Modeling Task Complexity in Crowdsourcing

Jie Yang, Judith Redi,
Gianluca Demartini, Alessandro Bozzon

Task Properties in Crowdsourcing

- Studying ***task properties*** is key for addressing core crowdsourcing problems such as *task assignment*, *worker retention* and *reliability* ***(Yang and Bozzon 2016)***

Jie Yang, and Alessandro Bozzon. On the improvement of quality and reliability of trust cues in micro-task crowdsourcing. In TRUSTINCW Workshop at ACM WebSci, 2016.

Task Properties in Crowdsourcing

- Studying *task properties* is key for addressing core crowdsourcing problems such as *task assignment*, *worker retention* and *reliability* *(Yang and Bozzon 2016)*

- Work exists on properties such as *compensation* and *execution time*, which are typically related (piecework?)

Jie Yang, and Alessandro Bozzon. On the improvement of quality and reliability of trust cues in micro-task crowdsourcing. In TRUSTINCW Workshop at ACM WebSci, 2016.

Task Properties in Crowdsourcing

- Studying *task properties* is key for addressing core crowdsourcing problems such as *task assignment*, *worker retention* and *reliability* *(Yang and Bozzon 2016)*

- Work exists on properties such as *compensation* and *execution time*, which are typically related (piecework?)

- What about **complexity**?

Jie Yang, and Alessandro Bozzon. On the improvement of quality and reliability of trust cues in micro-task crowdsourcing. In TRUSTINCW Workshop at ACM WebSci, 2016.

Complexity

—— one of the most important task properties

Complexity

—— one of the most important task properties

- Depends on

Complexity

—— one of the most important task properties

- Depends on
    - Intrinsic property of a task

Complexity

—— one of the most important task properties

- Depends on
  - Intrinsic property of a task
  - + Individual preferences of a doer

Complexity

—— one of the most important task properties

- Depends on
  - Intrinsic property of a task
  - + Individual preferences of a doer
- **Perceived task complexity** can influence the task selection strategy of workers, as well as the quality of their performance

Research Question

Can we measure (and predict) perceived task complexity based on task design characteristics?

Modeling Task Complexity

Observe subjective perception of task complexity

- Instantiate a bunch of different tasks
- Ask workers to carry them out and evaluate their complexity

Map features into subjective perception

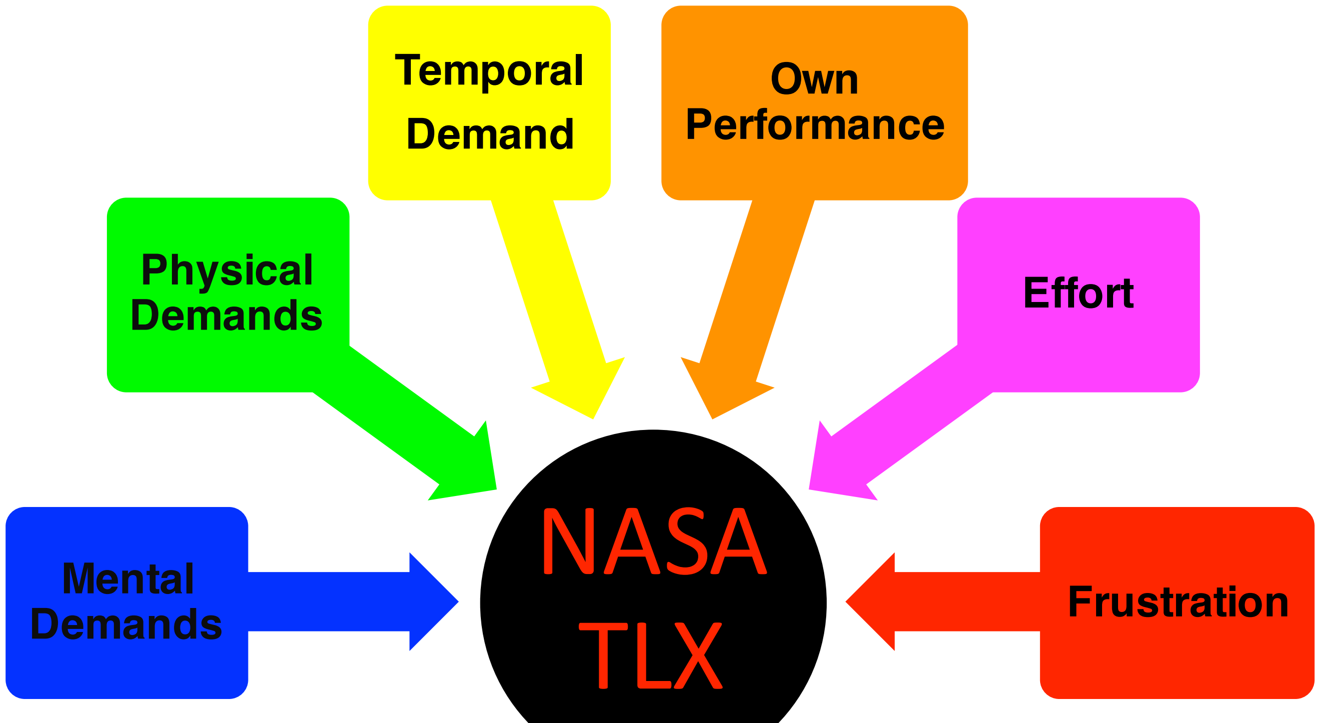
Regression

Design a set of features computable from the task

Metadata
Semantics
Visual

Subjective Task Complexity Evaluation

- Perception of the level of complexity **associated with** the action of performing of a task
- Crowdsourcing experiment to **measure** subjective task complexity
- NASA Task Load Index (TLX): **complexity factors** and **weight**
  - Overall complexity = weighted sum of all complexity factors



Tasks

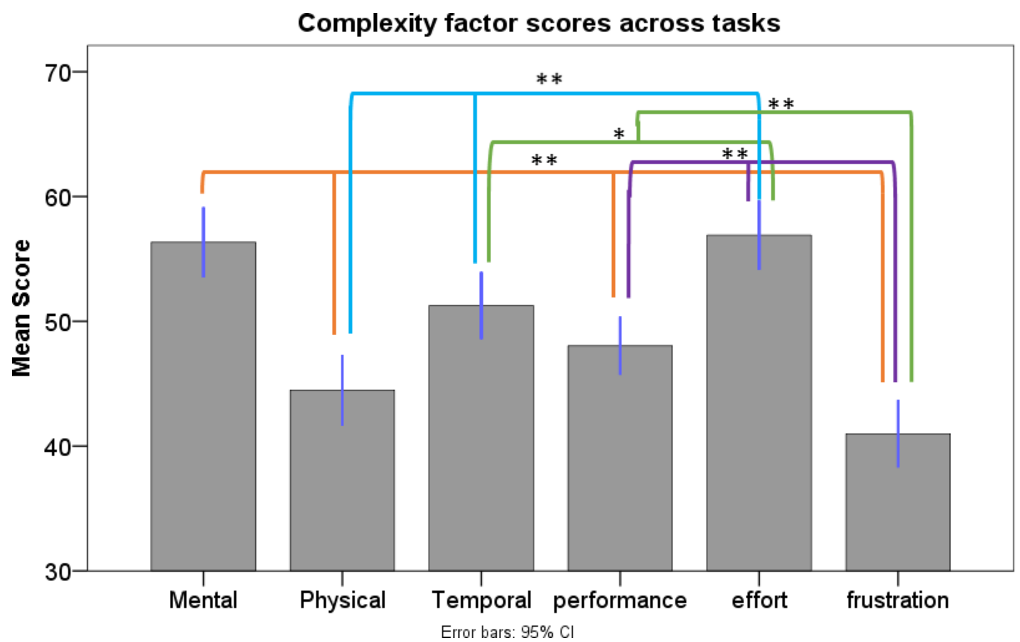
- Dataset of 61 real mTurk Tasks
  - Crawled through extension of the mTurk-tracker to retrieve metadata, formatting (JS, CSS) and MM content if applicable
  - One week observation: from each requester 1 task per type

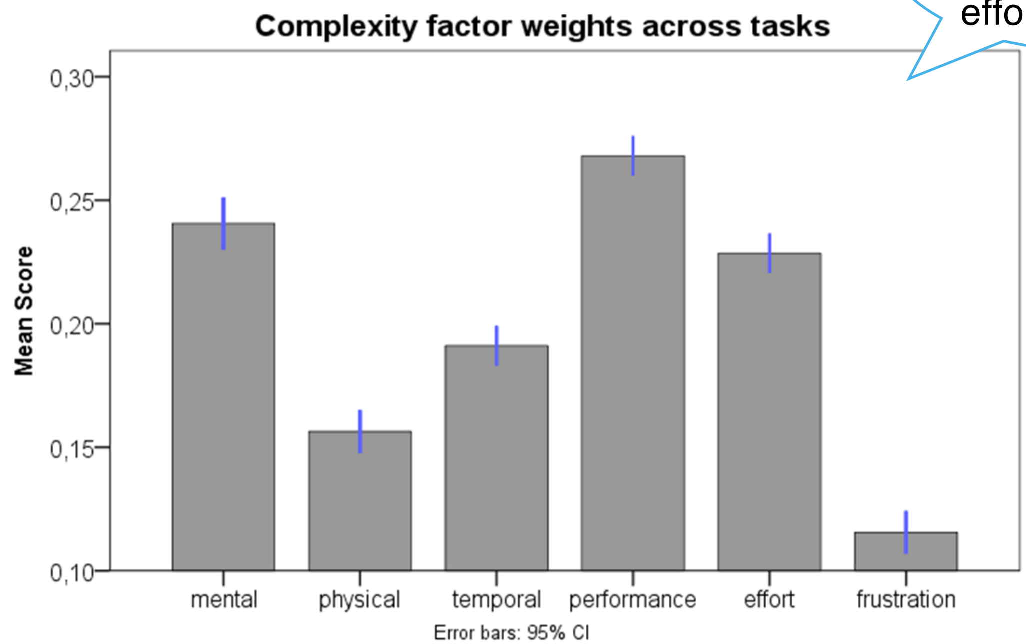
Task Type	Count	Percentage
Survey (SU)	4	7%
Content Creation (CC)	19	31%
Content Access (CA)	4	7%
Interpretation and Analysis (IA)	17	28%
Verification and Validation (VV)	2	3%
Information Finding (IF)	14	23%
Other	1	2%

Experimental Setup

- **Protocol**: Task execution **+** TLX (referred to the task just executed)
  - Evaluation Tasks were **re-instantiated** in CrowdFlower
  - TLX was **appended** at the end of concluded tasks
- Tasks were executed and evaluated by min **13** and max **16** workers (903 evaluations in total)
- Filtering
  - 3 control questions, 2 mistakes = out (15% of the evaluations discarded)
  - Completion time, too long or too short = out (6% of the evaluations discarded)

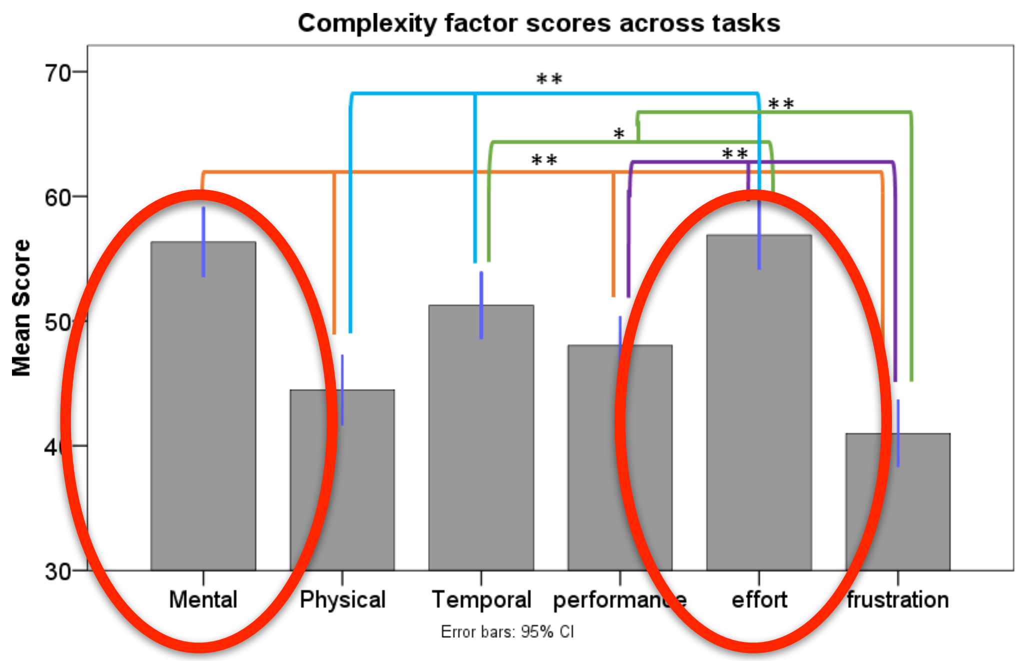
Perceived Task Complexity

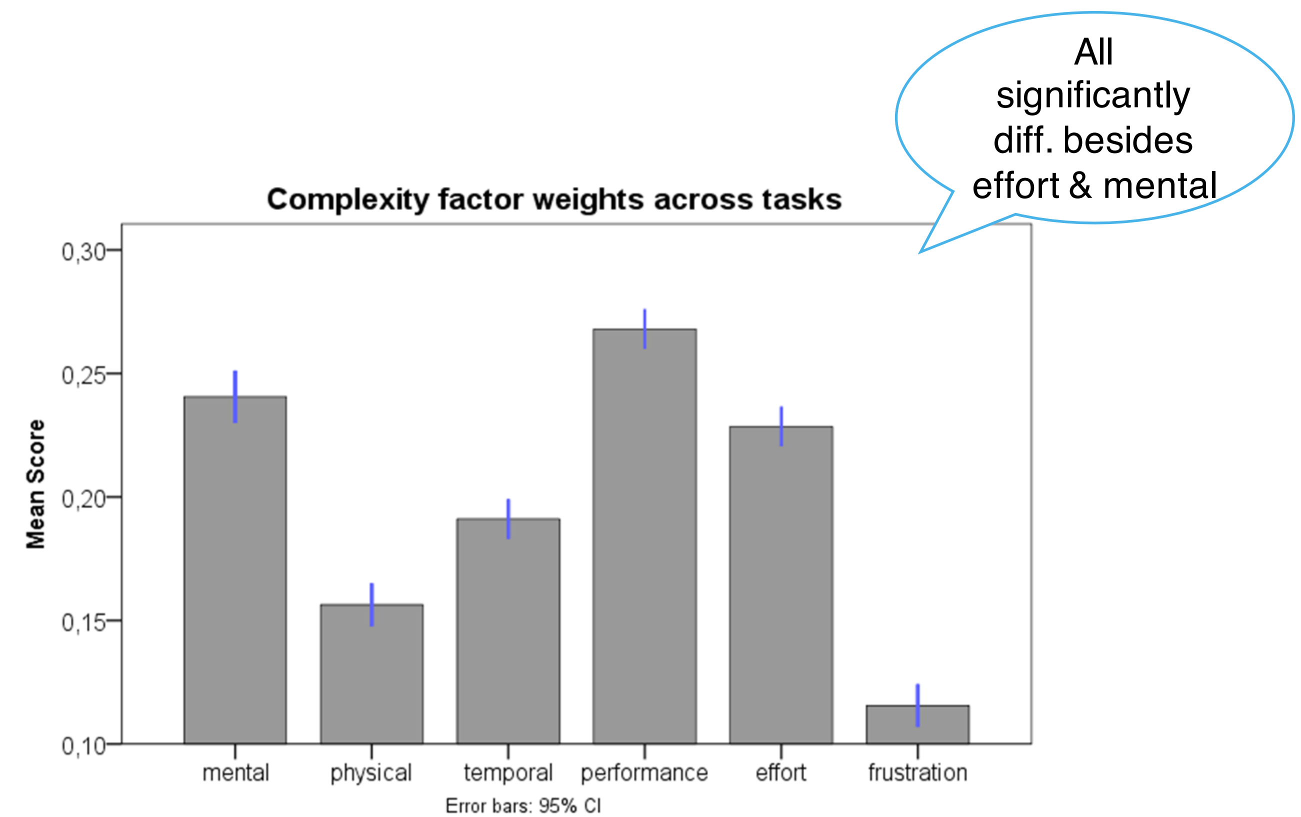




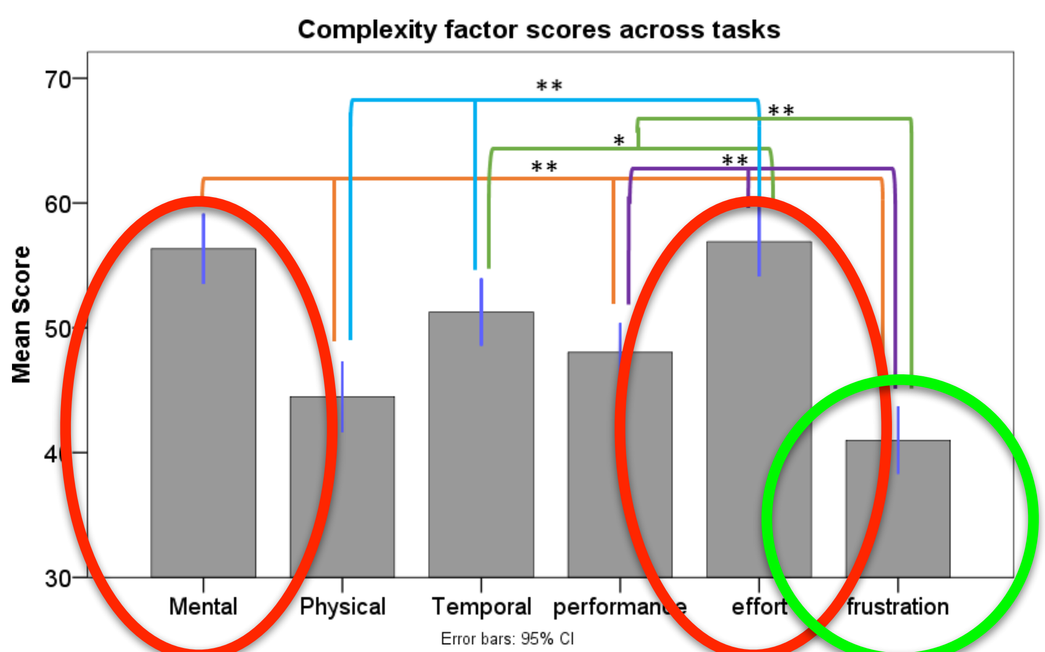
All significantly diff. besides effort & mental

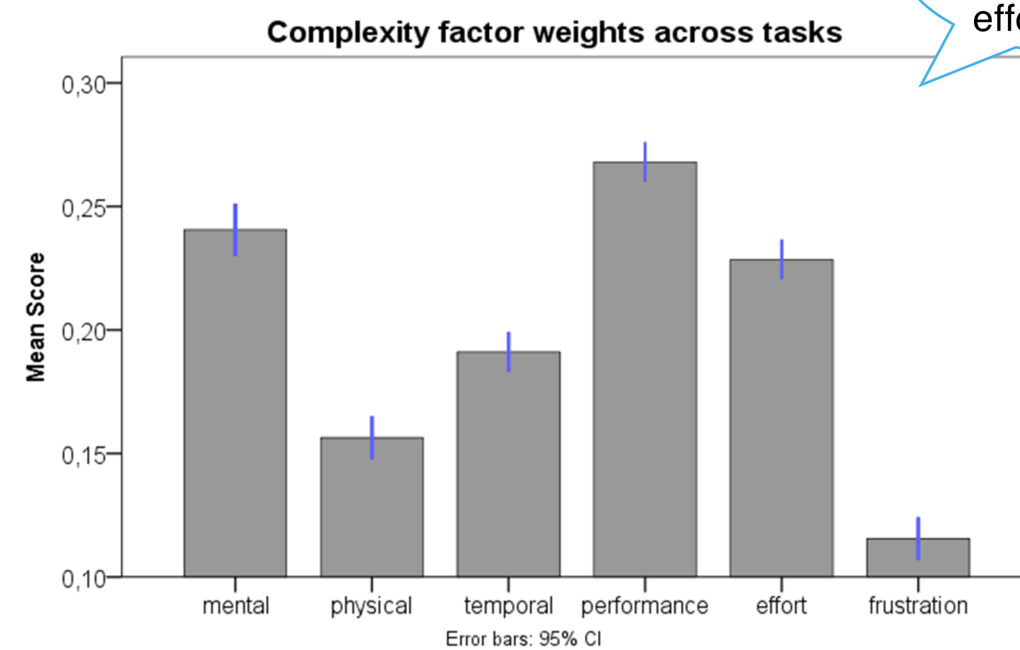
Perceived Task Complexity





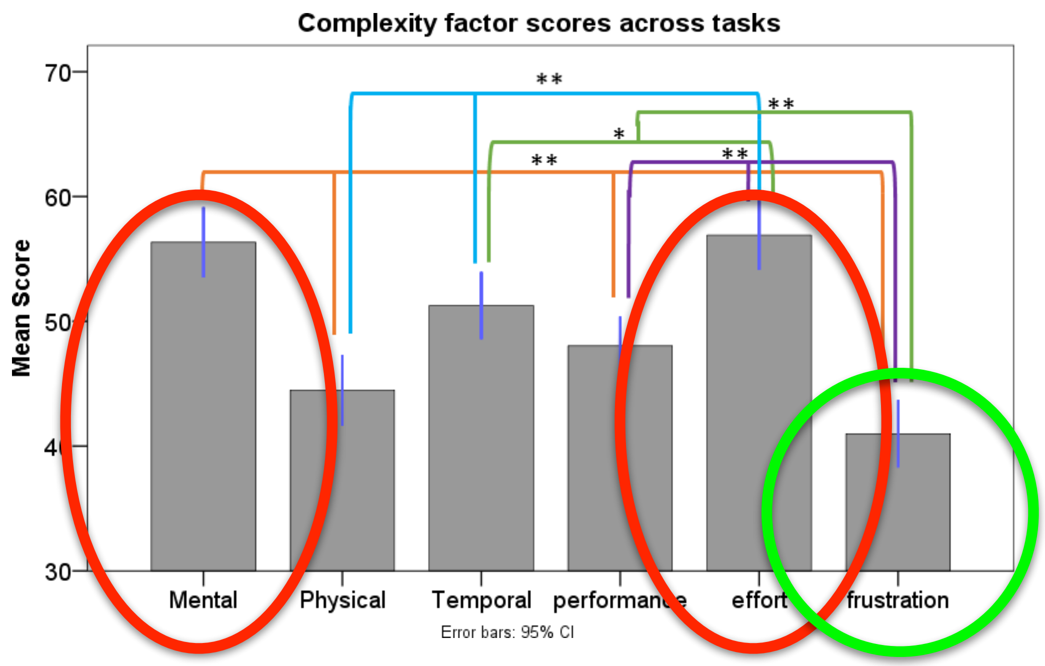
Perceived Task Complexity

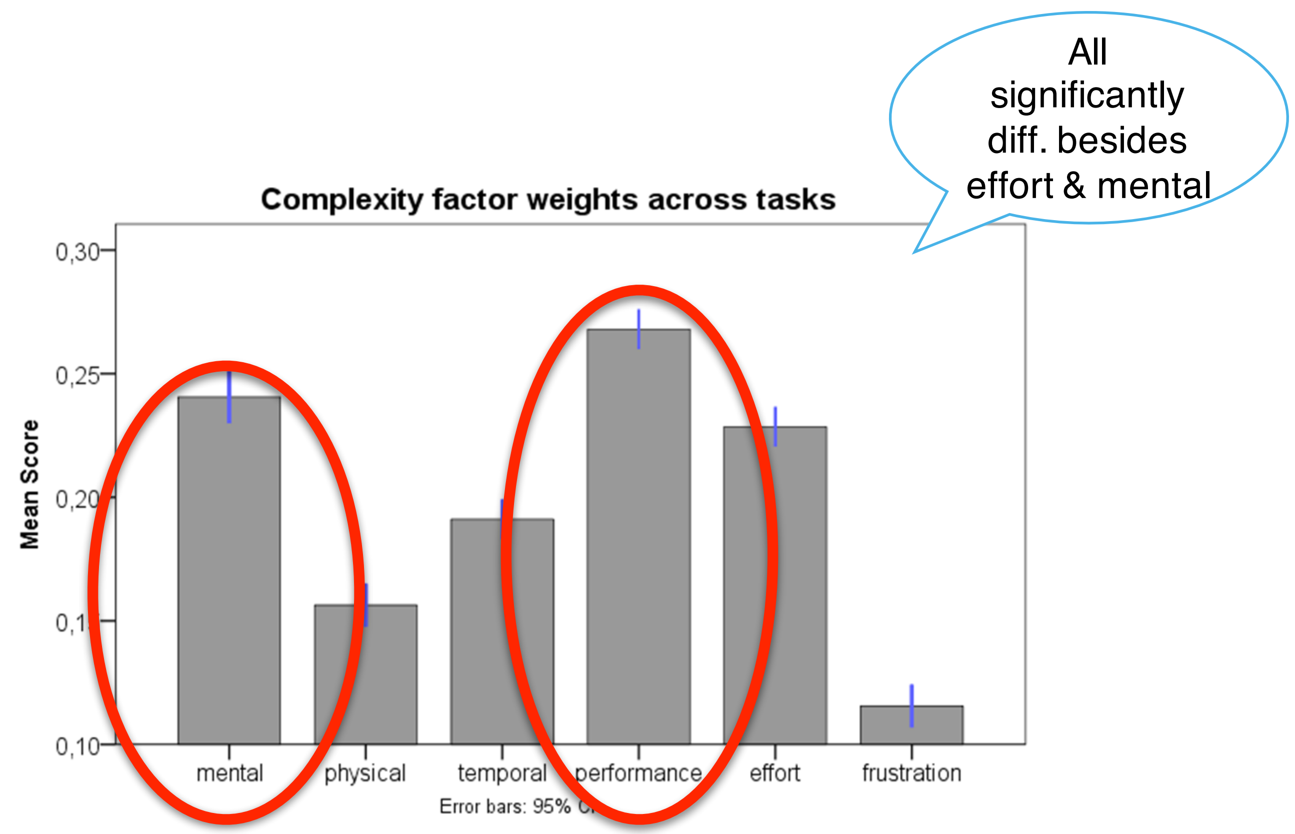




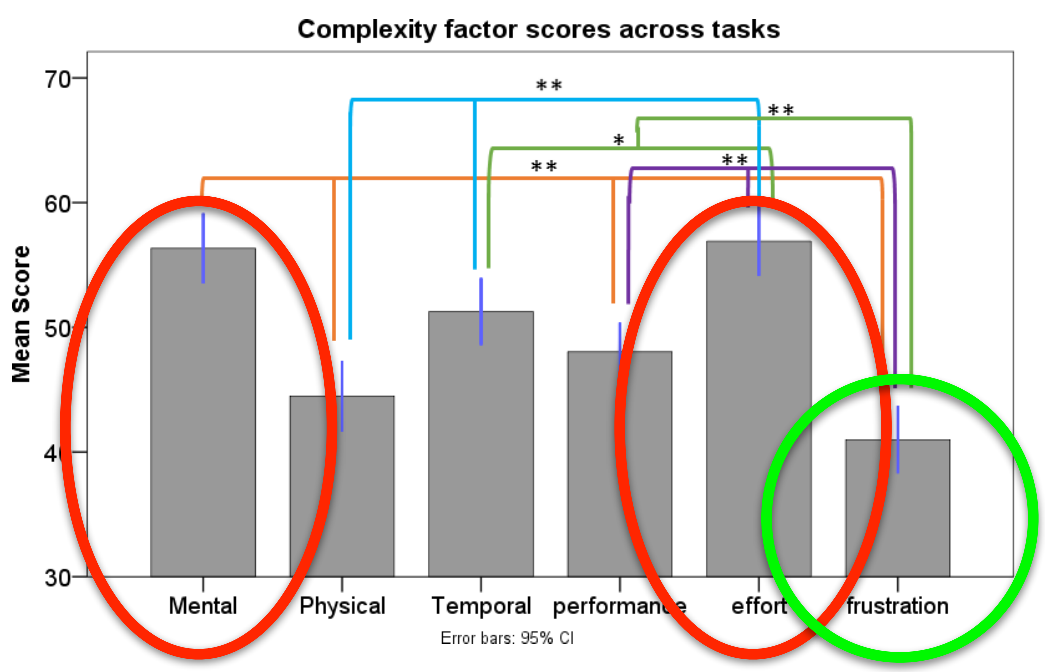
All significantly diff. besides effort & mental

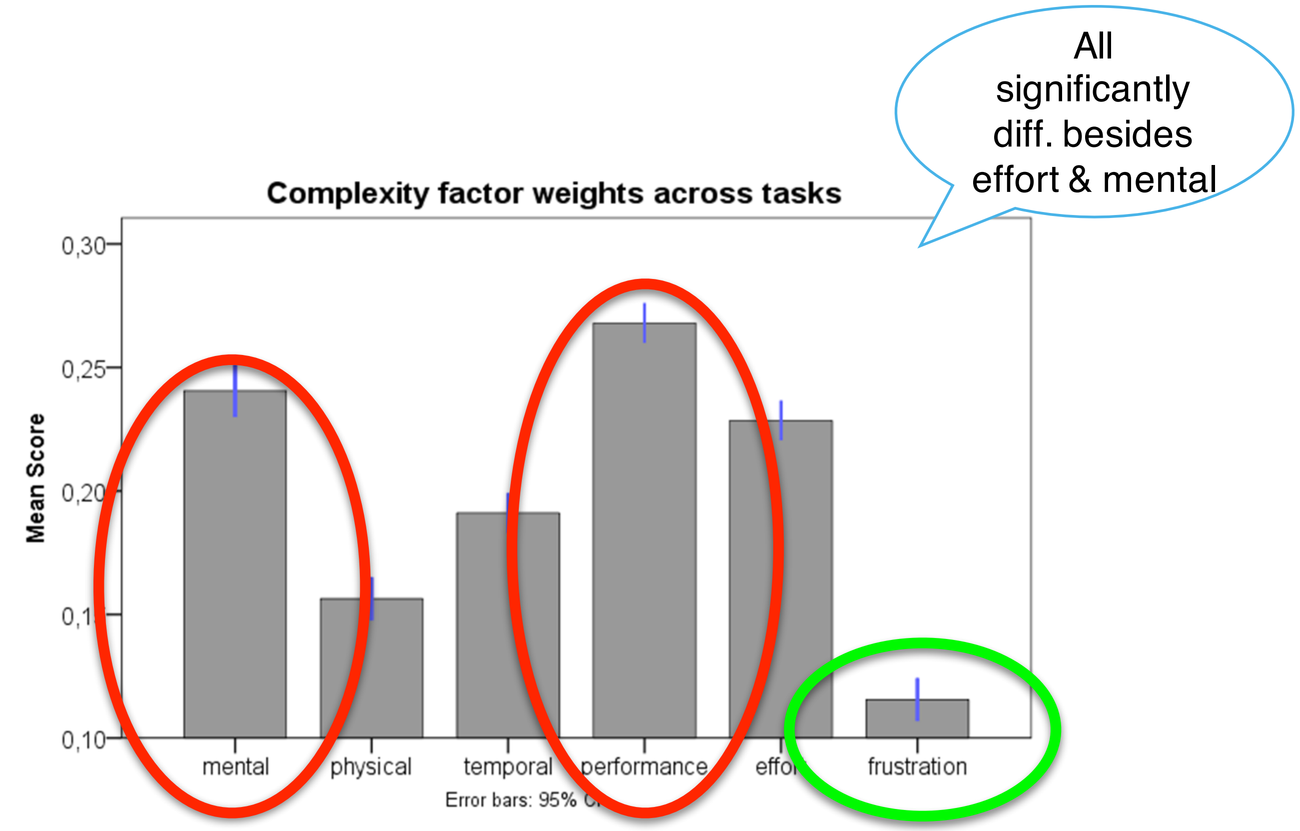
Perceived Task Complexity



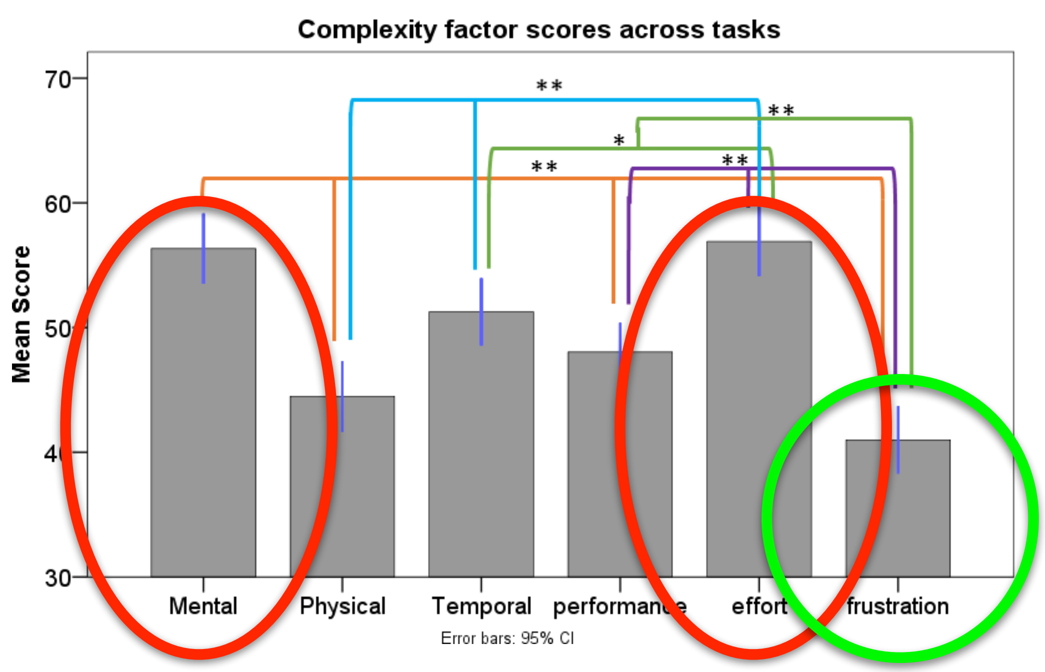


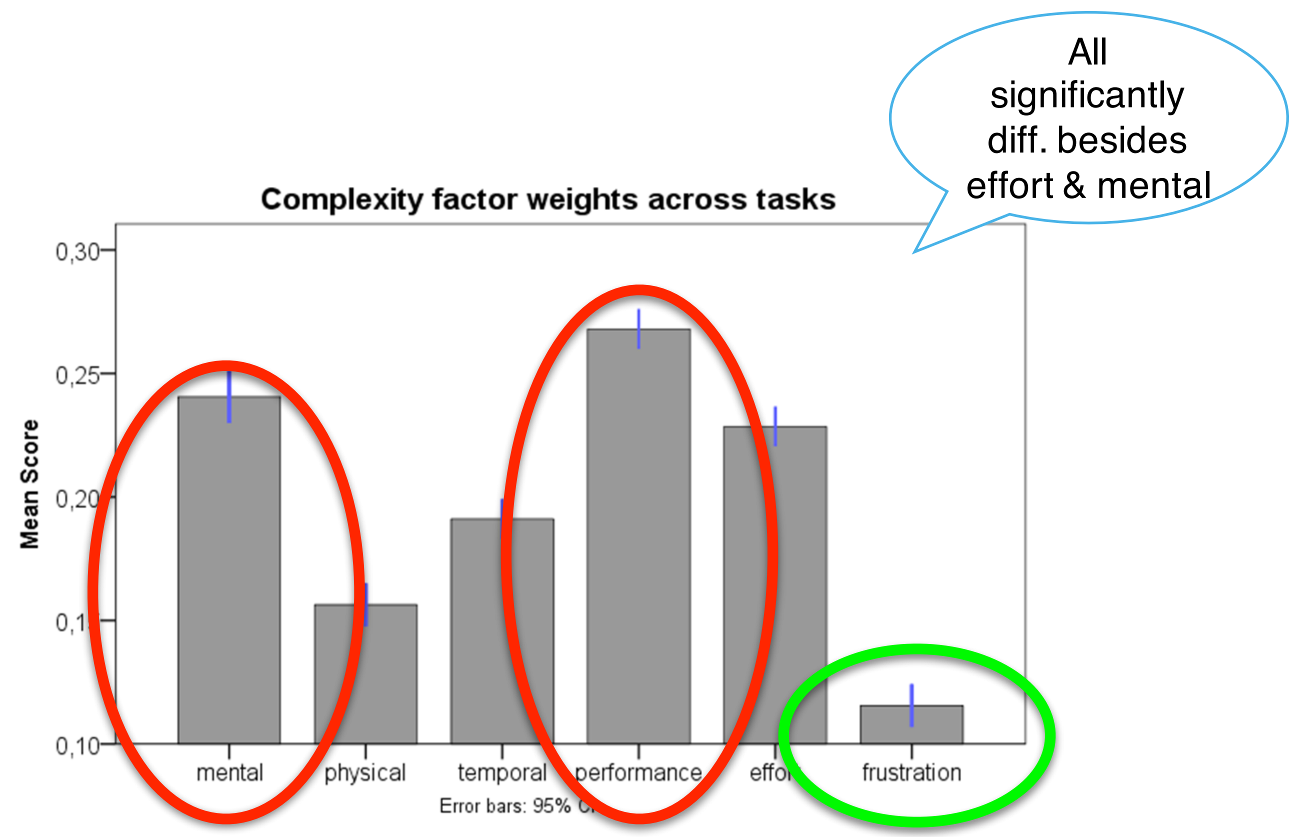
Perceived Task Complexity





Perceived Task Complexity





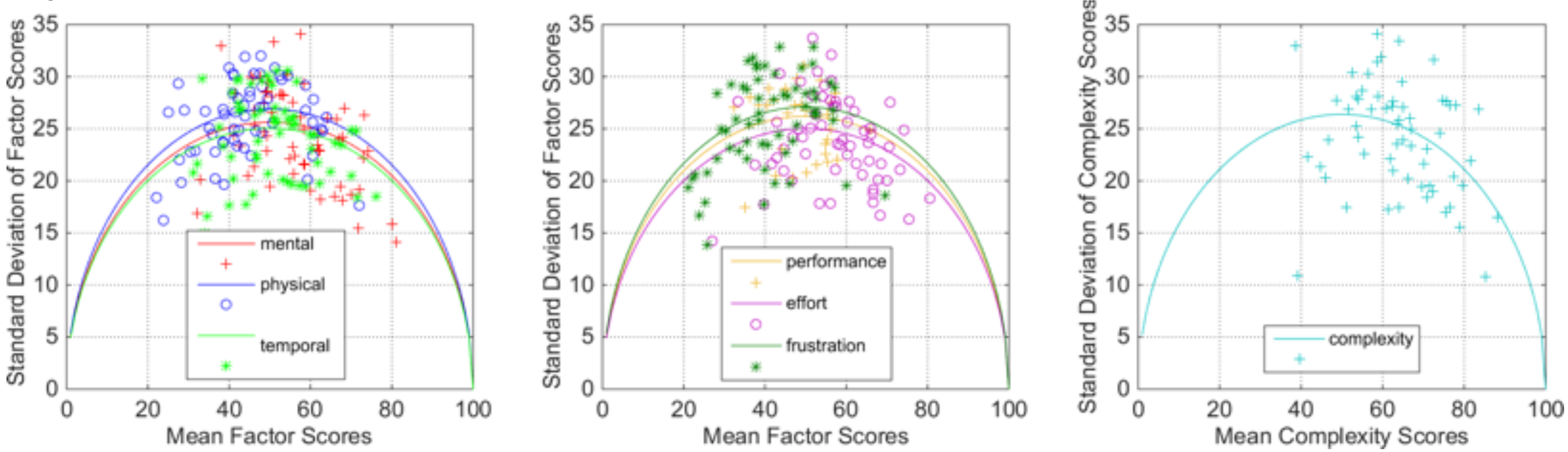
Mental demands and effort are mostly perceived task complexity factors; and workers care about their performance

Reliability of Scores

—— SOS Analysis

- SOS hypothesis: Mean Opinion Score (MOS, e.g. mean complexity, or complexity factor score) and the spreads of individual scores (SOS) are linked by a squared relationship
  - Useful in subjective assessment involving a pool of participants scoring the same item
  - **Alpha: variability in evaluations**

Factor	Complexity	Mental	Physical	Temporal	Performance	Effort	Frustration
Alpha	0.2785	0.2627	0.2745	0.2897	0.2507	0.2503	0.2937

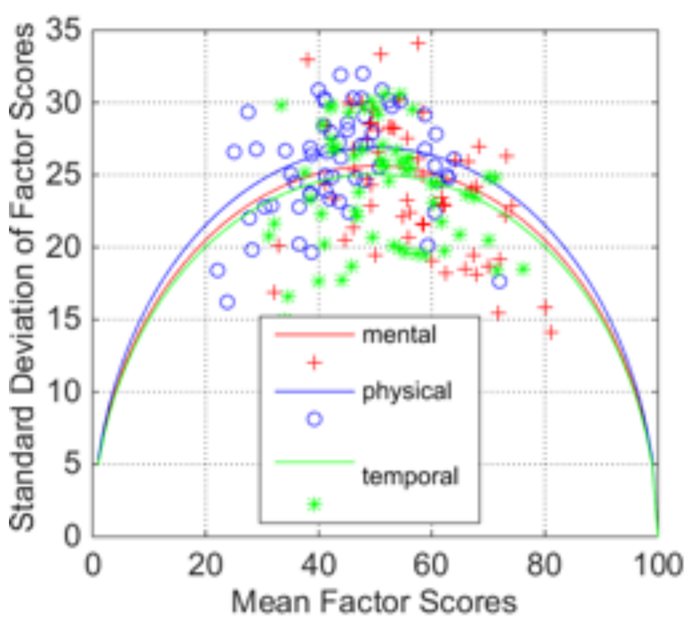


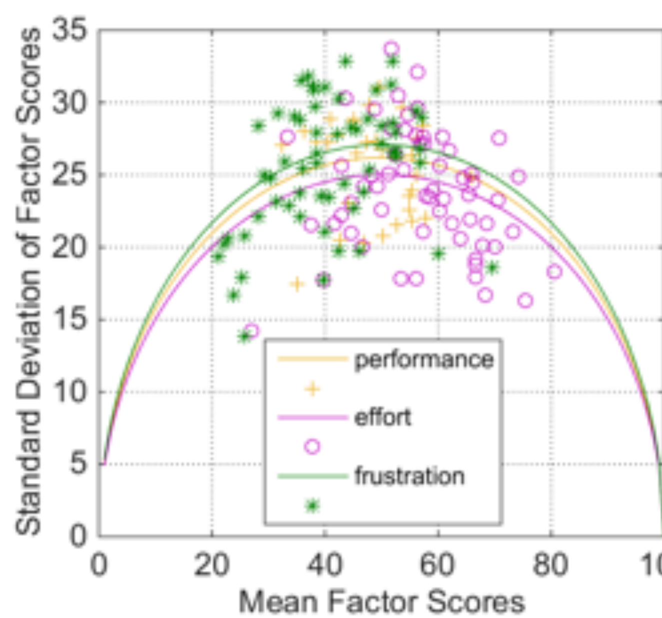
Reliability of Scores

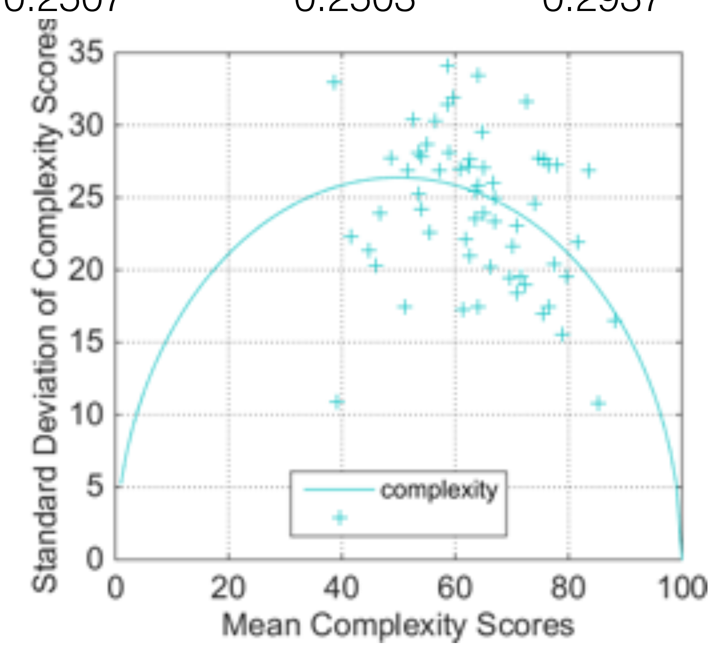
—— SOS Analysis

- SOS hypothesis: Mean Opinion Score (MOS, e.g. mean complexity, or complexity factor score) and the spreads of individual scores (SOS) are linked by a squared relationship
  - Useful in subjective assessment involving a pool of participants scoring the same item
  - **Alpha: variability in evaluations**

Factor	Complexity	Mental	Physical	Temporal	Performance	Effort	Frustration
Alpha	0.2785	0.2627	0.2745	0.2897	0.2507	0.2503	0.2937







Task complexity can be coherently perceived by workers

Modeling Complexity: Task Features

Metadata (9)	Semantics (1440)	Visual (47)
• Title length • Description length • Required worker location • Required Approval rate • Allotted time • Reward • Initial hits	• Amount of words • Amount of links • Amount of images • Unigrams • Topics (LDA) • Keywords	• Body text percentage • No. style files • No. text groups • No. Image Areas • Emphasized body text % • Colorfulness • Color histogram

Modeling Complexity: Regression

Ground truth:

63.78 ± 11.46

MFLR:

LR with **dimension reduction**

Feature Set	Regression Models			
	Linear	Lasso	MFLR	Random Forest
Metadata	13.37 ± 4.18	13.16 ± 4.24	—	9.94 ± 1.68
Visual	14.86 ± 4.01	12.50 ± 2.07	9.97 ± 1.28	10.21 ± 1.15
Content	12.87 ± 1.64	9.97 ± 1.27	**9.18 ± 1.83**	10.00 ± 1.47
Content LDA	10.34 ± 1.84	9.23 ± 1.44	—	11.80 ± 1.18

Modeling Complexity: Regression

Ground truth:

63.78 ± 11.46

MFLR:

LR with **dimension reduction**

Feature Set	Regression Models			
	Linear	Lasso	MFLR	Random Forest
Metadata	13.37 ± 4.18	13.16 ± 4.24	—	9.94 ± 1.68
Visual	14.86 ± 4.01	12.50 ± 2.07	9.97 ± 1.28	10.21 ± 1.15
Content	12.87 ± 1.64	9.97 ± 1.27	**9.18 ± 1.83**	10.00 ± 1.47
Content LDA	10.34 ± 1.84	9.23 ± 1.44	—	11.80 ± 1.18

Task complexity can be robustly (low std) predicted with relatively small error

Modeling Complexity: Regression

Ground truth:

63.78 ± 11.46

MFLR:

LR with **dimension reduction**

Feature Set	Regression Models			
	Linear	Lasso	MFLR	Random Forest
Metadata	13.37 ± 4.18	13.16 ± 4.24	—	9.94 ± 1.68
Visual	14.86 ± 4.01	12.50 ± 2.07	9.97 ± 1.28	10.21 ± 1.15
Content	12.87 ± 1.64	9.97 ± 1.27	**9.18 ± 1.83**	10.00 ± 1.47
Content LDA	10.34 ± 1.84	9.23 ± 1.44	—	11.80 ± 1.18

Task complexity can be robustly (low std) predicted with relatively small error

Content features with proper dimension reduction result in best performance

Most Significant Features

Visual Feature	Imp.	Semantic Features	Imp.
visualAreaCount	3.35	**linkCount**	2.42
hueAvg	0.09	**wordCount**	1.37
		keyword: *audio*	0.09
		keyword: *transcribe*	0.07
		keyword: *writing*	0.06
imageAreaCount	-0.27	**unigram: *clear***	-0.06
colourfulness1	-0.63	**unigram: *identify***	-0.07
scriptCount	-1.52	**unigram: *date***	-0.09
valAvg	-1.71	**keyword: *easy***	-0.10
cssCount	-1.82	**imageCount**	-1.01

Most Significant Features

More visual items lead to higher task complexity perceived by workers

		keyword: *audio*	0.09
		keyword: *transcribe*	0.07
		keyword: *writing*	0.06
imageAreaCount	-0.27	**unigram: *clear***	-0.06
colourfulness1	-0.63	**unigram: *identify***	-0.07
scriptCount	-1.52	**unigram: *date***	-0.09
valAvg	-1.71	**keyword: *easy***	-0.10
cssCount	-1.82	**imageCount**	-1.01

Most Significant Features

More visual items lead to higher task complexity perceived by workers

		keyword: *audio*	0.09

A better design of the task presentation (CSS) and more interactive components (JS) could decrease the complexity

scriptCount	-1.52	**unigram: *date***	-0.09
valAvg	-1.71	**keyword: *easy***	-0.10
cssCount	-1.82	**imageCount**	-1.01

Most Significant Features

More visual items lead to higher task complexity perceived by workers

		keyword: ***audio***	0.09
[redacted]			

A better design of the task presentation (CSS) and more interactive components (JS) could decrease the complexity

scriptCount	-1.52	**unigram:** ***date***	-0.09

Complexity is reflected from the point of view of required actions to be performed by workers (e.g. transcribe), task type (e.g. writing), and content matter (e.g. audio).

Applying Complexity to Throughput Prediction

- Task throughput, i.e. completion rate
  - Dominated by **batch size** (Difallah et al. 2015)
  - Workers select tasks with many HITs to maximise reward opportunities
- Control for batch size, then apply **complexity**
  - Help most in the predicting the throughput of small tasks

Applying Complexity to Throughput Prediction

- Task throughput, i.e. completion rate
  - Dominated by **batch size** (Difallah et al. 2015)
  - Workers select tasks with many HITs to maximise reward opportunities
- Control for batch size, then apply **complexity**
  - Help most in the predicting the throughput of small tasks

suggesting that complexity could help explaining the task selection strategy

Conclusions & Discussions

- **We can**, to some extent, measure and predict task complexity from task properties
- Can this help to:
  - Inform better *task design*?
  - Inform *task recommendation*?
  - *Estimate reliability* in task completion?

Thank you!

Jie Yang, Judith Redi,
Gianluca Demartini, Alessandro Bozzon

j.yang-3@tudelft.nl

Delft University of Technology

The University Of Sheffield.