level results, a sample of between 150 and 200 participants was recommended (Multimedia Appendix 1) [26]. Meanwhile, the formulas by Orme [27] and Kyne [28] indicated a sample size of 167 respondents was robust for this study (Multimedia Appendix 1).

For participants to rank all possible combinations of 4 attributes and their 3 attribute levels (the full factorial plan), it would require them to complete 3^4, that is, 81, possible choice tasks, which would likely cause survey fatigue. Therefore, we used the orthogonal main effects plan, which was constructed using the oa.design() function from the R package (see Multimedia Appendix 1 for the orthogonal array), to reduce the number of choice tasks to 9 while ensuring that every attribute level appears with equal frequency across the survey. To avoid unrealistic profiles in which all good and bad levels appear in the same profile or choice task, orthogonal array levels (1, 2, 3) in Multimedia Appendix 1 were mapped nonsequentially across some attribute-level categories rather than strictly adhering to ordinal rankings (Table 1) [29]. This profile construction method allows the estimation of all main effects of the attributes and their levels (Multimedia Appendix 1) [15,29-31]. It is important to note that this fractional factorial design assumes additive utility, which means that the preference for each attribute level can be measured independently, assuming that these attribute levels do not influence each other's preferences when presented together in the same profile.

The final questionnaires comprised two sections: (1) 8 multiple-choice questions regarding participants'

demographics, awareness, practices, and perceptions of automation in SGLRs, adapted from Scott et al [7], and (2) 9 BWS choice tasks/profiles (Multimedia Appendix 1). The order of the choice tasks was randomized to prevent order bias using Qualtrics' built-in block randomization feature [32]. The questionnaires were piloted, and improvements were made based on participants' comments [24].

Data Analysis

Data from participants who completed the full questionnaire were included in the analysis. Incomplete responses were excluded from the analysis. Questions on respondents' demographics, awareness, and perceptions of automated SGLR were analyzed descriptively. The analysis for the 9 BWS questions is described as follows.

Dataset Construction

After the raw dataset from the Qualtrics platform was exported and cleaned, each respondent's best and worst selections across nine tasks were converted into long-format (row-stacked) datasets using the bws.dataset() function in the *support.BWS2* R package (Hokkaido University) [30]. A long-format dataset requires the attribute levels to be coded as 1 for the best attribute level, –1 for the worst attribute level, and 0 for the remaining attribute levels. Each task consists of 4 attribute levels, yielding 12 possible pairs (4 ×3), resulting in 108 rows of data per respondent (9 tasks × 12 pairs). The cleaned dataset was subsequently summarized descriptively [33] and visualized to characterize its skewness and tails (Multimedia Appendix 1).

Baseline Model—Multinomial Logistic Regression

Subsequently, the paired model, which selects the pair with the most significant utility difference as the best and worst attribute levels, was used to measure preferences for the attribute levels among respondents in the BWS survey [15]. The baseline model for the preferred features of automated SGLRs was estimated using multinomial logistic (MNL) regression. The components of the utility function are as follows:

$$V_{ni} = \beta_{brief_exp}X_{ni,brief_exp} + \beta_{detailed_exp}X_{ni,detailed_exp} + \beta_{no_exp}X_{ni,no_exp} + \beta_{mod_prec}X_{ni,mod_prec} + \beta_{high_prec}X_{ni,high_prec} + \beta_{low_prec}X_{ni,low_prec} + \beta_{one_human}X_{ni,one_human} + \beta_{two_human}X_{ni,two_human} + \beta_{no_human}X_{ni,no_human} + \beta_{mgmt_accept}X_{ni,mgmt_accept} + \beta_{every_accept}X_{ni,every_accept},$$

where V_{ni} is the systematic utility of alternative i for individual n, and β is the estimated parameter (utility coefficient) for the attribute level X. To allow for a direct comparison of the relative utility of levels across attributes, rather than just within attributes, the estimates for all attribute levels are calculated relative to one another on a single scale [34]. This was achieved by omitting the attribute level with the lowest utility, that is, the acceptance of automated SGLR output by peer colleagues [34].

Heterogeneity Analysis—Mixed Logit Regression

To explore subgroup heterogeneity, we fitted mixed logit (MXL) models with normal, uniform, and triangular distributions, using identical fixed effects (the same set of regressors) [18,35] and 500 Halton draws [36]. To account for repeated-choice tasks within participants, robust SEs were clustered at the individual-respondent level [37].

Models were compared using the Akaike information criterion (AIC), and the lowest-AIC specification (ie, MXL model with uniform distribution, AIC=6440) was adopted as the base for heterogeneity analyses across participants' current positions and age groups (Multimedia Appendix 1). Subgroup moderation was specified by adding interaction terms between base attribute dummies and demographic factors, which are the age groups and the profession groups [35].

Ethical Considerations

The study was registered in Malaysia's National Medical Research Register (NMRR ID-24-02828-JCW(IIR)) and was approved by the Medical Research Ethics Committee (RNW ID-26-01182-FU8). The survey was conducted in accordance with the CHERRIES (Checklist for Reporting Results of Internet E-Surveys) checklist (Checklist 1). The questionnaire was distributed between November 2024 and May 2025 via QR codes or hyperlinks through the authors' professional networks (eg, social media and colleagues). Whenever possible, it was expanded through snowball sampling and a purposively sampled cohort, using the Qualtrics platform. Respondents who agreed to participate were required to access the survey using a QR code or a hyperlink. All participants were permitted to proceed to answer the questionnaire only after they had read the digital Participant Information Sheet and completed the digital consent form (by selecting the "Yes, I consent" button). No personal information was collected during the process. Participants were not compensated for participating in this study.

Respondent Recruitment

As the sample size was near-minimal (ie, 150), as recommended by Chapman and Rodden [26], and some participants were recruited via nonprobability snowball sampling, both MNL and MXL models served as descriptive and exploratory utility-indexing tools for this specific cohort rather than for formal population-level statistical inference. All calculations were performed using R (version 4.5.1; R Foundation for Statistical Computing) [38].

Results

Practices and Perceptions Regarding the Adoption and Preferred Features of SGLR Produced by Automation Tools

A total of 231 respondents attempted the questionnaire, of whom 168 (72.7%) provided complete responses (Table 2).

Respondents' age ranged between 19 and 61 years, with a mean age of 36.4 (SD 7.9) years. The majority of respondents' current practices were located in Asia (n=136, 81.0%; 123 based in Malaysia), and most comprise researchers, clinicians, or health care professionals, with only 2 being undergraduate students aged 19 and 23 years, who were classified as the "Other" group (Table 2).

Table 2. Respondents' demographics, awareness, practices, and perceptions on automation in systematic gray literature reviews (SGLRs; N=168).

Domain, variable, and description	Values
Demographics	
Age (y), mean age (SD)	36.4 (7.9)
Current country of practice, n (%)	
Asia	136 (81.0)
Europe	15 (8.9)
North America	10 (6.0)
Oceania	2 (1.2)
South America	5 (3.0)
Current position, n (%)	
Researcher	74 (44.0)
Clinician/health care professional	56 (33.3)
PhD candidate/postdoctorate	18 (10.7)
Health care policymaker	8 (4.8)
Academician	7 (4.2)
Other	5 (3.0)
Practices, n (%)	
The primary motivation for reading a systematic review among the participants[a]	
As part of the literature review for a proposal or research project	119 (70.8)
As part of the manuscript/review synthesis process	88 (52.4)
As part of the evidence to inform clinical decision-making	78 (46.4)
As part of the evidence to inform policy decision-making	57 (33.9)
As part of the guideline development process	49 (29.2)
As part of the health technology assessment process	21 (12.5)
No specific reason	5 (3.0)
For learning purposes	1 (0.6)
Type of systematic review read/consumed by the participants[a], n (%)	
Clinical trials/interventions	101 (60.1)
Health policy and systems research	92 (54.8)
Patient preferences and values	42 (25.0)
Economic evaluations	36 (21.4)
Health utilities	30 (17.9)
Diagnostic test accuracy	26 (15.5)
Etiology	15 (8.9)
Prognostic	15 (8.9)
Public health and epidemiology	6 (3.6)
Basic and laboratory sciences	3 (1.8)
Research methodologies	3 (1.8)
Health care quality and management	2 (1.2)
Awareness, n (%)	
Automation tools that participants were aware of[b]	
ChatGPT	125 (74.4)
EndNote, Mendeley, or other referencing managers	122 (72.6)
Review Manager (RevMan)	59 (35.1)
Covidence	45 (26.8)
Rayyan	31 (18.5)

Domain, variable, and description	Values
PICO Portal	13 (7.7)
Abstrackr	12 (7.1)
DistillerSR	10 (6.0)
EPPI-Reviewer	9 (5.4)
RevMan HAL	8 (4.8)
RobotAnalyst	7 (4.2)
RobotReviewer	5 (3.0)
Other systematic review/literature screening tools[c]	9 (5.4)
Other LLM[d] tools	3 (1.8)
None of the above	8 (4.8)
Perceptions on automation in SRs[e], n (%)	
Concerns about AI-generated SGLRs[a]	
Fears that automation may overlook studies or introduce biases	129 (76.8)
Challenges in assessing the quality of included studies without human oversight	127 (75.6)
Concerns about the transparency of the methods and algorithms used in automation	98 (58.3)
Organizations such as Cochrane Reviews still have yet to recommend some of these tools	66 (39.3)
Machine learning automation tools may hallucinate or generate false information	2 (1.2)
Cost and usability concerns	1 (0.6)
Current opinion on AI-generated SGLRs	
Open-minded: I haven't formed an opinion yet, and want to see what the evidence says	77 (45.8)
Optimistic: Systematic gray literature reviews that use SGLR automation tools in their protocol are promising, and I would like to adopt them	46 (27.4)
Excited: Automation is the future of systematic gray literature review	23 (13.7)
Cautious: I am skeptical of systematic gray literature reviews that use SGLR automation tools in their protocol, and think more people need to adopt it first before I can form an opinion	18 (10.7)
Concerned: I am worried about adopting SGLRs that use SGLR automation tools in their protocol, and I think it might be ill-advised by the automation.	4 (2.4)

[a]Multiselect questions.
[b]Other AI/LLM tools: Perplexity, Gemini, and Grok.
[c]Other systematic review/literature screening tools: Colandr, ExaCT, Scispace, mySRL, SRA, GRADEpro, SciSpace, and Python-based crawler.
[d]LLM: large language model.
[e]SR: systematic review.

Most of the respondents (119/168, 70.8%) often used SRs as sources of information for research, and half of them used SRs for clinical decision-making. Most use cases of SRs were related to clinical trials and interventions as well as health policy and systems. Less than a quarter of respondents used SRs addressing topics such as patient preferences and values, economic evaluation, and health utilities.

About three-quarters of participants were aware of ChatGPT (n=125, 74.4%; OpenAI), an LLM, and reference managers (n=122, 72.6%; eg, EndNote, Clarivate Analytics and Mendeley, Elsevier). Awareness was lower for systematic review software, including RevMan (n=59, 35.1%; The Cochrane Collaboration), Covidence (n=45, 26.8%; Veritas Health Innovation), and Rayyan (n=31, 18.5%; Rayyan Systems Inc). While most respondents were aware of ChatGPT, their concerns about automated tools operating without human oversight primarily centered on methodological integrity. Respondents frequently cited the risk of relevant studies or research being overlooked (n=129, 76.8%) and the potential for bias (n=127, 75.6%). Furthermore, many respondents (n=98, 58.3%) expressed concern about the lack of transparency in automation methods, followed by concern about the lack of endorsement of their use by reputable organizations (n=66, 39.3%; Table 2).

When exploring general views on SGLRs produced by automation tools, attitudes were largely cautiously optimistic, that is, 46% (n=77) of respondents maintained an open-minded position, awaiting further empirical evidence. Meanwhile, 41.1% (n=69) reported being optimistic or excited about the technology and were considering its adoption (Table 2).

When compared to acceptance by peer colleagues only, an MNL analysis of the most preferred attribute levels for automated SGLRs indicates that high precision was the dominant factor, achieving the highest values across all measured metrics (best-worst [BW] score: 269, and M-logit estimate B=2.26, OR 9.60, 95% CI 7.72-11.93; P<.001), followed by the inclusion of a detailed explanation of algorithms in automated SGLRs (BW score: 218, B=2.06, OR 7.85, 95% CI 6.35-9.7; P<.001). Regarding workflow, the involvement of 2 or more humans (BW score: 118, B=1.54, OR 4.68, 95% CI 3.84-5.72; P<.001) was preferred over a collaborative approach involving one human and one robot,

though both were viewed positively. Conversely, the analysis identified the least desired features. The lowest preference scores were associated with low colleague acceptance, a total lack of human involvement, and the absence of an explanation (Table 3).

Table 3. Best and worst counts and multinomial logistic regression (MNL) estimates across attribute levels[a].

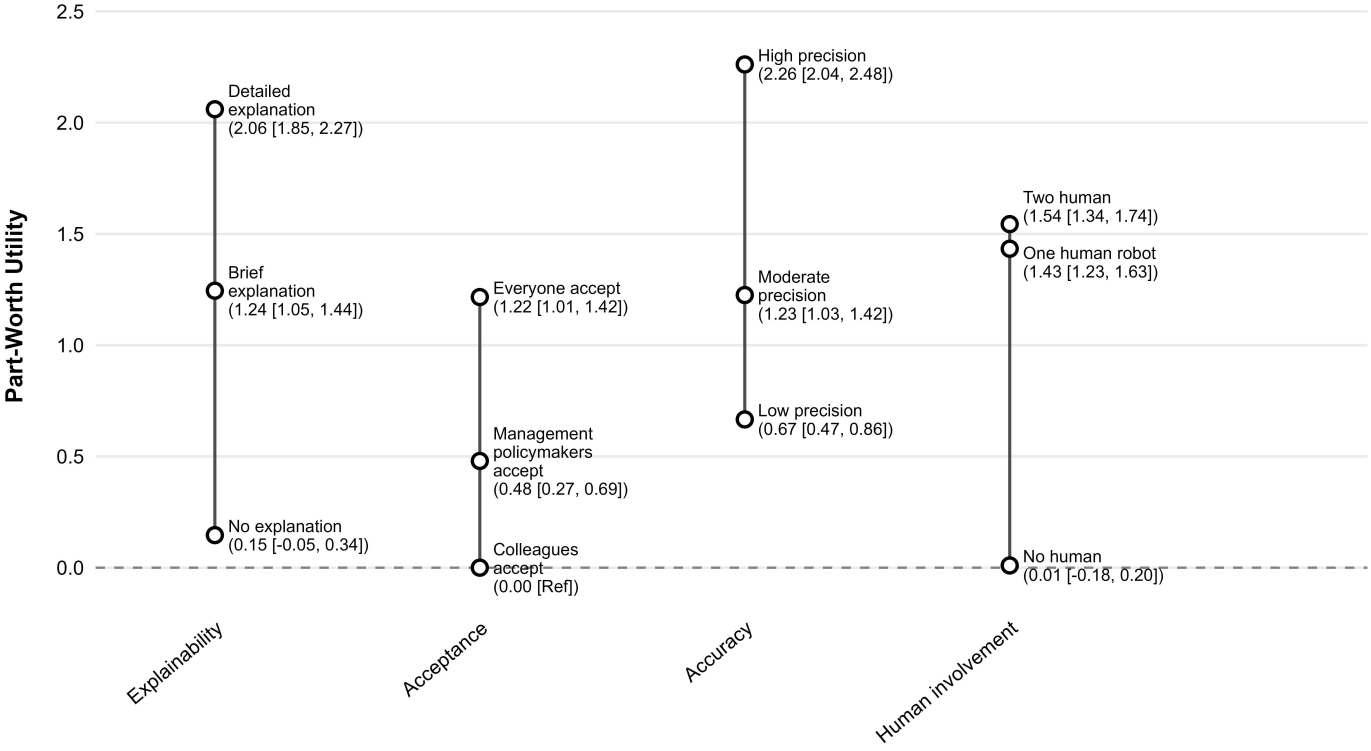
Attribute	Attribute level	Count approach					Modeling approach		
		Best count	Worst count	Difference of BW[b] count (BW score)	SD of BW score	Scaled BW score	Estimate, *B* (SE)	Odds ratio (95% CI)	*P* value
Accuracy	High precision	293	24	269	1.282	100	2.262 (0.111)	9.601 (7.72-11.93)	<.001
Transparency and explainability	Detailed explanation	259	41	218	1.554	71.933	2.061 (0.108)	7.851 (6.35-9.7)	<.001
Human involvement	≥2 humans are involved	179	61	118	1.491	49.027	1.544 (0.102)	4.684 (3.84-5.72)	<.001
Human involvement	One human and one robot are involved	171	85	86	1.435	40.594	1.434 (0.102)	4.194 (3.43-5.13)	<.001
Transparency and explainability	Brief explanation	148	96	52	1.578	35.536	1.245 (0.101)	3.473 (2.85-4.24)	<.001
Accuracy	Moderate precision	113	45	68	1.175	45.353	1.225 (0.102)	3.405 (2.79-4.16)	<.001
Acceptance	Everyone acceptance	143	141	2	1.741	28.822	1.216 (0.105)	3.375 (2.75-4.15)	<.001
Accuracy	Low precision	74	123	–49	1.461	22.199	0.666 (0.099)	1.947 (1.60-2.37)	<.001
Acceptance	Top management/policymakers acceptance	44	198	–154	1.466	13.492	0.480 (0.106)	1.616 (1.31-1.99)	<.001
Transparency and explainability	No explanation	42	217	–175	1.522	12.591	0.146 (0.098)	1.158 (0.95-1.4)	.14
Human involvement	No human involved	19	220	–201	1.346	8.411	0.010 (0.098)	1.010 (0.83-1.22)	.92
Acceptance	Colleagues' acceptance (reference)	27	261	–234	1.238	9.205	0	1	—[c]

[a]The table is sorted in descending order by the estimates from multinomial logistic regression.
[b]BW: best-worst.
[c]Not applicable.

(page number not for citation purposes)

The largest separation on the utility scale illustrates the attribute levels that had the most significant impact on participant preference [34]. Considerable increases in preference for automated SGLRs were observed with improvements in selected features, including increasing precision from moderate to high, providing a brief explanation of AI involvement rather than none, and using a collaborative workflow with one human and one robot reviewer rather than no human involvement, for example, moving from "moderate precision" to "high precision," from "no explanation" to a "brief explanation" of AI involvement, from "no human reviewer is involved" to a collaborative "one human reviewer and one robot involvement" workflow. In contrast, increasing the acceptance attribute from "colleagues only" to "everyone" resulted in a comparatively smaller change in preference. The utility gained from full acceptance was similar only to the second level of other attributes, such as brief explanations, moderate precision, or one human-one robot involvement. This indicates that participants prioritize higher levels of explainability and accuracy over maximizing widespread acceptance (Figure 1).

Figure 1. Utility weights given to the various attribute levels of an automated systematic gray literature review (SGLR) from the readers' perspective. Point estimates and 95% CIs from the multinomial logit model are shown in parentheses (estimate, 95% CI).



Heterogeneity

The MXL model using a uniform distribution provided the best fit for the data (AIC=6440; Table B1 in Multimedia Appendix 1), outperforming alternative distributional specifications in the heterogeneity analysis. For the reference subgroup in the MXL interaction model (ie, the baseline age group and position category), features such as brief explanations, acceptance by everyone, the involvement of at least 2 humans, and human-robot collaboration were also preferred but to a lesser degree. Conversely, respondents consistently disliked moderate precision and acceptance by top management or policymakers (Table B2 in Multimedia Appendix 1).

Random parameter estimates provided evidence of heterogeneity between subgroups, with academicians showing a much higher preference for detailed explanations (B=1.36, OR 3.91, 95% CI 1.51-10.17) than those in other professions. At the same time, policymakers and clinicians exhibited inconsistent preference patterns. Similarly, respondents aged 18 to 29 years exhibited the highest relative utility weights for detailed explanations (B=0.60, OR 1.81, 95% CI 1.29-2.55) and acceptance by everyone (B=0.35, OR 1.41, 95% CI 1.05-1.90) compared to those in other age groups. The youngest respondents were more likely to prioritize everyone's acceptance than respondents in other age groups (Table B2 in Multimedia Appendix 1).

Discussion

Principal Findings

The advent of generative AI has created new opportunities to accelerate evidence generation for health systems and policy research by automating SGLRs with AI tools. While the automation of conventional SRs was made possible a decade ago, the automation of SGLRs and their perception remains relatively unknown. This study found that the SGLR readers prioritized 3 critical attribute levels for automated

SGLRs: high precision and sensitivity as the most critical features, followed by a detailed explanation of the AI algorithms used, and the involvement of at least 2 independent human reviewers, when compared to acceptance only by peer colleagues. The demand for such standards is most likely due to the unique challenges of reviewing GL, which is often unindexed, unstructured, and lengthy, making it significantly harder to search, comprehend, and extract relevant data [2,3]. Furthermore, the heterogeneity of GL limits the use of standardized quality appraisal tools. These findings suggest that overcoming these inherent challenges by enhancing accuracy and explainability aligns closely with user preferences for automated SGLRs. Our findings may also apply to conventional SRs as the inclusion of GL is recommended in SRs [39].

Comparison With Prior Work

Findings of this study concur with previous research on SRs for guideline development by Arno et al [40] and Arno [41], which found that the quality of guidelines was identified as a nonnegotiable attribute, even in the face of time or cost savings afforded by automation. Furthermore, many respondents referred to SRs for research or clinical decision-making, where there is a strong preference for high accuracy, reflecting concerns that imprecision in automated SGLRs could directly compromise patient safety and research outcomes.

More than three-quarters and half of respondents were concerned about the accuracy and the transparency of automated SGLRs. The BWS findings, which indicated that respondents prioritized high precision and sensitivity comparable to manual methods, detailed explainability, and a human-in-the-loop approach as the most preferred features for automated SGLRs, further support this. Similarly, Arno et al [40] and Arno [41] reported that transparency and the ability to evaluate research methods independently are critical for establishing the trust required in health guideline development. This concurs with the Joint AI Methods Group's position statement on the use of AI in evidence synthesis where methodological transparency is one of the key emphases in the statement [42]. Despite these concerns regarding accuracy and transparency, 86% (n=146) of respondents remained open-minded or optimistic about the technology, suggesting a favorable inclination toward automated SGLRs when these preferred features are incorporated.

The introduction of a single human reviewer in the automated process accounts for the substantial increase in utility from 0.01 (no human reviewer is involved) to 1.43 (Figure 1). This highlights the importance of human involvement from the reader's perspective. The mere difference in the part-worth utilities of these 2 features (at least 2 human involvement and the human-robot involvement) suggests that incorporating even a single human reviewer produces a considerable increase in relative preference compared to a fully automated process. Arno et al [40] and Arno [41] also emphasized the importance of human involvement, though the authors did not specify the required number of human reviewers. Similarly, the Joint AI Methods Group's position statement sets human oversight as an indispensable component in SR [42].

Even though acceptance by everyone was the best of the 3 social acceptance features in the survey, it ranked seventh out of the 12 features. This indicates that the social acceptance domain had less influence on readers' preferences. We postulated that this outcome might be due to the trust readers place in the journal's established peer-review process before publication. Therefore, they are more willing to trust the output of an automated tool, provided its results are accurate with the AI algorithm clearly explained [43]. From the perspective of the UTAUT theory, social influence is an important factor during the initial phase of technology adoption, although its importance might gradually lessen as users establish familiarity with the tool's performance.

Heterogeneity analysis suggests that academicians might have high expectations for the explainability of the automated SGLR. This concurs with the principles in SR, where transparency and methodological rigor cannot be compromised [44,45]. While midcareer professionals (aged 30-39 y) demonstrated the lowest marginal requirement for detailed explanation of AI algorithms, both the age groups of 18 to 29 years and 40 to 49 years required significantly higher levels of explanation. The authors postulate that those in their early careers may use the detailed outputs of automation tools as a learning resource to understand the logic of evidence synthesis. This aligns with the UTAUT, which indicates that exfort expectancy and the need for facilitating conditions are higher for less-experienced users [17]. In contrast, participants in the age group of 40 to 49 years demonstrated a heightened requirement for methodological transparency. We hypothesized that trusting AI, especially in older age groups, is not an inherent trait but is built through cognitive evaluation [46], as cognitive trust is established when a user can rationally assess the system's competence and reliability [46]. With detailed explanations, the age group of 40 to 49 years can understand how these tools work. Creating this mental model allows these experienced researchers to calibrate their trust toward the tools, ensuring that their reliance on the tool is based on its actual performance rather than on unfounded trust [46].

Limitations

One key strength of the study is its geographically diverse sample spanning nearly every continent, thereby providing a broader perspective for readers. However, most respondents are from a single country (ie, Malaysia), with some recruited through nonprobability snowball sampling. Consequently, this might introduce a confounding effect due to respondent characteristics (eg, institutional culture or prior familiarity with AI). Therefore, the findings, especially the heterogeneity analysis, which was also derived from a relatively small sample size, should be interpreted as exploratory preference patterns within this specific cohort rather than a formal, population-level statistical inference [18].

Even though the data analyzed were derived from 168 respondents who completed the whole questionnaire, and the attribute-level balance of the orthogonal main effects plan array was preserved, the estimated utilities are based on an additive assumption: a respondent's evaluation of one item (eg, precision level) was assumed not to be influenced by the co-occurrence of another item (eg, level of human oversight) in the same choice task. Therefore, the estimated utilities represent simplified average preference weights. These unmodeled interactions cannot be ruled out and might affect the main effects or utility weights.

Furthermore, the BWS score showed a skewed distribution (Multimedia Appendix 1). To address this, we selected the lowest-AIC uniform-distribution variant for the MXL regression, which provides the best balance between goodness-of-fit to the data and parsimony, thereby minimizing the risk of overfitting. The survey achieved a 72.7% completion rate, consistent with the 44.0% to 78.5% completion rates reported in a meta-analysis of DCEs [47]. The high cognitive burden and lack of incentive in completing the surveys associated with BWS and DCEs may explain the noncompletion [48].

To reduce the cognitive burden on respondents, the authors combined sensitivity and precision to represent the accuracy of automated SGLRs. Consequently, a strong relative utility weight observed for the highest accuracy level (90%-100%) may partly reflect a general preference for high-performance anchors rather than an explicit trade-off with other attribute levels. Nonetheless, these estimates suggest that higher accuracy might be associated with a greater preference for an automated SGLR among respondents in this sample, a feature that developers of SGLR automation tools might focus on during development. Based on the limitations described above, a future multicountry study is needed to directly evaluate multiattribute interactions and capture the views of different regions.

Conclusions

This study demonstrates that respondents' stated preferences for automated SGLRs follow clear patterns. Highly valued features included high precision and sensitivity comparable to manual methods, transparent and detailed explanations of the AI algorithms, and human involvement through at least 2 human reviewers, suggesting that readers preferred verifiable accuracy and accountability. This is consistent with the findings of existing studies on SRs. Therefore, to foster trust and encourage reader adoption, developers of automation tools may emphasize features such as accuracy, algorithmic transparency, and human-in-the-loop processes in development workflows.

Acknowledgments

The authors thank the Director General of Health, Malaysia, for permission to publish this paper. The authors also express their gratitude to Fun Weng Hong, Sondi Sararaks, Hazwa Harith, and Melody Soong Yin Yin for their support throughout the process.

Generative AI Use

Grammarly was used for editing, grammar checking, and refining the final draft of the manuscript.

Funding

The authors declare that no funding has been obtained for this study.

Data Availability

Data are provided as supplementary material. Other data will be made available upon reasonable request.

Authors' Contributions

Conceptualization: ZYA, WHL, SWHL
Data curation: ZYA, SM, SWHL
Formal analysis: ZYA, WHL, SWHL
Investigation: ZYA, SWHL
Methodology: ZYA, WHL, SWHL
Supervision: WHL
Validation: ZYA, WHL, SWHL
Visualization: ZYA, SM, SWHL
Writing – original draft: ZYA, WHL, SWHL
Writing – review and editing: ZYA, SM, WHL, SWHL

Conflicts of Interest

None declared.

Multimedia Appendix 1

Supplementary methods, sample size calculation, statistical models, score distributions, and experimental design.
[DOCX File (Microsoft Word File), 218 KB-Multimedia Appendix 1]

Checklist 1

CHERRIES checklist.
[DOCX File (Microsoft Word File), 27 KB-Checklist 1]

References

1. Tricco AC, Antony J, Zarin W, et al. A scoping review of rapid review methods. BMC Med. Sep 16, 2015;13(1):224. [doi: 10.1186/s12916-015-0465-6] [Medline: 26377409]
2. Godin K, Stapleton J, Kirkpatrick SI, Hanning RM, Leatherdale ST. Applying systematic review search methods to the grey literature: a case study examining guidelines for school-based breakfast programs in Canada. Syst Rev. Oct 22, 2015;4(1):138. [doi: 10.1186/s13643-015-0125-0] [Medline: 26494010]
3. Yang AZ, Hong FW, et al. Searching for grey literature for rapid compilation of Malaysia's health systems response during the COVID-19 pandemic: challenges and lessons learnt. J Health Transl Med. 2023;26(2):81-88. [doi: 10.22452/jummec.vol26no2.9]
4. Saleh AA, Ratajeski MA, Bertolet M. Grey literature searching for health sciences systematic reviews: a prospective study of time spent and resources utilized. Evid Based Libr Inf Pract. 2014;9(3):28-50. [doi: 10.18438/b8dw3k] [Medline: 25914722]
5. de la Torre-López J, Ramírez A, Romero JR. Artificial intelligence to automate the systematic review of scientific literature. Computing. Oct 2023;105(10):2171-2194. [doi: 10.1007/s00607-023-01181-x]
6. Khraisha Q, Put S, Kappenberg J, Warraitch A, Hadfield K. Can large language models replace humans in systematic reviews? Evaluating GPT-4's efficacy in screening and extracting data from peer-reviewed and grey literature in multiple languages. Res Synth Methods. Jul 2024;15(4):616-626. [doi: 10.1002/jrsm.1715] [Medline: 38484744]
7. Scott AM, Forbes C, Clark J, Carter M, Glasziou P, Munn Z. Systematic review automation tools improve efficiency but lack of knowledge impedes their adoption: a survey. J Clin Epidemiol. Oct 2021;138:80-94. [doi: 10.1016/j.jclinepi.2021.06.030] [Medline: 34242757]
8. Text generation. OpenAI Developers. URL: https://platform.openai.com/docs/guides/text?api-mode=chat#prompt-engineering [Accessed 2026-10-22]
9. Uploading files and audio to ChatGPT. OpenAI. URL: https://help.openai.com/en/articles/8555545-file-uploads-faq? [Accessed 2026-10-22]
10. Syriani E, David I, Kumar G. Screening articles for systematic reviews with ChatGPT. J Comput Lang. Aug 2024;80:101287. [doi: 10.1016/j.cola.2024.101287]
11. Wang S, Scells H, Zhuang S, Potthast M, Koopman B, Zuccon G. Zero-shot generative large language models for systematic review screening automation. In: Goharian N, Tonellotto N, He Y, Lipani A, McDonald G, Macdonald C, Ounis I, editors. Advances in Information Retrieval 46th European Conference on Information Retrieval, ECIR 2024, Glasgow, UK, March 24–28, 2024, Proceedings, Part I. Springer; 2024:403-420. [doi: 10.1007/978-3-031-56027-9_25]
12. Mühlbacher AC, Zweifel P, Kaczynski A, Johnson FR. Experimental measurement of preferences in health care using best-worst scaling (BWS): theoretical and statistical issues. Health Econ Rev. Dec 2016;6(1):5. [doi: 10.1186/s13561-015-0077-z] [Medline: 26822869]
13. Bäckström M, Björklund F. Social desirability in personality inventories: symptoms, diagnosis and prescribed cure. Scand J Psychol. Apr 2013;54(2):152-159. [doi: 10.1111/sjop.12015] [Medline: 23252410]
14. Finn A, Louviere JJ. Determining the appropriate response to evidence of public concern: the case of food safety. J Public Policy Mark. Sep 1992;11(2):12-25. [doi: 10.1177/074391569201100202]
15. Flynn TN, Louviere JJ, Peters TJ, Coast J. Best-worst scaling: what it can do for health care research and how to do it. J Health Econ. Jan 2007;26(1):171-189. [doi: 10.1016/j.jhealeco.2006.04.002] [Medline: 16707175]
16. Günther OH, Kürstein B, Riedel-Heller SG, König HH. The role of monetary and nonmonetary incentives on the choice of practice establishment: a stated preference study of young physicians in Germany. Health Serv Res. Feb 2010;45(1):212-229. [doi: 10.1111/j.1475-6773.2009.01045.x] [Medline: 19780851]
17. Venkatesh V, Morris MG, Davis GB, Davis FD. User acceptance of information technology: toward a unified view. MIS Q. Sep 1, 2003;27(3):425-478. [doi: 10.2307/30036540]
18. Xu RH, Zhou L, Wong ELY, Wang D, Xiang GC, Xu C. A best-worst scaling survey of medical students' perspective on implementing shared decision-making in China. BMC Med Educ. Dec 2, 2020;20(1):486. [doi: 10.1186/s12909-020-02406-9] [Medline: 33267806]
19. Wranik WD, Székely RR, Mayer S, Hiligsmann M, Cheung KL. The most important facilitators and barriers to the use of health technology assessment in Canada: a best-worst scaling approach. J Med Econ. 2021;24(1):846-856. [doi: 10.1080/13696998.2021.1946326] [Medline: 34162282]

20. de Andrés-Sánchez J, Gené-Albesa J. Explaining policyholders' chatbot acceptance with an unified technology acceptance and use of technology-based model. J Theor Appl Electron Commer Res. 2023;18(3):1217-1237. [doi: 10.3390/jtaer18030062]
21. Norton JD, Zeng C, Bayliss EA, et al. Ethical aspects of physician decision-making for deprescribing among older adults with dementia. JAMA Netw Open. Oct 2, 2023;6(10):e2336728. [doi: 10.1001/jamanetworkopen.2023.36728] [Medline: 37787993]
22. Pasco PMD. Physician user perspectives in the practice of telemedicine in the Philippines. J Int Soc Telemed eHealth. 2016;4:1-9. URL: https://journals.ukzn.ac.za/index.php/JISfTeH/article/view/172 [Accessed 2026-10-22]
23. Andrews JE, Ward H, Yoon J. UTAUT as a model for understanding intention to adopt AI and related technologies among librarians. J Acad Librarianship. Dec 2021;47(6):102437. [doi: 10.1016/j.acalib.2021.102437]
24. Ang ZY, Shakirah MS, Melody SYY, Fun WH, Shaun LWH. Developing and validating instruments to identify preferred features of systematic grey literature review automation tools. Presented at: The 12th National Public Health Conference in conjunction with 26th NIH Scientific Conference Embracing the New Era: Advancing Public Health Through AI and Digitalisation; Jul 8-10, 2025; The Everly Putrajaya, Putrajaya, Malaysia. URL: https://www.pppkam.org.my/wp-content/uploads/2025/07/abstract-free-paper.pdf? [Accessed 2026-10-22]
25. Yusoff MSB. ABC of Content Validation and Content Validity Index Calculation. EIMJ. 2019;11(2):49-54. URL: https://eduimed.usm.my/EIMJvol11no2.html [doi: 10.21315/eimj2019.11.2.6]
26. Chapman C, Rodden K. Quantitative User Experience Research: Informing Product Decisions by Understanding Users at Scale. Apress; 2023. ISBN: 9781484292679
27. Orme B. MaxDiff analysis: simple counting, individual-level logit, and HB. Sawtooth Software, Inc; 2009. URL: https://content.sawtoothsoftware.com/assets/8e69929d-a089-4b93-a9f3-d4c64d156642? [Accessed 2026-10-22]
28. Kyne D. What is best-worst scaling? (examples, methods, free tools). OpinionX. May 9, 2024. URL: https://www.opinionx.co/blog/best-worst-scaling [Accessed 2026-10-22]
29. Louviere JJ, Flynn TN, Marley AAJ. Best-Worst Scaling: Theory, Methods and Applications. Cambridge University Press; 2015. ISBN: 9781107043152
30. Aizaki H, Fogarty J. An R package and tutorial for case 2 best–worst scaling. J Choice Model. Sep 2019;32. [doi: 10.1016/j.jocm.2019.100171]
31. Aizaki H, Fogarty J. Chapter 4: an illustrative example of case 2 best–worst scaling. Non-market valuation with R. URL: https://nmvr.skr.jp/04-bws2.html? [Accessed 2026-10-23]
32. Ho TQA, Engel L, Le LKD, et al. Discrete choice experiment versus best–worst scaling: an empirical comparison in eliciting young people's preferences for web-based mental health interventions. Patient. Jul 2025;18(4):357-372. [doi: 10.1007/s40271-025-00739-5] [Medline: 40314883]
33. Uy EJB, Bautista DC, Xin X, Cheung YB, Thio ST, Thumboo J. Using best-worst scaling choice experiments to elicit the most important domains of health for health-related quality of life in Singapore. PLoS One. 2018;13(2):e0189687. [doi: 10.1371/journal.pone.0189687] [Medline: 29420564]
34. Franco MR, Howard K, Sherrington C, et al. Eliciting older people's preferences for exercise programs: a best-worst scaling choice experiment. J Physiother. Jan 2015;61(1):34-41. [doi: 10.1016/j.jphys.2014.11.001] [Medline: 25499647]
35. Hensher DA, Greene WH. The mixed logit model: the state of practice. Transportation (Amst). May 2003;30(2):133-176. [doi: 10.1023/A:1022558715350]
36. Cheung KL, Mayer S, Simon J, et al. Comparison of statistical analysis methods for object case best-worst scaling. J Med Econ. Jun 2019;22(6):509-515. [doi: 10.1080/13696998.2018.1553781] [Medline: 30482068]
37. Flynn TN, Louviere JJ, Peters TJ, Coast J. Estimating preferences for a dermatology consultation using best-worst scaling: comparison of various methods of analysis. BMC Med Res Methodol. Nov 18, 2008;8:76. [doi: 10.1186/1471-2288-8-76] [Medline: 19017376]
38. The R Project for Statistical Computing. URL: http://r-project.org/? [Accessed 2026-10-22]
39. Higgins J, Thomas J, Chandler J, et al, editors. Cochrane Handbook for Systematic Reviews of Interventions, Version 6.3. Cochrane; 2022. URL: https://www.cochrane.org/authors/handbooks-and-manuals/handbook/current [Accessed 2026-10-22]
40. Arno A, Elliott J, Wallace B, Turner T, Thomas J. The views of health guideline developers on the use of automation in health evidence synthesis. Syst Rev. Jan 8, 2021;10(1):16. [doi: 10.1186/s13643-020-01569-2] [Medline: 33419479]
41. Arno AD. The adoption and effectiveness of automation in health evidence synthesis [Dissertation]. University College London; 2022. URL: https://discovery.ucl.ac.uk/id/eprint/10141555/1/Arno%20-%20Automation%20in%20Health%20Evidence%20Synthesis.pdf [Accessed 2026-10-22]

42. Flemyng E, Noel-Storr A, Macura B, et al. Position statement on artificial intelligence (AI) use in evidence synthesis across Cochrane, the Campbell Collaboration, JBI, and the Collaboration for Environmental Evidence 2025. Environ Evid. Oct 31, 2025;14(1):20. [doi: 10.1186/s13750-025-00374-5] [Medline: 41174758]
43. Renner M, Lins S, Söllner M, Thiebes S, Sunyaev A. Achieving trustworthy artificial intelligence: multi-source trust transfer in artificial intelligence-capable technology. Presented at: 42nd International Conference on Information Systems (ICIS 2021); Dec 12-15, 2021; Austin, Texas, United States. URL: https://publikationen.bibliothek.kit.edu/1000138300 [Accessed 2026-10-23]
44. O'Mara-Eves A, Thomas J, McNaught J, Miwa M, Ananiadou S. Using text mining for study identification in systematic reviews: a systematic review of current approaches. Syst Rev. Jan 14, 2015;4(1):5. [doi: 10.1186/2046-4053-4-5] [Medline: 25588314]
45. Page MJ, McKenzie JE, Bossuyt PM, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. BMJ. Mar 29, 2021;372:n71. [doi: 10.1136/bmj.n71] [Medline: 33782057]
46. Glikson E, Woolley AW. Human trust in artificial intelligence: review of empirical research. Acad Manag Ann. Jul 2020;14(2):627-660. [doi: 10.5465/annals.2018.0057]
47. Watson V, Becker F, de Bekker-Grob E. Discrete choice experiment response rates: a meta-analysis. Health Econ. Jun 2017;26(6):810-817. [doi: 10.1002/hec.3354] [Medline: 27122445]
48. Hess S, Hensher DA, Daly A. Not bored yet—revisiting respondent fatigue in stated choice experiments. Transp Res Part A Policy Pract. Mar 2012;46(3):626-644. [doi: 10.1016/j.tra.2011.11.008]

Abbreviations

AIC: Akaike information criterion
BW: best-worst
BWS: best-worst scaling
CHERRIES: Checklist for Reporting Results of Internet E-Surveys
DCE: discrete choice experiment
EE: effort expectancy
GL: gray literature
LLM: large language model
MNL: multinomial logistic
MXL: mixed logit
NMRR: National Medical Research Register
OR: odds ratio
SGLR: systematic gray literature review
SR: systematic review
UTAUT: unified theory of acceptance and use of technology

Edited by Amaryllis Mavragani; peer-reviewed by Marcin Kozak, Taofeek Yusuff; submitted 10.Apr.2026; final revised version received 18.Aug.2026; accepted 07.Sep.2026; published 16.Nov.2026

Please cite as:
Ang ZY, MdSharif S, Lim WH, Lee SWH
Eliciting Reader Preferences for Automated Systematic Gray Literature Reviews: Best-Worst Scaling Study
JMIR Form Res 2026;10:e97813
URL: https://formative.jmir.org/2026/1/e97813
doi: 10.2196/97813

© Zen Yang Ang, Shakirah MdSharif, Wern Han Lim, Shaun Wen Huey Lee. Originally published in JMIR Formative Research (https://formative.jmir.org), 16.Nov.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.