Modelling user-experience and frequency judgement

Paul van Schaik

p.van-schaik@tees.ac.uk
http://sss-studnet.tees.ac.uk/psychology/staff/Paul_vs/index.htm

27/3/2014, Department of Psychology, University of Fribourg

Outline

- User-experience (UX) and UX-modelling
- Direction of causality of UX-modelling
- Empirical studies from inference perspective
- Frequency judgement
- Experimental studies and modelling

User-experience (UX)

- "Users' judgement of product quality arising from their experience of interaction, and the product qualities which engender effective use and pleasure" (Sutcliffe, 2010)
- Interactive products do not only deliver functional benefits, they promote experiences too
- Users' intention to (re)live positive experiences is an important driver of technology use
- Instrumental and non-instrumental factors in UX (Thüring & Mahlke, 2007)
  - Usability may strongly contribute to negative experiences, if it does not reach a satisfactory level expected by users
  - However, in order to achieve positive experiences, high levels of non-instrumental factors (e.g. positive affect) are needed
- UX models – determinants of positive experiences

Imagine you want to enhance your voice-over-calls with a high-definition image. By coincidence, a local shop makes an exceptional offer (in terms of ‘value for money’) of a multifunctional (‘all-singing-all-dancing’) webcam. Will you accept? The problem is to predict whether or to what extent the product would meet your needs. As you have no hands-on experience, you visit the shop to see for yourself what the product looks like in reality and to get further information from the helpful staff. However, you are not allowed to open the attractive transparent box in which the seductive product patiently awaits your expenditure. You simply cannot try the product before buying it. Therefore, in effect, you try to ‘guess’ – or infer – the product’s reliability, usefulness and ease of use from the specific pieces of information that you find relevant.

Direction of causality in UX-modelling

- Specific-to-general inference/induction
- General-to-specific inference/deduction

Specific-to-general inference

- Overall assessments or attitudes are ‘built’ from the careful consideration, weighting and integration of specific attributes (e.g. usability, aesthetics)
- UX models related to computational, multi-attribute theories of decision-making
- Examples
    - UX model (Hassenzahl, 2003, 2004)
    - Components-of-UX model (Thüring & Mahlke, 2007)
    - Environmental-psychology model of UX (Porat &Tractinsky, 2012)
    - Also van Schaik and Ling (2008, 2011)
- However, should not be taken as the major or even the only inference process!

General-to-specific inference

- Related to non-computational approach to decision-making
- Supported by wealth of evidence (Gigerenzer & Gaissmaier, 2011)
- People use relatively simple strategies
- People infer momentarily hard-to-assess product attributes, even when information is absent or limited
- Inference rules, based on lay theory
- Hassenzahl and Monk's (2010) inference model of UX

General-to-specific inference (2)

- Example 1: price-quality rule
- Example 2: halo effect (“I like it, it must be good on all attributes”),

  so potentially incorrect model specification from results if specific-to-general inference is assumed
- Crucial are (1) notion of inference and (2) careful consideration of how assessments are potentially made in different situations
- No theoretical justification without these

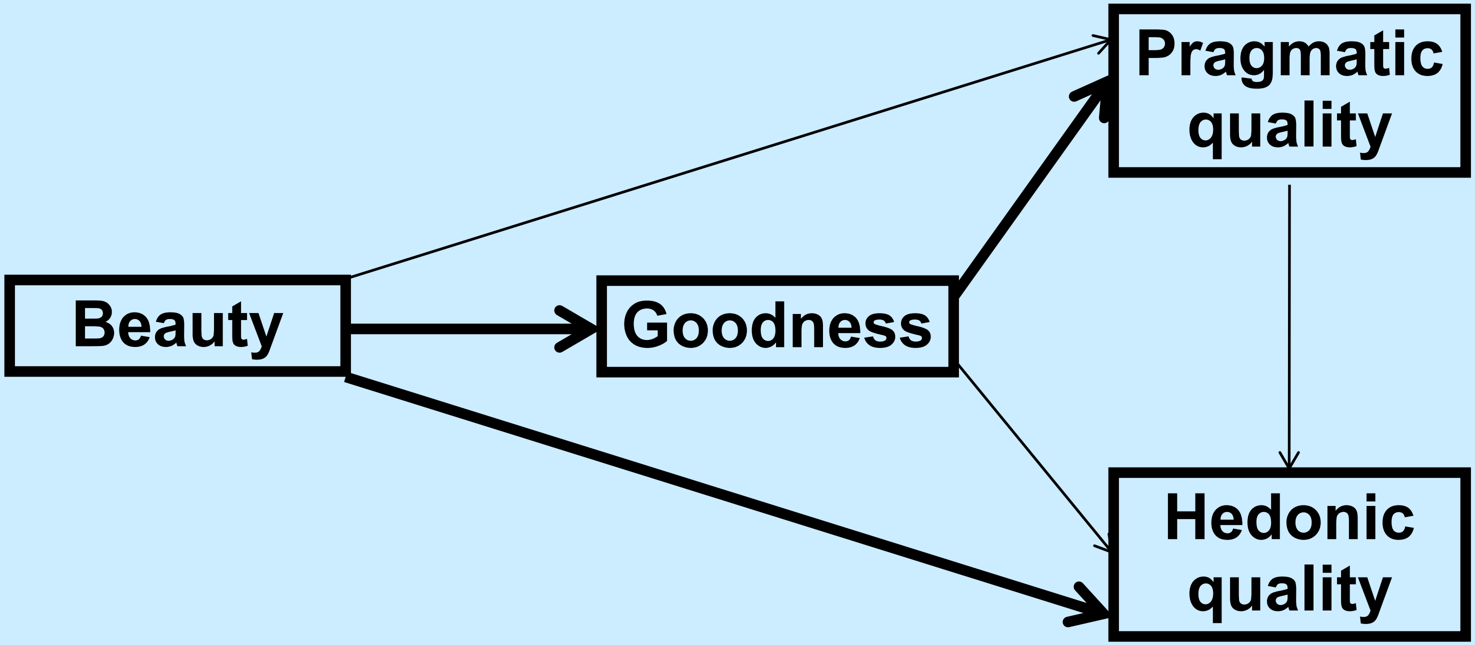
Three studies of UX from an inference perspective

Schaik, P. van, Hassenzahl, M. & Ling, J. (2012). Modeling user-experience from an inference perspective. *ACM Transactions on Computer-Human Interaction*, *19*(2), Article 11.

Inference of UX from a wider perspective: Kruglanski et al.'s (2007) unified framework for conceptualizing and studying judgement as inference

Aims

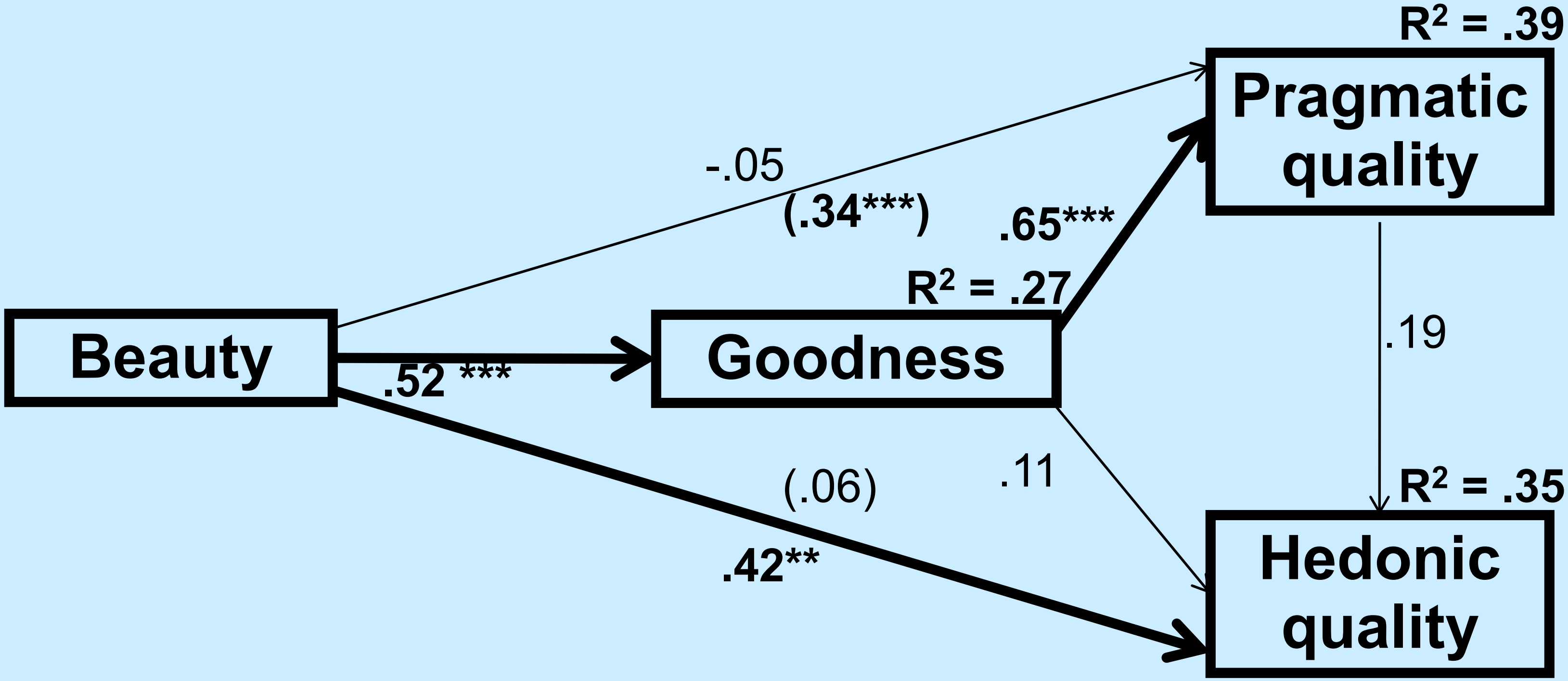
1. Replicate Hassenzahl and Monk's (2010) inference model
2. Explore potential effects of hands-on experience on the model
3. Explore how well the inference model works across different types of experience



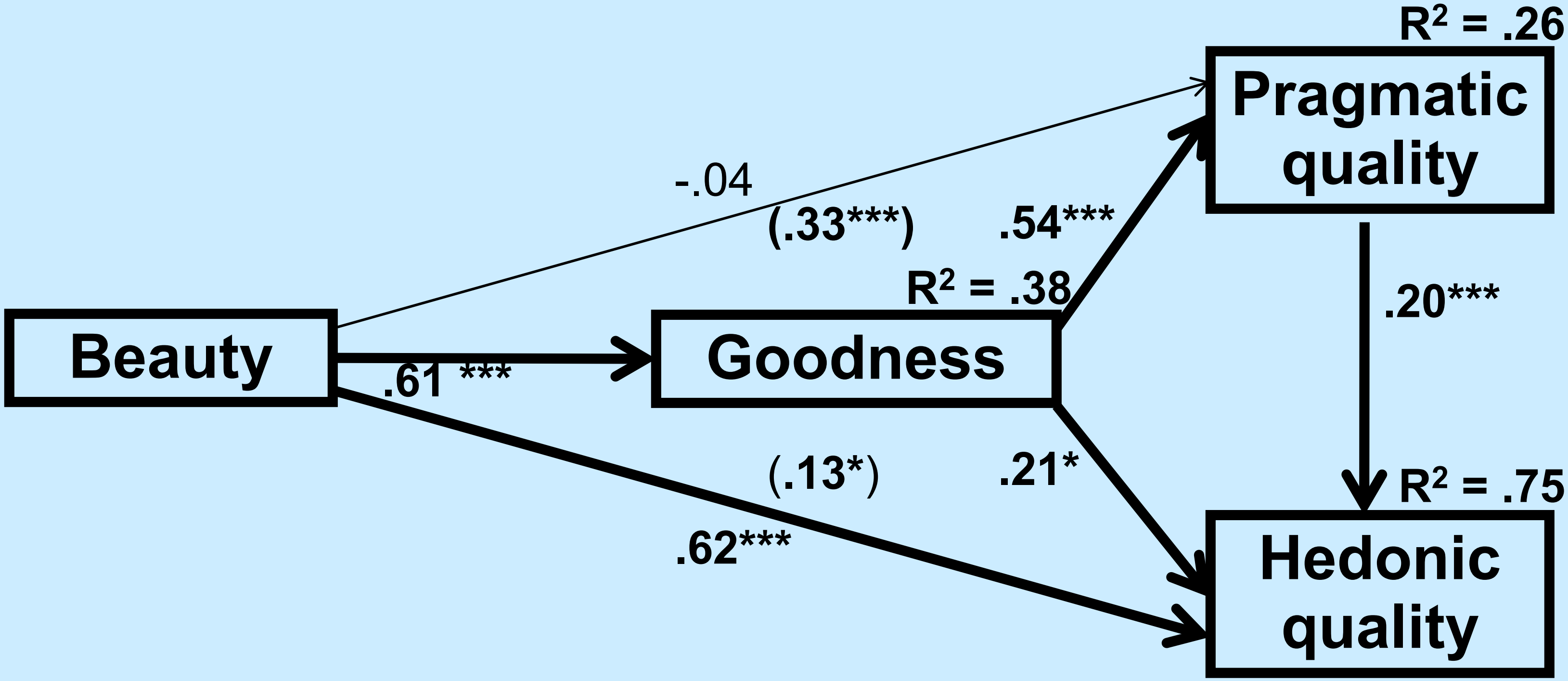
Experiment 1: action mode

- N = 94 undergraduate students (73 females, mean age = 24, SD = 9)
- Wikipedia users
- AttrakDiff2 questionnaire
- Phase 1: viewing screenshots of Wikipedia; then UX rating
- Phase 2: exploring Wikipedia; then UX rating
- Data analysis: PLS path modelling

Before use


R² = .39
Pragmatic quality
-.05
(.34***)
.65***
R² = .27
.19
Beauty
Goodness
.52 ***
(.06)
.11
R² = .35
Hedonic quality
.42**

After use

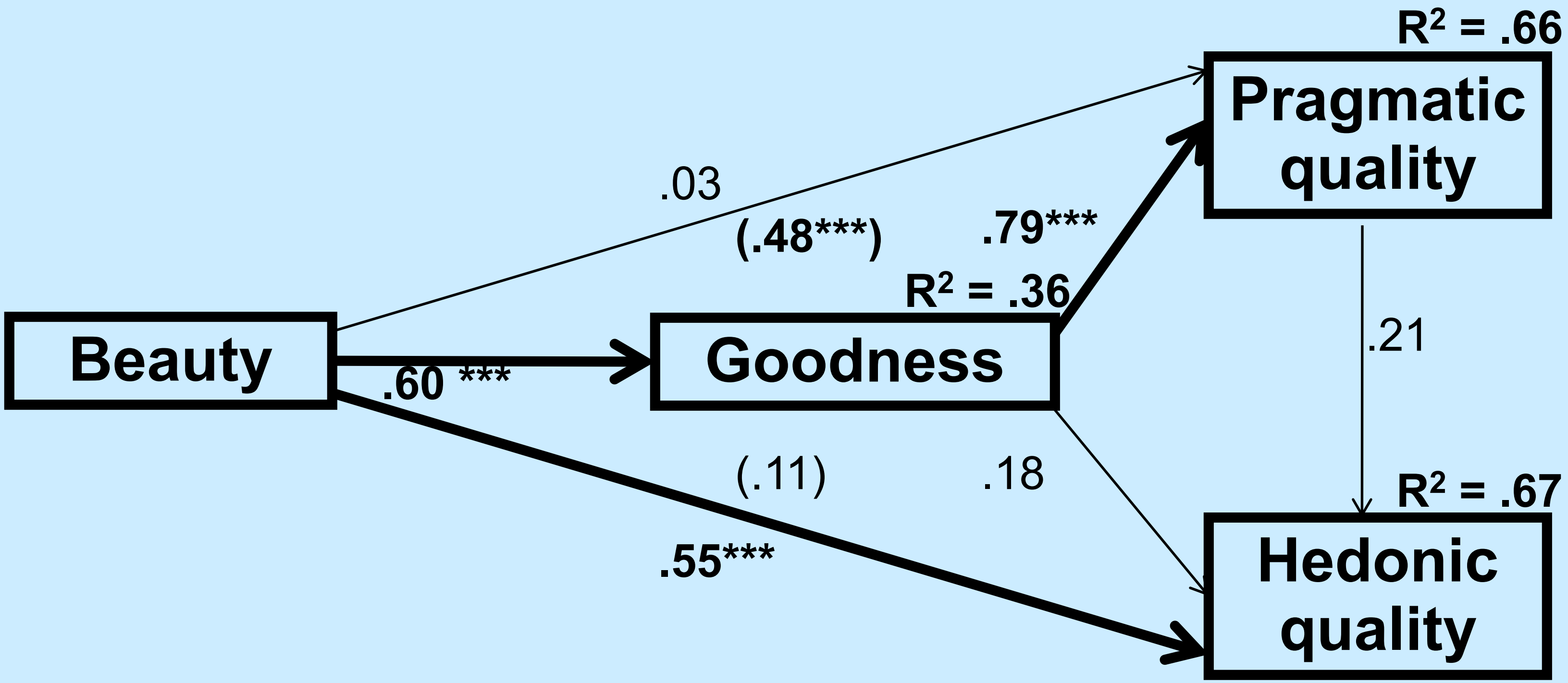


R² = .26
Pragmatic quality
-.04
(.33***)
.54***
R² = .38
.20***
Beauty
Goodness
.61 ***
(.13*)
.21*
R² = .75
.62***
Hedonic quality

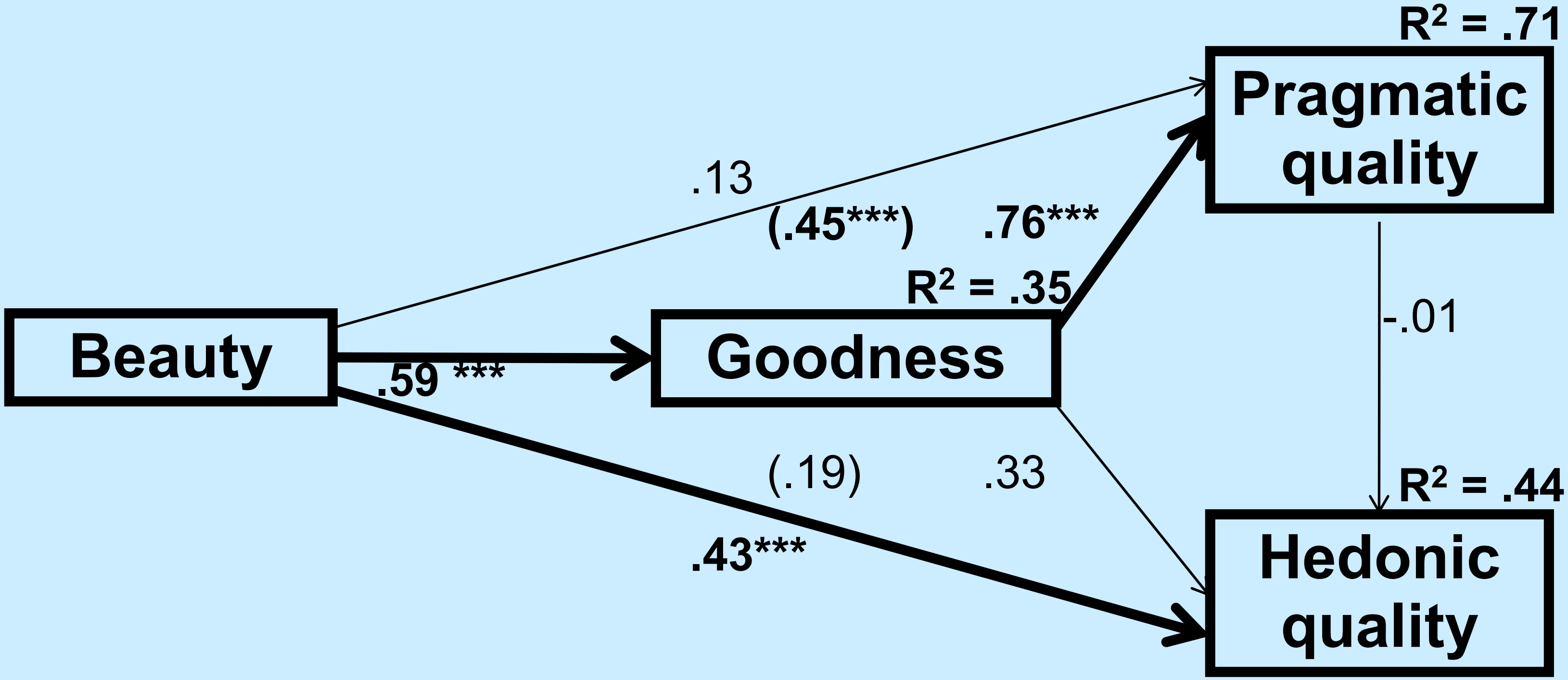
Experiment 2: goal mode

- N = 66 undergraduate students (49 females, mean age = 24, SD = 8)
- Web users
- AttrakDiff2 questionnaire
- Phase 1: viewing screenshots of Manchester City Council site; then UX rating
- Phase 2: retrieving information from site; then UX rating
- Data analysis: PLS path modelling

Before use


R² = .66
Pragmatic quality
.03
(.48***)
.79***
R² = .36
.21
Beauty
Goodness
.60 ***
(.11)
.18
R² = .67
.55***
Hedonic quality

After use



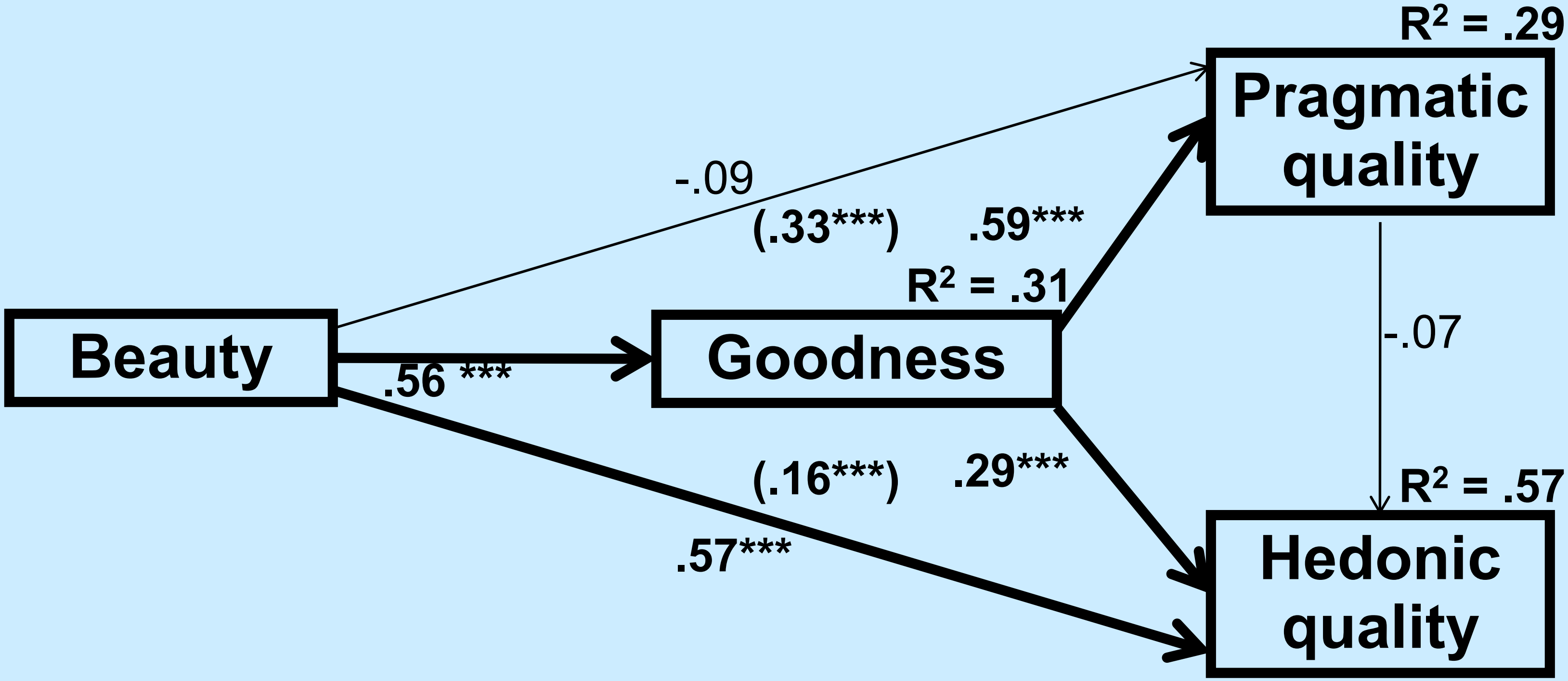
R² = .71
Pragmatic quality
.13
(.45***)
.76***
R² = .35
-.01
Beauty
Goodness
.59 ***
(.19)
.33
R² = .44
.43***
Hedonic quality

Experiment 3: goal mode with varied complexity

- 2-by-2 experimental design (task complexity [2]; artefact complexity [2])
- N = 127 undergraduate students (102 females, mean age = 23, SD = 8)
- Web users
- AttrakDiff2 questionnaire
- Phase 1: viewing screenshots of university course website; then UX rating
- Phase 2: retrieving information; UX rating
- Data analysis: PLS path modelling

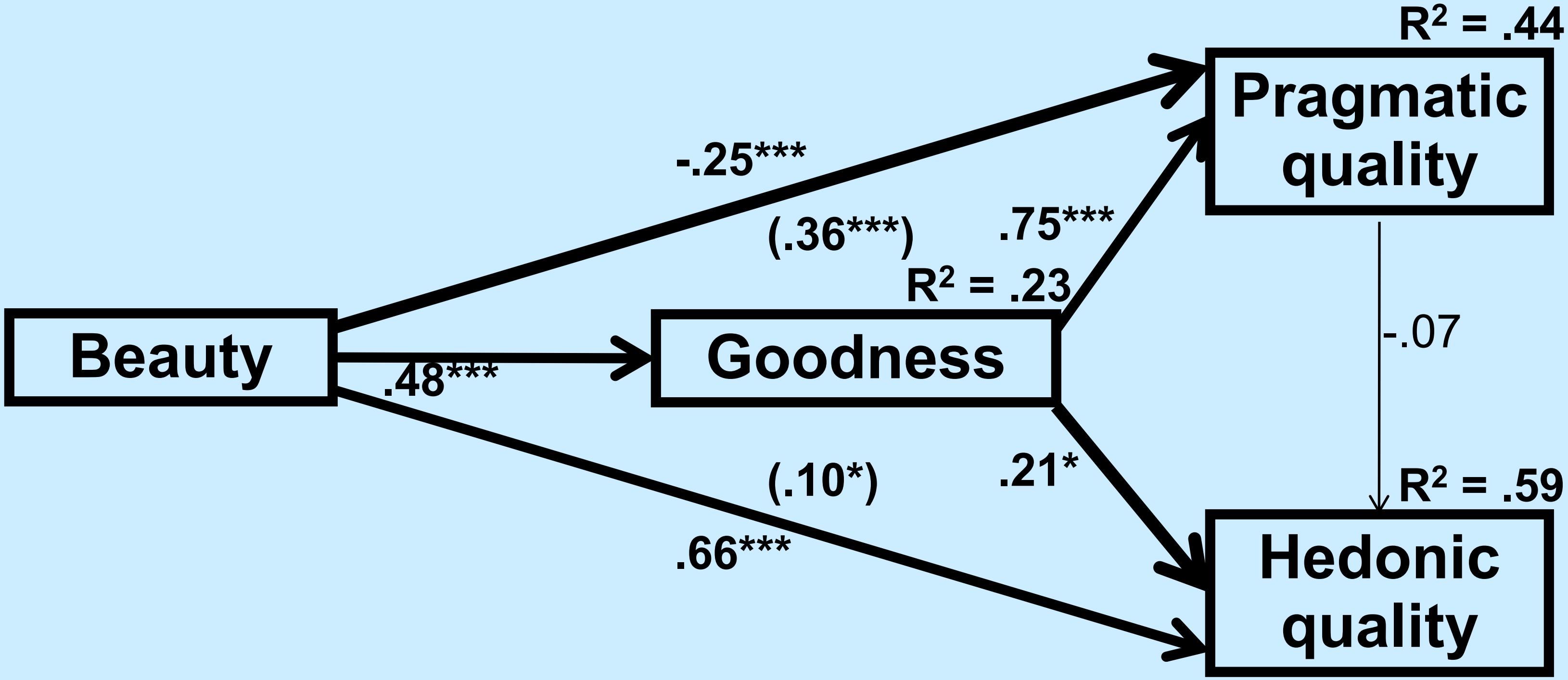
Web site versions

Before use

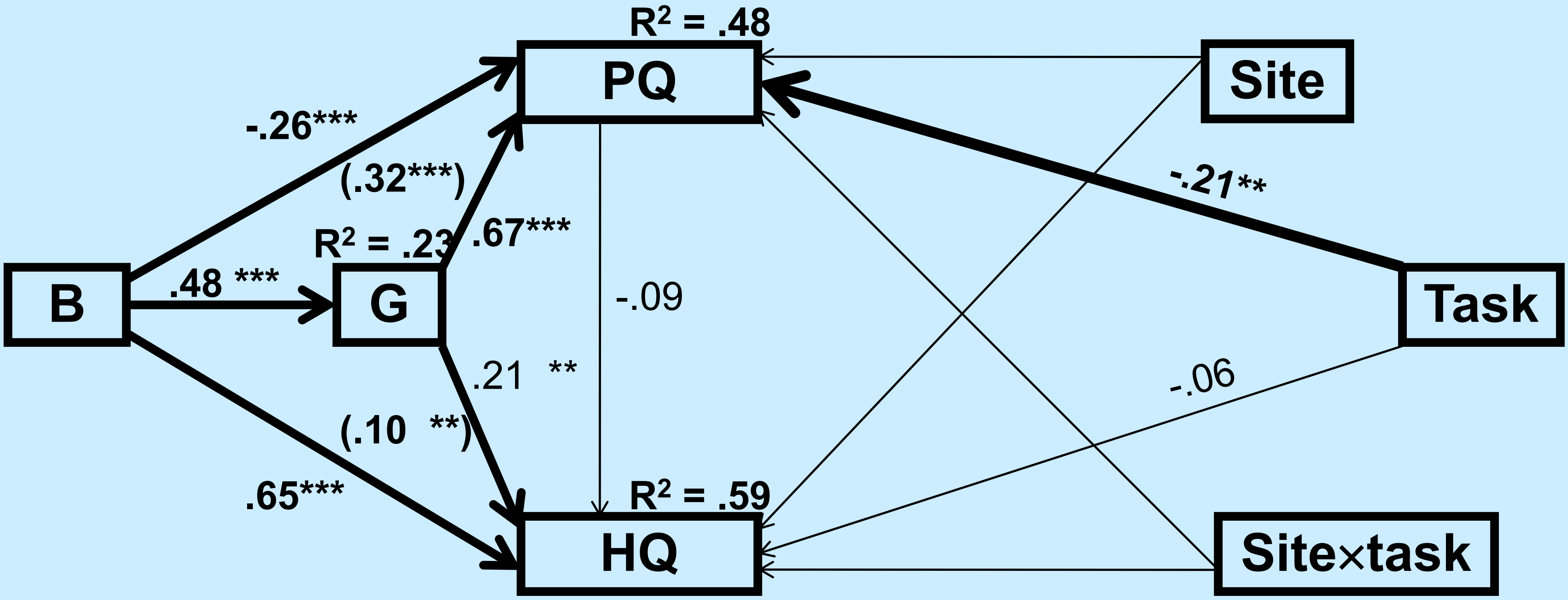


R² = .29
Pragmatic quality
-.09
(.33***)
.59***
R² = .31
-.07
Beauty
Goodness
.56 ***
(.16***)
.29***
R² = .57
.57***
Hedonic quality

After use


R² = .44
Pragmatic quality
-.25***
(.36***)
.75***
R² = .23
-.07
Beauty
Goodness
.48***
(.10*)
.21*
R² = .59
.66***
Hedonic quality

After use


R² = .48
PQ
Site
-.26***
(.32***)
-.21**
R² = .23
.67***
B
.48 ***
G
-.09
Task
.21 **
-.06
(.10 **)
.65***
R² = .59
HQ
Site×task

Discussion (Aim 1)

- Three studies supported our specific inference model
- Beauty and overall evaluation were highly correlated, confirming the longstanding inference rule of "What is beautiful is good" (Dion et al., 1972)
- Effect of beauty on hedonic quality was primarily direct (probabilistic consistency as an inference rule), but
- Effect of beauty on pragmatic quality was primarily indirect (evaluative consistency as an inference rule), in other words, mediated by goodness

Discussion (Aims 2 and 3)

- Evidence for inference rules when hands-on experience was experimentally controlled
- Evidence for the suggested inference rules

1. across two types of task (goal mode and action mode)
2. within different products (Wikipedia, council website, university course website) and
3. even when task complexity and artefact complexity were systematically varied
4. Our findings thus increase external validity

Discussion(3)

- **Beauty and pragmatic quality: compensatory inference**
- **Beauty and hedonic quality: evaluative and probabilistic consistency combined**
- **Pragmatic quality and hedonic quality:**
  - independence between pragmatic and hedonic quality may be less strong when the focus is on the action itself (‘action mode’; Hassenzahl, 2003) rather than on achieving goals
  - This is because in such a situation, the interaction itself could to some extent be a source of pleasure

Inference of UX from a wider perspective

- **Computational versus non-computational models**
- **Kruglanski et al.'s (2007) unified framework for conceptualizing and studying judgment as inference**
- **Information sources**
  - Impression from the presentation of a product
  - Hands-on experience from of subsequent interaction with the product
  - Memory of previous product experience
- **Judgement parameters of inference-based judgement**
  - Informational relevance
  - Task demands
  - Cognitive resources
  - Motivation: both non-directional (effort) and directional (bias)

Conclusion

- UX-modelling
  - develop cumulative knowledge
  - basis for UX-engineering
- Flexibility of model specification on theoretical and practical grounds is essential
- Direction of causality is crucial
- Example 1: cognitive-experiential model when task performance is important (van Schaik & Ling, 2012a, 2012b)
- Example 2: general-to-specific UX inference (van Schaik et al., 2012)

Modelling frequency judgement (1)

- **How do people make judgements based on sequential information?**
- **Frequency judgement**
  - Enumeration
  - Ease-of-retrieval heuristic
  - Automatic encoding

Modelling frequency judgement (2)

- **Research question:**
  do people use a frequency strategy that uses information about the sequence pattern?
- **Rationale:**
  in making frequency judgements, people are constrained by limitations of information-processing and memory; tend to use strategies that minimise cognitive load

Published research

- Kusev, P., Ayton, P. Schaik, P. van, Tsaneva-Atanasova, K., Stewart, N. & Chater, N. (2011). Judgments relative to patterns: how temporal sequence patterns affect judgments and memory. *Journal of Experimental Psychology: Human Performance and Perception*, *37*, 1874-1886.
- Kusev, P., Tsaneva-Atanasova, K., Schaik, P. van & Chater, N. (2012). Modelling judgement of sequentially presented categories using weighting and sampling without replacement. *Behavior Research Methods*, *44*, 1129-1134.

Some sequences

XOXOOOOXXXOOOOOXXXXOOXXXOXOX

XOXOXXXOOXXXXOOOOOXXXOOOOXOX

LHHHHHHHLHHLHHLHHLLLLLLLLLHL

HHHHHHHHHHHHHHHLLLLLLLLLLLLLLLLLLLLLHHHHHH

LLLLLLHHHHHHHHHHHHHHHHHHHHHLLLLLLLLLLLLLLL

GRGGGGGGGGGRRGRRGRRGRRRRRRG

CTCTTTTTTTTTTTCCCCCCCCCCTCTC

RGRRRRRRRRRGGRGGRGGRGGGGGGR

TCTCCCCCCCCCCCTTTTTTTTTTCTCT

GRGRRGRGRGRGGGGGGR

GRGRGRGRGRGGGGGGGGGGRRRRRRRRRR

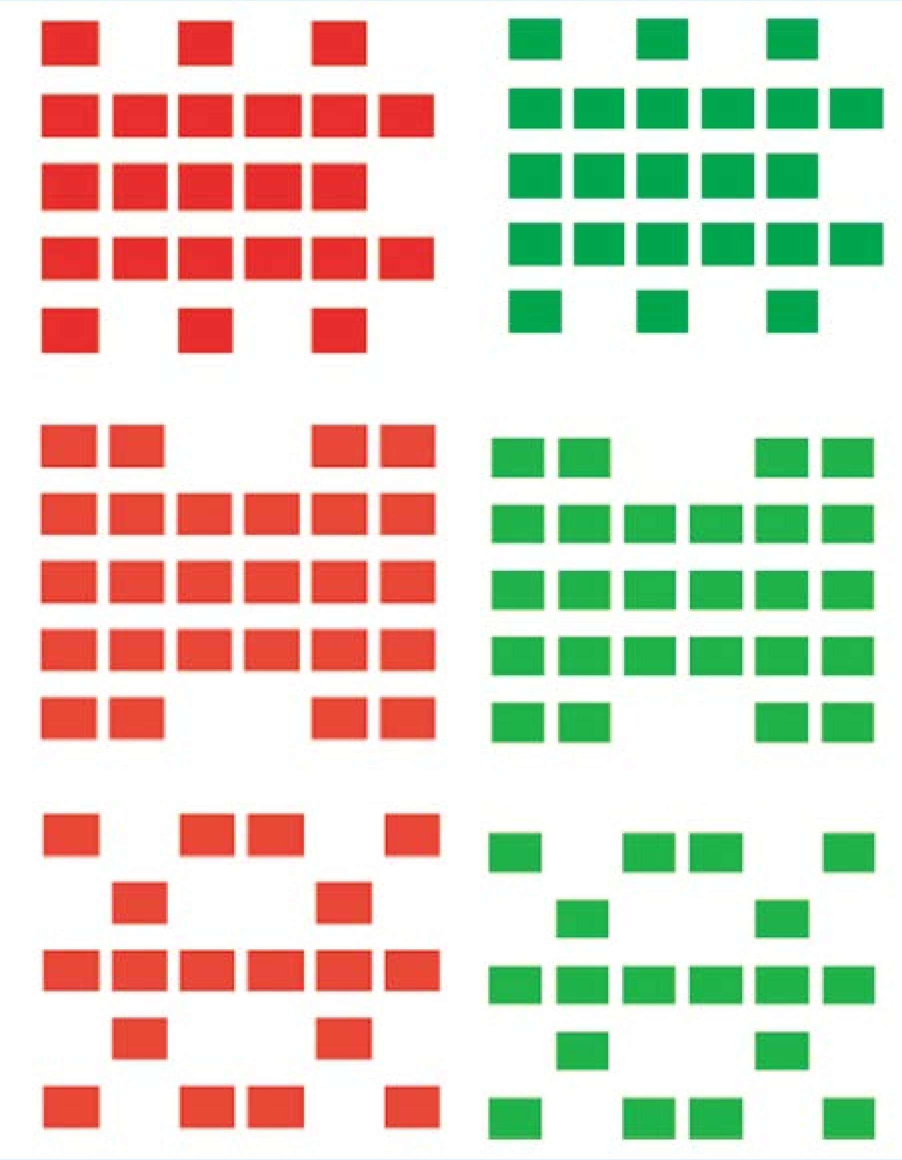
CACCCCCCCACCACCACCAAAAAAAAAAAC

Experiment series

- Visual modality, category defined by colour: Experiments 1, 2CD, 3AB, 4AB, 5AB
- Auditory modality, category defined by sound frequency: Experiments 2AB
- Visual modality, category defined by shape: Experiment 2D
- Visual modality, category defined by concept: Experiment 6

Experiment 1 (1)

- Sequences: random alternation of stimuli from two categories (red, green)
- Sequence length: 30
- Response: frequency judgement



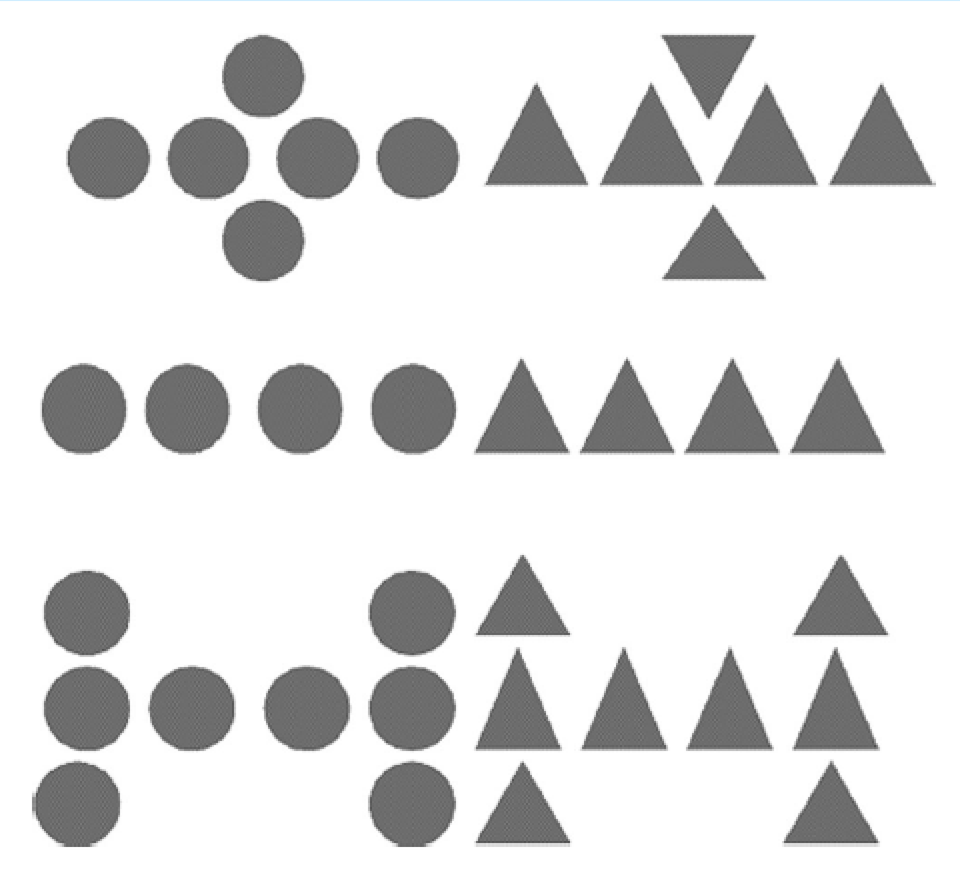
Experiment 1 (2)

N	Sequence characteristic	Identification of sequence characteristic	Value of sequence characteristic
1	First-run category	RGRRRRGRGRRGGGGGGGGGRRRRGRRGGR	RED
2	Last-run category	RGRRRRGRGRRGGGGGGGGGRRRRGRRGGR	GREEN
3	Length of the first run	RGRRRRGRGRRGGGGGGGGGRRRRGRRGGR	4
4	Length of the last run	RGRRRRGRGRRGGGGGGGGGRRRRGRRGGR	2
5	Number of runs (first-run category)	RGRRRRGRGRRGGGGGGGGGRRRRGRRGGR	4
6	Number of runs (non–first-run category)	RGRRRRGRGRRGGGGGGGGGRRRRGRRGGR	2
7	Single stimuli (first-run category)	RGRRRRGRGRRGGGGGGGGGRRRRGRRGGR	3
8	Number of single stimuli (non–first-run category)	RGRRRRGRGRRGGGGGGGGGRRRRGRRGGR	4
9	Number of single stimuli (first-run category) before the first run	RGRRRRGRGRRGGGGGGGGGRRRRGRRGGR	1
10	Number of single stimuli (non–first-run category) before the first run	RGRRRRGRGRRGGGGGGGGGRRRRGRRGGR	1
11	Number of single stimuli (first-run category) after the last run	RGRRRRGRGRRGGGGGGGGGRRRRGRRGGR	1
12	Number of single stimuli (non–first-run category) after the last run	RGRRRRGRGRRGGGGGGGGGRRRRGRRGGR	0
13	Number of runs (last-run category)	RGRRRRGRGRRGGGGGGGGGRRRRGRRGGR	2
14	Number of runs (non–last-run category)	RGRRRRGRGRRGGGGGGGGGRRRRGRRGGR	4
15	Number of single stimuli (last-run category)	RGRRRRGRGRRGGGGGGGGGRRRRGRRGGR	4
16	Number of single stimuli (non–last-run category)	RGRRRRGRGRRGGGGGGGGGRRRRGRRGGR	3
17	Average run length of first-run category	RGRRRRGRGRRGGGGGGGGGRRRRGRRGGR	3
18	Average run length of non–first-run category	RGRRRRGRGRRGGGGGGGGGRRRRGRRGGR	5.5
19	Average run length of last-run category	RGRRRRGRGRRGGGGGGGGGRRRRGRRGGR	5.5
20	Average run length of non–last-run category	RGRRRRGRGRRGGGGGGGGGRRRRGRRGGR	3
21	Charateristic 9 + Charateristic 10	RGRRRRGRGRRGGGGGGGGGRRRRGRRGGR	2
22	Charateristic 11 + Charateristic 12	RGRRRRGRGRRGGGGGGGGGRRRRGRRGGR	1
23	Sequence length	RGRRRRGRGRRGGGGGGGGGRRRRGRRGGR	30

Experiment 1 (3)

- **Analysis:**
  stepwise multiple regression to determine best predictor of judgement
- **Result:**
  only significant predictor is first-run category, $F(1, 76) = 23.30$, $R^2 = 0.24$, $p < 0.001$

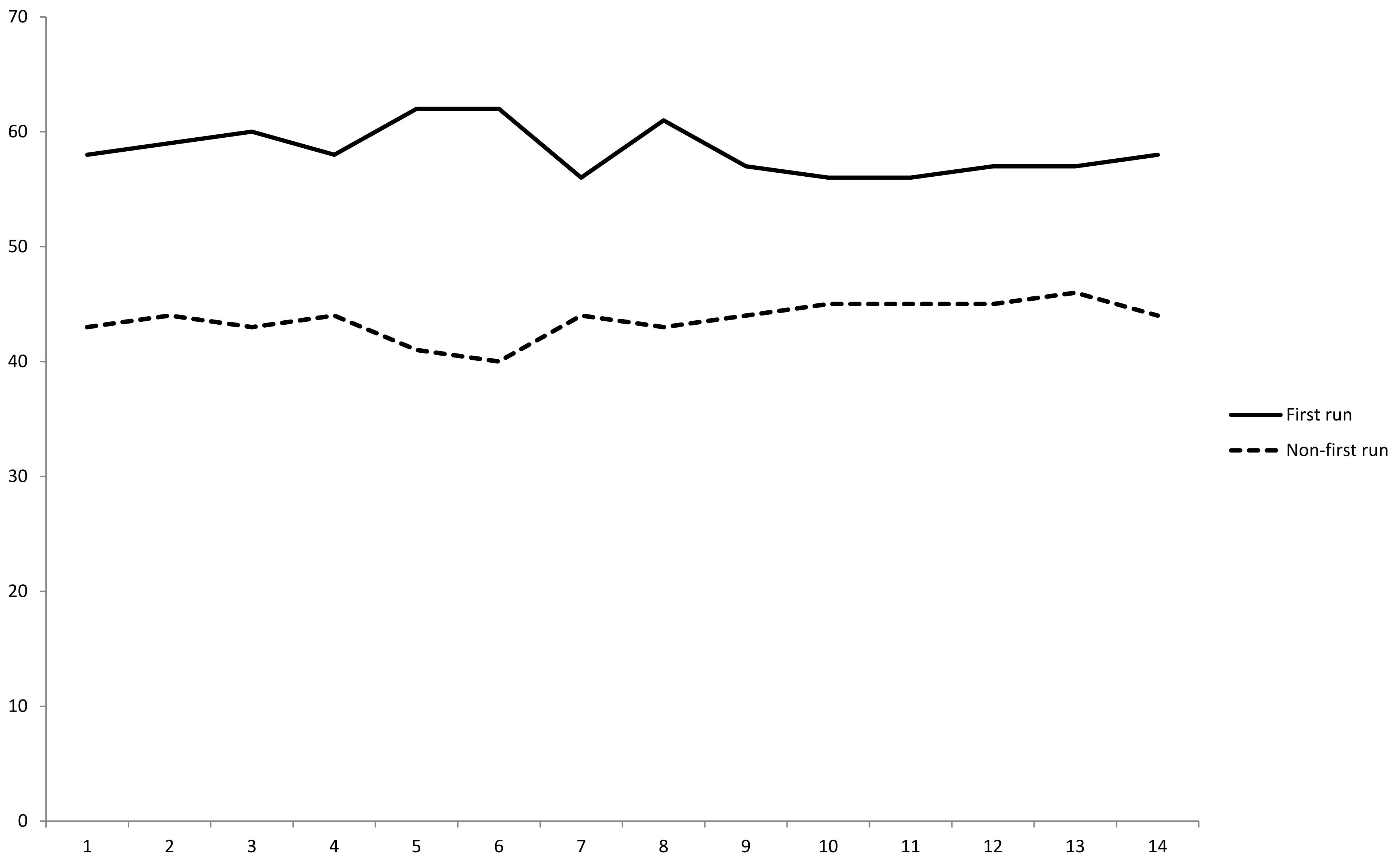
Experiment 2D



Experiments 2 and 3

Experiment	Sequence	First-run Mean	First-run *SD*	First-run 95% CI	Non–first-run Mean	Non–first-run *SD*	Non–first-run 95% CI
		Category					
		First-run			Non–first-run		
Experiment	Sequence	Mean	*SD*	95% CI	Mean	*SD*	95% CI
2A	LHHHHHHHLHHLHHLHHLLLLLLLLLHL	58	12	[53, 63]	43	7	[40, 46]
2A	HLLLLLLLHLLHLLHLLHHHHHHHHHLH	59	12	[54, 65]	44	9	[41, 48]
2B	LLLLLLLLLLLLLLLHHHHHHHHHHHHHHHHHHHHLLLLLL	60	12	[53, 66]	43	7	[39, 47]
2B	LLLLLLHHHHHHHHHHHHHHHHHHHHLLLLLLLLLLLLLLL	58	10	[52, 63]	44	8	[38, 48]
2B	HHHHHHHHHHHHHHHLLLLLLLLLLLLLLLLLLLLHHHHHH	62	13	[55, 70]	41	9	[36, 46]
2B	HHHHHHLLLLLLLLLLLLLLLLLLLLHHHHHHHHHHHHHHH	62	12	[55, 69]	40	9	[35, 45]
2C	GRGGGGGGGGGRRGRRGRRGRRRRRRRG	56	7	[53, 60]	44	8	[40, 49]
2C	RGRRRRRRRRRGGRGGRGGRGGGGGGGR	61	12	[54, 67]	43	11	[36, 49]
2D	CTCTTTTTTTTTTTCCCCCCCCCCTCTC	57	9	[52, 61]	44	7	[40, 47]
2D	TCTCCCCCCCCCCCTTTTTTTTTTCTCT	56	8	[52, 60]	45	9	[39, 49]
3A	GRGGGGGGGGGRRGRRGRRGRRRRRRRG	56	7		45	9	
3A	RGRRRRRRRRRGGRGGRGGRGGGGGGGR	57	7		45	12	
3B	CTCTTTTTTTTTTTCCCCCCCCCCTCTC	57	9		46	8	
3B	TCTCCCCCCCCCCCTTTTTTTTTTCTCT	58	11		44	10	

Note. Numbers are respondents' estimates of the percentage of stimuli in the first-run and the non–first-run categories. L = low frequency; H = high frequency; G = green; R = red; C = circle; T = triangle.



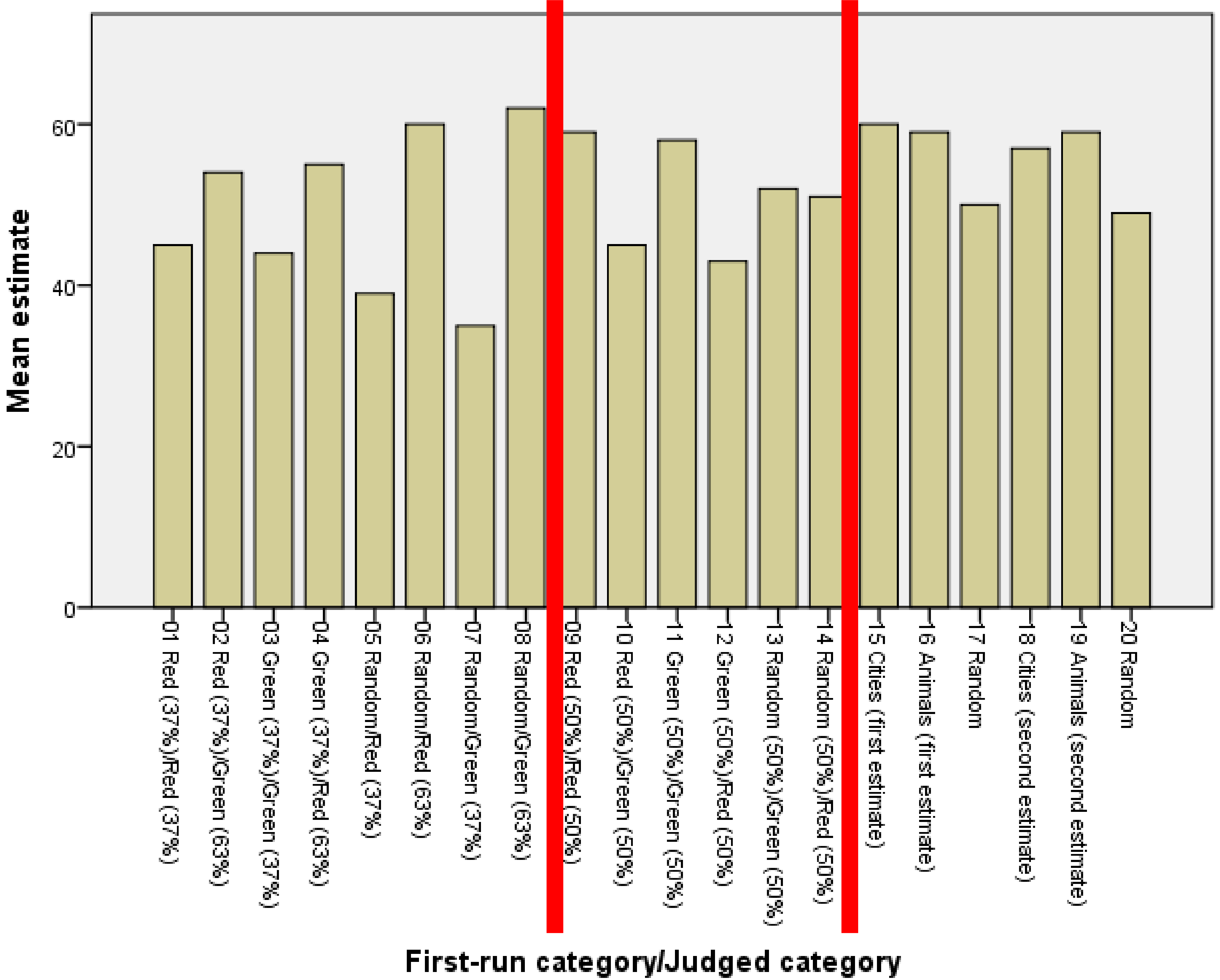
70
60
50
40
30
20
10
0
1
2
3
4
5
6
7
8
9
10
11
12
13
14
First run
Non-first run

Experiments 4, 5 and 6

Sequences and Test of Frequency Estimate Against the Actual Proportion (Experiments 4, 5, and 6)

Experiment	Sequence	First-run category	Judged category	Mean	*SD*	*t*	*df*	*p*	*r*
4A	GRGRRGRGRGRGGGGGGGR	Red (37%)	Red (37%)	45	9	3.43	16	<.01	.65
4A	GRGRRGRGRGRGGGGGGGR	Red (37%)	Green (63%)	54	8	−3.99	16	<.01	.71
4A	RGRGGRGRGRGRRRRRRRG	Green (37%)	Green (37%)	44	9	3.02	16	<.01	.60
4A	RGRGGRGRGRGRRRRRRRG	Green (37%)	Red (63%)	55	9	−3.54	16	<.01	.66
4B		Random	Red (37%)	39	6	1.04	16	.31	.25
4B		Random	Red (63%)	60	8	−1.49	16	.15	.35
4B		Random	Green (37%)	35	6	−1.25	16	.23	.30
4B		Random	Green (63%)	62	10	−0.41	16	.68	.10
5	RGRGRGRGRGRRRRRRRRRGGGGGGGGG	Red (50%)	Red (50%)	59	12	3.36	29	<.01	.61
5	RGRGRGRGRGRRRRRRRRRGGGGGGGGG	Red (50%)	Green (50%)	45	9	−2.64	29	<.01	.52
5	GRGRGRGRGRGGGGGGGGGRRRRRRRRR	Green (50%)	Green (50%)	58	9	4.00	29	<.01	.68
5	GRGRGRGRGRGGGGGGGGGRRRRRRRRR	Green (50%)	Red (50%)	43	8	−4.36	29	<.001	.71
5		Random (50%)	Green (50%)	52	11	0.78	29	.45	.18
5		Random (50%)	Red (50%)	51	9	0.65	29	.52	.15
6	CACCCCCCCACCACCACCAAAAAAAAAAAC	Cities (first estimate)		60[a]	8	6.77	29	<.001	.77
6	ACAAAAAAAACAACAACAACCCCCCCCCCA	Animals (first estimate)		59[a]	9	5.93	29	<.001	.73
6		Random		50[b]	12	−0.20	29	.85	.04
6	CACCCCCCCACCACCACCAAAAAAAAAAAC	Cities (second estimate)		57[a]	10	4.03	29	<.001	.59
6	ACAAAAAAAACAACAACAACCCCCCCCCCA	Animals (second estimate)		59[a]	9	5.41	29	<.001	.70
6		Random		49[b]	12	−0.50	29	.62	.09

Note. Means are estimated percentages of the categories. G = green; R = red; C = city; A = animal.
[a] Estimated percentage of the category with first run. [b] Estimated percentage of the response category.



Mean estimate
0
20
40
60
01 Red (37%)/Red (37%)
02 Red (37%)/Green (63%)
03 Green (37%)/Green (37%)
04 Green (37%)/Red (63%)
05 Random/Red (37%)
06 Random/Red (63%)
07 Random/Green (37%)
08 Random/Green (63%)
09 Red (50%)/Red (50%)
10 Red (50%)/Green (50%)
11 Green (50%)/Green (50%)
12 Green (50%)/Red (50%)
13 Random (50%)/Green (50%)
14 Random (50%)/Red (50%)
15 Cities (first estimate)
16 Animals (first estimate)
17 Random
18 Cities (second estimate)
19 Animals (second estimate)
20 Random
First-run category/Judged category

Experiment 6 – dissociation (1)

- **3(sequence)×2(category of estimate) ANOVA**
  - DV frequency estimate1/2: only sequence significant
  - DV recall city/animal: all effects non-significant

- **Frequency estimate stable from FE1 to FE2, $ICC = 0.89$, $p < 0.001$**
- **r(recall FR category, recall non-FR category) = 0.60, $p < 0.001$**
- **r(recall, frequency estimate): all $p > 0.05$ or $p >> 0.05$**

Experiment 6 – dissociation (2)

- 3(sequence)×(2)(recall category) ANOVA: significant interaction

Category of recall	First-run category: City	Animal	Random
City			
M	52	64	59
SD	20	19	21
Animal			
M	60	59	60
SD	20	17	21

Note. Numbers are percentages of correct recalls.

- r^2(total recall, recall FR category) = 0.20
- r^2(total recall, recall LR category) = 0.64

Modelling the first-run effect

- Probability of Category A appearing for the first time within the first run in Position x, $P_{n_x}{}^{(1)} = \Pr\{A_{n_x}{}^{(1)}\} = \frac{n_x}{N_x}$

- $$P_{n_x}{}^{(i)} = \frac{\Gamma(n_x+1)\Gamma(N_x+1-i)}{\Gamma(n_x+1-i)\Gamma(N_x+1)}$$

- Frequency of initial repeated items in the first run makes a larger contribution to frequency estimate than do later items

$$FJ_l = 100 \times \sum_{i=1}^{l} \frac{1}{i^w} P_{n_x}{}^{(i)}$$

- w = 1.73

Conclusion

- **First-run effect: frequency is overestimated of a given category of event when category is the first repeated category in sequence**
- **Dissociation between frequency judgement and recall**

  Dissociation does not fit established theoretical accounts – no direct relationship between memory and judgement

- **Results are consistent with results of 60 years of research:**
  people are unable to classify or estimate objects independently of preceding context
- **Simple mathematical model accounts for the first-run effect**

General conclusion

- **Judgements are constrained by limitations of information-processing and memory**
- **Judgements may be modelled by relatively simple rules (e.g., Kruglanski, Gigerenzer)**
  - Evaluative consistency
  - Probabilistic consistency
  - Compensatory inference
  - First-run (with decreasing influence of later items)

Questions?