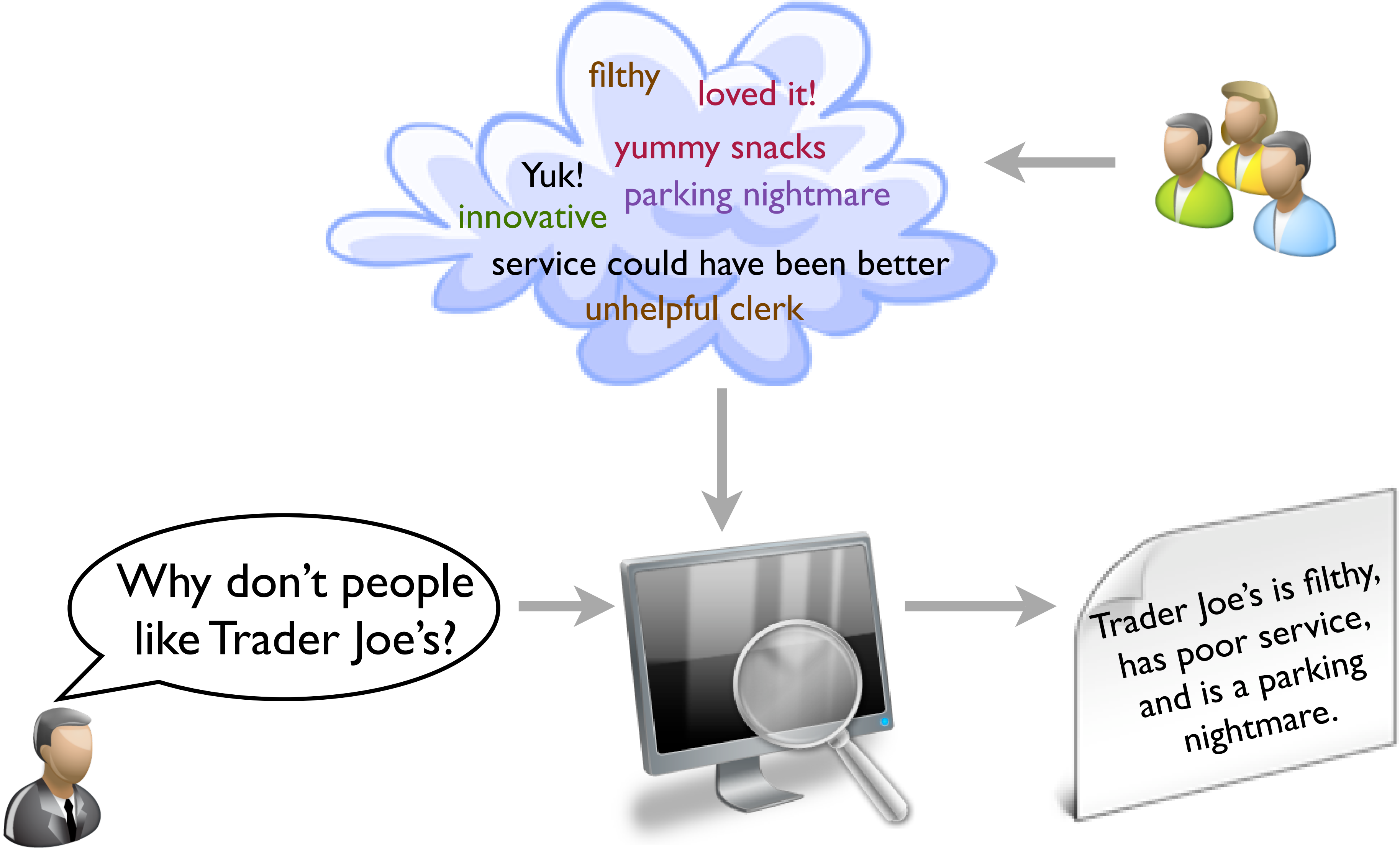
Overview of the TAC 2008 Opinion Question Answering and Summarization Tasks

Hoa Trang Dang
National Institute of Standards and Technology

Overview

- Pipelined opinion QA/summarization tasks
- Document collection, question series
- QA task and evaluation measures
- Summarization task and evaluation measures
- Conclusion

Opinion QA/Summarization Task



filthy

loved it!

yummy snacks

Yuk!

parking nightmare

innovative

service could have been better

unhelpful clerk

Why don't people like Trader Joe's?

Trader Joe's is filthy, has poor service, and is a parking nightmare.

Disclaimer: Opinions expressed about a particular company or product do not necessarily reflect those of NIST or its sponsors.

TREC/TAC Opinion Task

- Pipelined opinion IR/QA/Summarization tasks:

1. TREC Blog Track: Opinion task
2. TAC opinion QA task: return different aspects of opinion (holder, target, support) with a particular polarity in response to opinion question
    - evaluated 17 runs from 9 teams
3. TAC opinion summarization task: summarize answers to complex (“squishy”) questions
    - evaluated 36 runs from 19 teams

Document Collection

- Blog06 collection (Dec 6, 2005 - Feb 21, 2006)
  - 3.2 million permalink “docs” from 100K blogs
- QA: Answers to all questions must be supported by documents in Blog06
  - *Optionally* given top 50 docs retrieved by Prise for each target
- Summarization: Summary of relevant documents in Blog06
  - Optionally given answer-snippets from humans and QA systems

Question Series

- 50 series of questions -- same targets and assessors as opinion task in TREC Blog Track
- Series is an abstraction of a “user session”
- Each series is about a specified target
  - Person, Organization, Product, Issue, ...
- Goal is to gather opinions about target
- Series contains 2-4 questions
- Questions could depend on previous answers
- Questions tagged as to type (rigid list, squishy list)

Example Question Series

TARGET 1018: "MythBusters"

1018.1	RIGID LIST	Who likes Mythbuster’s?
1018.2	SQUISHY LIST	Why do people like Mythbuster’s?
1018.3	RIGID LIST	Who do people like on Mythbuster’s?

TARGET 1047: "Trader Joe’s"

1047.1	RIGID LIST	Who likes Trader Joe’s?
1047.2	SQUISHY LIST	Why do people like Trader Joe’s?
1047.3	RIGID LIST	Who doesn’t like Trader Joe’s?
1047.4	SQUISHY LIST	Why don’t people like Trader Joe’s?

Rigid Lists vs. Squishy Lists

- Rigid (“named” entities)
  - entities are disjoint
  - small number of ways of referring to the same entity
  - boundaries of referring expression well-defined
- Squishy (complex concepts)
  - concepts can overlap, subsume each other
  - many different ways of expressing the same concept
  - boundaries of concept descriptors not well-defined

Example Rigid List Question + Response

1047.1 RIGID LIST Who likes Trader Joe's?

BLOG06-4201	Peggy Archer
BLOG06-5961	david Ford
BLOG06-5961	Michelle
BLOG06-9274	thalassa_mikra
BLOG06-2189	FoodMonkey
BLOG06-6816	trackingtraderjoes
BLOG06-6816	http://www.trackingtraderjoes.com/index.rdf#
BLOG06-4201	http://filmhacks.blogspot.com/atom.xml#

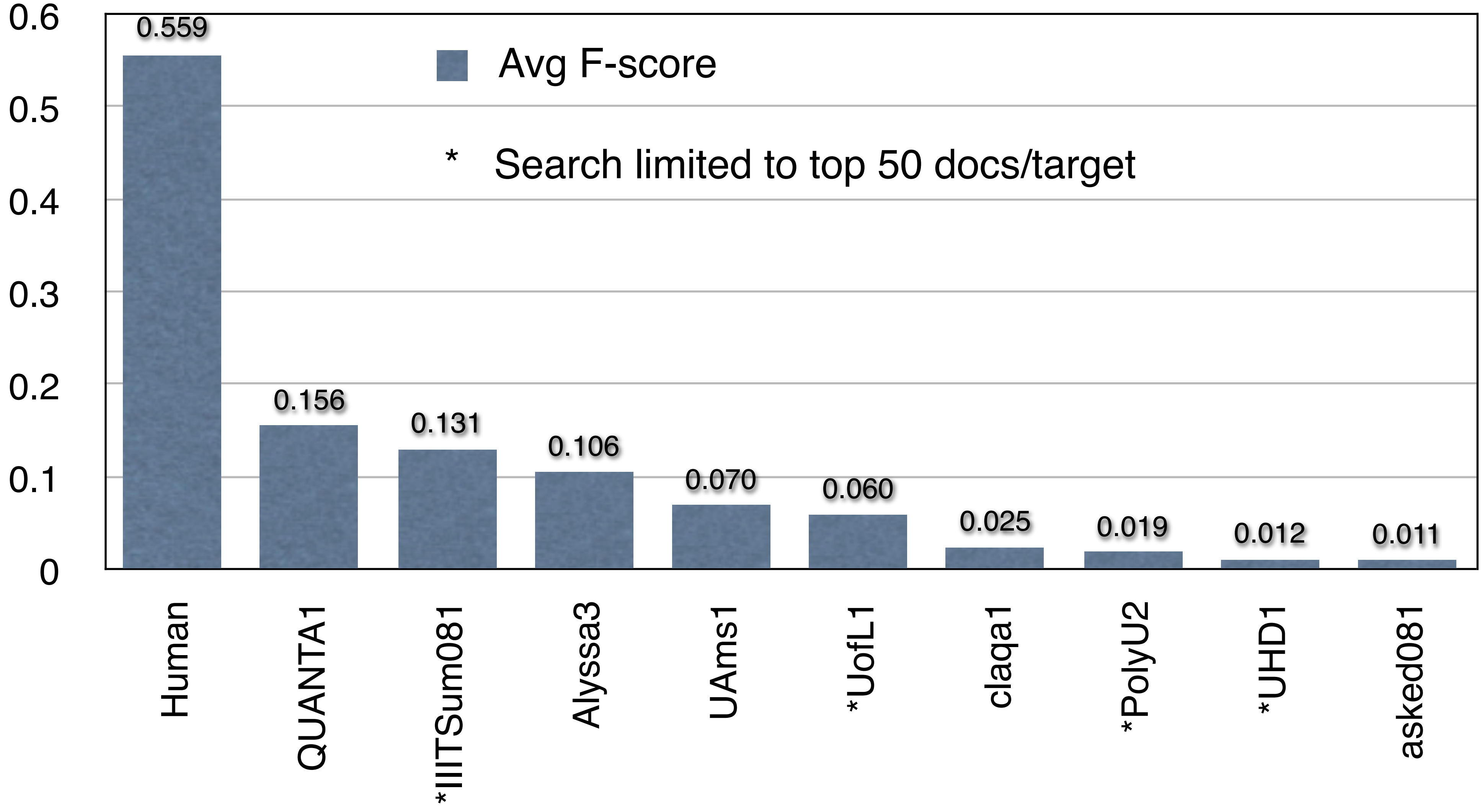
Rigid List Component

- Questions seek list of entities with a particular property *("Who likes Trader Joe's")*
- Response is set of [docid, answer-string] pairs
- Human assessors judged each pair as one of:
  - wrong, unsupported, inexact, correct
- Equivalent correct answer-strings (names for the same entity) count as a single entity
- 90 rigid list questions, 1-40 entities per list (median 8)

Rigid List Scoring

- Final list of known correct entities (found by assessor and/or system)
- Precision = # correct entities found / # answer-strings returned
- Recall = # correct entities found / # known correct entities
- Combine precision and recall: $F = (2*P*R)/(P+R)$
- Rigid List Score = F score of rigid list question
- Rigid List component score is average F over 90 rigid list questions

Rigid List Component Results



Average F score of best run for each team

Example Squishy Question and Response

1047.2 SQUISHY LIST Why do people like Trader Joe’s?

BLOG06-3227 Trader Joes is your destination if you prefer Industrial wines (unlike Whole Foods).

BLOG06-2494 Everytime I walk into a Trader Joes it's a fun filled experience, and I always learn something new....

BLOG06-4400 Sure, we have our natural food stores, but they are expensive and don't have the variety that Trader Joe's has.

BLOG06-2494 Then I went to Trader Joe's and they have all the good stuff for cheap.

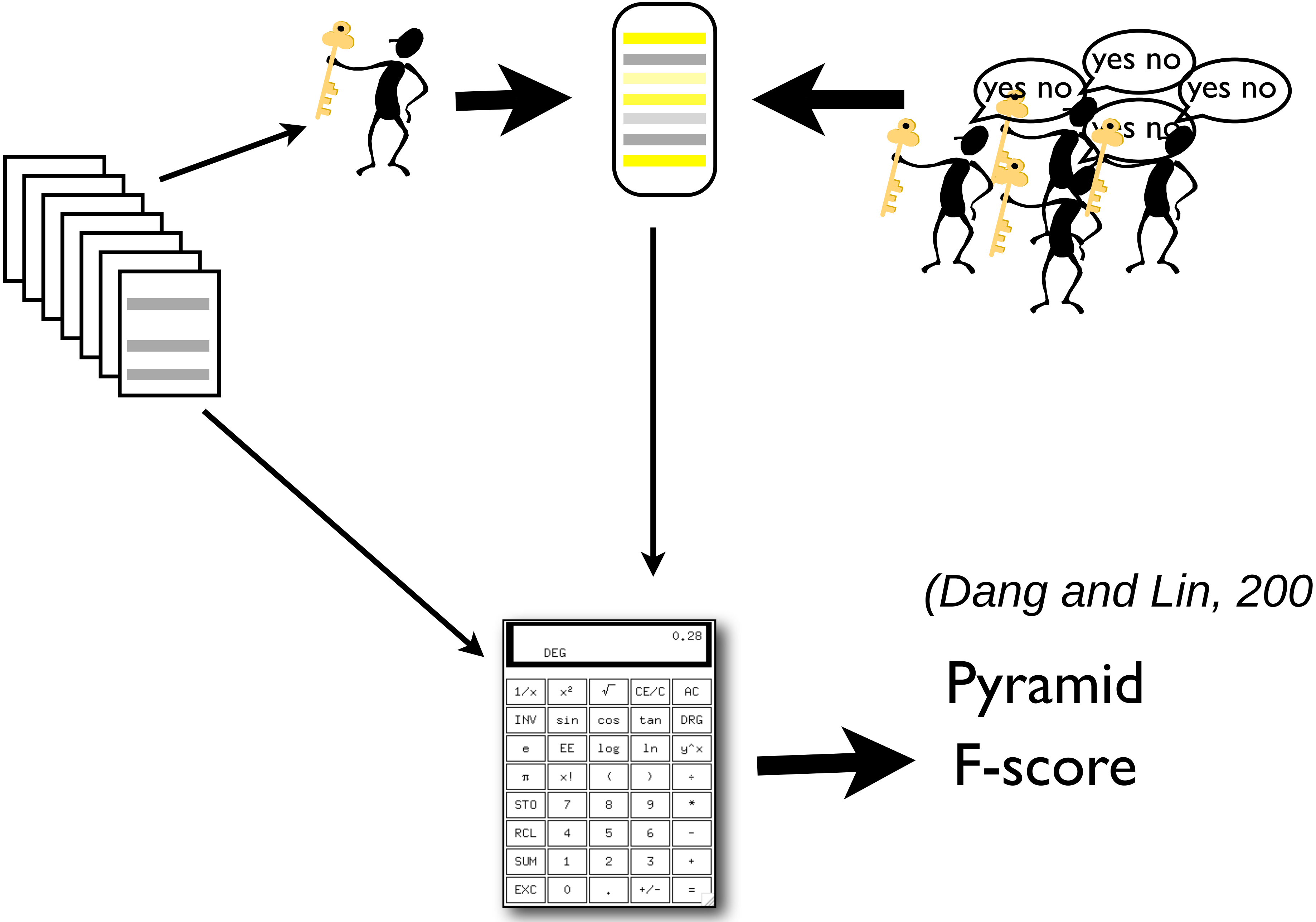
“Squishy” List Component

- Response is a set of [docid, answer-string] pairs
- Response should contain information nuggets answering question *(“Why do people like Trader Joe’s”)*
- Primary assessor determines set of information nuggets that a good response should contain
  - distinction between *vital* and *okay* nuggets
- Primary assessor marks which nuggets appear in system response

“Squishy List” Scoring (Nugget Pyramids)

- Using assessor judgments, compute nugget recall and approximation of nugget precision (a function of response length)
- Score for question is F(beta=3), which gives more weight to recall than precision
- Pyramid F-score (Lin and Demner-Fushman, 2006)
  - 10 judgments of vital/okay from 9 different assessors, using nugget list from primary assessor
  - Nugget weight is fraction of judgments of vital for the nugget, normalized so maximum nugget weight is 1.0

Squishy List Evaluation



(Dang and Lin, 2007)

Pyramid

F-score

Example Nugget Pyramid

Why don’t people like Trader Joe’s?	0	1	2	3	4	5	6	7	8	9	wt
long waits in line		1	1	1	1	1	1	1	1	1	9 / 9
moldy food on shelves		1		1	1	1	1	1	1	1	8 / 9
erodes local businesses	1			1	1	1	1	1	1	1	8 / 9
drops products without warning	1		1	1		1		1	1	1	7 / 9
store smells foul				1	1	1	1	1	1	1	7 / 9
organic dairy items not available				1		1	1	1	1	1	6 / 9
parking lot is crowded	1		1		1	1		1		1	6 / 9
don’t like store decor	1				1		1		1	1	5 / 9
canned chicken stew smelled foul				1		1	1		1	1	5 / 9
employees have body odor					1	1	1	1		1	5 / 9
employees are hippies					1	1	1			1	4 / 9
Santa Fe store seriously lacking							1	1		1	3 / 9
rude people in the parking lot								1		1	2 / 9
not expanding fast enough									1	1	2 / 9
employees are tree huggers						1				1	2 / 9

Calculation of Pyramid F-Score

a # of nuggets in response

r sum of weights of nuggets in response

R sum of weights of nuggets in answer key

l # of non-whitespace characters in response

C character allowance per nugget (C = 100)

Allowance: $A = C * a$

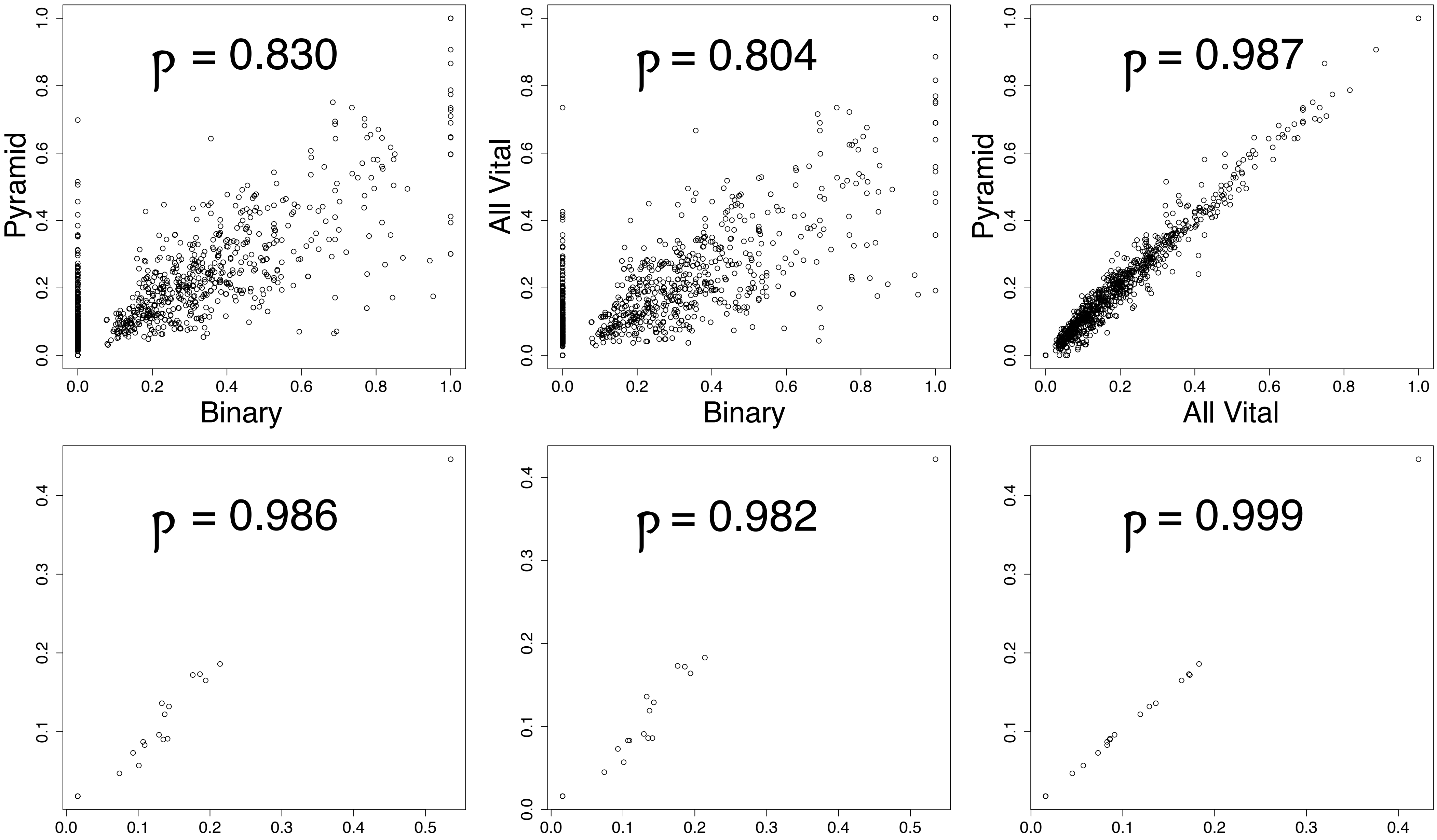
Recall: $\mathrm{R} = r/R$

Precision: $\mathrm{P} = 1$ if $l < A$

$\mathrm{P} = 1 - ((l - A) / l)$ otherwise

F-score = $(ß^2 + 1) * \mathrm{R} * \mathrm{P} / (ß^2 * \mathrm{P} + \mathrm{R})$, $ß = 3$

Correlation between Nugget Weighting Methods



Multiple Comparison of Runs

- 2-way ANOVA: F-score ~ question + run
- Multiple comparison of runs, Tukey's HSD criterion, experiment-wise Type I error <= 5%
- Binary F: 50 pairs different

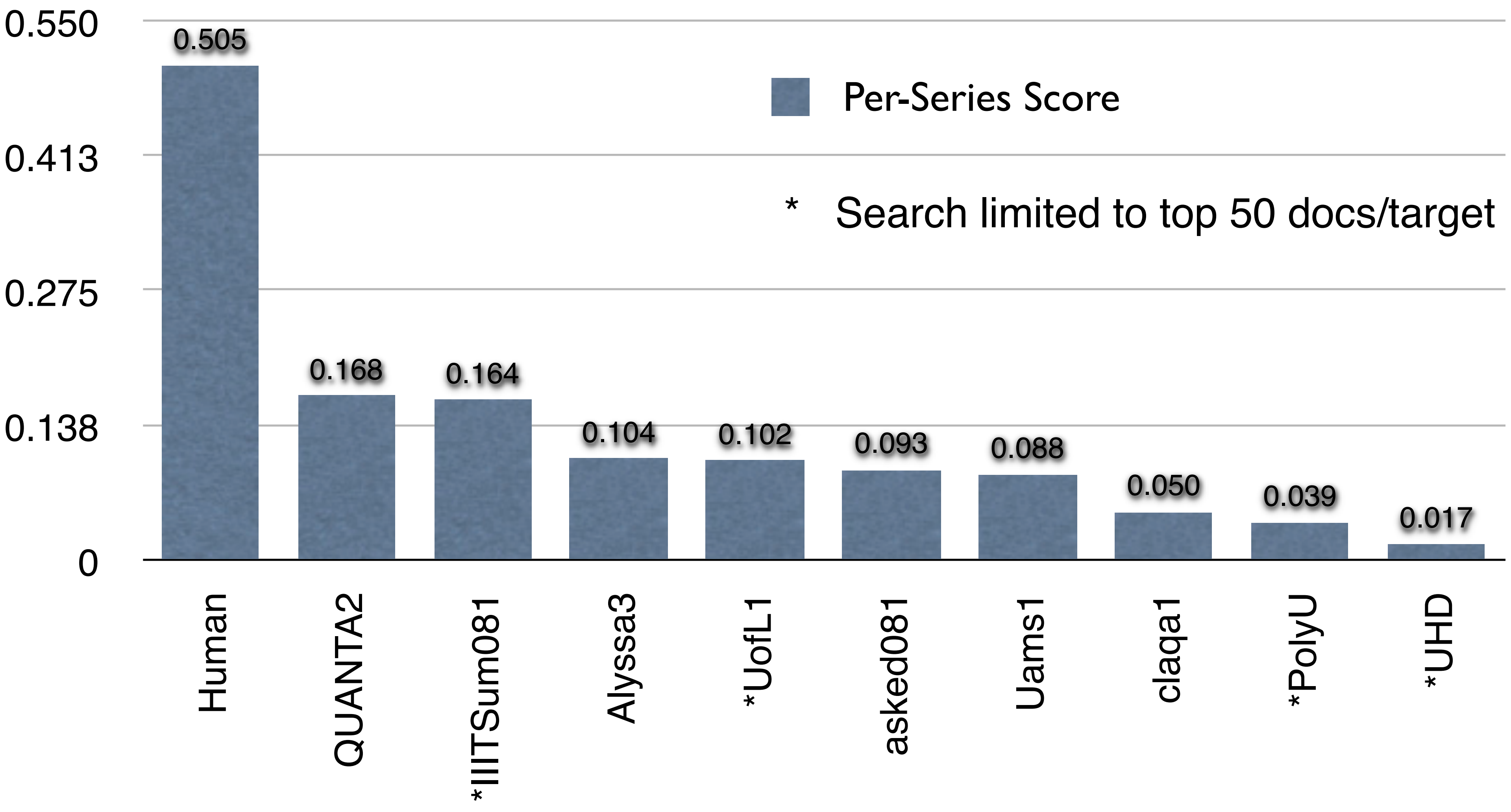
- Pyramid F: 82 pairs different

```
Assessors    0.4461  A
IIITSum081   0.1864    B
asked081     0.1732    B C
QUANTA2      0.1715    B C
IIITSum082   0.1654    B C
QUANTA1      0.1362    B C D
asked082     0.1323    B C D
UofL1        0.1216      C D E
UAms1        0.0958        D E F
Alyssa2      0.0909        D E F
Alyssa3      0.0897        D E F
Alyssa1      0.0869        D E F
UAms2        0.0830        D E F
claqa1       0.0734          E F G
PolyU1       0.0572            F G
PolyU2       0.0465            F G
UHD2         0.0178              G
UHD1         0.0178              G
```

- All Vital F: 83 pairs different

```
Assessors  0.4220  A
IIITSum081 0.1825    B
QUANTA2    0.1726    B C
asked081   0.1717    B C
IIITSum082 0.1642    B C
QUANTA1    0.1360    B C D
asked082   0.1291    B C D E
UofL1      0.1191      C D E
UAms1      0.0908        D E F
Alyssa2    0.0864        D E F
Alyssa3    0.0860        D E F
UAms2      0.0827        D E F
Alyssa1    0.0826        D E F
claqa1     0.0730          E F
PolyU1     0.0568            F G
PolyU2     0.0454            F G
UHD2       0.0160              G
UHD1       0.0160              G
```

Combined (Rigid, Squishy) Score Results



Per-Series combined score for best run for each team

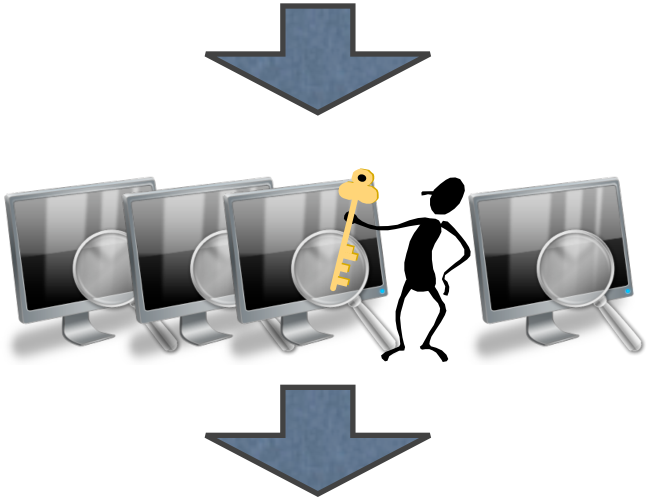
Discussion: Opinion Squishy Lists

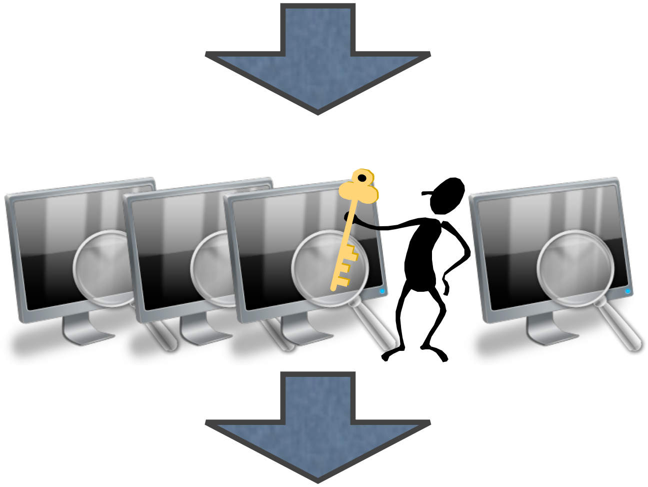
- Both rigid list and squishy list tasks are difficult
- Different pyramid weights vs same (all vital) weight makes almost no difference in relative system score *even at the individual question level*
- Different nugget weights reflect human preferences but systems don’t differ in their ability to optimize for these preferences

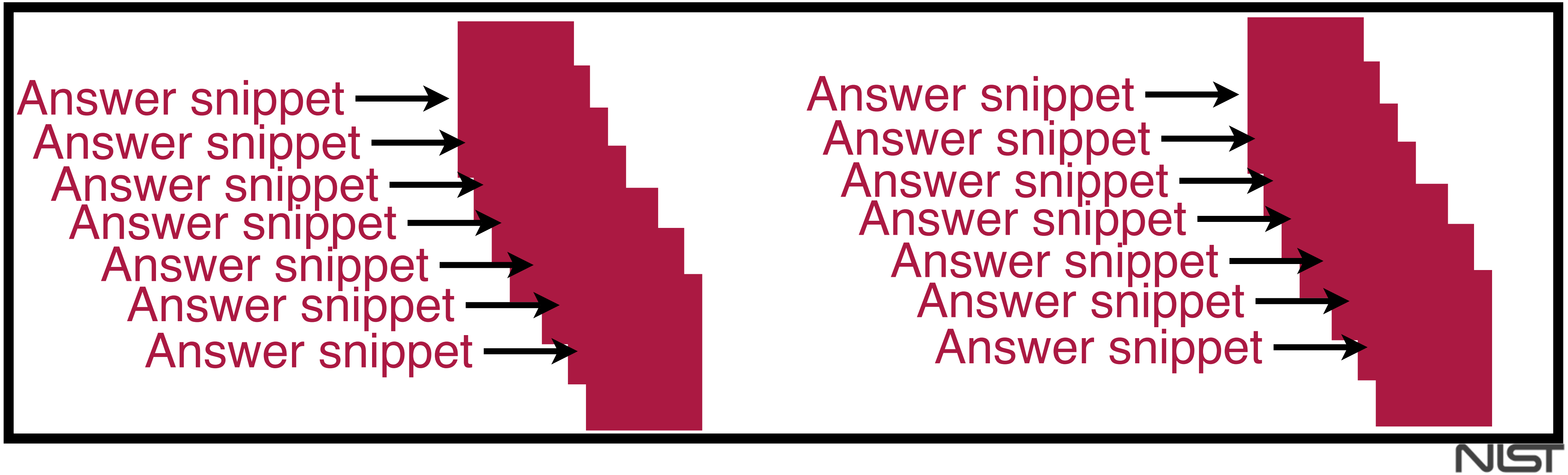
From Answer Snippets to Answer Summary

Why do people like Trader Joe's?

Why don't people like Trader Joe's?







Answer snippet
Answer snippet
Answer snippet
Answer snippet
Answer snippet
Answer snippet
Answer snippet

Answer snippet
Answer snippet
Answer snippet
Answer snippet
Answer snippet
Answer snippet
Answer snippet

Summarization Task

- Input:
  - Target, 1-2 squishy questions
  - Documents known to have answers
  - *Optional* answer snippets in each document
- Output:
  - One fluent summary per target, that summarizes the answers to all the squishy questions for the target
    - Allow at most 7K non-white-space characters per question for the target (i.e., upper limit of either 7K or 14K characters)

Input Characteristics

- 22 targets evaluated
- usually 2 questions per target
- Avg documents per target: 24 (min 9; max 39)
- Avg document length (char): 76K (4K; 23,200K)
- Avg snippets per target: 57 (19; 125)
- Avg snippet length (nws char): 149

- Avg nuggets per question: 16 (2; 35)
- Avg number of snippets per nugget: 2.67 (1; 28)

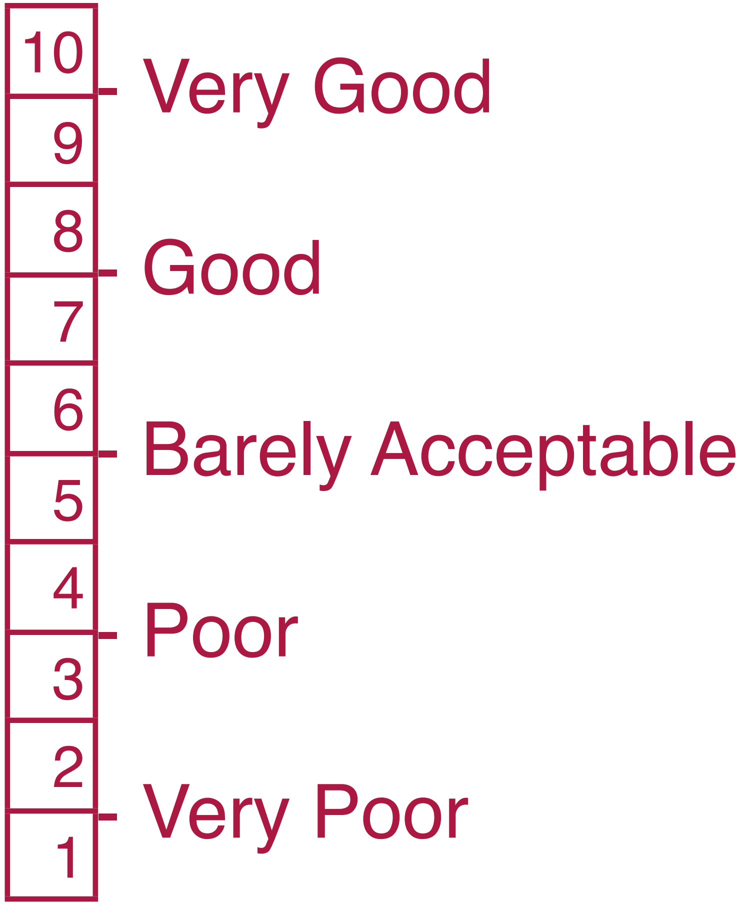
Summary Evaluation

- Content: Pyramid F-score, Beta=1
  - Pyramid from combined nuggets list of both questions, weights normalized by max vital count of nugget in combined list
- Readability
- Overall Responsiveness (“What would I pay for this summary of the answers to my questions?”)

Readability, Overall Responsiveness

1. Grammaticality
2. Non-redundancy
3. Structure/coherence
4. Overall Readability
5. Overall responsiveness

   (Content + Readability)



10
9
8
7
6
5
4
3
2
1

Very Good

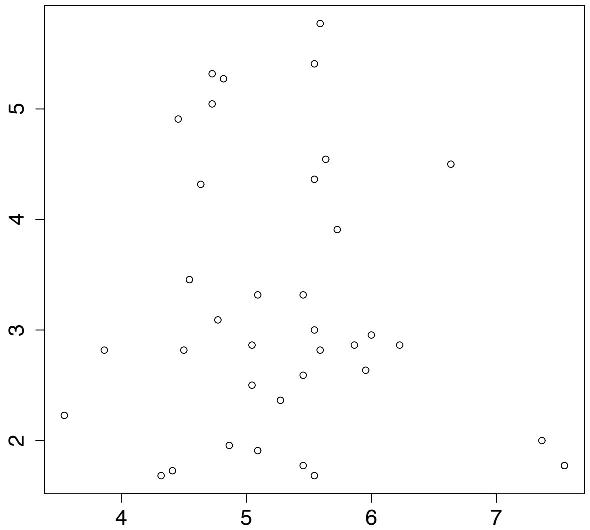
Good

Barely Acceptable

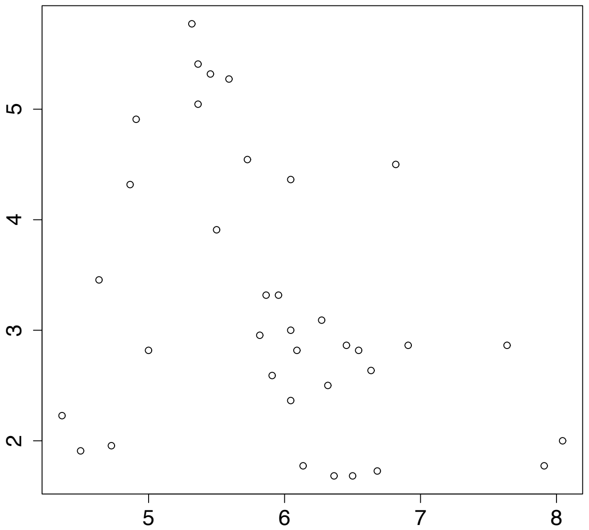
Poor

Very Poor

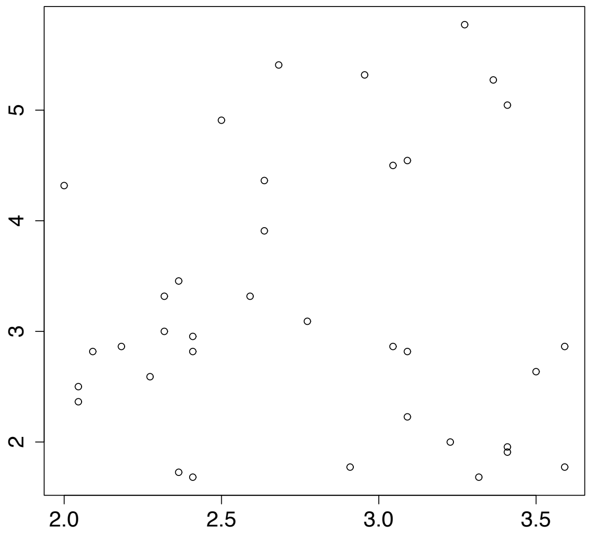
Overall Responsiveness vs Linguistic Quality



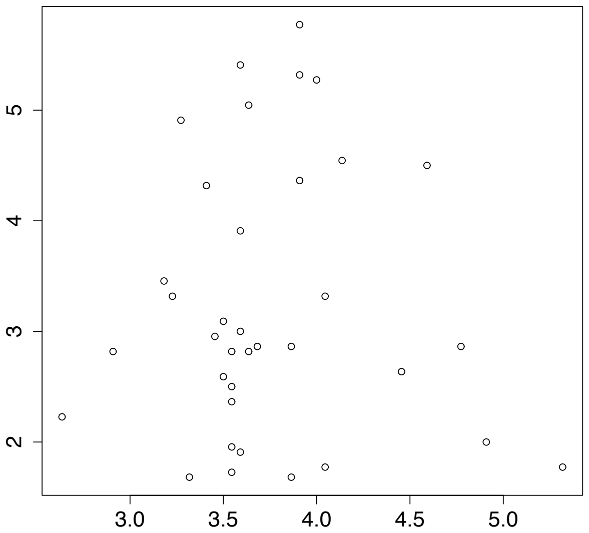
Grammaticality (5.3)



Non-Redundancy (6.0)

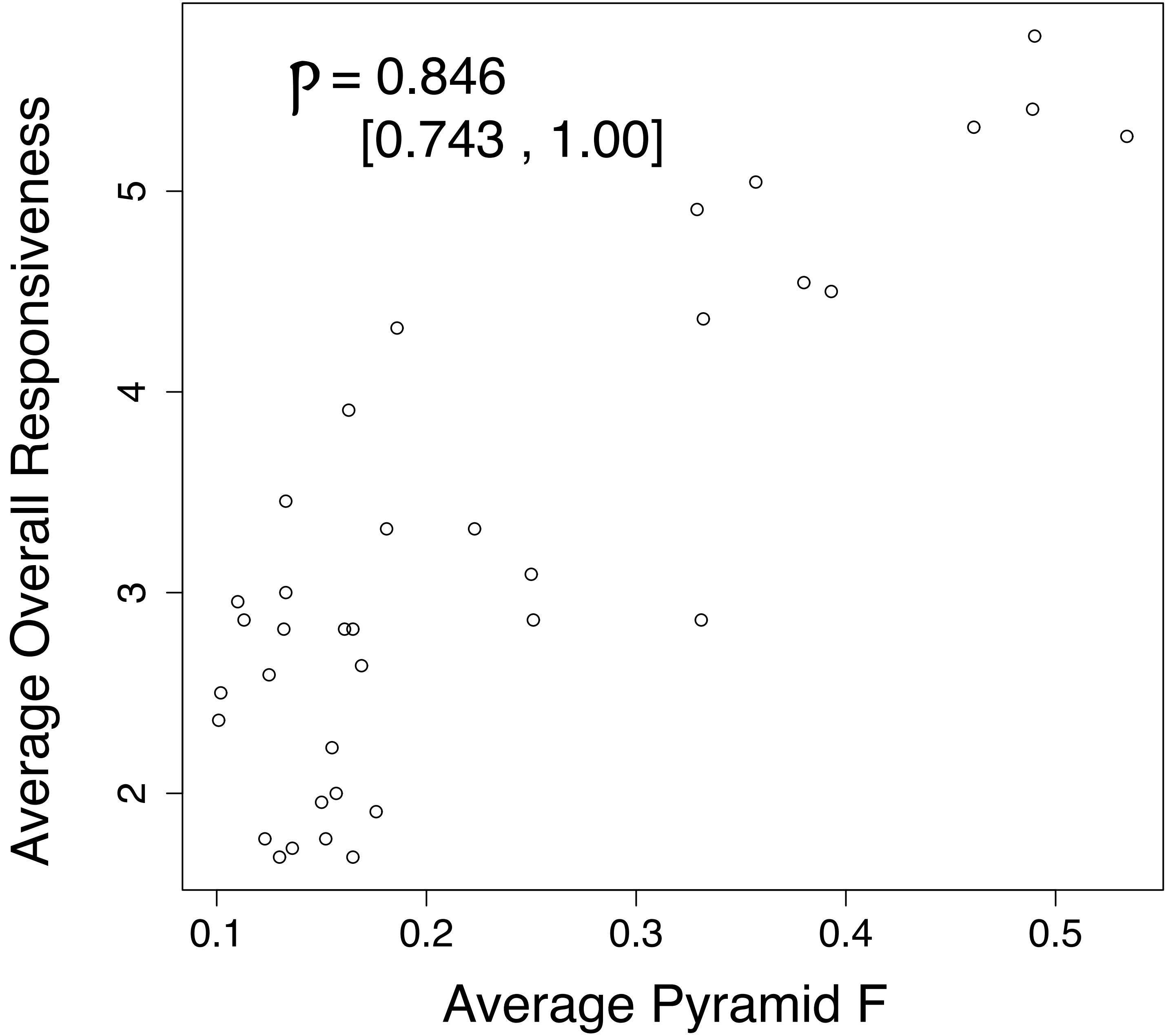


Structure/Coherence (2.8)

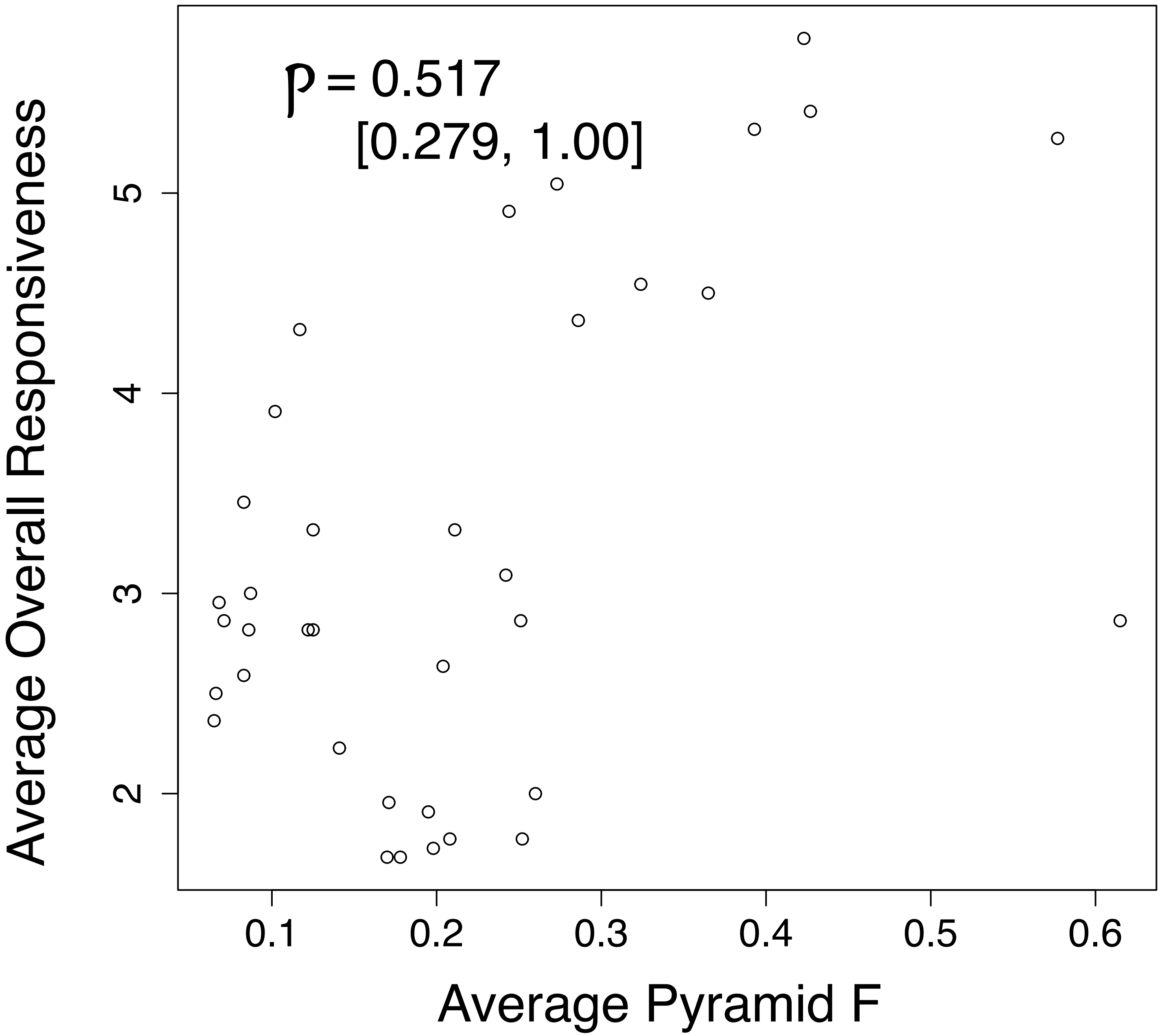


Overall Readability (3.7)

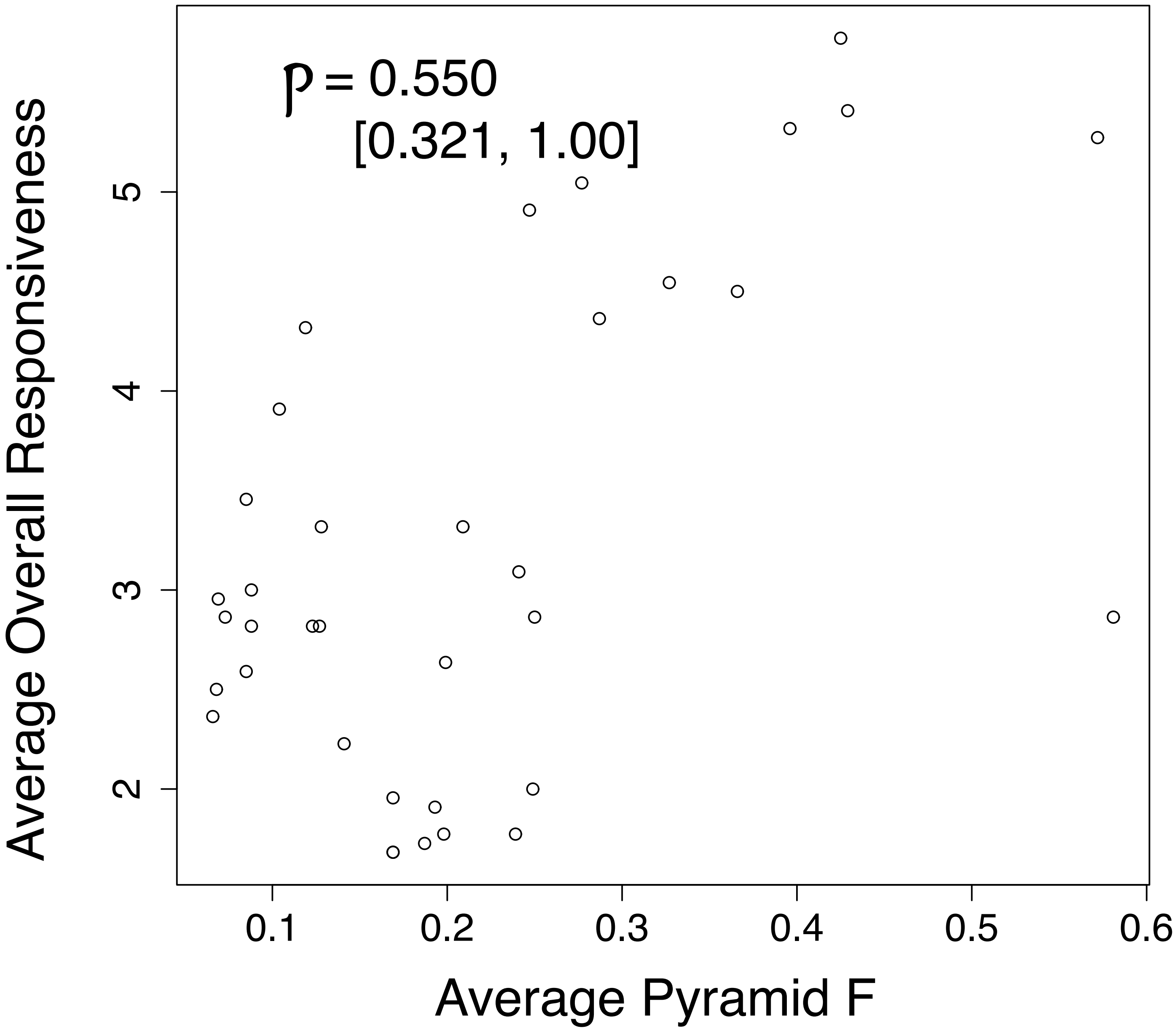
Overall Responsiveness vs. F (Beta = 1.0)



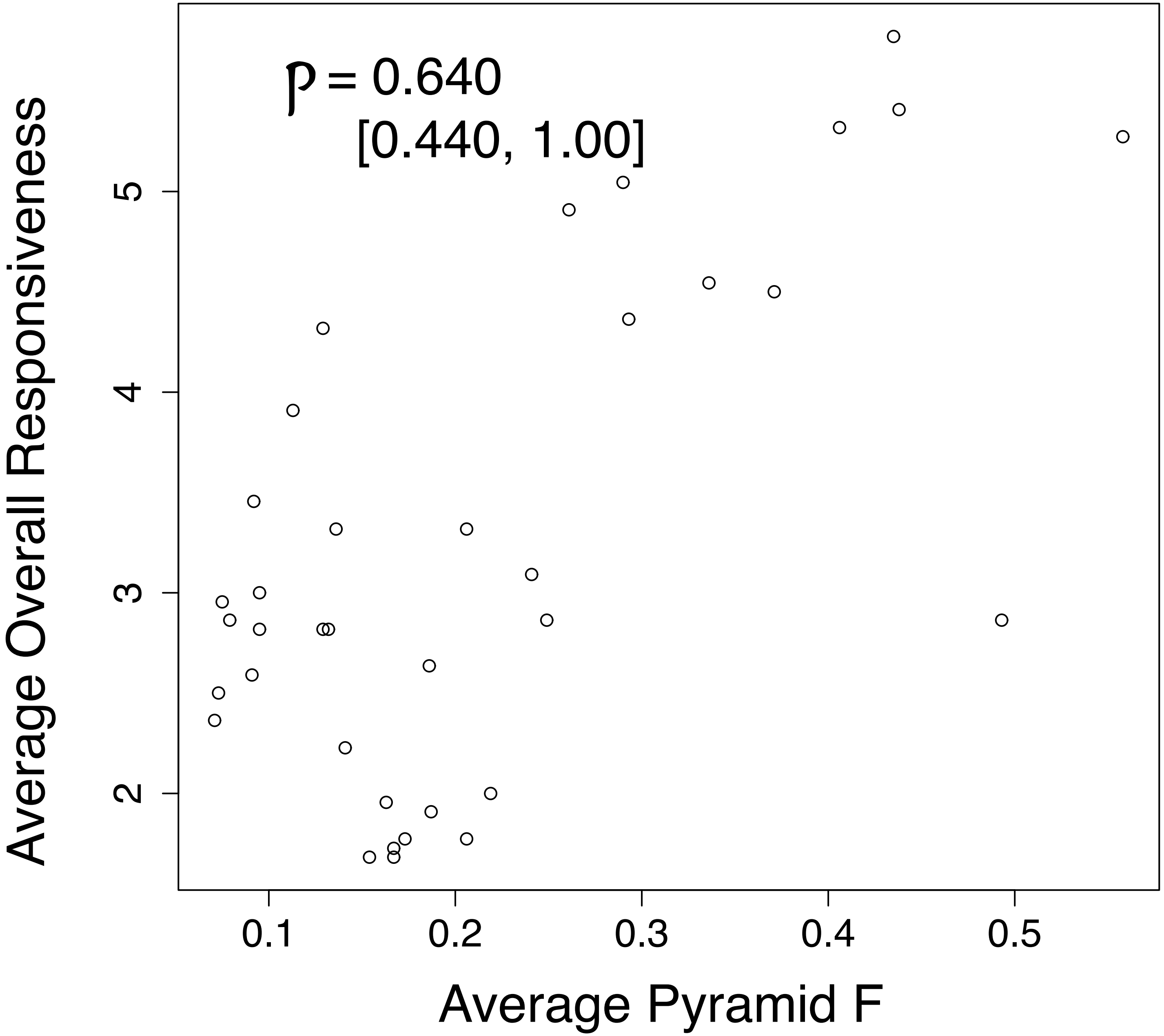
Overall Responsiveness vs. F (Beta = 0.1)



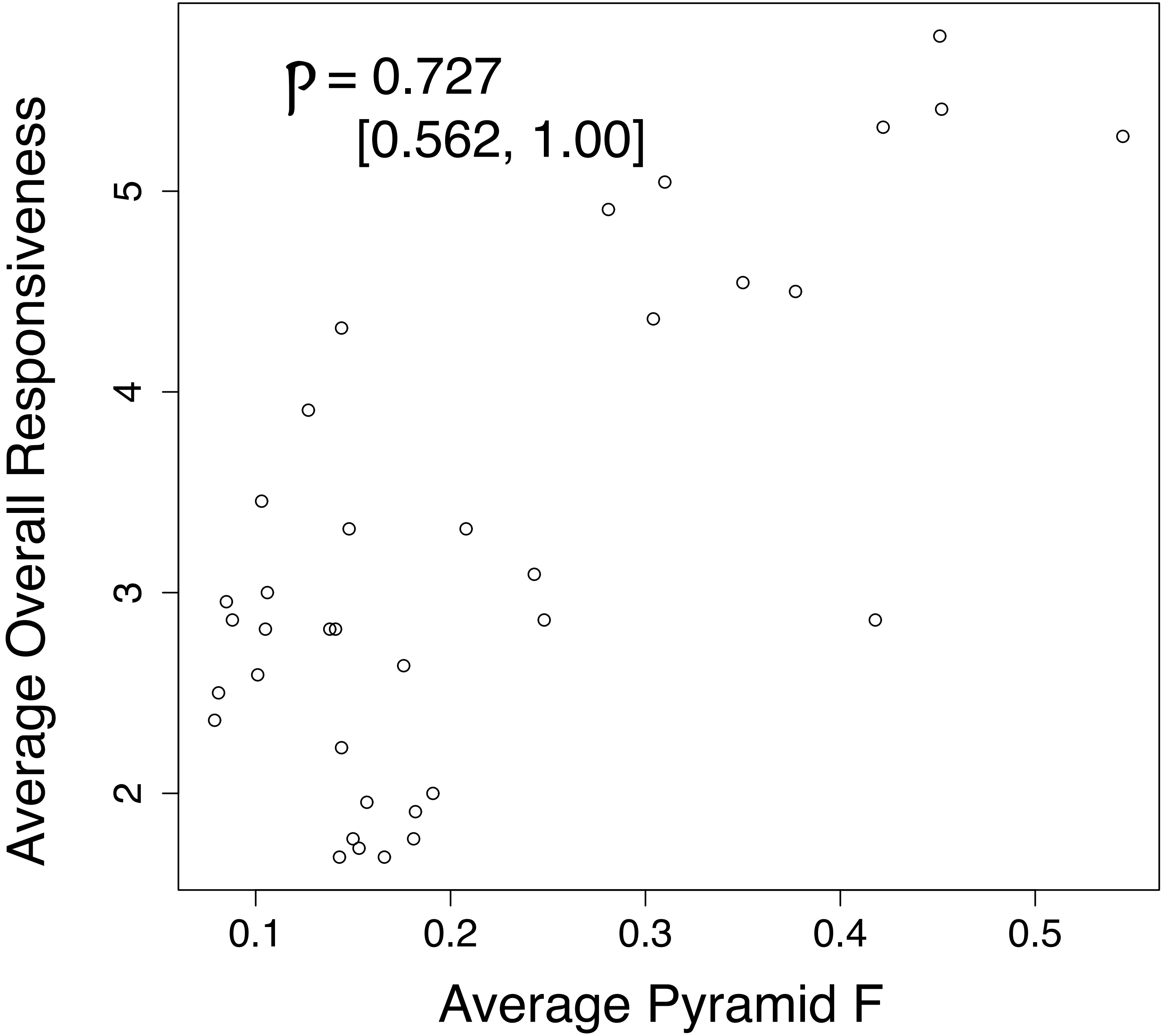
Overall Responsiveness vs. F (Beta = 0.2)



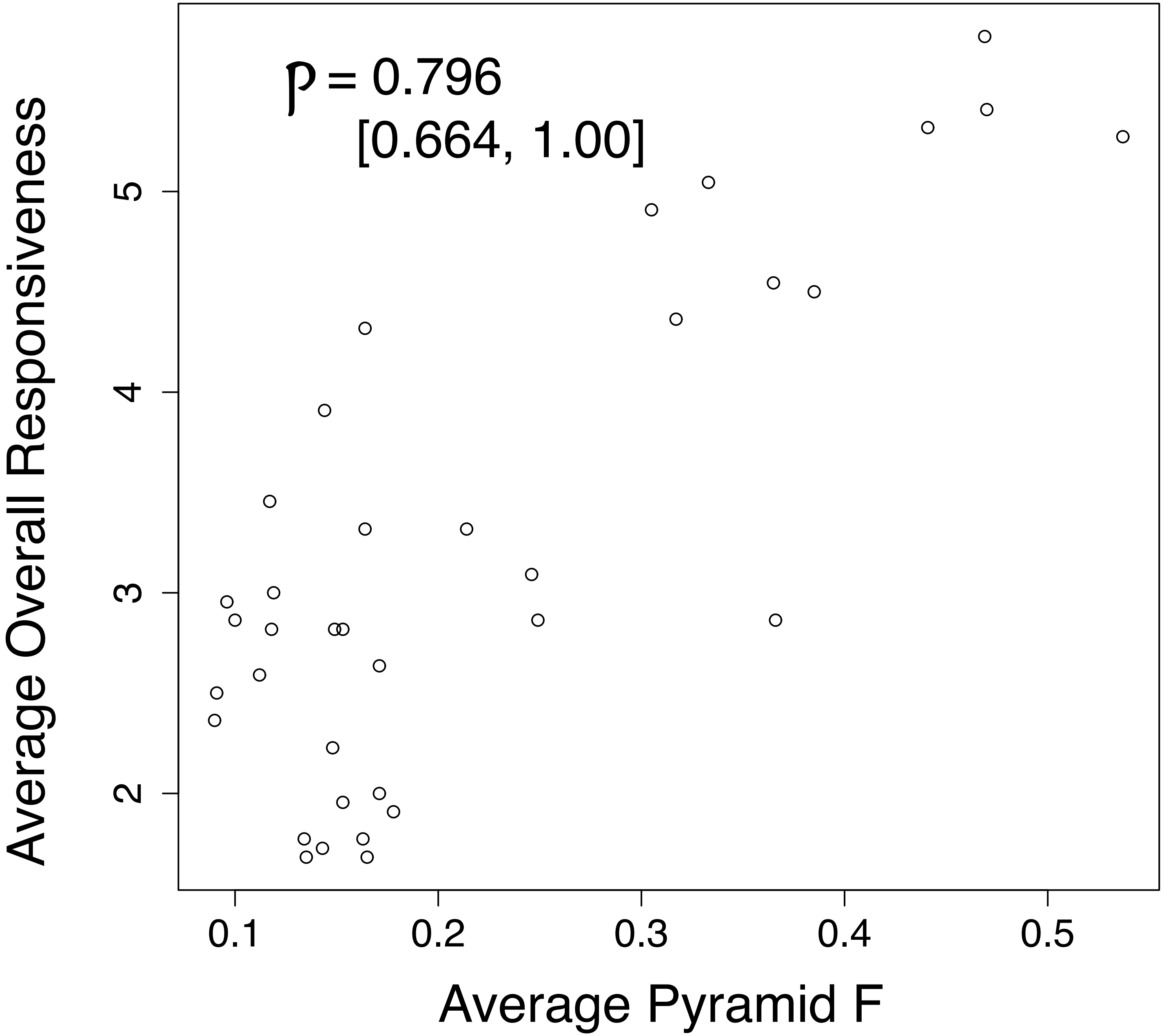
Overall Responsiveness vs. F (Beta = 0.4)



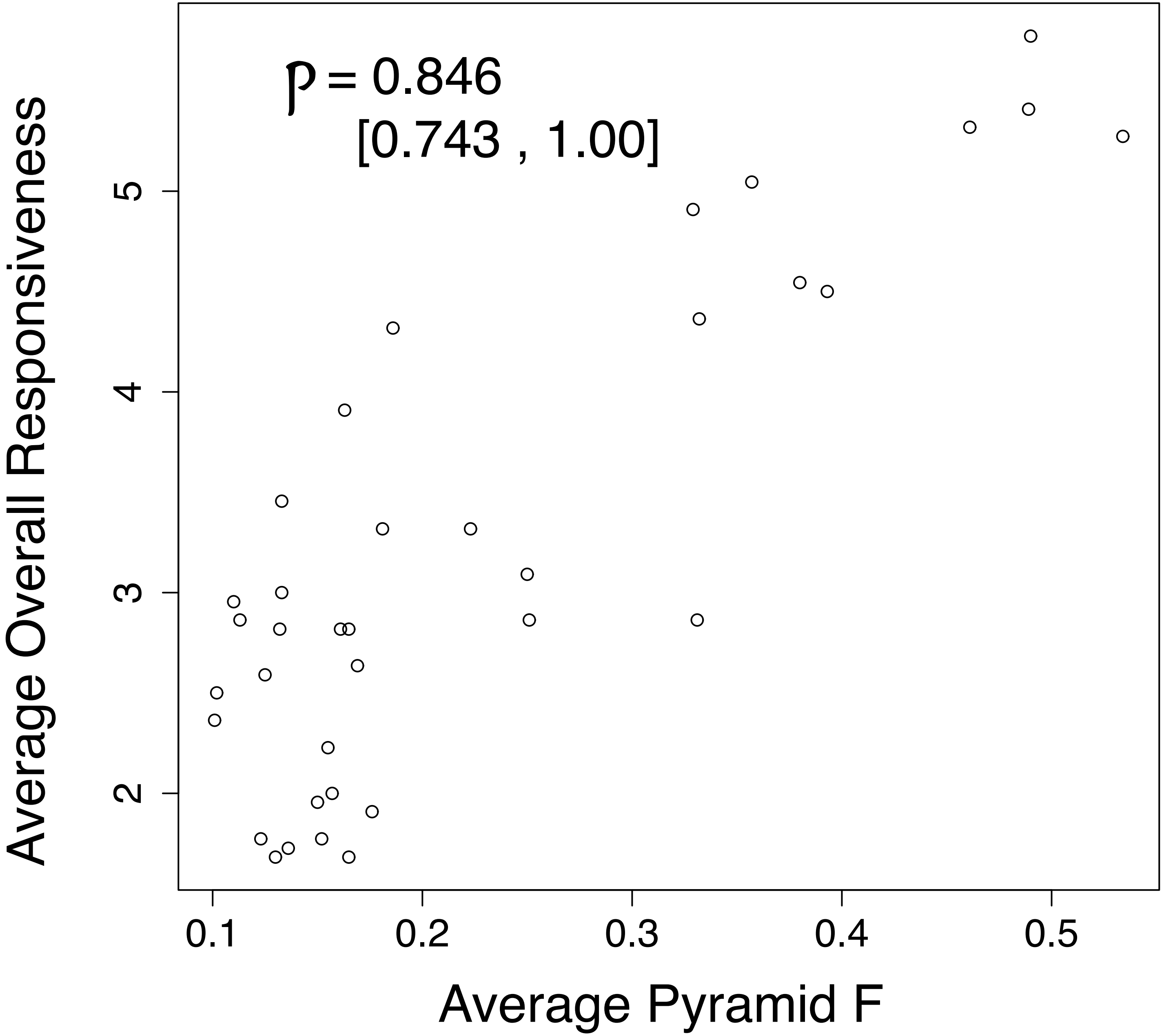
Overall Responsiveness vs. F (Beta = 0.6)



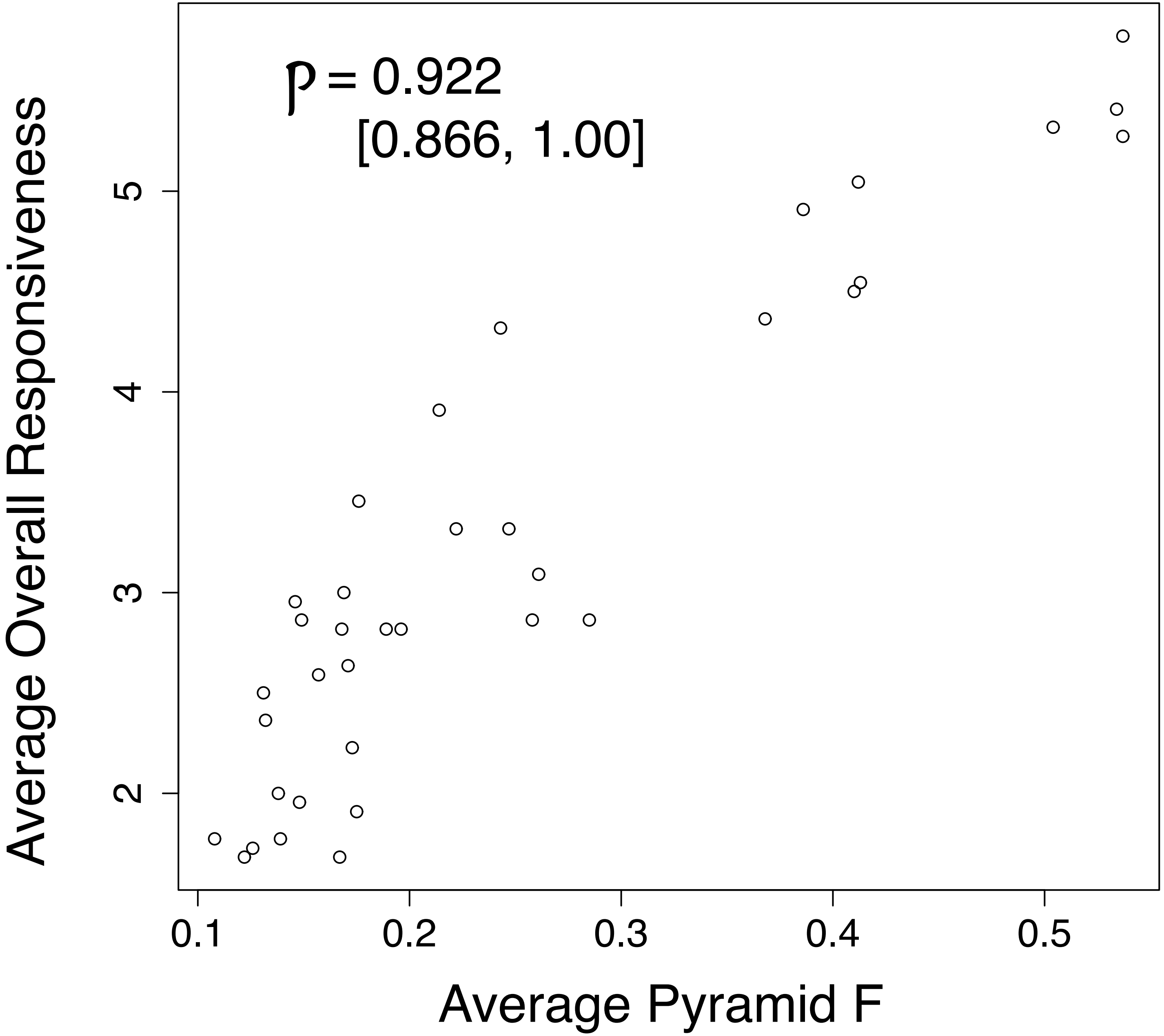
Overall Responsiveness vs. F (Beta = 0.8)



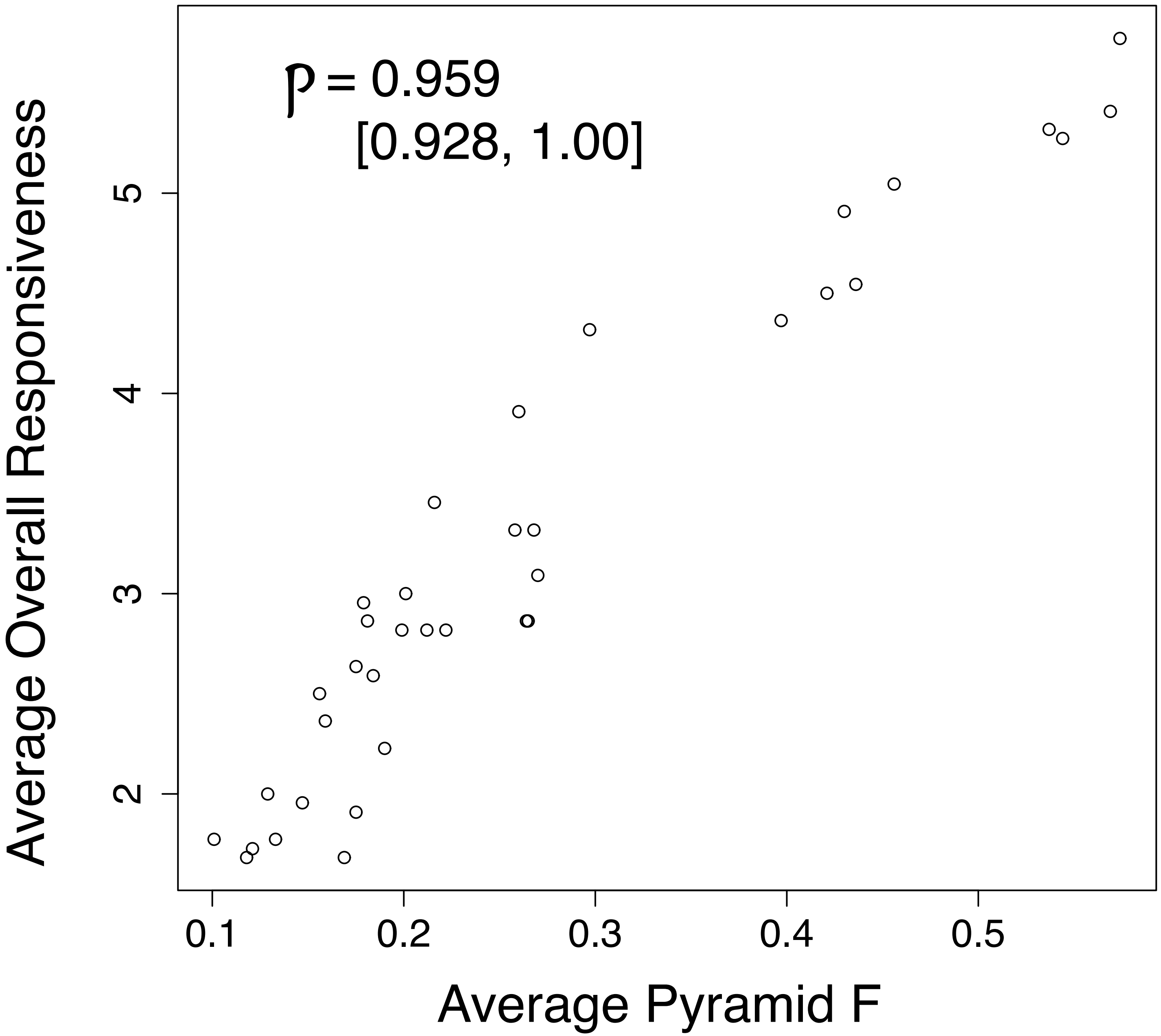
Overall Responsiveness vs. F (Beta = 1.0)



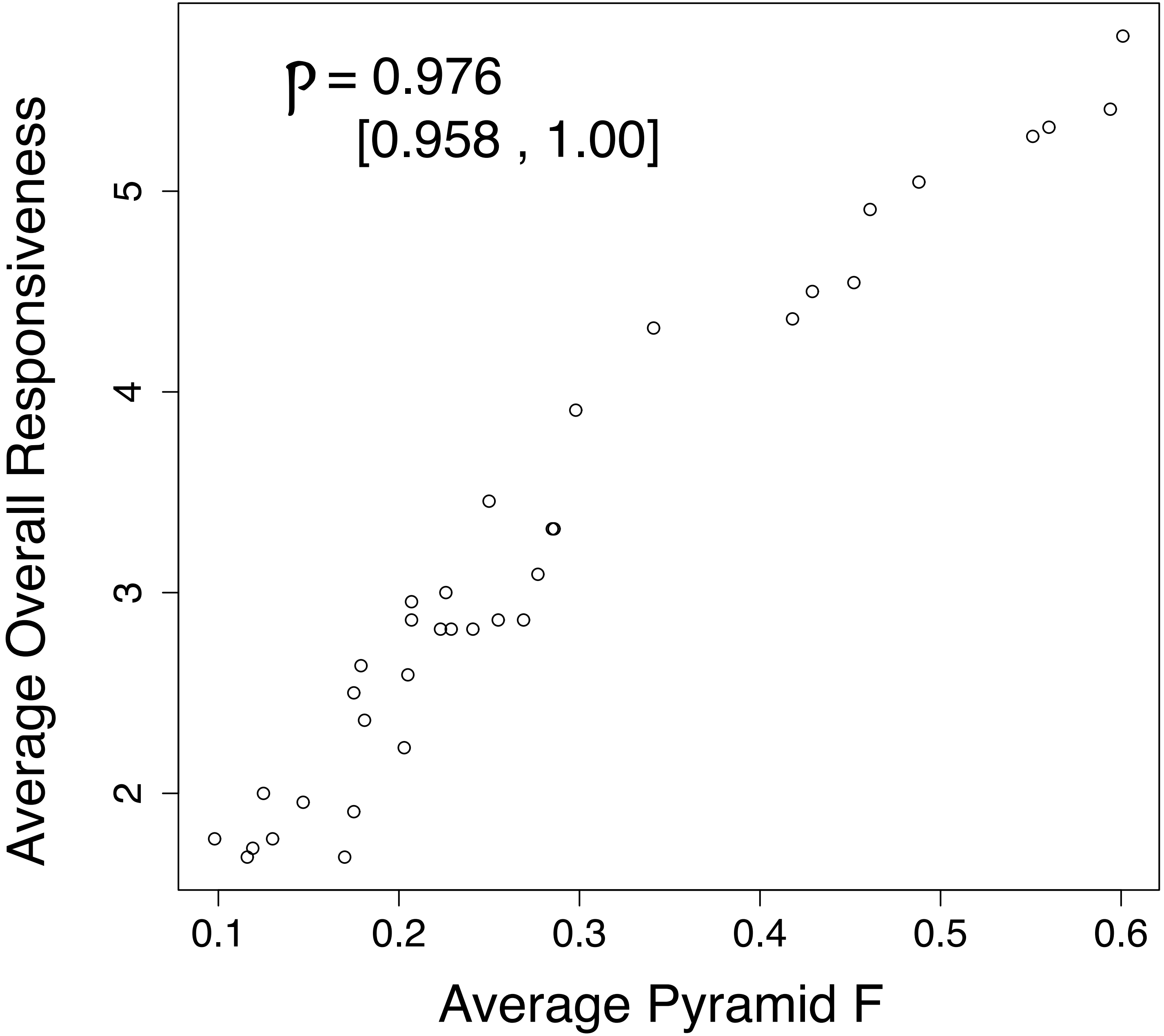
Overall Responsiveness vs. F (Beta = 1.5)



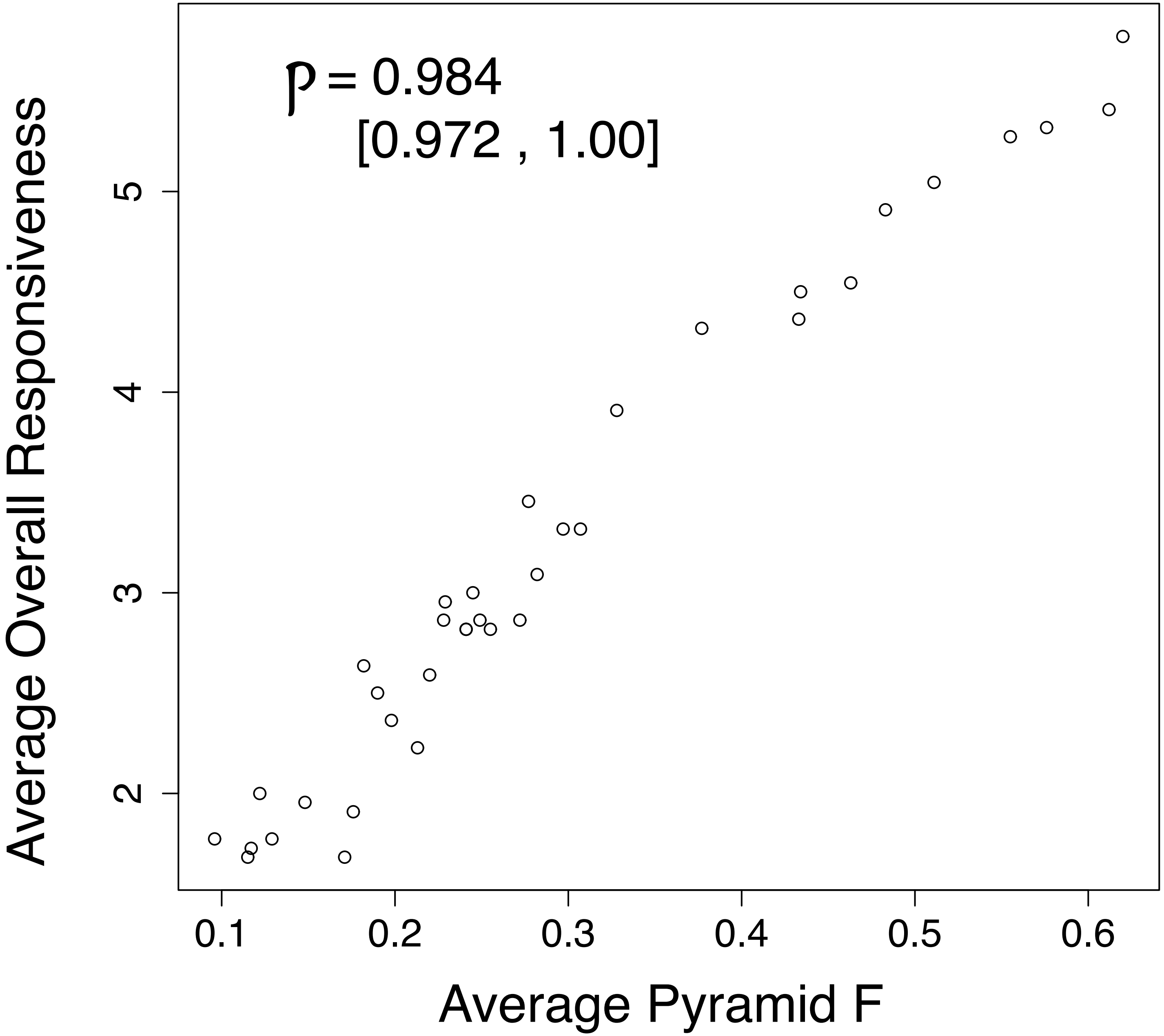
Overall Responsiveness vs. F (Beta = 2.0)



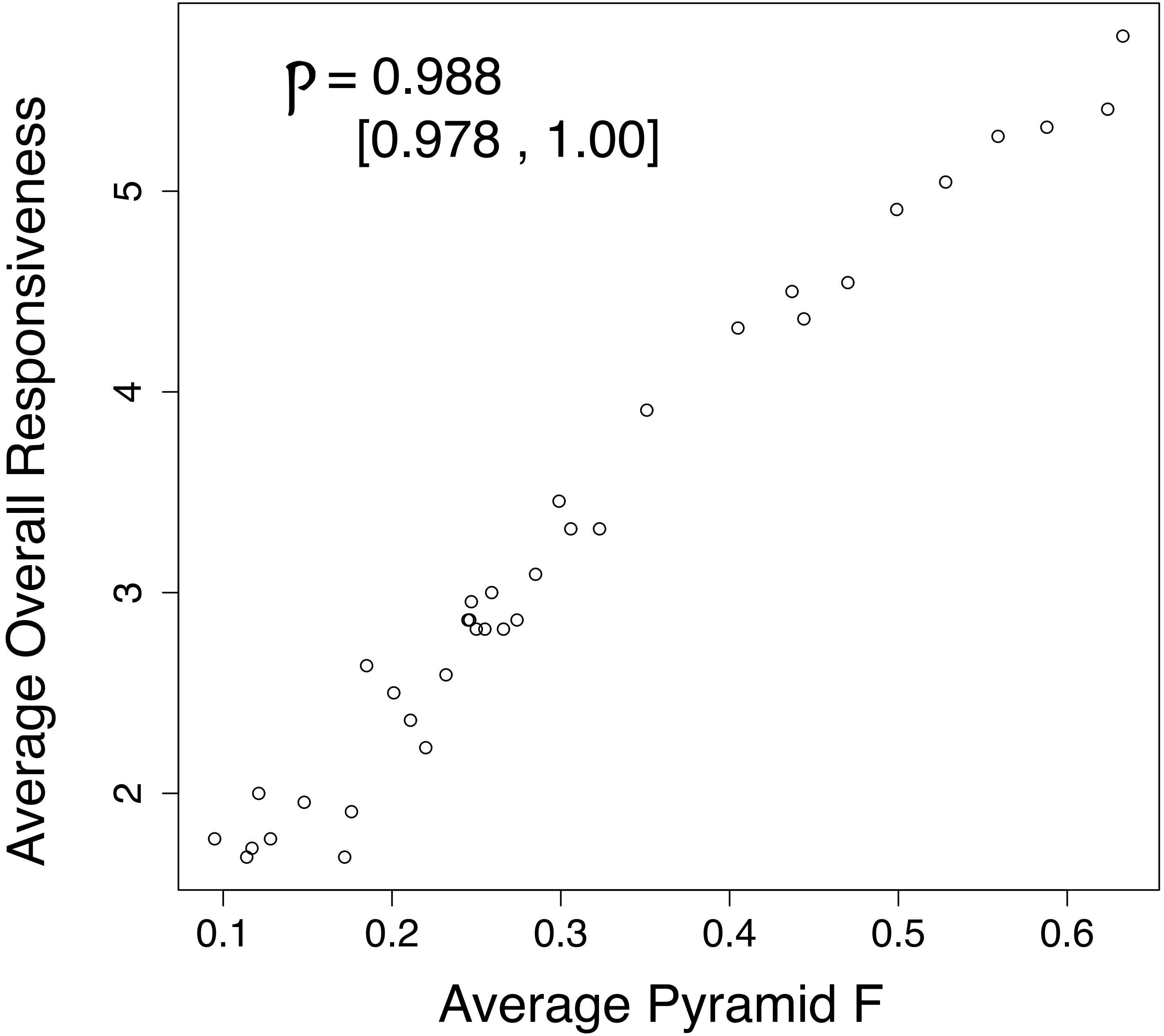
Overall Responsiveness vs. F (Beta = 2.5)



Overall Responsiveness vs. F (Beta = 3.0)

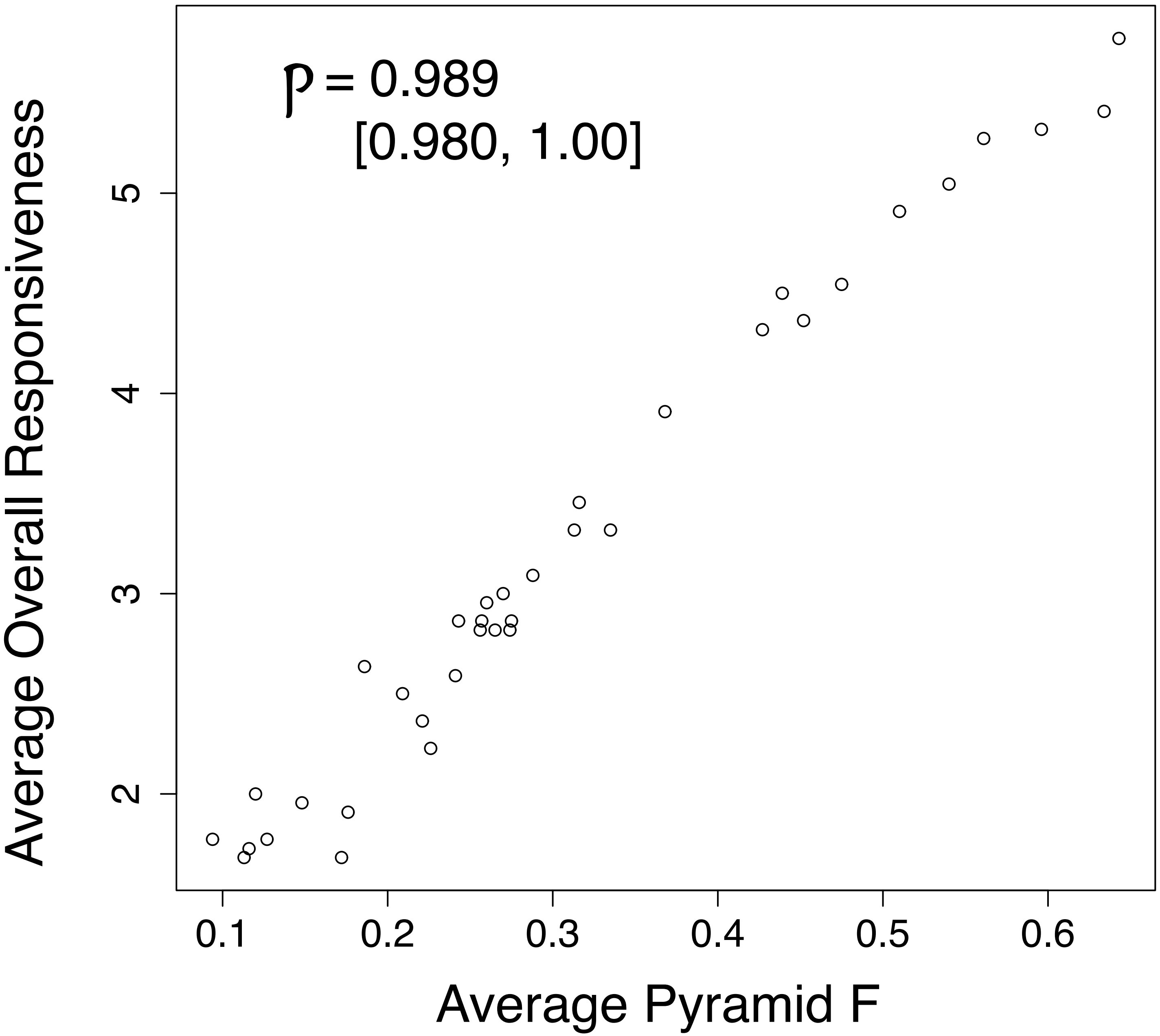


Overall Responsiveness vs. F (Beta = 3.5)

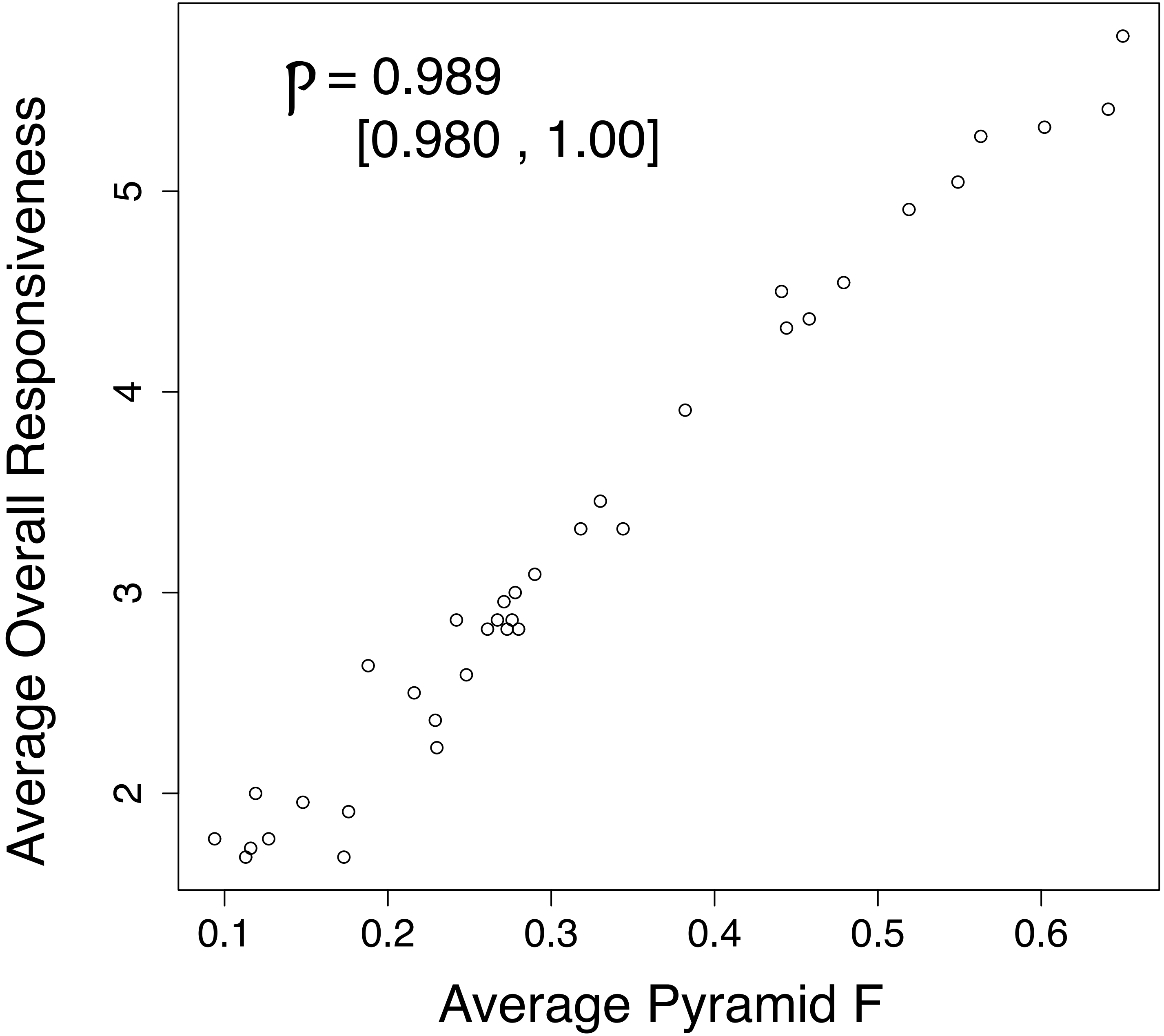


NIST

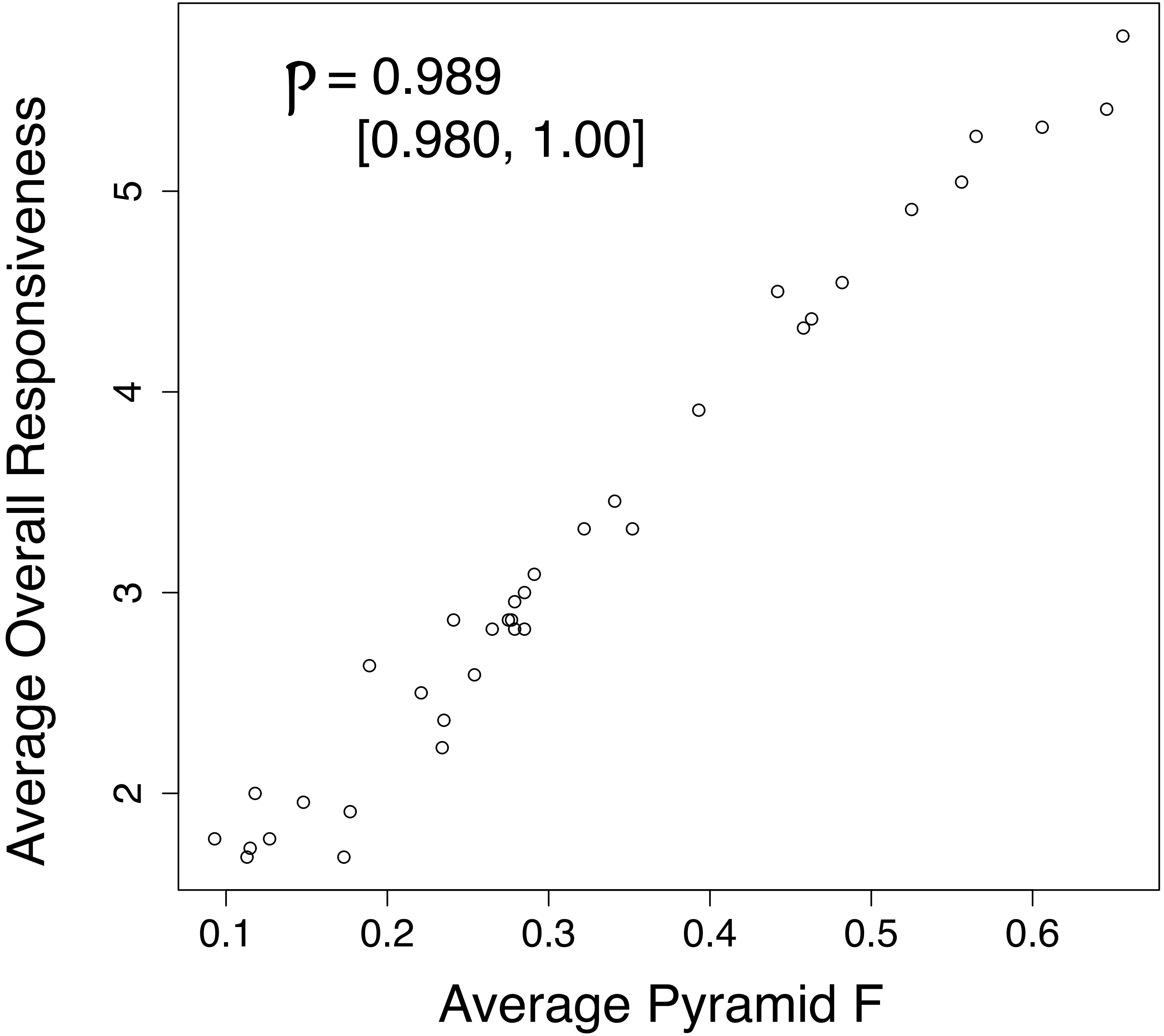
Overall Responsiveness vs. F (Beta = 4.0)



Overall Responsiveness vs. F (Beta = 4.5)



Overall Responsiveness vs. F (Beta = 5.0)



Discussion: Summarization

- Blog source text itself is low on readability
    - Most extractive summaries have barely acceptable grammaticality, non-redundancy
    - Most extractive summaries have poor coherence and overall readability
- Overall responsiveness highly correlated with pyramid F-score (Beta = 3...5); content dominates overall responsiveness
  - Recall is (still) more important than precision

Conclusion

- Large interest in opinion QA/summarization task
- Pilot task suggests possible modifications to future tasks
  - QA evaluation: vital/okay distinction may not be required for current opinion QA systems
  - Summarization task:
    - One summary per question
    - [fixed summary length, no credit given for shorter summaries]
- May repeat task if sufficient interest and resources