Mitigation Approach for Certain Biases in Survey Scale Input within AI Control and Decision Systems

Steve Chan
VTIRL, VT/DE-STEA
Orlando, USA
stevechan@ieee.org

***Abstract*—For the realm of Artificial Intelligence (AI), Transparency, Explainability, and Accountability (TEA) have become a critical triumvirate. Central to TEA are reliability and validity. However, classical assessment instruments may fall short in this regard. The purpose of this paper was to formulate a prospective TEA-centric AI-facilitated assessment instrument that better contends with quantitative fallacy and human bias issues that are often not addressed. Various hybridizations of Semantic Difference Scales and Object Measures (OMs) were experimented with for the purposes of enhancing the involved Multi-Attribute Decision-Making/Multi-Objective Decision-Making Subjective Measure/OM counterpoisings, particularly on the Human-Informed Repertoire of Experience side of the AI Control and Decision System — which impacts the machine-centric Decision Engineering/Decision-Making side — thereby effectuating a more robust System TEA (STEA) paradigm.**

Keywords—decision engineering, control and decision systems, multi-criteria decision-making, socio-technical system, artificial intelligence, machine learning, Likert scale, non-Likert scales, transparency, explainability, accountability, justification logic.

I. Introduction

Artificial Intelligence (AI) for Control and Decision Systems (CDS) has undergone intense scrutinization in recent years. AI-related bias has become a significant issue, and movements, such as Explainability in AI (XAI), as well as standards, such as the IEEE Standard for Transparency of Autonomous Systems (P7001) along with the Association for Computing Machinery (ACM) "Principles for Algorithmic Transparency and Accountability" have risen to prominence. The U.S. National Institute of Standards and Technology (NIST) has advanced matters, such as via Special Publication 1270: "Towards a Standard for Identifying and Managing Bias in Artificial Intelligence" (which is linked to NIST's work on its AI Risk Management Framework). As Winfield alludes, this era of System Transparency Specification (STS) and System Transparency Assessment (STA) have segued to a heightened expectation of epistemic Transparency, Explainability, and Accountability (TEA), for System TEA (STEA), in the architecture of AI CDS (and a movement away from "black box" constructs) [1]. Winfield also notes that "P7001 recognises that AI technology cannot be separated from the larger Socio-Technical System [STS] of which it is a component" [1]. The STS rubric, which spans "*humans*, technology, and the environs" has experienced a surge in *humans*-in-the-loop "to mitigate against 'non-perfect' algorithmic and AI/Machine Learning (ML)-centric paradigms, particularly for 'high-stakes tasks'" [2]. Along this Human-Computer Interaction (HCI) vein, NIST reports that AI bias can stem from this STS arena of "*human* biases and systemic, institutional biases as well" [3]. The *humans*-in-the-loop paradigm is particularly crucial for certain borderline areas of "Decision Engineering/Decision-Making (DE/DM)." At this nexus, the AI CDS is neither in a state of Lower Ambiguity High Uncertainty (LAHU) (i.e., given a prior repertoire of experience, the tolerance for uncertainty is higher) nor High Ambiguity Low Uncertainty (HALU) (i.e., given the lack of prior experience, the tolerance for uncertainty is lower and more data is desired to decrease uncertainty) [4]. For the apriori experiences, the AI CDS may consider prior table-topping, simulations, and serious gaming exercises as constituting part of the repertoire of experience, so the subjective human feedback provided during After Action Reviews (AARs) can potentially profoundly impact the involved AI CDS DE/DM.

On the machine-centric DE/DM side, the AI CDS will likely involve a Multi-Criteria Decision-Making (MCDM) element, which Fattoruso and others note as consisting of Multi-Attribute Decision-Making (MADM) and Multi-Objective Decision-Making (MODM) [5]. While MODM typically contends with a myriad of objectives ("often conflicting") and seeks to ascertain an optimal solution set among "undetermined continuous alternatives," MADM typically contends with a single objective and organizes/sorts/ranks to determine the optimal solution among "a finite set of discrete alternatives" [4]. Ideally, the MADM/MODM construct is architected/designed to robustly leverage various Subjective Measures (SMs)/Objective Measures (OMs), as an overall well-counterpoised MADM/MODM SM/OM construct can better address and enhance the robustness of STEA.

On the Human-Informed Repertoire of Experience (HIRE) side, Choi has noted that "unanticipated communication barriers" (e.g., a disjointedness "between the investigator and respondents" due to varied interpretations) can yield inaccurate results, and Westland has noted exemplars, such as that of "bias and informativeness of Likert metrics have been the center of recent" dialectic [6][7]. From a TEA vantage point, Taherdoost has noted that the Likert "scale validity may be difficult to demonstrate and there is a lack of reproducibility" [8]; along this vein, on the Likert scale validity front, Magdolen had found, in a prior study, that "participants less often chose the extreme answer categories" and central tendency bias may ensue [9]. McLeod notes that "prior research has shown that using Likert scales can be problematic," particularly for AI

CDS, and biases include, among others, social desirability bias, acquiescence bias, etc. [10]. Yet, as "high-stakes tasks" typically involve table-topping exercises, simulations, and serious gaming, which are often assessed by AAR-related Likert scale surveys, the carefully counterpoised MADM/MODM SM/OM [AI CDS] architectural construct on the machine-centric DE/DM side is likely susceptible to the prospective bias issues on the STS HIRE side [11][12]. Accordingly, this paper presents a prospective mitigation approach for certain biases on the STS HIRE side of AI CDS to advance STEA matters. To make the content more accessible for the reader, a glossary of terms is provided in Table I below.

TABLE I. GLOSSARY OF TERMS FOR KEY CONCEPTS

Acronym	*Terms for associated Key Concepts*
AI	Artificial Intelligence
CDS	Control and Decision System
CFA	Confirmatory Factory Analysis
CFI	Comparative Fix Index
CLT	Central Limit Theorem
CRITIC	CRiteria Importance through Intercriteria Correlation
CN	Critical N
DE	Decision Engineering
DEA	Data Envelopment Analysis
DES	Dissociative Experiences Scale
DM	Decision-Making
EFA	Exploratory Factor Analysis
EM	Expectation Maximization
ESM	Experience-Sampling Method
ETAS	Expanded Tellegen Absorption Scale
FQ	Flow Questionnaire
FSS	Flow State Scale; also Flow Short Scale
GFPS	General Flow Proness Scale
HALU	High Ambiguity Low Uncertainty
HCI	Human-Computer Interaction
HIRE	Human-Informed Repertoire of Experience
ITQ	Immersive Tendencies Questionnaire
KMT	Krejcie and Morgan Table
LAHU	Lower Ambiguity High Uncertainty
MADM	Multi-Attribute Decision-Making
MCDM	Multi-Criteria Decision-Making
MCSDS	Marlowe-Crowne Social Desirability Scale
MODM	Multi-Objective Decision-Making
MODTAS	Modified-Tellegen Absorption Scale
OM	Objective Measure
PANAS	Positive and Negative Affect Schedule
PROMETHEE	Preference Ranking Organization Method for Enrichment Evaluation
PXI	Player Experience Inventory
RMSEA	Root Mean Square Error of Approximation
RMSR	Root Mean Square Residual
RWS	Real-World Scenario
SDB	Social Desirability Bias
SDS	Semantic Difference Scale or Semantic Differential Scale
SEM	Structural Equation Modeling
SM	Subjective Measure
SOTA	State-of-the-Art
STA	System Transparency Assessment
STEA	System TEA
STS	Socio-Technical System; also, System Transparency Specification
TAS	Tellegen Absorption Scale
TEA	Transparency, Explainability, and Accountability
TLI	Tucker-Lewis Index
TOPSIS	Technique of Order Preference by Similarity to an Ideal Solution
VAP	Value-Added Proposition
VAS	Visual-Analogue Scale
XAI	Explainability in AI

Section I delineates the impetus of the paper. Section II provides pertinent background information. Section III presents theoretical foundations, experimental considerations, and the setup. Section IV concludes with a preliminary discussion, reflections, and posited future work.

II. BACKGROUND

On the STS HIRE side, the involved table-topping exercises, simulations, and serious gaming are informed by the dimensions of, among others, immersion, flow, and absorption. These are described in subsections A1 through A3, and classical "gold standard" scales are listed as well [13].

A. Dimensional Aspects

1) Immersion

Wirth describes immersion as the "capacity to induce the feeling of actually being a part of" the emulated environment, and Banos describes it as "the experience of becoming engaged in the game-playing experience while retaining some awareness of one's surroundings" [13]. Witner's Immersive Tendencies Questionnaire (ITQ), which is oftentimes implemented as a Likert scale, is widely utilized as an instrument [14].

2) Flow

Brockmyer describes flow as "being one with the activity," while Moneta notes that "having a specific goal and an immediate performance feedback structure increase the likelihood of flow, and being in a flow state seems to enhance learning" [13]. The Csikszentmihalyi Flow Questionnaire (FQ) as well as the Larson-Csikszentmihalyi Experience-Sampling Method (ESM) tend to utilize a Likert scale or a Visual-Analogue Scale (VAS) [15]; along this vein, other derivative flow questionnaires, such as the General Flow Proness Scale (GFPS), Flow State Scale (FSS), Flow Short Scale (also FSS), etc. also tend to utilize the Likert scale [16][17][18].

3) Absorption

Irwin describes [psychological] absorption as constituting the degree of "engagement in the present experience" [13]. Absorption can fulfill two functional roles. In the first case, it can be either an enhancement or detractor to the flow state; for example, Funk notes that negative affect includes "anxiety and frustration," and Brockmyer notes that "negative affective states are antithetical to the flow state" [13]. In the second case, absorption may be shaped by motivation, which can involve self and/or peers. Lucas notes that "game-playing competition with peers is often a primary motivator," and the degree of motivation can intensify/de-intensify absorption [13]; in this case, absorption can be considered as distinct and disparate from flow [13]. Classically, absorption is often gauged via, among other scales, the Tellegen Absorption Scale (TAS) (which is a "34-item true-false measure" Likert scale), the Modified-Tellegen Absorption Scale (MODTAS) (which is a 25-item Likert scale), the Positive and Negative Affect Schedule (PANAS) (a 20-item Likert-based measure,

comprised of a 10-item Negative Affect Scale and a 10-item Positive Affect Scale), the Expanded Tellegen Absorption Scale (ETAS) (which is a 20-item Likert scale), etc. [13][19][20][21]. Also, given that psychological absorption is a form of nonpathological dissociation, the Dissociative Experiences Scale (DES) as well as the DES-II version, which are often implemented as a Likert scale, are also leveraged [22].

B. Biases

As can be gleaned, the "gold standards" utilized by the cited dimensions, among others, center upon the potentially problematic (from the TEA vantage point) Likert scale. These include the following biases among others:

1) Central Tendency Bias

Wang, Liu, and others note that Real-World Scenario (RWS) data "often exhibit a long-tailed" distribution [23][24]. However, the increasing utilization of *humans*-in-the-loop and *human* input via modalities, such as those previously described involving Likert scales, has resulted in the phenomenon of central tendency bias, according to Sabolic, Akbari, and others [25][26]. This can potentially obscure/obfuscate the long-tail realities of the involved RWS and, potentially, adversely impact the "E" of TEA.

2) Social Desirability Bias (SDB)

Teh notes that SDB "is one of the oldest forms of response bias" [27]. In essence, respondents often provide "socially acceptable" responses, such as in accordance with prevailing societal trends. SDB can be somewhat accounted for via the Marlowe-Crowne Social Desirability Scale (MCSDS), which can measure SDB [28]; however, this may be sub-optimal.

3) Acquiescence Bias

Friborg and others note that "Likert-based response formats may introduce acquiescence bias" [29]. To mitigate against this phenomenon, the Likert responses can be "transformed into negations" [29]; yet these "transformations may introduce errors, as negatives of positive constructs may appear contra-intuitive" (with the ensuing unintended consequences) [29].

C. Mitigation Enhancements

Accordingly, Friborg notes that a semantic differential response format (a.k.a., Semantic Differential Scale or SDS) may serve as "an alternative to negations for reducing the acquiescence bias" [29]. Affirming this, McLeod posits that the "fast form" approach to the SDS is psychometrically equivalent to the Likert scale and that the use of a SDS can alleviate certain of the Likert scale biases [10]. The SDS can also be conjoined with certain advantages of the slider rating scale; in particular, Xiao notes that the slider "can lower the tendency of neutral responses compared to a Likert ([e.g.,] five-point) with explicit midpoints" [30]. This is a key point, as the theoretical Value-Added Proposition (VAP) of a bipolar Likert scale is to capture the bipolar result (e.g., extremely positive/negative); however, the propensity for central tendency bias tends to negate this VAP. Cape asserts that "slider scales afford a more accurate reflection of respondents' true attitudes and opinions" [31]. Wang has conducted explorations along this vein by combining, among other elements, SDS and Technique of Order Preference by Similarity to an Ideal Solution (TOPSIS) implementations [32][33]. This paper extrapolates upon that type of exploration to better address Choi's concern that Likert scales can yield inaccurate results. Ji et al. put it nicely: "AI alignment aims to make AI systems behave in line with human intentions and values. "As AI systems grow more capable, so do risks from misalignment" [34]. This paper addresses the alignment issue to assist with STEA by better capturing the human intent.

Apart from the AI alignment issues, there are also prospective quantitative fallacy issues; according to NIST, "the idea that quantitative measures are better and more objective than other observations is known as the McNamara Fallacy," which should be considered alongside other issues, such as that of quantitative exactitude (e.g., measurement accuracy, precision of the tools, etc.) [3]. Taherdoost had also raised the issue of "lack of reproducibility"/repeatability for the Likert scale, thereby also illuminating the issue of quantitative accuracy, which encompasses precision (i.e., "a measure of repeatability"), accuracy (i.e., "a measure of nearness to the true value" or "the nearness [of an estimate] to the true value"), and correctness (i.e., "the desired precision" [35][36][37][38]. NIST Special Publication 1270 also notes that "processes for documenting potential sources of human bias are essential but often overlooked elements for characterizing AI model Transparency[,…] Explainability," and Accountability (i.e., TEA) "in addition to addressing AI bias and fairness" at the system level [3]. For this reason, mitigation advancements, to address this described challenge, are crucial. Perhaps, Boateng said it best: "As science advances and novel research questions are put forth, new scales become necessary" [39]. Of the 3 core scale evaluation tests (e.g., validity, reliability, dimensionality), first, "validity refers to a lack of systematic error" and "precision is a lack of random error" [40]; Taherdoost has pointed out that, with regards to TEA, validity "may be difficult to demonstrate" for the Likert scale [8]. Second, according to Louangrath, who cites Glasser, "reliability is defined as consistency in results from repeated measurements" [41]. Louangrath's findings were that non-Likert type scales are more reliable; in essence, non-Likert type scales in the form of (0,1,2,3) tend to have the highest level of reliability and validity, and Likert-type scales in the form of (1,2,3,4,5),(1,2,3,4,5,6,7) as well as (1,2,3,4,5,6,7,8,9,10) tend to fail both reliability and validity tests [41]. As "reliability and validity are the cornerstones of scientific research," Louangrath points out that Likert-type scales can, potentially, fail 2 of the 3 core scale evaluation tests [41].

III. Experimentation

A. Theoretical Foundations

It was previously noted that AI CDS may consider prior table-topping, simulations, and serious gaming exercises (collectively, "gaming") as constituting apriori experience. Almac notes that "a player's subjective interaction with a digital game is referred to as player experience" [42]. A prevalent player experience survey is that of Abeele's Player Experience Inventory (PXI), which delineates between the "functional and psychosocial consequences of gaming" [42]. Functional aspects include "audio-visual appeal" (as well as "ease of control, progress feedback" … "clarity of goals, and

challenge") [42]. Psychosocial aspects include "immersion" (as well as "mastery, curiosity"…"autonomy, and meaning") [42]. The "audio-visual aesthetics of games can be defined as the music, sound effects and graphics, and animation that are used" [42]. While some, such as Jarvinen, emphasize the "graphical elements" (e.g., with the categories of "photorealism, caricaturism, and abstractionism"), axiomatically, music and sound effects also have a substantial role in the immersion facet of PXI, as noted by Sepulveda [43]. Others, such as Qin, have opined that the backstory/narrative can also greatly contribute to immersion by successfully breeding "empathy, and familiarity" [44]. Ermi concurs with the significance of empathy, particularly "emphathis[ing] with the characters" [45]. For the involved experimentation, photorealism, music/sound effects, and backstory/narrative were underscored; these qualities well lend to immersion, which when combined with motivation, impact "presence" and flow [46]; along this vein, "presence" marks the "early stage of…engagement" [47]. Phan asserts that presence, engagement, immersion, flow, and absorption, among other dimensions, are key facets for PXI [48].

B. Experimental Considerations and Rationale

The underlying intent of this paper was to formulate and validate a prospective new TEA-centric AI-facilitated assessment instrument for the scenario described in Sections IIA and IIIA. To optimize the robustness and quality of the instrument, current State-of-the-Art (SOTA) practices were adhered to; this included: (1) item pool generation (to formulate the "item pool for the scale"), (2) MADM/MODM SM/OM review of the item pool, (3) scale development with participants from three groups of players (i.e., Starting Trainee/Novice, Midcore/Intermediate, Hardcore/Expert), (4) Exploratory Factor Analysis (EFA) (a certain sample set of players were surveyed, wherein EFA honed in on the factors to be utilized and reduced the size of the item pool), and (5) Confirmatory Factory Analysis (CFA) (an additional independent sample set of players were surveyed, via an EFA-based revised scale, so as to "further validate the scale"); incidentally, CFA is a form of Structural Equation Modeling (SEM). Then, from the ecosystem, a myriad of survey instruments as well as prevailing heuristics were considered, as exampled in Table II below.

TABLE II. MYRIAD OF SURVEY INSTRUMENTS AND PREVAILING HEURISTICS CONSIDERED

Developers/Authors	***Instrument/Design Heuristics***
Papaloukas [49]	Heuristics for New Genre Games
Senap [50]	Heuristics Evaluation for Educational Games
Leutner, Raghavan [51][52]	Cognitive Absorption Scale (via gamification of cognitive tests)
Guo, Qin [53][54]	Player Immersion in Computer Game Narrative Questionnaire
Fu, Sweetzser [55][56]	Game Flow
Desurvire, Li [57][58]	Game Approachability/Playability Principles
Rukmono, Oprins [59][60]	Instructional Game Evaluation Framework

The mean hours of engagement (per week) for the serious gaming exercises were stratified as follows: (1) less than 1 hour, (2) 1 to 3 hours, (3) 4 to 9 hours, (4) 10 to 19 hours, (5) 20 to 35 hours, and (6) more than 35 hours. Non-responses or "Not Applicable (N/A)" responses were equated as missing values. Expectation Maximization (EM) was utilized to contend with these missing values, as EM has a certain robustness, especially when missing values > 10% and are not necessarily random [48]. Kyriazos and DeVellis note that "the factor pattern developed by a large-scale factor analysis is probably more stable than that based on a small sample size," and a sample size >= 300 *for field testing* (e.g., in December 2023, the sample size set was still at 52, so additional time into 2024 was needed to increase the sample size appropriately) was deemed to be "good enough" and stable (e.g., sample sizes < 300 "tend to diverge") [61][62]. The generally recognized gold standard of Thompson's *Exploratory and Confirmatory Factor Analysis* opines that any sample size greater than 300 is considered satisfactory, such as for "instrument development and validation" (e.g., the questionnaire is a prototypical instrument in "survey research") [63][64]. Comrey and Lee as well as Clark affirm (e.g., for the substantive portion of factor analysis paradigms, a sample size of 300 should suffice, and key results derived from the Rasch Model include findings that "more misfitting items were identified in the simulations with a sample size of 400"), and their asserted exemplar sample sizes for factor analyses (as well as those from some other investigators) are presented in Table III (modeled after Gunawan's work) [64][65].

TABLE III. EXEMPLAR SAMPLE SIZE RECOMMENDATIONS

Recommended Sample Sizes	***Investigators***
200-300	Guadagnoli and Velicer [66]
>=200	Hair et al. [67]
>=200-1000	Nevitt and Hancock [68]
300	Clark and Watson [69] VanVoorhis and Morgan [70] White [71]
200->=1000, wherein "50=very poor, 100-poor, 200=fair, **300=good**, 500=very good, >=1,000=excellent" [64][72]	Comrey and Lee [72]
400	Aleamoni [73]

The asserted minimum exemplar sample sizes, such as 300 (bolded), are aimed at achieving less biased posits. This can be furthered via exemplar recommendations regarding "item to response ratio (*p: N*)" (i.e., on average, the number of responses per item), as shown in Table IV, and "estimated parameter to sample size ratio (*q: N*)," as shown in Table V [64].

TABLE IV. EXEMPLAR ITEM-TO-RESPONSE RATIO RECOMMENDATIONS

Recommended Item-to-Response Ratios	***Investigators***
1:3 to 1:20	Mundfrom [74]; Hair et al. [75]
1:5	Suhr [76]
1:10	Boateng [39]; Nunnally [77]; White [71]
1:15	Hair et al. [75]
1:30	Memon et al. [78]; Brown and Greene [79]; Forsberg and Rantala [80]
1:70 to 1:119	Maxwell [81]

Referencing Table IV, Memon and Roscoe note that 1:30 (bolded) "should be considered the minimum" whether for the

group or its constituent subgroups (by way of example, for a 10 item study, using the 1:30 ratio, the sample size would equate to 300, which is nearing — for the purpose herein — the gold standard of 384 put forth by the Krejcie and Morgan Table or KMT) [78]. The rationale is predicated upon the Central Limit Theorem (CLT), wherein it is posited that "the distribution of the sample means approaches (or tends to approach) a normal distribution as the sample size increases" [78].

TABLE V. EXEMPLAR PARAMETER-TO-SAMPLE SIZE RECOMMENDATIONS

Recommended Parameter-to-Sample Size Ratios	***Investigators***
1:5 to 1:20	Klein [82]
1:10	Gunawan [64]; Jackson [83]
1:10 to 1:20	Jackson [83]
1:15 to 1:20	Hair et al. [75]
1:15 to 1:30	Osborne [84]; Pehazur [85]

Referencing Table V, Memon notes that Hair et al.'s recommendation of 1:15 to 1:20 (bolded) is preferred [78]; this equates to 15 to 20 "observations per independent variable" [78]. After all, ratios tending toward the 1:5 tend to overfit, and those tending toward the 1:20 or 1:30 tend to be more apropos. Then, extrapolating upon the notions of Jarvinen, Qin/Ermi, and Phan, nine factors derived from the EFA phase were utilized for the posited full model: (1) Visual Aesthetics (5 items), (2) Audio Aesthetics (5 items), (3) Narratives (8 items), (4) Resonance of Goals/Objectives (7 items), (5) Ambiguity/Uncertainty (e.g., Mystery) (8 items), (6) Core Mechanics (7 items), (7), Creative Freedom (7 items), (8) Engrossment (to encompass the PXI gamut) (7 items), and (9) Satisfaction (7 items). It could be discerned, via the preliminary results from Root Mean Square Error of Approximation (RMSEA), Root Mean Square Residual (RMSR), and Critical N (CN) (i.e., key "goodness-of-fit" indices), that the 9-factor model exhibited a "good fit" with the incoming observed sample set data. Alternative models (e.g., 8 factor, 7 factor, 6 factor, etc.), wherein (1) and (2) were combined for the 8 factor, (1), (2), and (3) were combined for the 7 factor, (6) and (7) were combined as well as (1), (2), and (3) for the 6 factor, etc., had neither lower RMSEA/RMSR values nor higher CN values (i.e., Hoelter's notion of the largest sample size for acceptance of the posit that a model is correct). As others have found, "the 9-factor model had the 'best fit'," as exhibited with its higher Comparative Fix Index (CFI) and Tucker-Lewis Index (TLI) values [48]. Consistent with Phan's findings, the involved chi-square difference tests yielded "statistically significant results" between the posited 9 factor and each of the alternative models; hence, the 9-factor model seemed the most apropos [48].

C. *Experimental Setup*

With the aforementioned factors established, further explorations, atop existing work in the literature (e.g., Wang's SDS/TOPSIS hybridization), could be conducted. OM methods that lent well to AI facilitation (e.g., in furtherance of an enhancement to the MADM/MODM SM/OM counterpoising, such as pertains to item pool generation, etc.) were experimented with. Exemplar OMs as well as approaches that allow for both SM/OM, including the referenced TOPSIS, are shown in Table VI.

TABLE VI. EXEMPLAR OMS FOR SDS HYBRIDIZATION

#	***Methods***	***MADM Method***	***OM-centric***
1	Technique of Order Preference by Similarity to an Ideal Solution (TOPSIS)	MADM [86]	OM [87]
2	CRiteria Importance through Intercriteria Correlation (CRITIC)	MADM [87]	OM [88]
3	Data Envelopment Analysis (DEA)	MADM [87]	OM [87]
4	Preference Ranking Organization Method for Enrichment Evaluation (PROMETHEE) (e.g., I and II)	MADM [89]	SM/OM [89][90]

As part of the experimentation, a bespoke experimental architectural construct was explored. Previously, the construct utilized was that of a rudimentary MADM/MODM SM/OM counterpoising, as shown on the left side of Fig. 1. The experimentation of this paper focused upon testing for the $MADM_{is}$ to $MADM_{ss}$, wherein *is* equates to "input set" and *ss* equates to "solution set" and the "MODM 'solution set' ($MODM_{ss}$) facilitates the MADM 'input set' ($MADM_{is}$) to MADM 'output solution set' ($MADM_{ss}$) progression," with OM Methods#1-4 (denoted in orange), as shown on the right side of Fig. 1 [91].

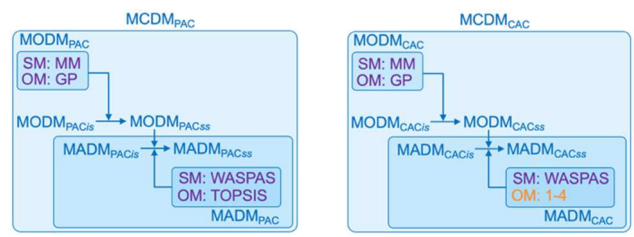

Fig. 1. Prior Architectural Construct (PAC) with TOPSIS (left) and Current Architectural Construct (CAC) with explicit OM Methods#1-4 (right)

Taking the three metrics of Performance (P), Consistency ("which is a useful indicator" for stability as well as "the underlying convergence paradigm"), and Flexibility ("for adaptation, hybridization, etc."), various comparative evaluations were conducted, and the interim findings are delineated in Table VII.

TABLE VII. OM METHOD HYBRIDIZATION WITH SDS BENCHMARKING

MODM SM	P	C	F
MM			

MADM SM	P	C	F
WASPAS			

MODM OM	P	C	F
GP			

MADM OM	P	C	F
TOPSIS			
CRITIC			
DEA			
PROMETHEE I, II			

Green denotes satisfactory performance, red denotes unsatisfactory performance, and orange signifies that the results were all within one standard deviation (and comparable relative to each other) [91]; the hybridizations yielded some interesting results, which are presented in the following Section IV.

IV. CONCLUSION

A. Preliminary Discussion

The ability to contend with both discrete and continuous paradigms is central. At the heart of DE/DM is MCDM and its constituent components of MODM ("undetermined continuous alternatives") and MADM (discrete alternatives"), which require continuous as well as discrete evaluations, respectively. As Likert scales do not contain "0," discrete testing is not possible (only "continuous distribution testing is); in contrast, non-Likert scales do indeed contain "0" and are amenable to both discrete and continuous distribution testing. Hence, non-Likert scale datasets lend toward a higher F and tend to segue to more robust results; accordingly, non-Likert scales can be more amenable to higher nuance and insight. An extension of Wang's experimentation shows that there are other MADM OM/OM hybridizations, apart from TOPSIS, that yield higher P, C, and F. For example, the hybrid usage of CRITIC and PROMETHEE II lends to an amalgam of much greater efficacy (the aggregate is in the green, as reflected in Table VII).

As Wang points out (particularly given just some of the biases discussed in Section IIB), hybridization of SDS with OMs (TOPSIS for Wang's case) is crucial. In fact, beyond the discussed biases (e.g., central tendency bias), there are compounding effects, such as via the reinforcing effects of "straightlining," which Magdolen describes as "selecting the same response category for each item," and it is a phenomenon, wherein respondents "easily run through the given statements without careful reading and understanding of the questioned content, especially when using online surveys" [9]. Similar aggravating effects include speeding, wherein the respondent is not taking sufficient time "to read, understand and answer the question" [9]. Moreover, a principal disadvantage of implementation via a Likert scale is that of "unequal responses" (i.e., "the distance between responses is not measurable;" hence, it may not be straightforward to rank the responses") [92]. Louangrath's experimentation notes that "non-Likert scale (0,1,2,3) is 92% reliable while the Likert-type of scale had 90, 89, and 88% reliability" [41]. Louangrath further asserts that validity testing showed that his involved "non-Likert scale was 93% reliable, while the Likert-type scale had 89, 61, and 57% precision" [41]. Hence, for this paper, a non-Likert approach, predicated upon Louangrath's findings, was adopted.

For the higher efficacy case, such as wherein CRITIC ascertains weights and PROMETHEE II undertakes the ranking, the inherent Likert disadvantages can be somewhat mitigated (e.g., the TEA-centric counterpoised MADM/MODM SM/OM mitigation approach exhibits efficacy). Planned future work includes further apropos quantitative experimentation; for example, the current use of Cronbach's alpha should be re-evaluated, as suggested by Louangrath: "the function of the Cronbach's alpha is to verify the consistency of answers among respondents. It does not verify whether the instrument itself is reliable" [41][93]. This and other points will be further investigated for future work in a qualitative as well as quantitative fashion. In addition, as certain heuristics were utilized for the forays herein, optimality considerations, such as with regards to sample size, item-to-response ratio, and parameter-to-sample size ratio, will be further explored. Exemplar heuristics included: (1) Tabachnick's generalization that "larger samples result in increased power" (i.e., the higher likelihood of a "correct decision"), and (2) Mundfrom's finding that "when the variables-to-factors ratio exceeds 6, the minimum sample size begins to stabilize regardless of the number of factors or the level of communality" [70][74][94]. Moreover, there is ongoing dialectic in the literature regarding validity testing. Researchers differ on the issue on whether to use one or both forms of factor analysis —EFA as well as CFA — and in what order. As the impetus herein was to develop a new measure, CFA followed EFA. Also, the notion of whether the exemplar sample size suffices for "significance tests," "likelihood ratio tests," and the like is better addressed by power analysis, and this will be furthered in future work.

B. Reflections

Overall, in the case of this paper, the impetus was to examine some of the reliability and validity deficiencies of classical assessment instruments (as well as various AI alignment issues) and to explore a novel TEA-centric AI-facilitated assessment instrument that better mitigates against quantitative fallacy as well as bias issues for an enhanced "high-stakes" STS-aware, STEA-centric AI CDS DE/DM architecture. Various SDS (e.g., the "fast form" approach to the SDS is psychometrically equivalent to the Likert scale, and the SDS use can alleviate certain of the Likert scale biases) and OM combinatorials were experimented with, as enhancing the MADM/MODM SM/OM counterpoising (e.g., for both the machine-centric DE/DM and HIRE sides of the AI CDS) segues to a more robust STEA paradigm. An experimental framework was presented, and various limitations as well as future directions for research were discussed.

REFERENCES

[1] A. Winfield et al, "IEEE P7001: A Proposed Standard on Transparency," *Front. Robot. AI*, vol. 8, pp. 1-11, Jul. 2021.

[2] B. Velden, "Explainable AI: Current Status and Future Potential," *Eur. Radiol.,* vol. 34, pp. 1187-1189, Aug. 2023.

[3] R. Schwartz, A. Viassilev, K. Greene, L. Perine, A. Burt, and P. Hall, "Towards a Standard for Identifying and Managing Bias in Artificial Intelligence," *NIST Special Publication 1270*, 2022.

[4] S. Chan, "AI-Facilitated Dynamic Threshold-Tuning for a Maritime Domain Awareness Module," *2024 IEEE Int. Conf. on Ind. 4.0, Artif. Intell., and Commun. Technol. (IAICT)*, Bali, Indonesia, Jul. 2024, pp. 192-198.

[5] G. Fattoruso, "Multi-Criteria Decision Making in Production Fields: A Structured Content Analysis and Implications for Practice," *J. Risk Financial Manag.*, vol. 15, pp. 1-21, Sep. 2022.

[6] B. Choi and A. Pak, "A Catalog of Biases in Questionnaires," *Prev Chronic Dis*., vol. 2, pp. 1-13, Dec 2004.

[7] J. Westland, "Information loss and bias in likert survey responses," *PLOS One,* vol. 17, pp. 1-17, Jul. 2022.

[8] H. Taherdoost, "What Is the Best Response Scale for Survey and Questionnaire Design; Review of Different Lengths of Rating Scale/ Attitude Scale/Likert Scale," *Int J. of Acad. Res. in Manage.* (IJARM), vol. 8, pp.1-12, Mar. 2019.

[9] M. Magdolen, S. Behrn, Jan Vallee, B. Chlond, and P. Vortisch, "Response bias in Likert-style psychological items – an example from a large-scale travel survey in China," *Transp. Res. Procedia*, vol. 76, pp. 349-360, Apr. 2024.

[10] A. McLeod, S. Pipp, and J. Wong, "Revisiting the Likert scale: can the fast form approach improve survey research?" *Int. J. of Behav. Accounting and Finance*, vol. 2, pp. 310-327, Jan. 2011.

[11] L. Alsulimani, "The feasibility of simulation-based high-stakes assessment in emergency medicine settings: A scoping review," *J. Educ. Health Promot*., vol. 10, pp. 1-10, Nov. 2021.

[12] J. Casar and M. Krátký, "Modelling and simulation of Ground Based Air Defence War Games," *2022 20th Int. Conf. on Mechatronics - Mechatronika (ME)*, Pilsen, Czech Republic, Dec. 2022, pp. 1-6.

[13] J. Brockmyer, C. Fox, K. Curtiss, E. McBroom, K. Burkhart, and J. Pidruzny, "The development of the Game Engagement Questionnaire: A measure of engagement in video game-playing," *J. of Exp. Social Psychol.*, vol. 45, Jul. 2009, pp. 624-634.

[14] A. Krassmann, M. Melo, B. Peixoto, D. Pinto, M. Bessa, and M. Bercht, "Learning in Virtual Reality: Investigating the Effects of Immersive Tendencies and Sense of Presence," in *Lecture Notes in Computer Science*: *Virtual, Augmented and Mixed Reality. Industrial and Everyday Life Applications*, vol. 12191, pp. 270-286, Jul. 2020.

[15] J. Fritz, M. Piccirillo, and L. Bringmann, "So You Want to Do ESM? 10 Essential Topics for Implementing the Experience-Sampling Method," *Adv. in Methods and Pract. in Psychol. Sci*., vol. 7, pp. 1-27, Sep. 2024.

[16] M. Elnes and H. Sigmundsson, "The General Flow Proneness Scale: Aspects of Reliability and Validity of a New 13-Item Scale Assessing Flow*," SAGE Open*, vol. 13, pp. 1-9, Feb. 2023.

[17] P. Hancock, A. Kaplan, J. Cruit, G. Hancock, K. MacArthur, and J. Szalma, "A meta-analysis of flow effects and the perception of time," *Acta Psychologica*, vol. 198, pp. 102836, Jul. 2019.

[18] M. Laakasuo, J. Palomaki, S. Abuhamdeh, O. Lappi, and B. Cowley, "Psychometric analysis of the flow short scale translated to Finnish," *Sci. Rep*., vol. 12., pp. 1-11, Nov. 2022.

[19] M. Girn, R. Spreng, D. Margulies, M. Elk, M. Lifshitz, "Trait absorption is not reliably associated with brain structure or resting-state functional connectivity," *Neuroimage: Rep.*, vol. 3, pp. 1-14, Jun. 2023.

[20] D. Acunzo, D. Terhune, A. Sharma, and C. Hickey, "Absorption and Dissociation Mediate the Relationship between direct verbal suggestibility and impulsivity/compulsivity," *Acta Psychologica*, vol. 231, pp. 1-10, Nov. 2022.

[21] A. Diaz-Garcia et al., "Positive and Negative Affect Schedule (PANAS): psychometric properties of the online Spanish version in a clinical sample with emotional disorders," *BMC Psychiatry*, vol. 20, pp. 1-13, Feb. 2020.

[22] A. Saggino et al. "Improving the psychometric properties of the dissociative experiences scale (DES-II): a Rasch validation study," *BMC Psychiatry*, vol 20., pp. 1-10, Jan. 2020.

[23] M. Wang, L. Zhou, Q. Li, and A. Zhang, "Open world long-tailed data classification through active distribution optimization," *Expert Syst. with Appl*., vol. 213, pp. 119054, Mar. 2023.

[24] Z. Liu, Z. Miao, X. Zhan, J. Wang, B. Gong, and S. Yu, "Large-Scale Long-Tailed Recognition in an Open World," *2019 IEEE/CVF Conf. on Comput. Vision and Pattern Recognit. (CVPR)*, Long Beach, CA, USA, Jan. 2019, pp. 2532-2541.

[25] D. Sabolic, M. Samuelson, and M. Magzan, "Obtaining and Interpreting Students' Attitudes – Some Methodological Considerations and a Case Study," *Strategies, Challenges and, Opportunities for Sustain. in Uncertain Environ.,* Dubrovnik, Croatia, Sep. 2022, pp. 1-7.

[26] K. Akbari, M. Eigruber, and R. Vetschera, "Risk attitudes: The central tendency bias, " *EURO J. on Decis. Process.*, vol. 12, pp. 1-13, Nov. 2023.

[27] W. Teh, et al. "Measuring social desirability bias in a multi-ethnic cohort sample: its relationship with self-reported physical activity, dietary habists, and factor structure," *BMC Public Health*, vol. 23, pp. 1-10, Mar. 2023.

[28] C. Lambert , S. Arbuckle, and R. Holden, "The Marlow-Crowne Social Desirability Scale outperforms the BIDR Impression Management Scale for identifying fakers," *J. of Res. in Personality*, vol. 61, pp. 80-86, Apr. 2016.

[29] O. Friborg, M. Martinussen, and J. Rosenvinge, "Libert-based vs. semantic differential-based scorings of positive psychological constructs: A psychometric comparison of two versions of a scale measuring resilience," *Pers. and Individual Differ*., vol. 40, pp. 873-884, Apr. 2006.

[30] Y. Xiao and H. Chen, "Mild responses style in Slider Scales compared to Likert Scales," *American Psychological Assoc. Conv. 2023 (APA 2023),* Washington D.C., U.S., Aug. 2023, pp. 1-2.

[31] C. Roster, L. Lucianetti, and G. Albaum, "Exploring Slider vs. Categorical Response Formats in Web-Based Surveys," *J. of Res. Pract.,* vol. 11, pp. 1-19, Jul. 2015.

[32] T. Wang, "A Novel Approach of Integrating Natural Language Processing Techniques with Fuzzy TOPSIS for Product Evaluation," *Symmetry*, vol. 14, pp. 1-29, Jan. 2022.

[33] M. Wang, X. Cheng, and Z. He, "Research on Multiple Affective Responses Design of Product based on Kansei Engineering and TOPSIS-AISM," *Math. Problems in Eng*., vol. 2022, pp. 1-12, May 2022.

[34] J. Ji et al., "AI Alignment: A Comprehensive Survey," *Arxiv:2310.19852*, pp. 1-102, May 2024.

[35] G. Nicholson and Chris Holmes, "A note on statistical repeatability and study design for high-throughput assays," *Statist. in Medicine*, vol. 36, pp. 1-9, Nov. 2016.

[36] J. Williams, "A language for measurements," *Quantifying Meas.: The Tyrrany of Numbers*, San Rafael, CA: Morgan and Claypoole Publishers, Oct. 2016, pp. 3-1 to 3-7.

[37] C. Budde et al., "On Correctness, Precision, and Performance in Quantitative Verification," *Lecture Notes in Comput. Sci.: Leveraging Appl. of Formal Methods, Verification and Validation*, vol. 12479, pp. 216-241, Oct. 2020.

[38] "NCSES Supplemental Guidelines for New Surveys or Major Revisions of Existing Surveys: Supplemental Guidelines," *National Science Foundation*, pp. 1-29, May 2020.

[39] G. Boateng, T. Neilands, E. Frongillo, H. Melgar-Quinonez, and S. Young, "Best Practices for Developing and Validating Scales for Health, Social, and Behavioral Research: A Primer," *Frontiers in Public Health*, vol. 6, pp. 1-18, Jun. 2018.

[40] M. Carlson and R. Morrison, "Study Design, Precision, and Validity in Observational Studies," *J. Palliat. Med*., vol. 12, pp. 77-82, Jan. 2009.

[41] P. Louangrath, "Reliability and Validity of Survey Scales," *Int. J. of Res. & Methodology in Social Sci*., vol. 4, p. 50-62, Mar. 2018.

[42] M. Almac, "Effect of Audio-Visual Appeal on Game Enjoyment: Sample from Turkey," *Acta Ludologica*., vol. 6, pp. 42-61, Dec. 2023.

[43] F. Sepulveda, "Visual Aesthetics in Video Games and their effects on player motivation," M.S. thesis, College of Arts, Media, and Design, Northeastern Univ., Boston, MA, 2020.

[44] O. Schaffer and X. Fang, "Digital Game Enjoyment: A Literature Review," *HCI in Games: First Int. Conf*., Orlando, FL, Jul. 2019, pp. 191-214.

[45] L. Ermi and Frans Mayra, "Fundamental Components of the Gameplay Experience: Analysing Immersion," *Digit. Games Res. Conf.2005, Changing Views: Worlds in Play*, Vancouver, British Columbia, Canada, Jan. 2005, pp. 1-15.

[46] D. Weibel and B. Wissmath, "Immersion in Computer Games: The Role of Spatial Presence and Flow," *Int. J. of Comput. Games Technol*., vol. 2011, pp. 1-14, Jan 2011.

[47] L. Michailidis, E. Balaguer-Ballester, and Sun He, "Flow and Immersion in Video Games: The Aftermath of a Conceptual Challenge," *Front. Psychol.,* vol. 9, pp. 1-8, Sep. 2018.

[48] M. Phan, "The Development and Validation of the Game User Experience Satisfaction Scale (GUESS)," *Human Factors: The J. of the Human Factors and Ergonomics Soc*., vol. 58, pp. 1217-1247, Sep. 2016.

[49] S. Papaloukas, K. Patriarcheas and M. Xenos, "Usability Assessment Heuristics in New Genre Videogames," *2009 13th Panhellenic Conf. on Inform.*, Corfu, Greece, Oct. 2009, pp. 202-206.

[50] N. Senap and R. Ibrahim, "A Review of Heuristics Evaluation Component for Mobile Educational Games," *Procedia Comp. Sci.,* vol. 161, pp. 1028-1035, 2019.

[51] F. Leutner, S. Codreanu, Suzanne Brink, and T. Bitsakis, "Game based assessments of cognitive ability in recruitment: Validity, fairness and test-taking experience," *Front. Psychol*., vol. 13, pp. 1-17, Jan. 2023.

[52] Raghavan, M., Barocas, S., Kleinberg, J., and Levy, K. (2020). "Mitigating bias in algorithmic hiring: evaluating claims and practices," *Proc. of the 2020 Conf. on Fairness, Accountability, and Transparency (FAT '20),* Barcelona, Spain, Jan. 2020, pp. 469–481.

[53] Z. Guo and C. Lo, "How Do Narrative Features Affect a Player's Immersion in Computer Games? The Analytical Hierarchy Process Approach," in *Lecture Notes in Comput. Sci.: Cross-Cultural Design – Appl. in Learn., Arts, Cultural Heritage, Creative Ind., and Virtual Reality,* vol.13312, pp. 468-483, Jun. 2022.

[54] H. Qin, P. Rau, and G. Salvendy, "Player Immersion in the Computer Game Narrative," in *Lecture Notes in Comput. Sci.: Entertainment Comput – ECEC 2007*, vol. 4740, pp. 458-461, Jan. 2007.

[55] F. Fu, R. Su, and S. Yu, "EgameFlow: A scale to measure learners' enjoyment of e-learning games," *Comput. & Educ*., vol. 52, pp. 101-112, Jan 2009.

[56] P. Sweetser and P. Wyeth, "GameFlow: A Model for Evaluating Player Enjoyment in Games," *Comput. in Entertainment*, vol 3., pp. 1-24, Jul. 2005.

[57] H. Desurvire and C. Wiberg, "User Experience Design for Inexperienced Gamers: GAP – Game Approachability Principles," *Human-Comput. Interact. Series: Game User Experience Evaluation*, vol. 2015, pp. 169-186, Jun. 2015.

[58] X. Li, Z. Zhang, and K. Stefanidis, "A Data-Driven Approach for Video Game Playability Analysis Based on Players' Review," *Inf*., vol. 12, pp. 1-18, Mar. 2021.

[59] S. Rukmono, F. Azif and M. Candra, "Designing an Educational Game Evaluation Framework Based on Game Mechanics," *9th Int. Conf.on Inf. and Commun. Technol. (ICoICT)*, Yogyakarta, Indonesia, Sep. 2021, pp. 98-103.

[60] E. Oprins, G. Boer-Visschedijk, M. Roozeboom, M. Dankbaar, "The game-based learning evaluation model (GEM): Measuring the effectiveness of serious games using a standardised method," *Int. J. of Technol. Enhanced Learn*., vol. 7, pp. 1-21, Jan. 2015.

[61] A. Dayimu, N. Simidjievski, N. Demiris, and J. Abraham, "Sample size determination for prediction models via learning-type curves," *Statist. in Med*., vol. 43, pp. 1-11, May 2024.

[62] R. Kumar and S. Singal, "Penstock material selection in small hydropower plants using MADM methods," *Renewable and Sustain. Energy Rev*., vol. 52, pp. 240-255, Dec. 2015.

[63] B. Thompson, *Exploratory and Confirmatory Factor Analysis: Understanding concepts and applications*, Washington DC, U.S.: APA, 2004, pp. 1-218.

[64] J. Gunawan, C. Marzilli, and. Y. Aungsuroch, "Establishing appropriate sample size for developing and validating a questionnaire in nursing research," *Belitung Nurs. J*., vol. 7, pp. 356-360, Oct. 2021.

[65] W. Chen, L. McLeod, T. Coles, "Rasch First? Factor First," *ISPOR 17th Ann. European Congr.*, Amsterdam, The Netherlands, Nov. 2014, pp. 1-2.

[66] E. Guadagnoli and W. Velicer, "Relation of sample size to the stability of component patterns," *Psychological Bulletin*, vol. 103, pp. 265-275, Mar. 1988.

[67] J. Hair, B. Black, B. Babin, and R. Anderson, "Overview of Multivariate Methods," in *Multivariate Data Analysis, 7th ed,* London: Pearson Hall, 2010, ch. 1, pp. 1-785.

[68] J. Nevitt and G. Hancock, "Performance of bootstrapping approaches to model test statistics and parameter standard error estimation structural equation modeling," *Struct. Equ. Model*., vol. 8, pp. 353-377, Nov. 2009.

[69] L. Clark and D. Watson, "Constructing validity: Basic issues in objective scale development," in *Methodological issues and strategies in clinical research, 4th ed.,* A. E. Kazdin (Ed.), Washington DC, U.S.: APA, 2016, pp. 187-203.

[70] C. VanVoorhis and B. Morgan, "Understanding the Power and Rules of Thumb for Determining Sample Sizes," *Tut. in Quantitative Methods for Psychology*, vol. 3, pp. 43-50, Sep. 2007.

[71] M. White, "Sample size in quantitative instrument validation studies: A sytematic review of articles published in Scopus, 2011," *Heliyon,* vol. 8, pp. 1-6, Dec. 2022.

[72] A. Comrey and H. Lee, *A first Course in Factor Analysis*, *2nd ed.*, New York, NY, U.S.: Psychology Press, 1992, pp. 1-142.

[73] L. Aleamoni, "The relation of sample size to the number of variables in using factor analysis techniques," *Educ. and Psychol. Meas.*, vol. 36, pp. 879-883, Dec. 1976.

[74] D. Mundfrom, D. Shaw, and T. Ke, "Minimum Sample Size Recommendations for Conducting Factor Analyses," *Int. J. of Testing*, vol. 5, pp. 159-168, Jun. 2005.

[75] J. Hair, W. Black, B. Babin, and R. Anderson, *Multivariate Data Analysis, 8th ed*., Boston: Cengage, 2018, pp. 1-813.

[76] D. Suhr, " Exploratory or Confirmatory Factor Analysis," *Proc. of the 31st Annu. SAS Users Group Int. Conf.,* Cary, NC., ?? 2006, pp. 1-16.

[77] J. Nunnally and I. Bernstein, *Psychometric Theory, 3rd ed*., New York, NY, U.S.: McGraw-Hill Companies, 1994, pp. 1-736.

[78] M. Memon, H. Ting, J. Cheah, R. Thurasamy, F. Chuah, and T. Cham, "Sample Size for Survey Research: Review and Recommendations," *J. of Appl. Struct. Equ. Model*., vol. 4, pp. i-xx, Jun. 2020.

[79] S. Brown and J. Green, "The Wisdom Development Scale: Translating the Conceptual to the Concrete," *J. of College Student Develop*.," vol. 47, pp. 1-19, Feb. 2006.

[80] A. Forsberg and A. Rantala, "The Being Taken Seriously Questionnaire—Development and Psychometric Evaluation of a PREM Measure for Person-Centeredness in a High-Tech Hospital Environment," *Int. J. of Environ. Res. and Public Health*, vol. 17, pp. 1-10, Apr. 2020.

[81] S. Maxwell, "Sample Size and Multiple Regression Analysis," *Psychol. Methods*, vol. 5, pp. 434-458, Dec. 2000.

[82] R. Kline, *Princ. and Pract. of Structural Equ. Model., 5th ed*., New York, NY, U.S.: Guilford Press, 2023, pp. 1- 494.

[83] A. Jackson, "Revisiting sample size and number of parameter estimates: Some support for the N:q hypothesis," *Struct. Equ. Model*., vol. 10, pp. 128-141, Nov. 2009.

[84] J. Osborne and A. Costello, "Sample size and subject to item ratio in principle components analysis," *Practical Assessment, Res. & Eval*., vol. 9, pp. 1-9, Jun. 2004.

[85] E. Pedhazur, *Multiple Regression in Behavioral Res.: Explanation and Prediction*, Fort Worth, TX U.S.: Harcourt Brace College Publishers, 1997, pp. 1-1058.

[86] D. Kacprzak, "A Novel Extension of the Technique for Order Preference by Similarity to Ideal Solution Method with Objective Criteria Weights for Group Decision Making with Interval Numbers," *Entropy*, vol. 23, pp. 1-20, Sep. 2021.

[87] A. Lubis, N. Khairina, and M. Riandra, "Comparison between Multiple Attribute Decision Making Methods through Objective Weighting Method in Determining Best Employee," *Int. J. of Innov. Res. in Comput. Sci. & Technol. (IJIRCST)*, vol. 11, pp. 2347-5552, Mar. 2023.

[88] E. Zavadskas, Z. Stevic, Z. Turskis, M. Tomasevic "A Novel Extended EDAS in Minkowski Space (EDAS-M) Method for Evaluating Autonomous Vehicles," *Stud. in Inform. and Control*, vol. 28, pp. 255-264, Oct. 2019.

[89] A. Chaudhuri and T. Sahu, "PROMETHEE-based hybrid feature selection technique for high-dimensional biodmedical data: application to Parkinson's disease classification," *Electron. Lett.*, vol. 56, pp. 1403-1406, Oct. 2020.

[90] M. Basilio, V. Pereira, and F. Yigit, "New Hybrid EC-Promethee Method with Multiple Iterations of Random Weight Ranges: Step-by-Step Applicadtion in Python," *MethodsX.*, vol. 13, pp. 1-16, Aug. 2024.

[91] S. Chan, "AI-Facilitated Selection of the Optimal Nondominated Solution for a Serious Gaming Information Fusion Module," *2024 IEEE Gaming, Entertainment, and Media Conf. (GEM)*, Turin, Italy, Jul. 2024, pp. 1-6,

[92] G. Sullivan and A. Artino "Analyzing and Interpreting Data from Likert-Type Scales," *J. Grad Med. Educ*., vol. 5, pp. 541-542, Dec. 2013.

[93] K. Taber, "The Use of Cronbach's Alpha When Developing and Reporting Research Instruments in Science Education," *Res. in Sci. Education*, vol. 48, pp. 1273-1296, Jun. 2017.

[94] B. Tabachnick and L. Fidell, *Using Multivariate Statistics, 7th ed*., New York, U.S.: Pearson, 2021, pp. 1-848.