Beyond Accuracy

Adaptive Decision Support Systems

Beyond accuracy

"Making recommendations better"

There are aspects beyond the algorithm that determine the quality of recommender systems

"Being accurate is not enough"

There are metrics beyond accuracy that should be used to build and evaluate recommender algorithms

"Explaining the user experience of recommender systems"

A framework to structure the evaluation of recommenders

“The algorithm accounts for only 5% of the commercial success of our recommender systems [...] The interactive components of a recommender account for about 50%”

HRI

Three pillars of Human Recommender Interaction (HRI)

The act of getting recommendations (dialogue)

Perception of the recommender over time (personality)

How the recommender meets the user's needs

Recommendations

Recommendations shouldn't just be correct, they should be notable (cf. serendipity) and useful

In system terms, this might require the recommender to not just be personalized, but to be bold and take risks

Affirmation is good, but prevent pigeon-holing

The system should adapt to changing needs and provide fresh recommendations

Process

The recommendation process must be transparent and usable

Inspire trust in users

Meet users' needs and expectations w.r.t. their task

Contributions

What does this paper contribute to the state-of-the-art?

Mindset: think about recommenders beyond the algorithm

Agenda-setting: outline goals and objectives of a recommender system (and research)

What else?

Criticisms

What is missing from this paper?

Practical design advice: how do we build a recommender to accomplish these goals?

Evaluation advice: How do we evaluate these qualities?

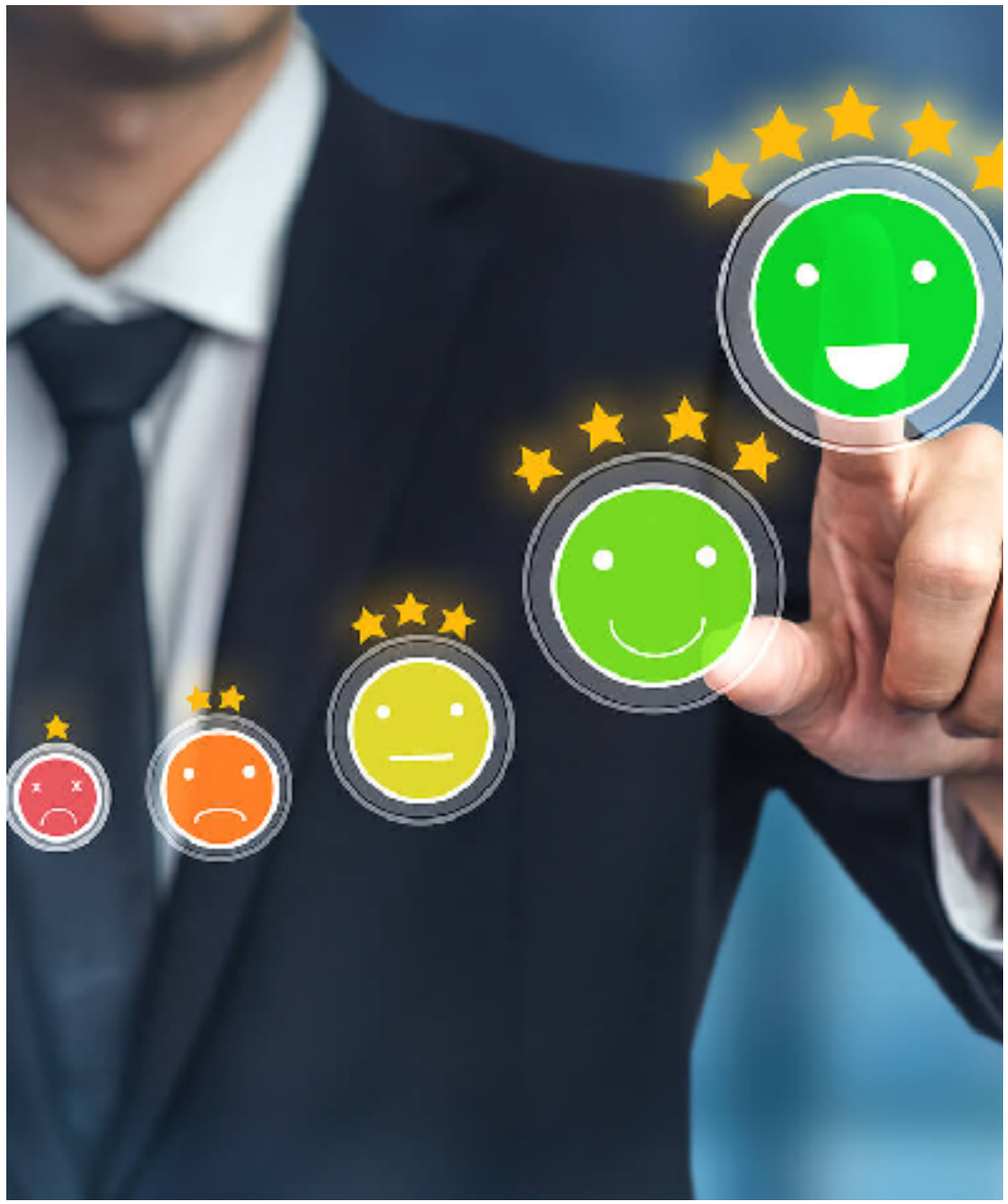
What else?


NBC
Left Field

Accuracy?

McNee et al. (2002): Item-Item CF provided the best predictions, but was rated least helpful by users!

Torres et al. (2004): CBF-separated CF had the lowest predictive accuracy, but resulted in the highest user satisfaction!



Beyond accuracy

Recognizes that accuracy is not the only way to evaluate a recommender

- Does not consider list composition
- Does not consider serendipity
- Does not consider user experiences and expectations

List composition

The recommendations are based on similarity

Remember w_{ij} and w_{uv}?

But they should also be diverse!

Compared to the rated items (non-obvious recommendations)

Compared to the other items in the list (this also reduces choice overload)

Over time (adaptability and freshness)

To some extent, diversity is the opposite of similarity!

List composition

If we consider diversity the opposite of similarity, you simply get the opposite of accuracy...

How can we reconcile diversity and similarity?

- Use different metrics: e.g. user-user or item-item similarity + genre or emotion diversity
- Diversify perpendicular to the user vector (see next slides)

List composition

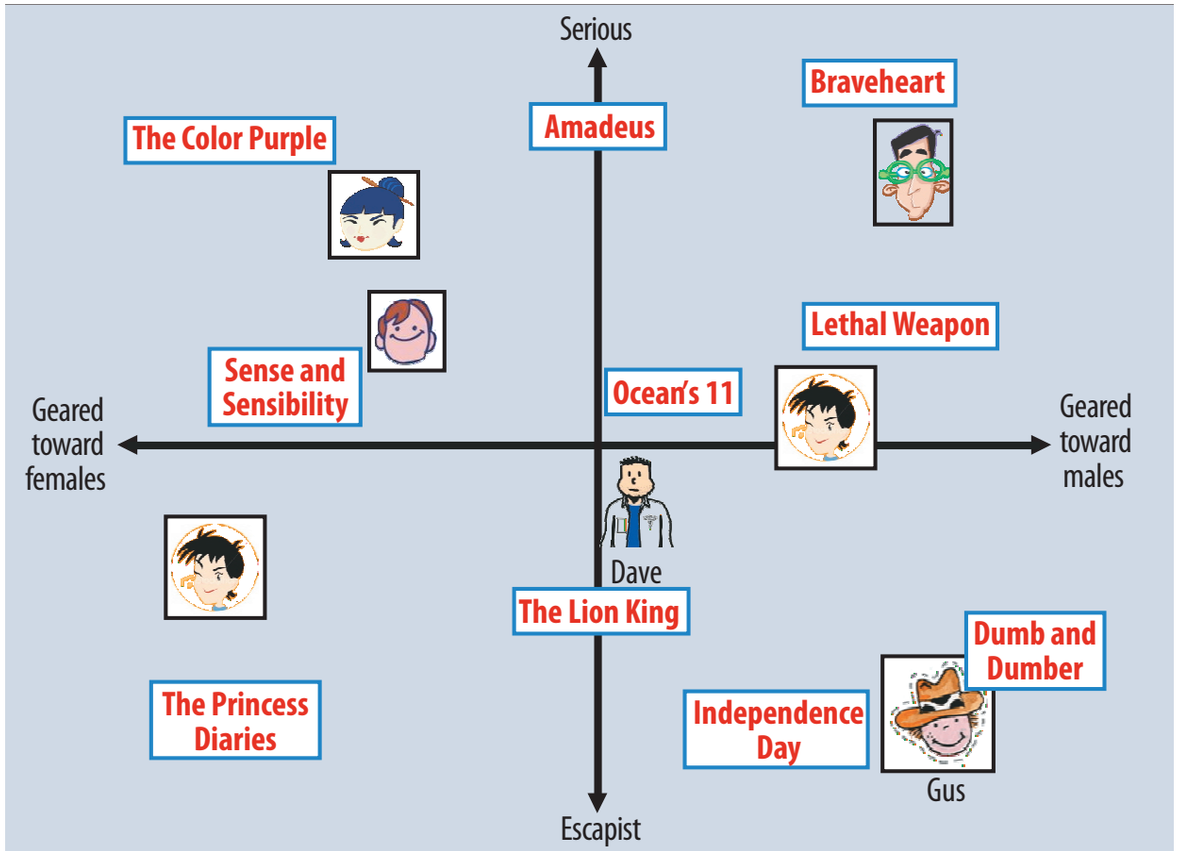
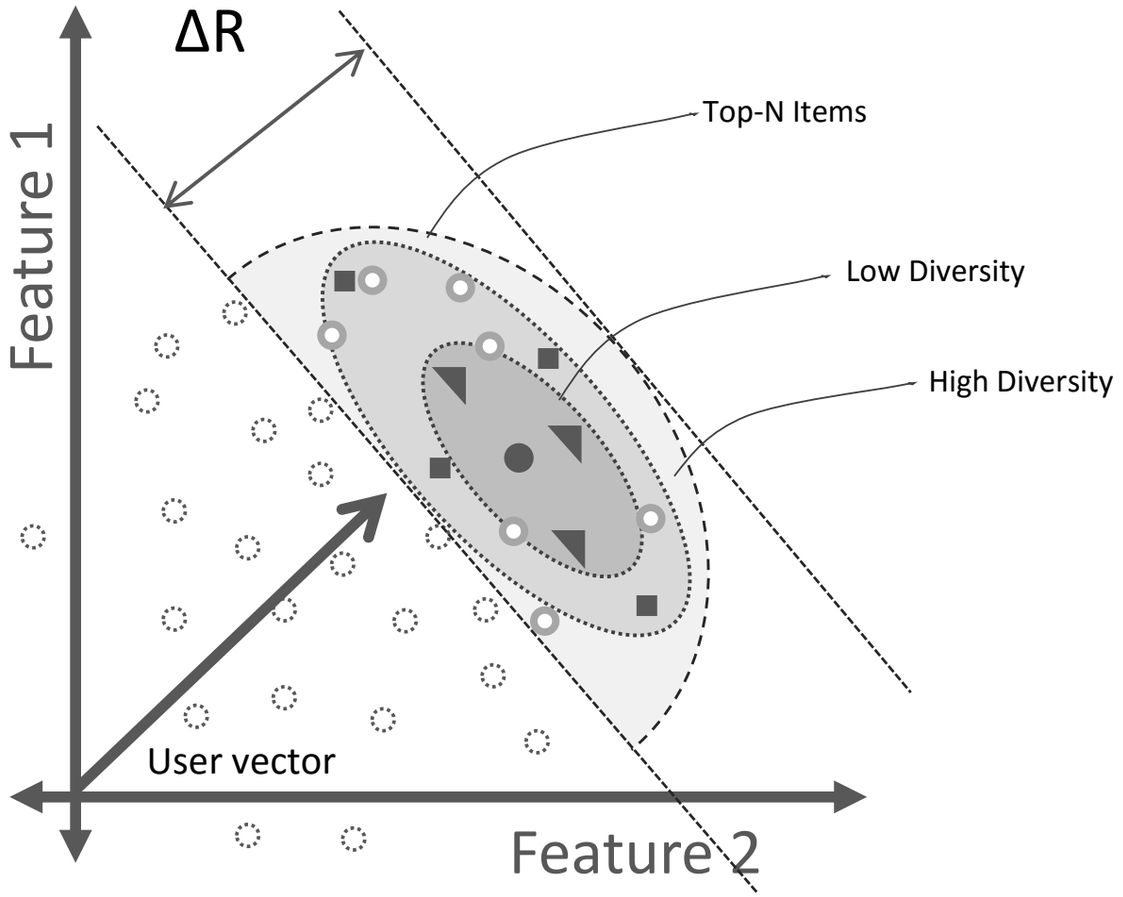


Figure 2. A simplified illustration of the latent factor approach, which characterizes both users and movies using two axes—male versus female and serious versus escapist.



Serendipity

Offline algorithms are tested on items that are already rated by the user

To some extent, you are optimizing/testing for “ratability”

How to produce (or evaluate) recommendations that are “non-obvious”?

e.g. Items in the “long tail”

The importance (and “flavor”) of this may depend on the task at hand!

User experience

User satisfaction does not always correlate with high recommender accuracy (McNee 2002, Ziegler 2005)

Why not?

- Importance of other recommendation qualities (diversity, serendipity, etc.)
- Importance of interaction mechanism usability (preference elicitation, transparency, etc.)
- Interpersonal (e.g. cultural) differences
- Situational (e.g. task-related) differences

Contributions

What does this paper contribute to the state-of-the-art?

Mindset: think about algorithms beyond accuracy

Agenda-setting: outline other metrics and objectives for optimize/test algorithms

What else?

Criticisms

What is missing from this paper?

Metrics: how do we measure diversity, serendipity, UX?

Reconciliation: If the metrics disagree, how do we reconcile them?

What else?

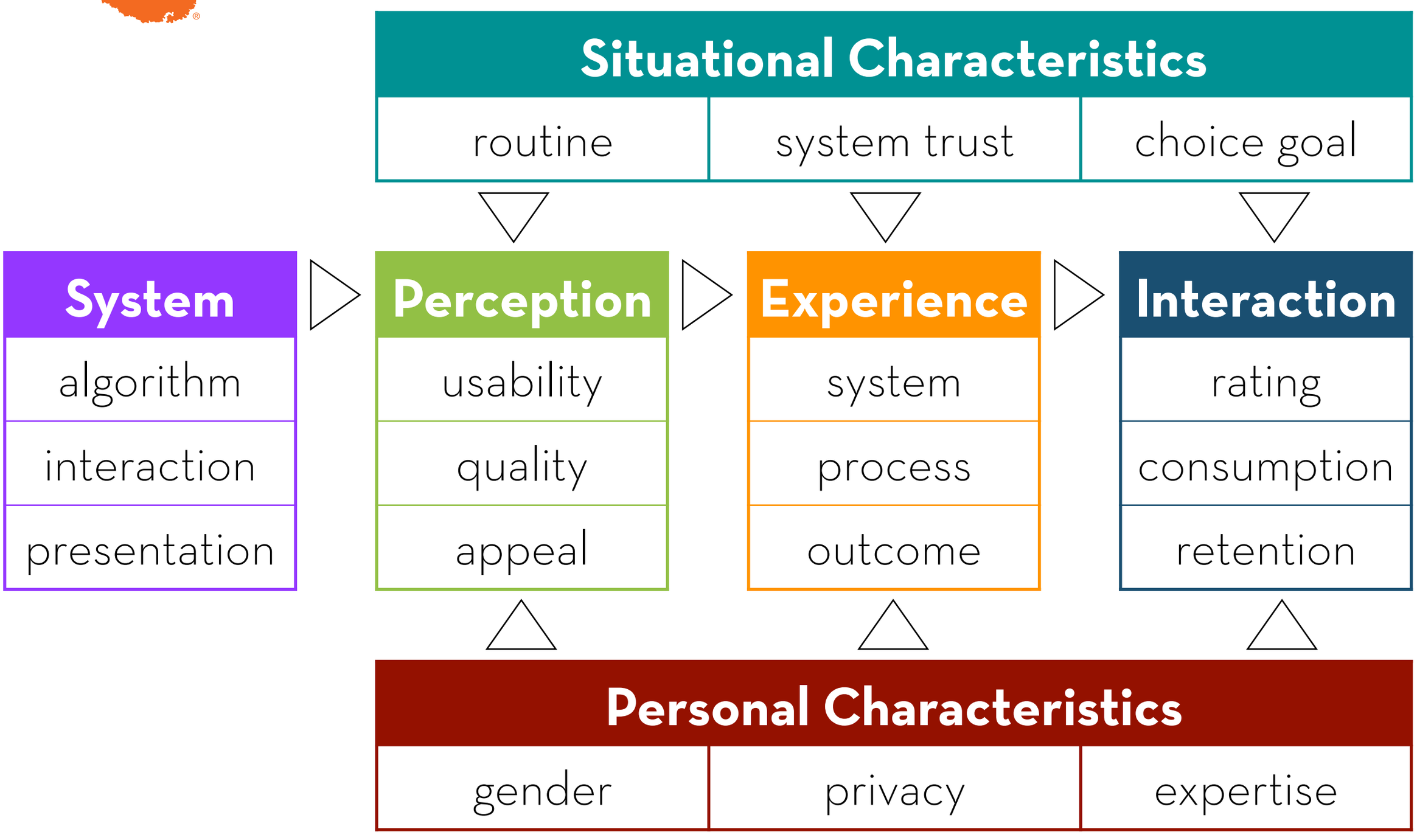
User experience

Present a framework to integrate user experience evaluations

Provide measurement scales to measure subjective aspects and experiences

Validate the framework in a series of field trials

Framework


Situational Characteristics
routine
system trust
choice goal
System
algorithm
interaction
presentation
Perception
usability
quality
appeal
Experience
system
process
outcome
Interaction
rating
consumption
retention
Personal Characteristics
gender
privacy
expertise

Framework

Situational Characteristics

routine	system trust	choice goal

System

algorithm
interaction
presentation

Objective System Aspects (OSA)

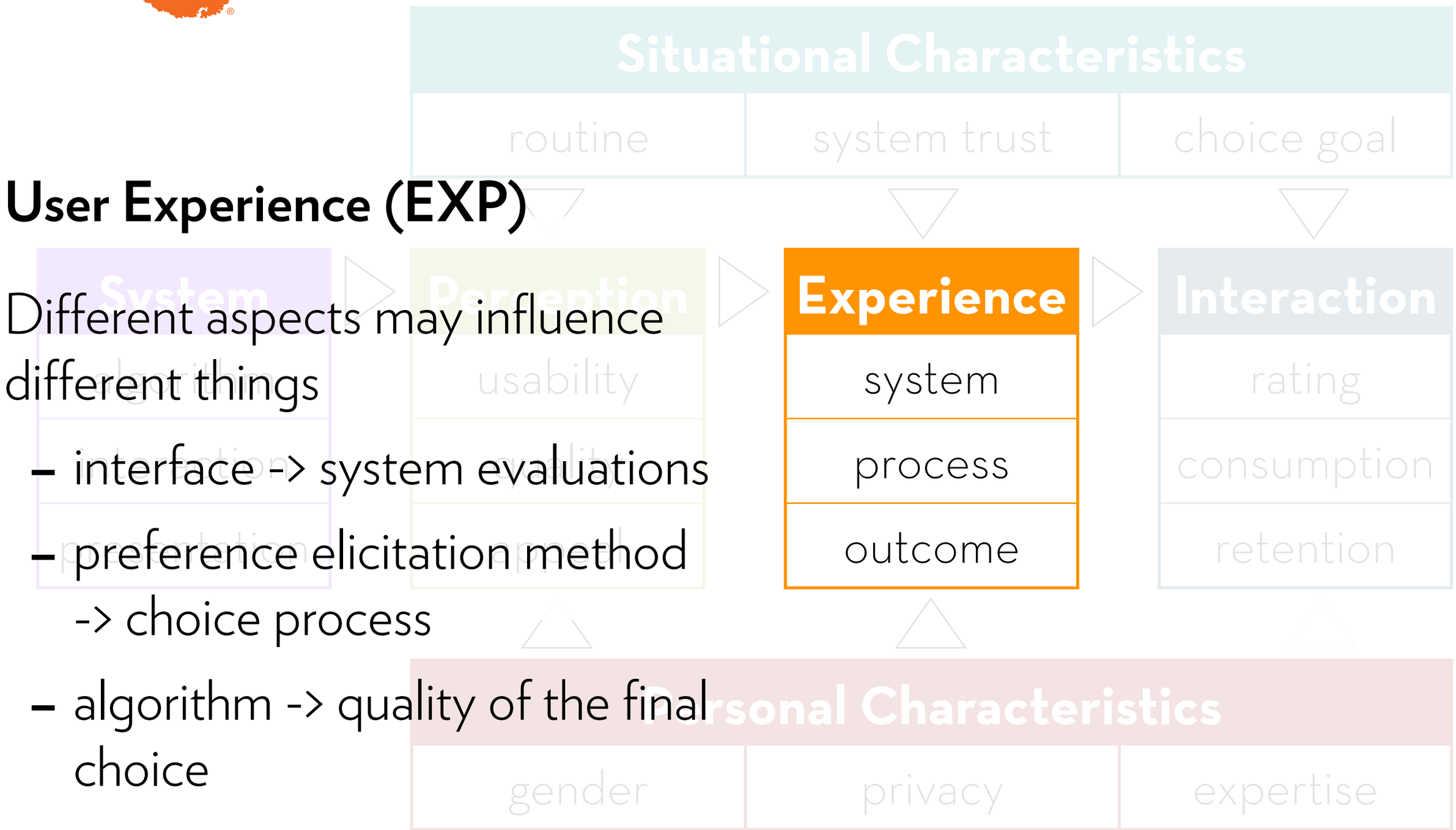
- visual / interaction design
- recommender algorithm
- presentation of recommendations
- additional features

Perception	Experience	Interaction
usability	system	rating
quality	process	consumption
appeal	outcome	retention

Personal Characteristics

gender	privacy	expertise

Framework



User Experience (EXP)

Different aspects may influence different things

- interface -> system evaluations
- preference elicitation method -> choice process
- algorithm -> quality of the final choice

Framework

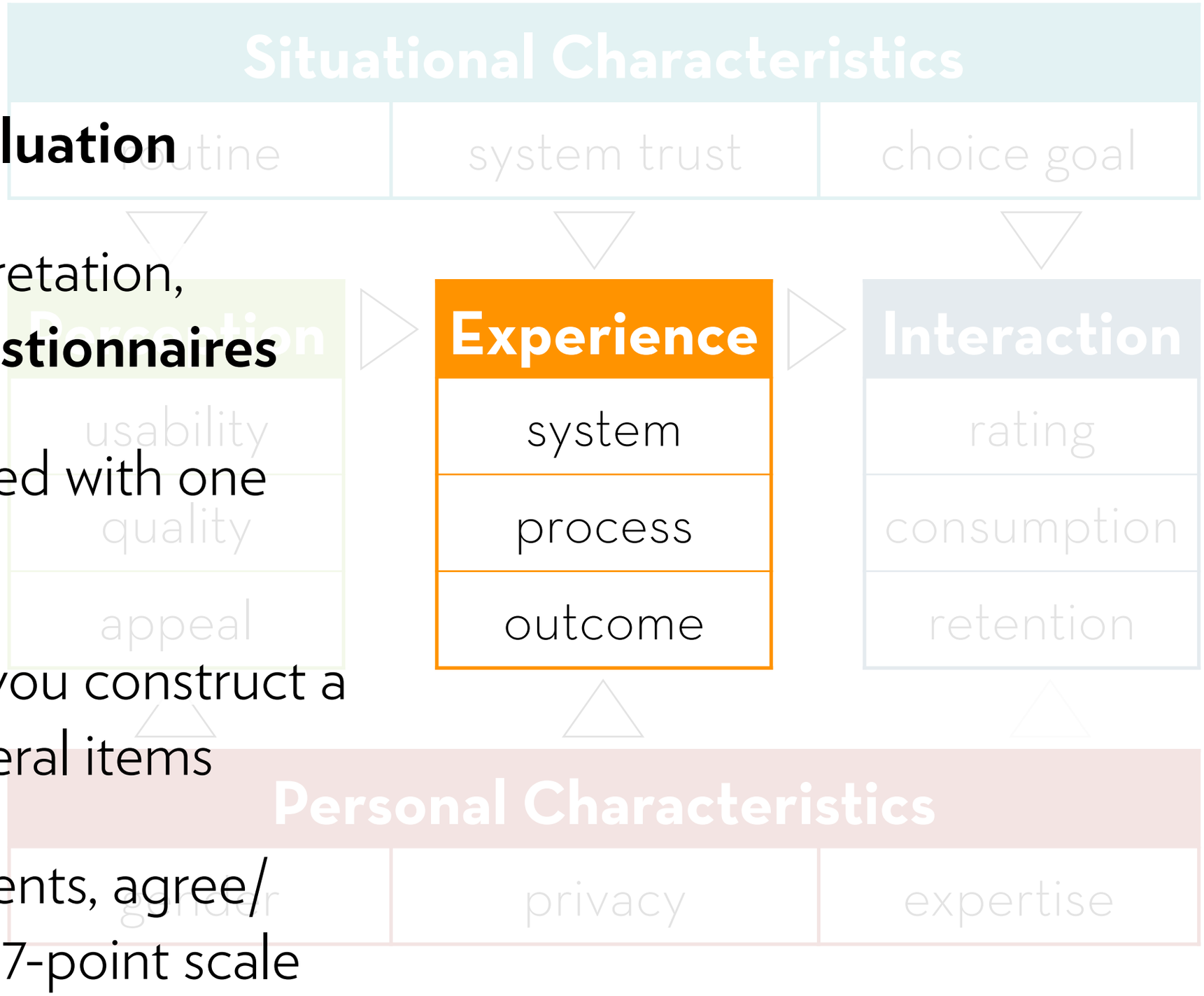
An attitudinal evaluation

self-relevant interpretation, measured with **questionnaires**

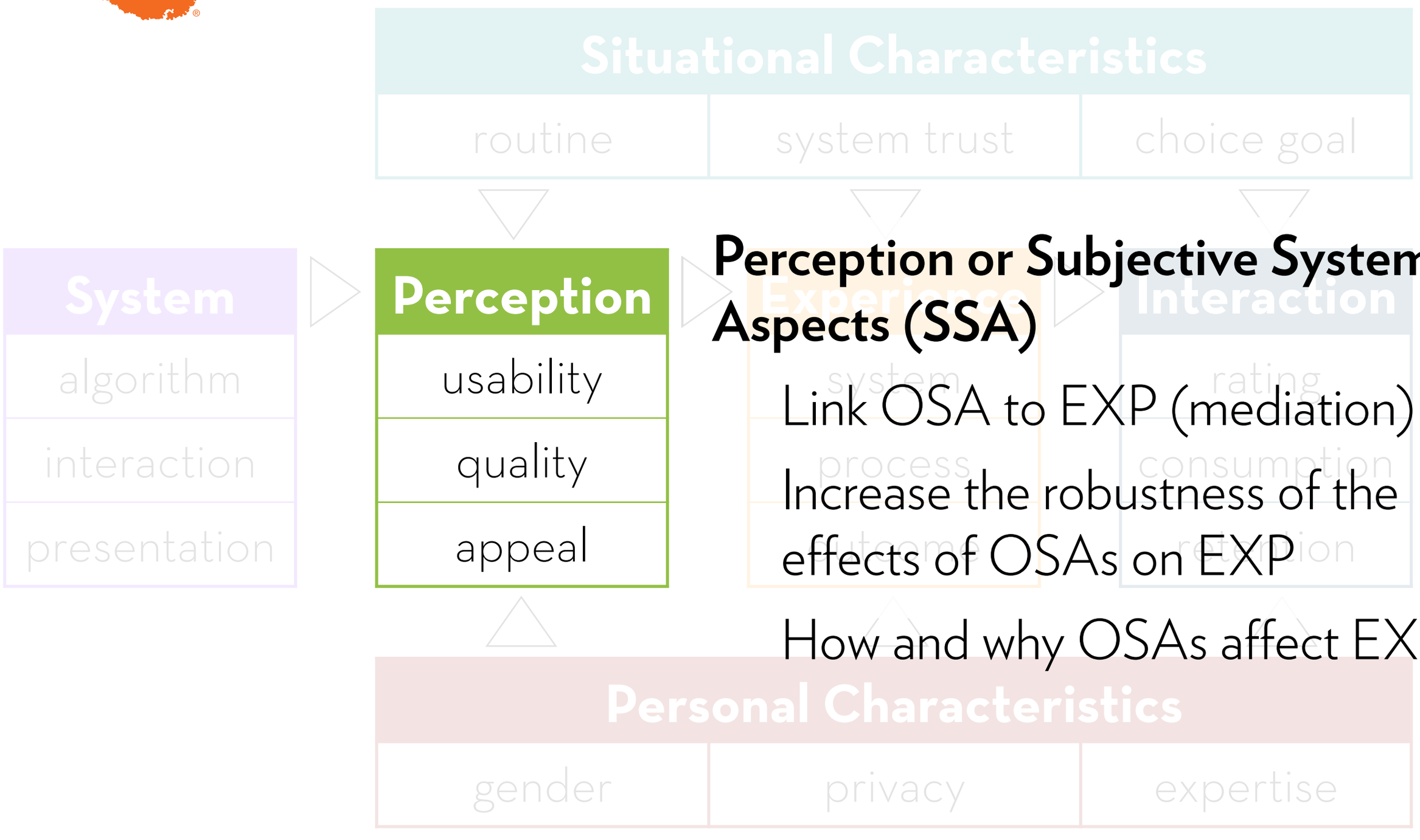
Cannot be measured with one question

Far more robust if you construct a scale based on several items

Usually 5-8 statements, agree/disagree on a 5- or 7-point scale



Framework



Perception or Subjective System Aspects (SSA)

Link OSA to EXP (mediation)

Increase the robustness of the effects of OSAs on EXP

How and why OSAs affect EXP

Framework

Personal characteristics (PC)

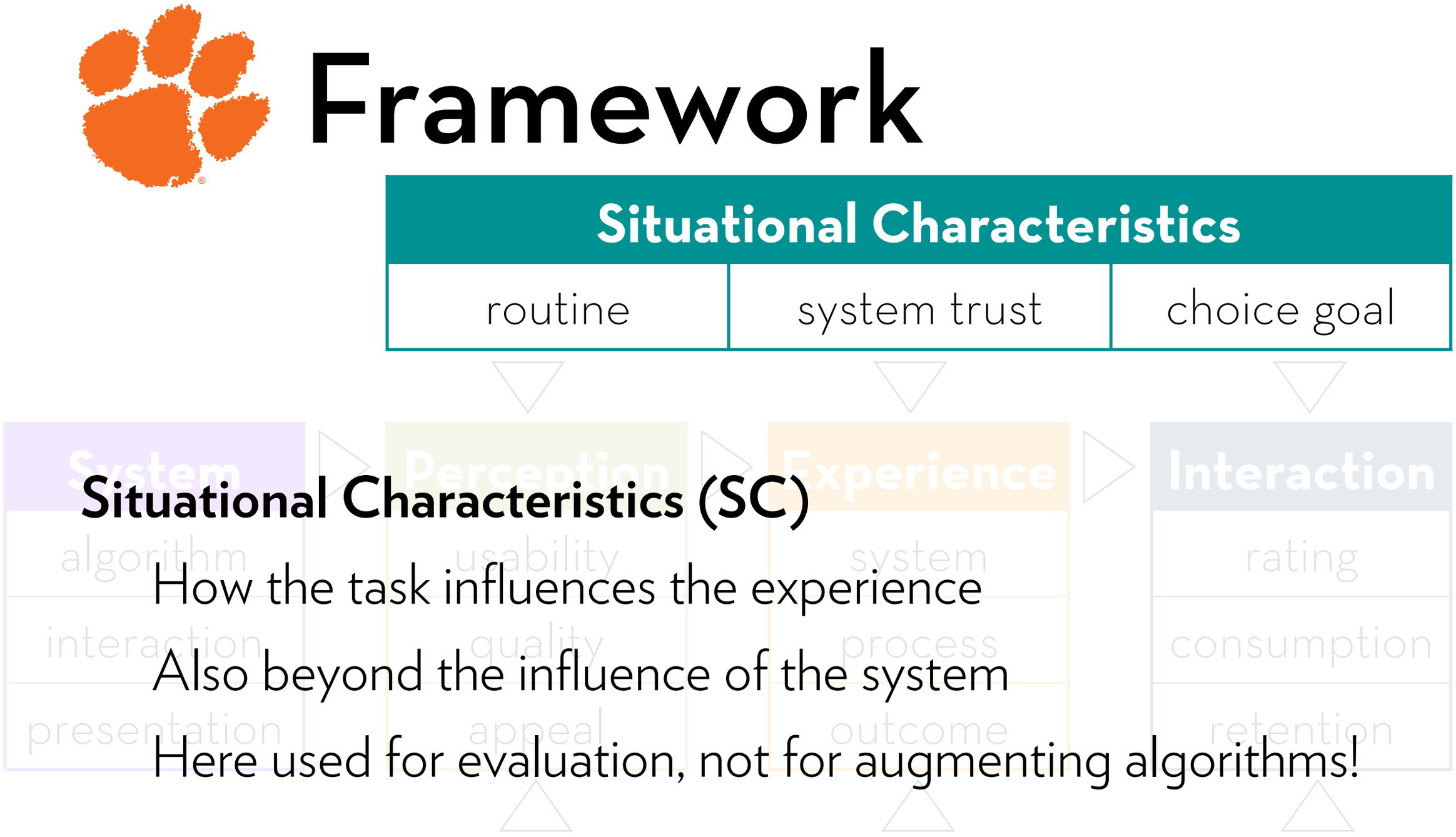
Users' general disposition

Beyond the influence of the system

Typically stable across tasks

Personal Characteristics		
gender	privacy	expertise

Framework



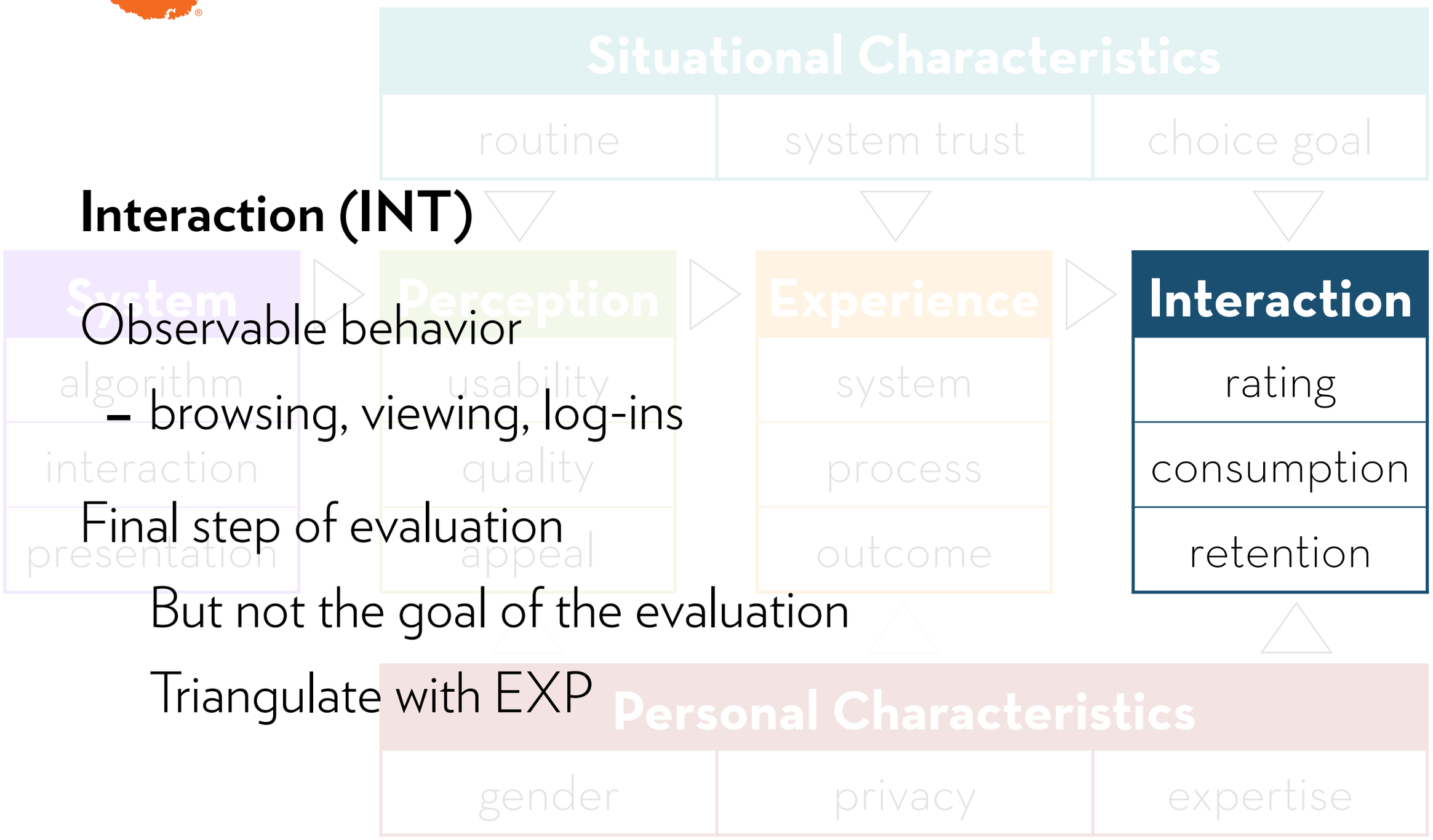
Situational Characteristics (SC)

How the task influences the experience

Also beyond the influence of the system

Here used for evaluation, not for augmenting algorithms!

Framework



Interaction (INT)

Observable behavior

- browsing, viewing, log-ins

Final step of evaluation

But not the goal of the evaluation

Triangulate with EXP

The full cycle

Link the objective interaction (purple) to objective system aspects (blue)

through a series of subjective constructs

Why? Because these constructs...

...are likely to **attenuate** and qualify the effects

...**explain** why users' behavior is different for different systems

This explanation is the main value of our framework

How to use the framework

What it is not

a **predefined metric** for standardized “performance” tests

What it is

conceptual definition of a generic **chain of effects**

helps researchers to explain **how and why** the user experience of recommender systems comes about

a generic **guideline** for in-depth empirical research

generator for **new hypotheses**

EMIC pre-trial (FT1)

Questions

Do users want to receive personalized recommendations **at all**?

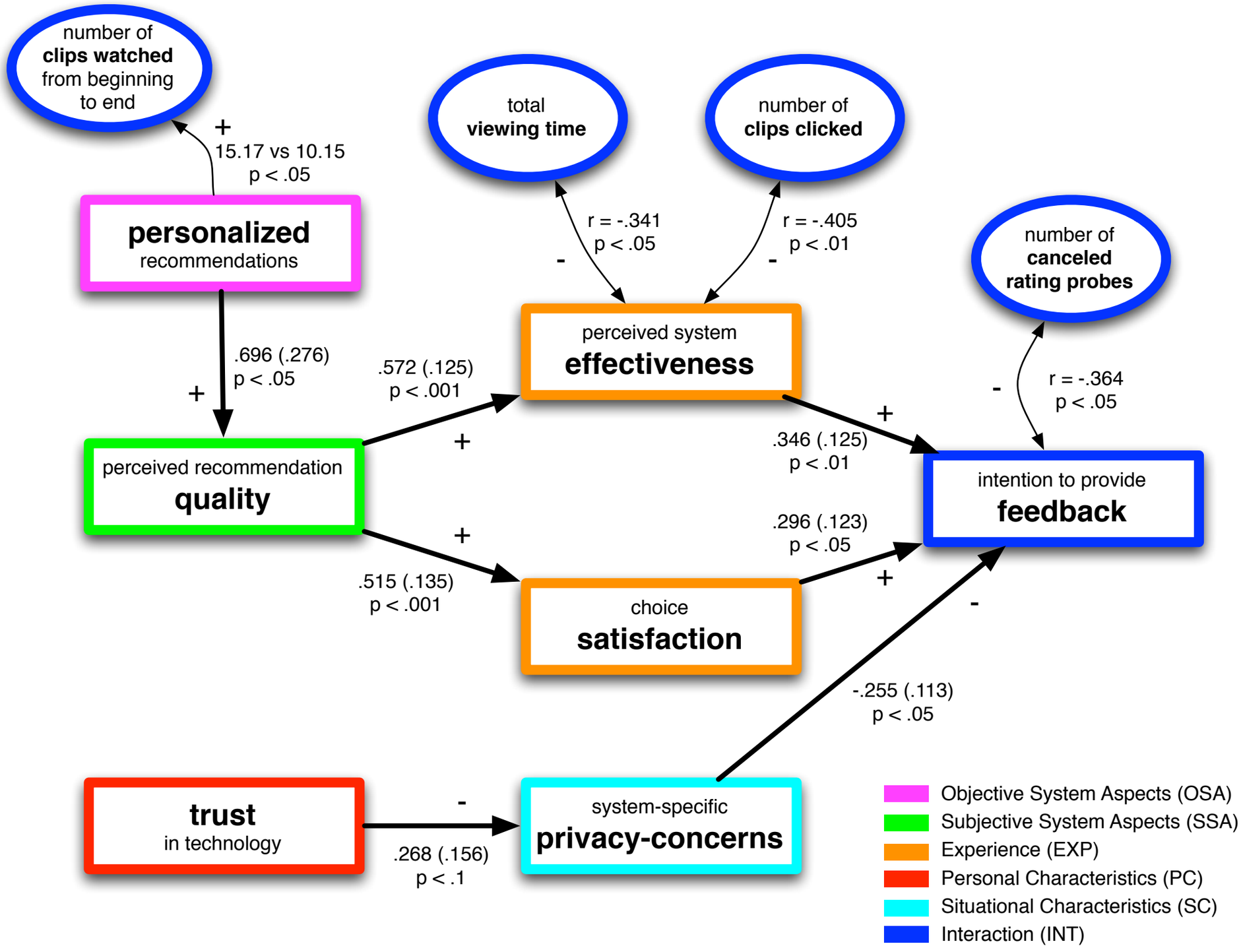
Will users provide adequate preference **feedback**?

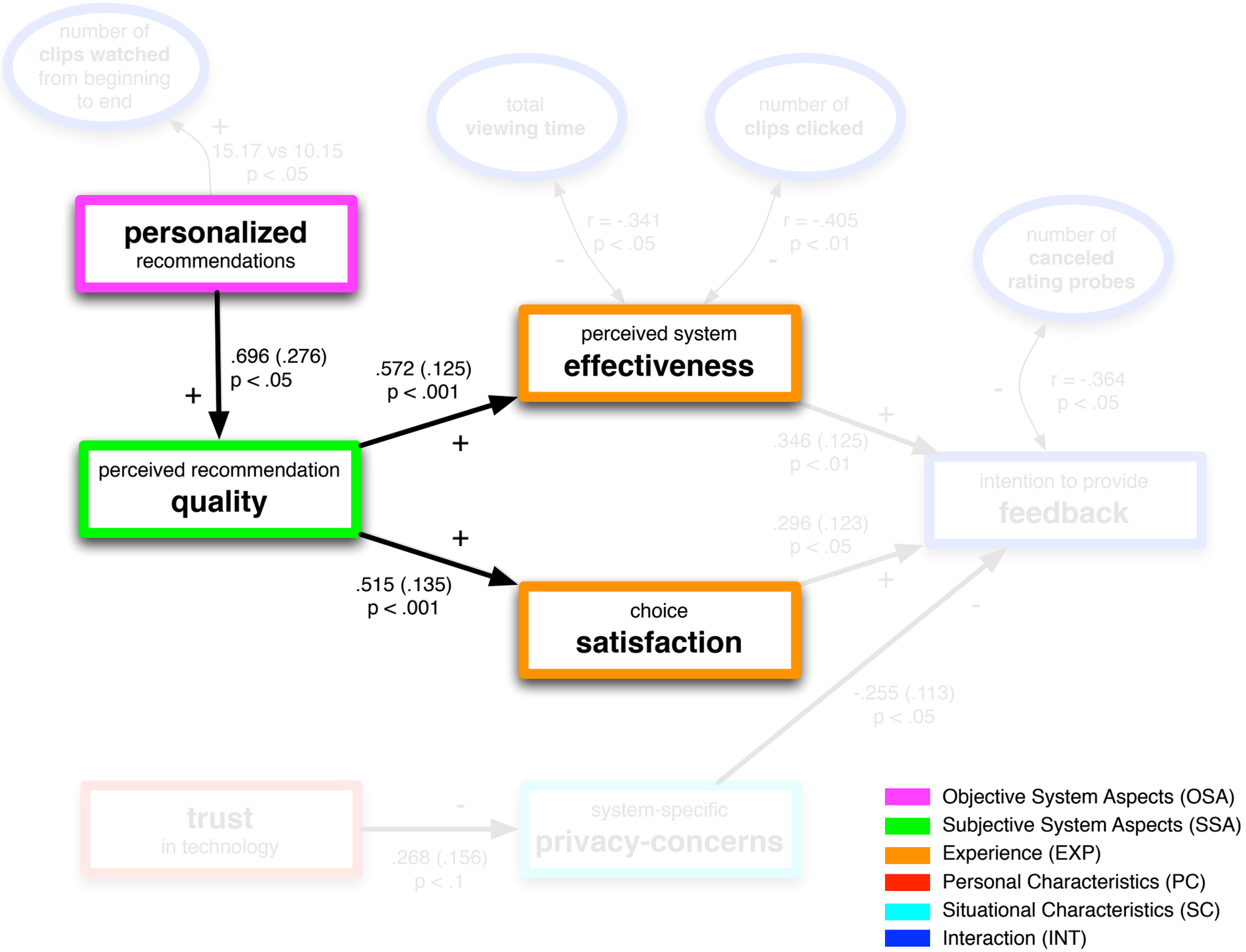
Setup

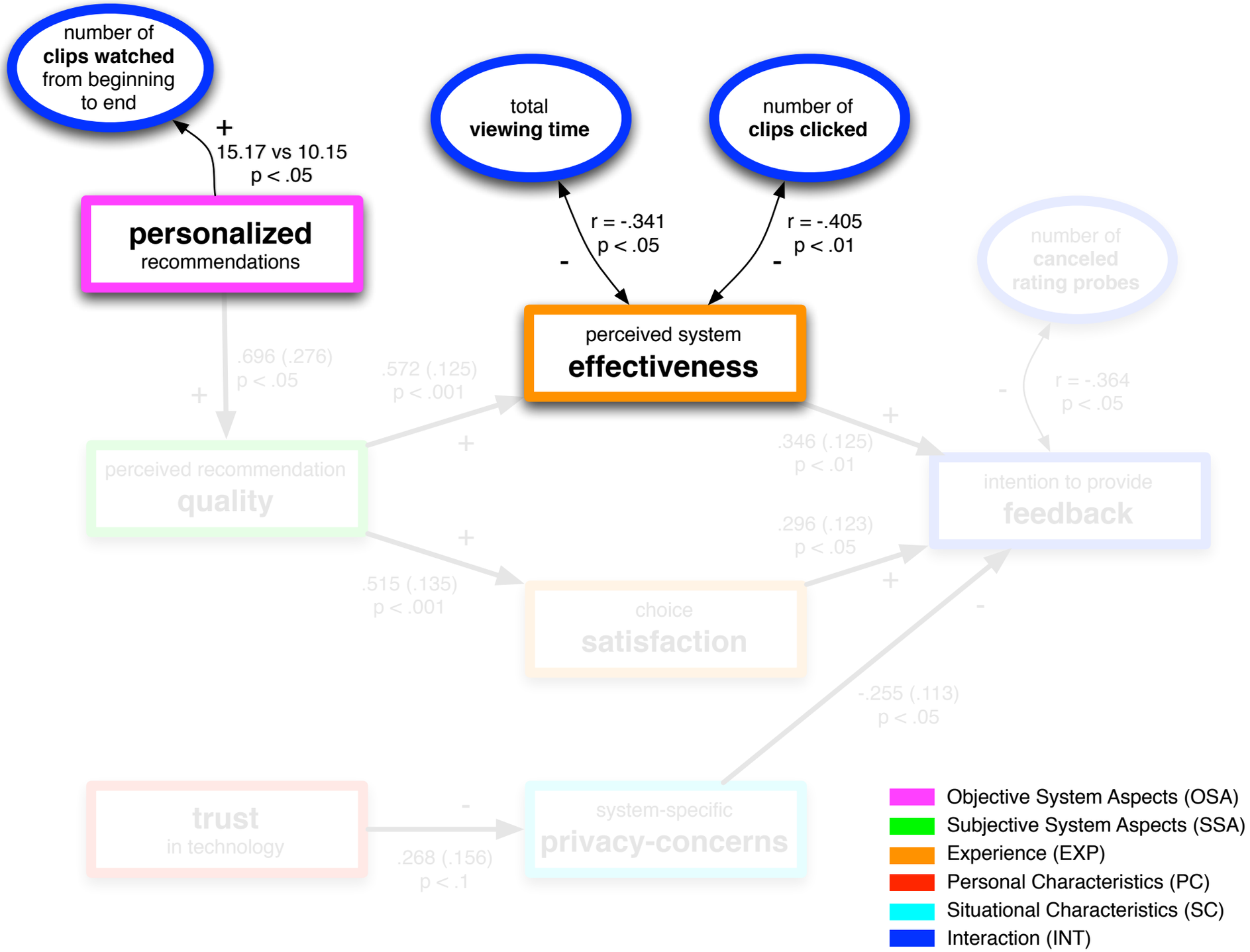
43 participants (between subjects)

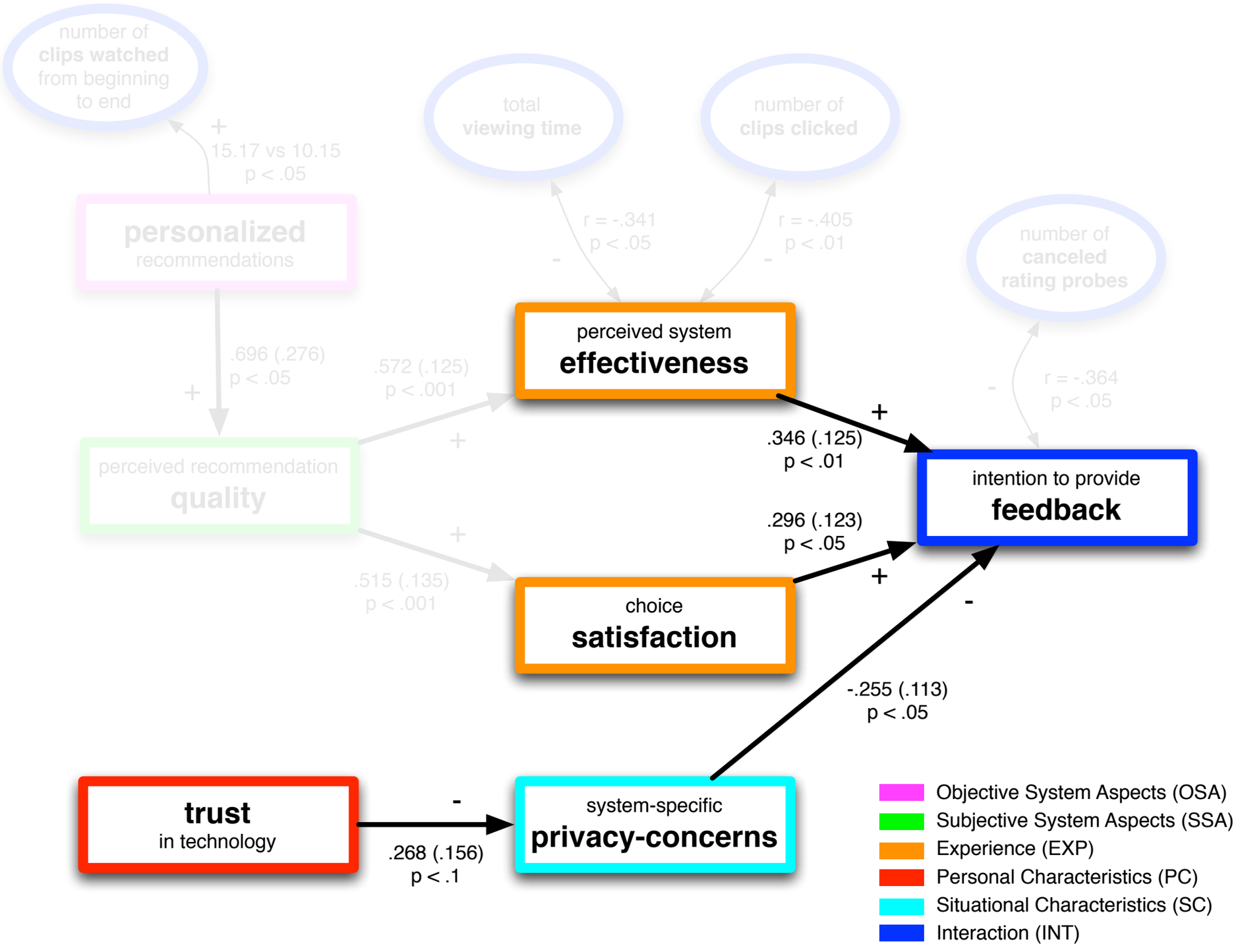
Random recommendations vs. personalized VSM algorithm recommendations (between subjects)

Measure quality perceptions, experience, privacy concerns, and intention to provide feedback; capture clickstream


number of
clips watched
from beginning
to end
+
15.17 vs 10.15
p < .05
personalized
recommendations
total
viewing time
number of
clips clicked
r = -.341
p < .05
-
r = -.405
p < .01
-
number of
canceled
rating probes
perceived system
effectiveness
.696 (.276)
p < .05
+
.572 (.125)
p < .001
+
-
r = -.364
p < .05
perceived recommendation
quality
+
.346 (.125)
p < .01
+
intention to provide
feedback
+
.296 (.123)
p < .05
+
.515 (.135)
p < .001
choice
satisfaction
+
-
-.255 (.113)
p < .05
trust
in technology
-
.268 (.156)
p < .1
system-specific
privacy-concerns
Objective System Aspects (OSA)
Subjective System Aspects (SSA)
Experience (EXP)
Personal Characteristics (PC)
Situational Characteristics (SC)
Interaction (INT)


number of
clips watched
from beginning
to end
+
15.17 vs 10.15
p < .05
total
viewing time
number of
clips clicked
r = -.341
p < .05
-
r = -.405
p < .01
-
number of
canceled
rating probes
personalized
recommendations
.696 (.276)
p < .05
+
.572 (.125)
p < .001
perceived system
effectiveness
-
r = -.364
p < .05
+
.346 (.125)
p < .01
perceived recommendation
quality
+
intention to provide
feedback
.296 (.123)
p < .05
+
.515 (.135)
p < .001
choice
satisfaction
+
-
-.255 (.113)
p < .05
trust
in technology
-
.268 (.156)
p < .1
system-specific
privacy-concerns
Objective System Aspects (OSA)
Subjective System Aspects (SSA)
Experience (EXP)
Personal Characteristics (PC)
Situational Characteristics (SC)
Interaction (INT)


number of
clips watched
from beginning
to end
+
15.17 vs 10.15
p < .05
personalized
recommendations
total
viewing time
number of
clips clicked
r = -.341
p < .05
-
r = -.405
p < .01
-
perceived system
effectiveness
number of
canceled
rating probes
r = -.364
p < .05
-
.696 (.276)
p < .05
+
.572 (.125)
p < .001
+
.346 (.125)
p < .01
+
perceived recommendation
quality
+
.515 (.135)
p < .001
choice
satisfaction
.296 (.123)
p < .05
+
intention to provide
feedback
-
-.255 (.113)
p < .05
trust
in technology
-
.268 (.156)
p < .1
system-specific
privacy-concerns
Objective System Aspects (OSA)
Subjective System Aspects (SSA)
Experience (EXP)
Personal Characteristics (PC)
Situational Characteristics (SC)
Interaction (INT)


number of
clips watched
from beginning
to end
+
15.17 vs 10.15
p < .05
total
viewing time
number of
clips clicked
r = -.341
p < .05
-
r = -.405
p < .01
-
number of
canceled
rating probes
personalized
recommendations
+
.696 (.276)
p < .05
.572 (.125)
p < .001
perceived system
effectiveness
-
r = -.364
p < .05
+
.346 (.125)
p < .01
perceived recommendation
quality
+
intention to provide
feedback
+
.296 (.123)
p < .05
+
.515 (.135)
p < .001
choice
satisfaction
+
-
-.255 (.113)
p < .05
trust
in technology
-
.268 (.156)
p < .1
system-specific
privacy-concerns
Objective System Aspects (OSA)
Subjective System Aspects (SSA)
Experience (EXP)
Personal Characteristics (PC)
Situational Characteristics (SC)
Interaction (INT)

BBC trial (FT4)

Questions

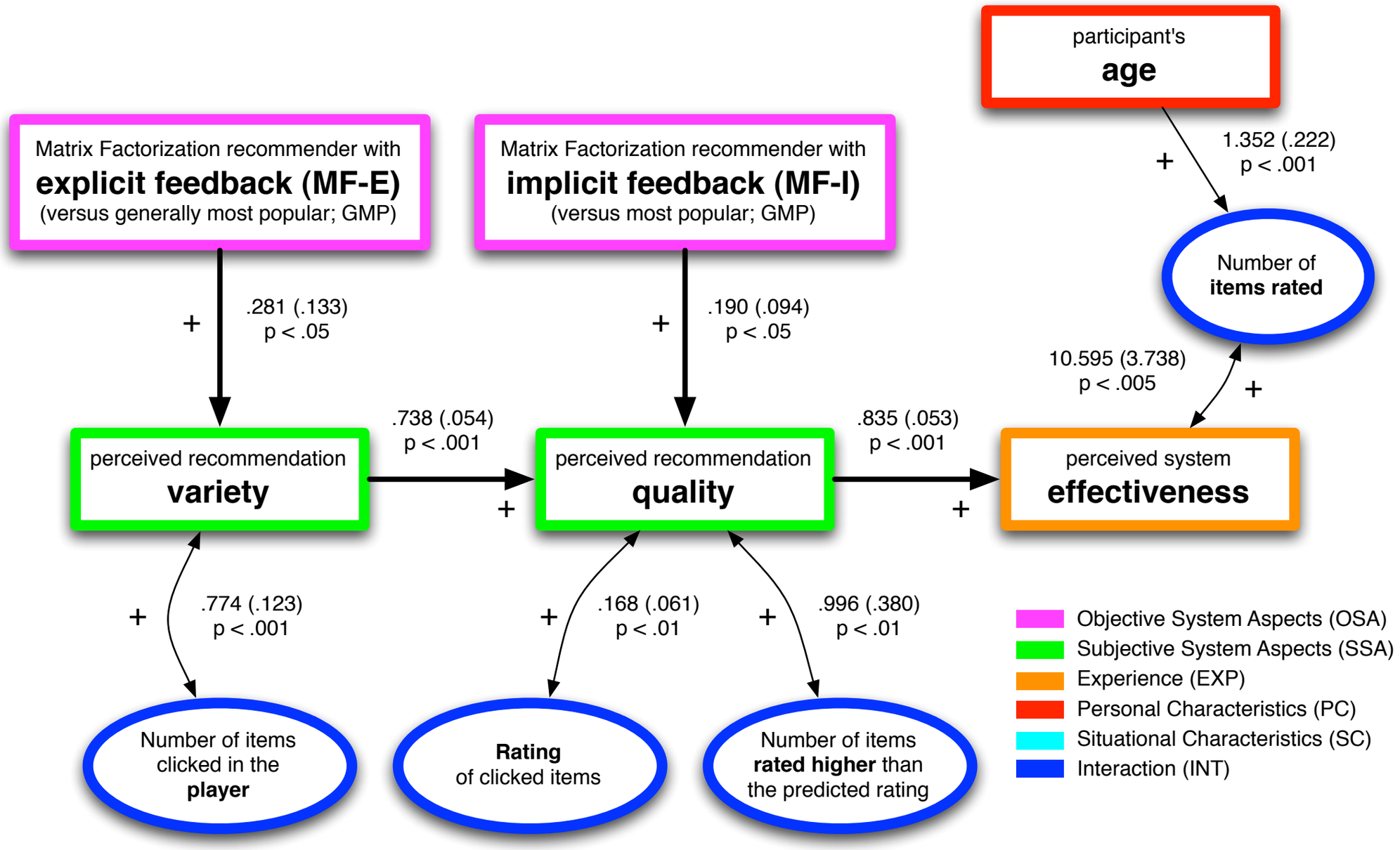
Can we find experience differences between non-personalized recommendations, an MF algo with **explicit feedback**, and an MF algo with **implicit feedback**?

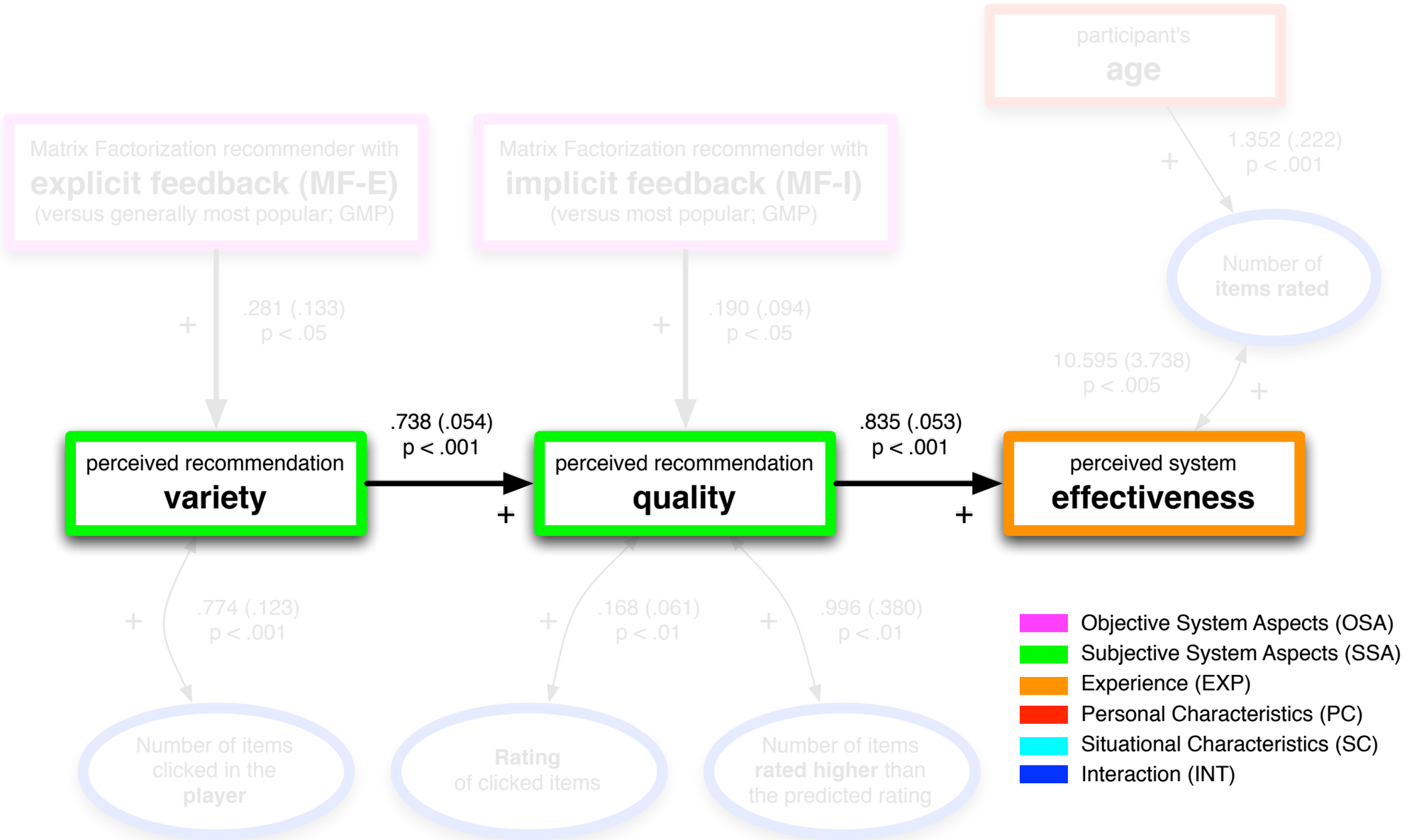
Setup

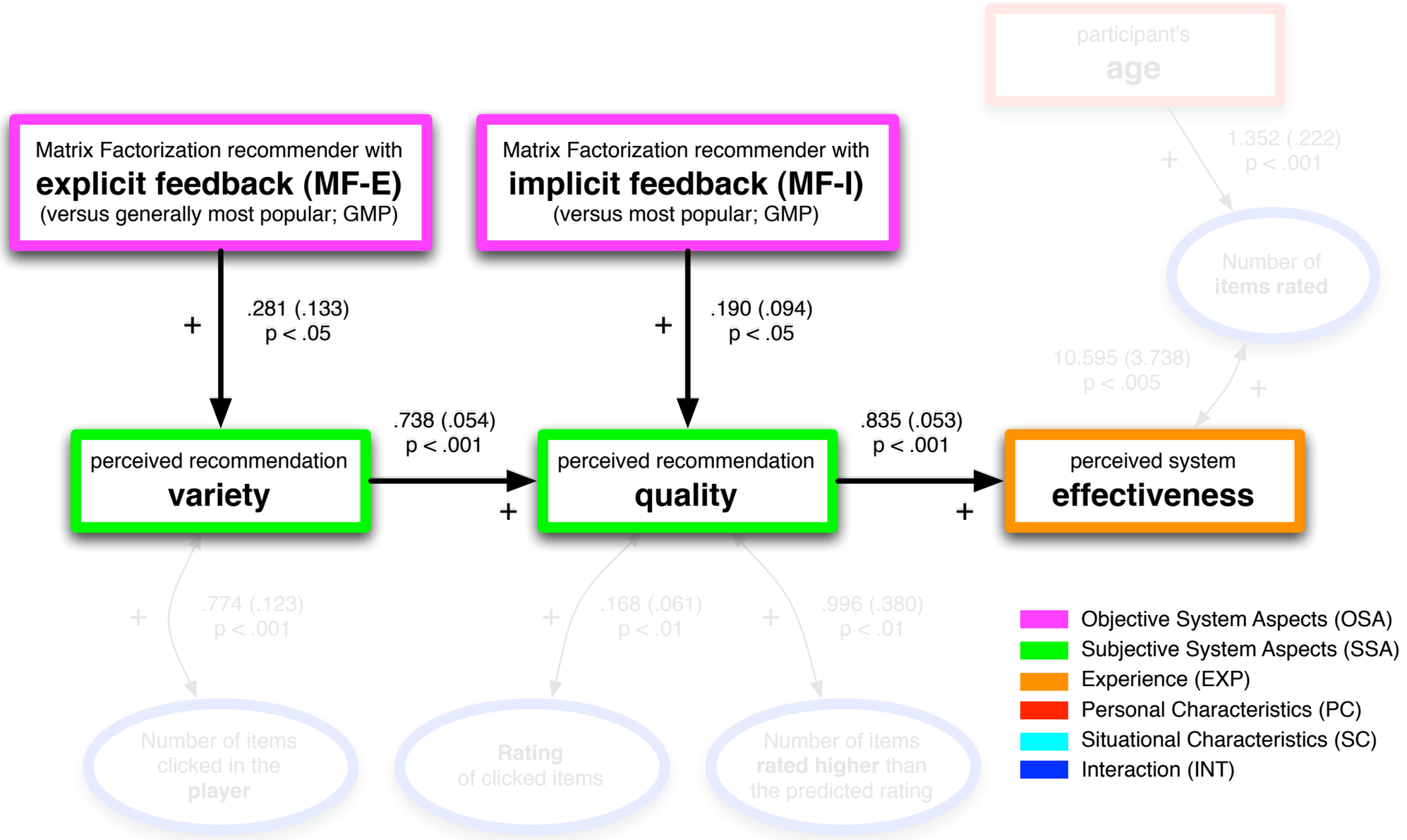
58 participants

Manipulate algorithm/PE method (within subjects)

Measure perceived quality and variety of recommendations, and experience; capture clickstream


Matrix Factorization recommender with
explicit feedback (MF-E)
(versus generally most popular; GMP)
Matrix Factorization recommender with
implicit feedback (MF-I)
(versus most popular; GMP)
participant's
age
1.352 (.222)
p < .001
+
Number of
items rated
10.595 (3.738)
p < .005
+
+
.281 (.133)
p < .05
+
.190 (.094)
p < .05
+
.738 (.054)
p < .001
.835 (.053)
p < .001
perceived recommendation
variety
perceived recommendation
quality
perceived system
effectiveness
+
+
+
.774 (.123)
p < .001
+
.168 (.061)
p < .01
+
.996 (.380)
p < .01
Number of items
clicked in the
player
Rating
of clicked items
Number of items
rated higher than
the predicted rating
Objective System Aspects (OSA)
Subjective System Aspects (SSA)
Experience (EXP)
Personal Characteristics (PC)
Situational Characteristics (SC)
Interaction (INT)


Matrix Factorization recommender with
explicit feedback (MF-E)
(versus generally most popular; GMP)
Matrix Factorization recommender with
implicit feedback (MF-I)
(versus most popular; GMP)
participant's
age
+
1.352 (.222)
p < .001
Number of
items rated
+
.281 (.133)
p < .05
+
.190 (.094)
p < .05
10.595 (3.738)
p < .005
+
perceived recommendation
variety
.738 (.054)
p < .001
+
perceived recommendation
quality
.835 (.053)
p < .001
+
perceived system
effectiveness
+
.774 (.123)
p < .001
+
.168 (.061)
p < .01
+
.996 (.380)
p < .01
Number of items
clicked in the
player
Rating
of clicked items
Number of items
rated higher than
the predicted rating
Objective System Aspects (OSA)
Subjective System Aspects (SSA)
Experience (EXP)
Personal Characteristics (PC)
Situational Characteristics (SC)
Interaction (INT)


Matrix Factorization recommender with
explicit feedback (MF-E)
(versus generally most popular; GMP)
Matrix Factorization recommender with
implicit feedback (MF-I)
(versus most popular; GMP)
participant's
age
1.352 (.222)
p < .001
+
Number of
items rated
10.595 (3.738)
p < .005
+
+
.281 (.133)
p < .05
+
.190 (.094)
p < .05
+
.738 (.054)
p < .001
.835 (.053)
p < .001
perceived recommendation
variety
perceived recommendation
quality
perceived system
effectiveness
+
+
+
.774 (.123)
p < .001
+
.168 (.061)
p < .01
+
.996 (.380)
p < .01
Number of items
clicked in the
player
Rating
of clicked items
Number of items
rated higher than
the predicted rating
Objective System Aspects (OSA)
Subjective System Aspects (SSA)
Experience (EXP)
Personal Characteristics (PC)
Situational Characteristics (SC)
Interaction (INT)

Contributions

The structure of the framework is validated

- With a few exceptions

We developed a number of measurement scales to measure SSAs, EXPs, PCs and SCs

- More have been added in subsequent work

We covered some gaps in existing research

- More details are in the paper itself

Questions…