TAMING THE TIGER: DEVELOPING A VALID AND RELIABLE ASSESSMENT SYSTEM IN PARTNERSHIP WITH FACULTY

Dr. Laura Hart

Dr. Teresa Petty

University of North Carolina at Charlotte

Two parts to our presentation:

1. Establishing our content validity protocol
2. Beginning our reliability work

Content Validity Protocol is available
http://edassessment.uncc.edu

Developing Content Validity Protocol

Setting the stage …

Stop and Share:

- What have you done to build this capacity at your institution for validity work?

(turn and share with a colleague: 1 min.)

Setting the Stage

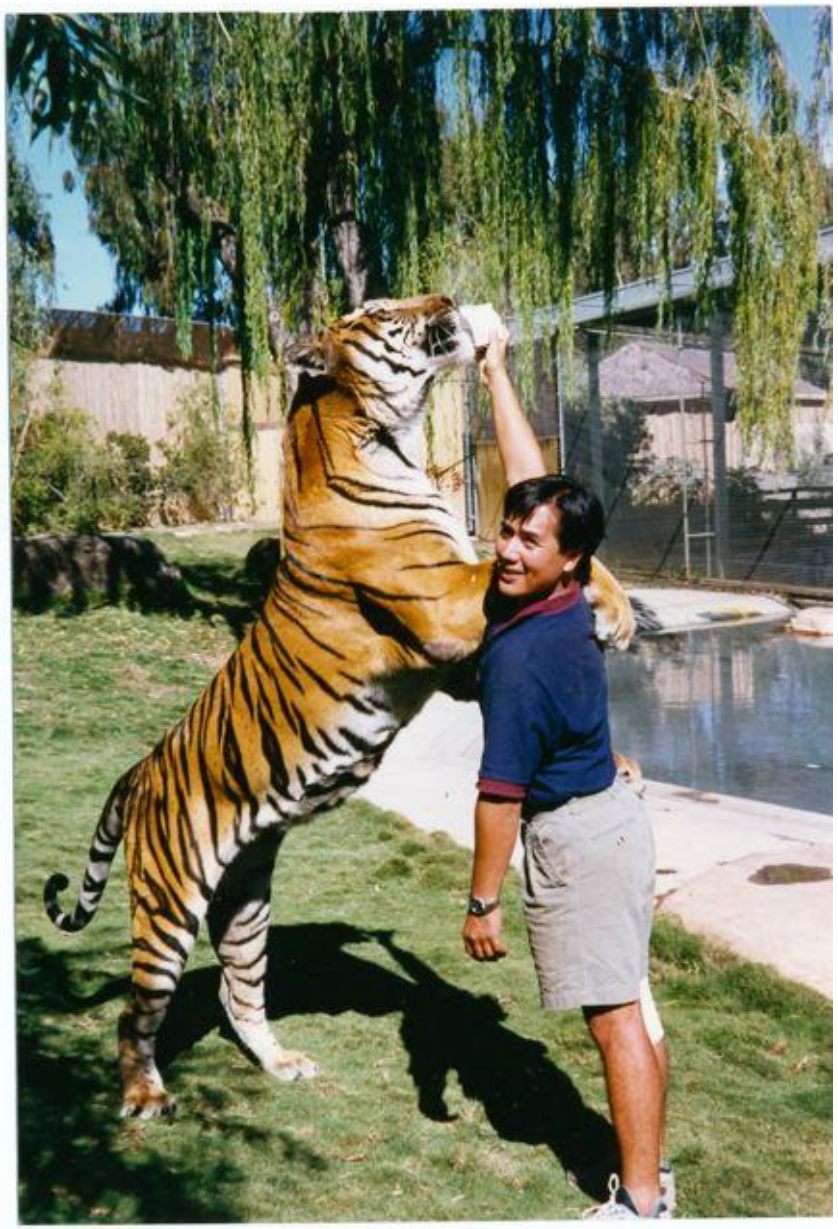
- Primarily with advanced programs where we had our “homegrown” rubrics
- Shared the message early and often: “It’s coming!” (6-8 months; spring + summer)
- Dealing with researchers → use research to make the case
- CAEP compliance was incidental → framed in terms of “best practice”
- Used expert panel approach -- simplicity
- Provided one-page summary of why we need this, including sources, etc.

Using the Right Tools

Using the Right Tools

- Started with CAEP Assessment Rubric / Standard 5
- Distilled it to a Rubric Review Checklist (“yes/no”)
- Anything that got a “no” → fix it
- Provided interactive discussion groups for faculty to ask questions – multiple dates and times
- Provided examples of before and after
  - Asked “which version gives you the best data?”
  - Asked “which version is clearer to students?”
- Created a new page on the website
- Created a video to explain it all

The “Big Moment”



The “Big Moment” – creating the response form

Example: Concept we want to measure: Content Knowledge

	Level 1	Level 2	Level 3
K2a: Demonstrates knowledge of content	• Exhibits lapses in content knowledge	• Exhibits growth beyond basic content knowledge	• Exhibits advanced content knowledge
K2b:Implements interdisciplinary approaches and multiple perspectives for teaching content	• Seldom encourages students to integrate knowledge from other areas	• Teaches lessons that encourage students to integrate 21st century skills and apply knowledge from several subject areas	• Frequently implements lessons that encourage students to integrate 21st century skills and apply knowledge in creative ways from several subject areas

The “Big Moment” – creating the response form

- Could create an electronic version or use pencil and paper
- Drafted a letter to use/include to introduce it
- Rated each item 1-4 (4 being highest) on
  - Representativeness of item
  - Importance of item in measuring the construct
  - Clarity of item
- Open ended responses to allow additional info

Talking to Other Tigers

Talking to Other Tigers (experts)

- Minimum of 7 (recommendation from lit review)
- 3 internal
- 4 external (including at least 3 community practitioners from field)
- Mixture of IHE Faculty (i.e., content experts) and B12 school or community practitioners (lay experts). Minimal credentials for each expert should be established by consensus from program faculty; credentials should bear up to reasonable external scrutiny (Davis, 1992).

Compiling the Results (seeing the final product)

Compiling the Results

- Submitted results to shared folder
- Generated a Content Validity Index (CVI) (calculated based on recommendations by Rubio et. al. (2003), Davis (1992), and Lynn (1986)):

  - $\frac{\text{The number of experts who rated the item as 3 or 4}}{\text{The number of total experts}}$

- A CVI score of .80 or higher will be considered acceptable.
- Working now to get the results posted online and tied to SACS reports

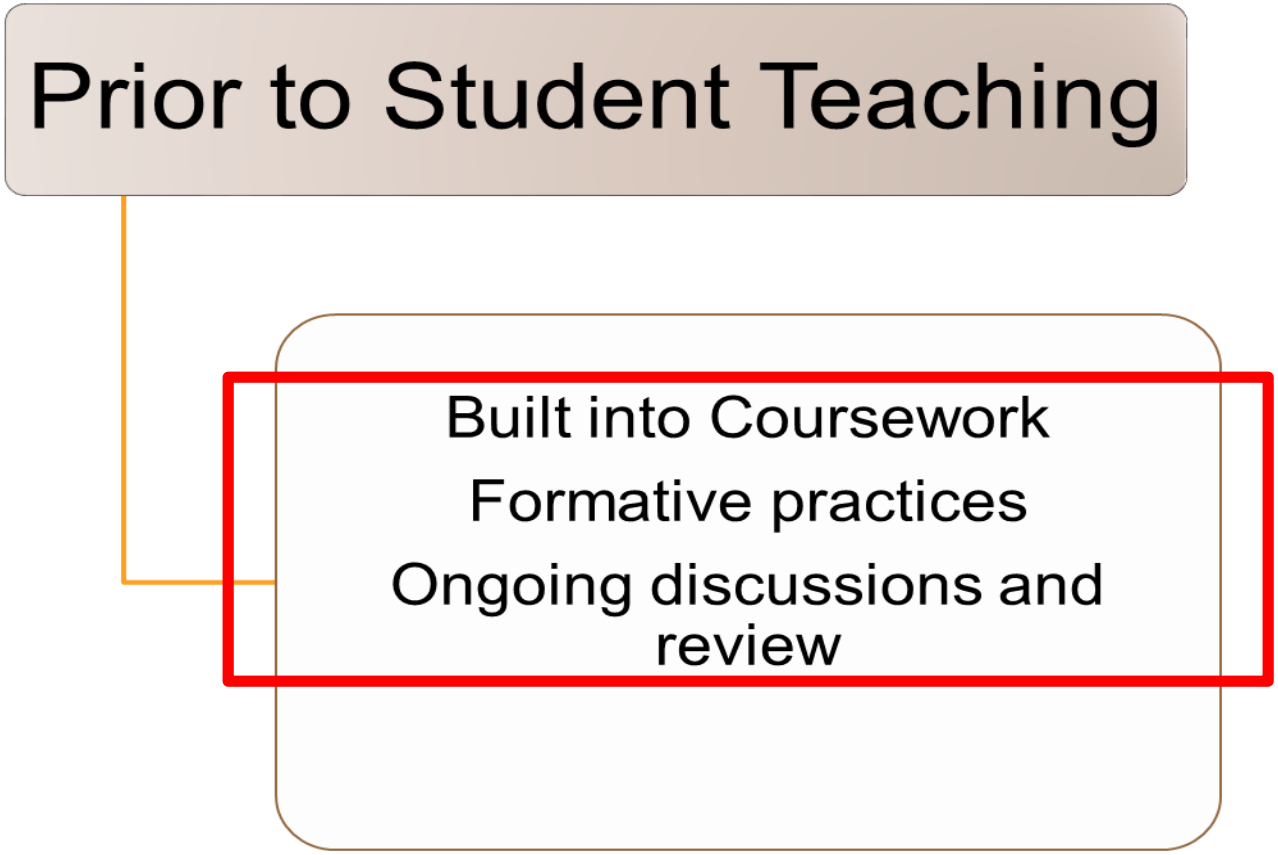
Stop and Share:

- Based on what you've heard, what can you take back and use at your EPP?

(Turn and talk: 1 minute)

Beginning Reliability Work

- Similar strategies as with Validity: “logical next step”
- Started with edTPA (key program assessment):



Focused on outcomes

- CAEP → incidental
- Answering programmatic questions became the focus:
  - Do the planned formative tasks and feedback loop across programs support students to pass their summative portfolios? Are there varying degrees within those supports (e.g., are some supports more effective than others)?
  - Are there patterns in the data that can help our programs better meet the needs of our students and faculty?
  - Are faculty scoring candidates reliably across courses and sections of a course?

Background: Building edTPA skills and knowledge into Coursework

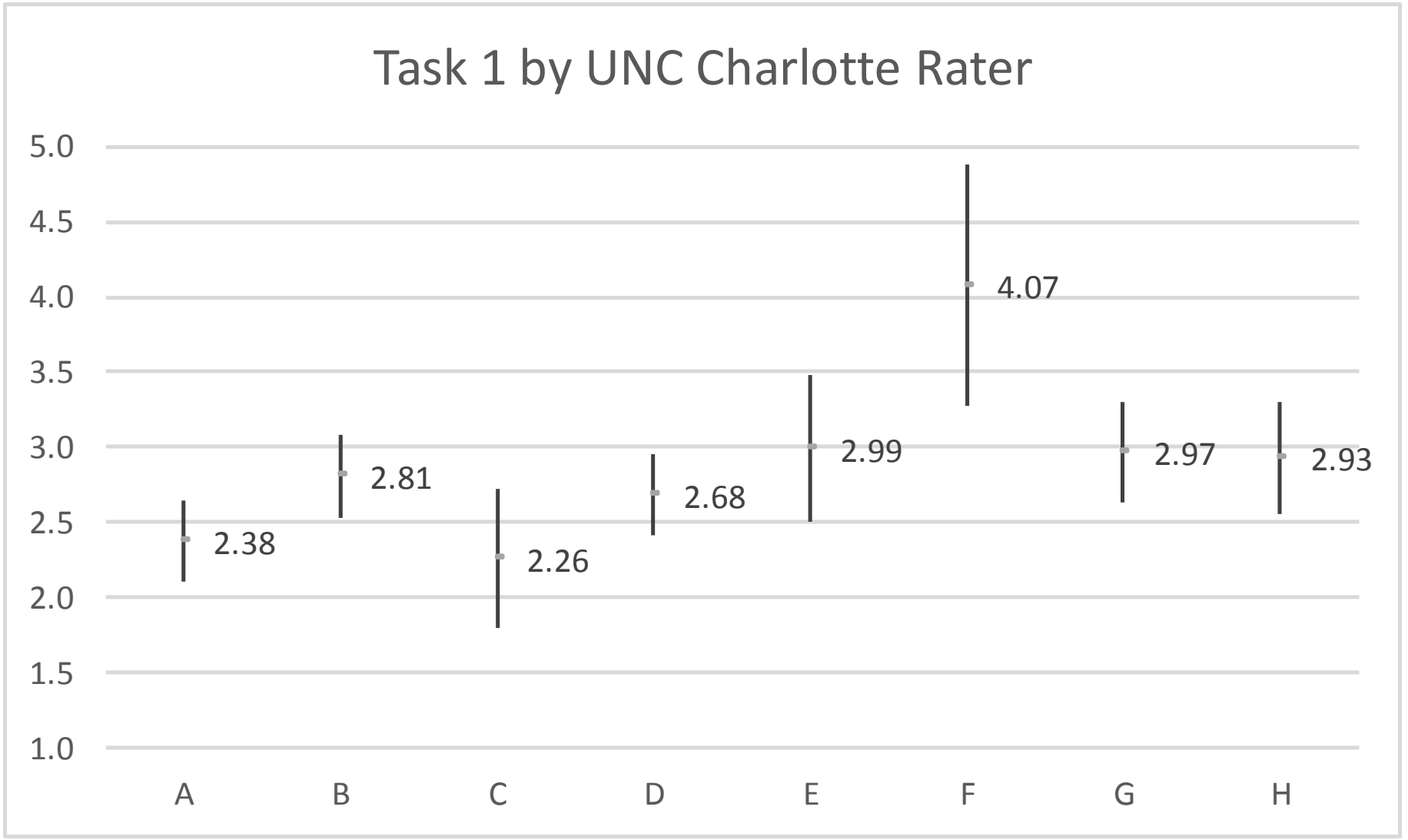
- Identified upper-level program courses that aligned with domains of edTPA (Planning, Implementation, Assessment)
- Embedded “practice tasks” into these courses
- Becomes part of course grade
- Data are recorded through TaskStream assessment system; compared later to final results
- Program wide support and accountability (faculty identified what “fit” into their course regarding major concepts within edTPA even if not practice task)

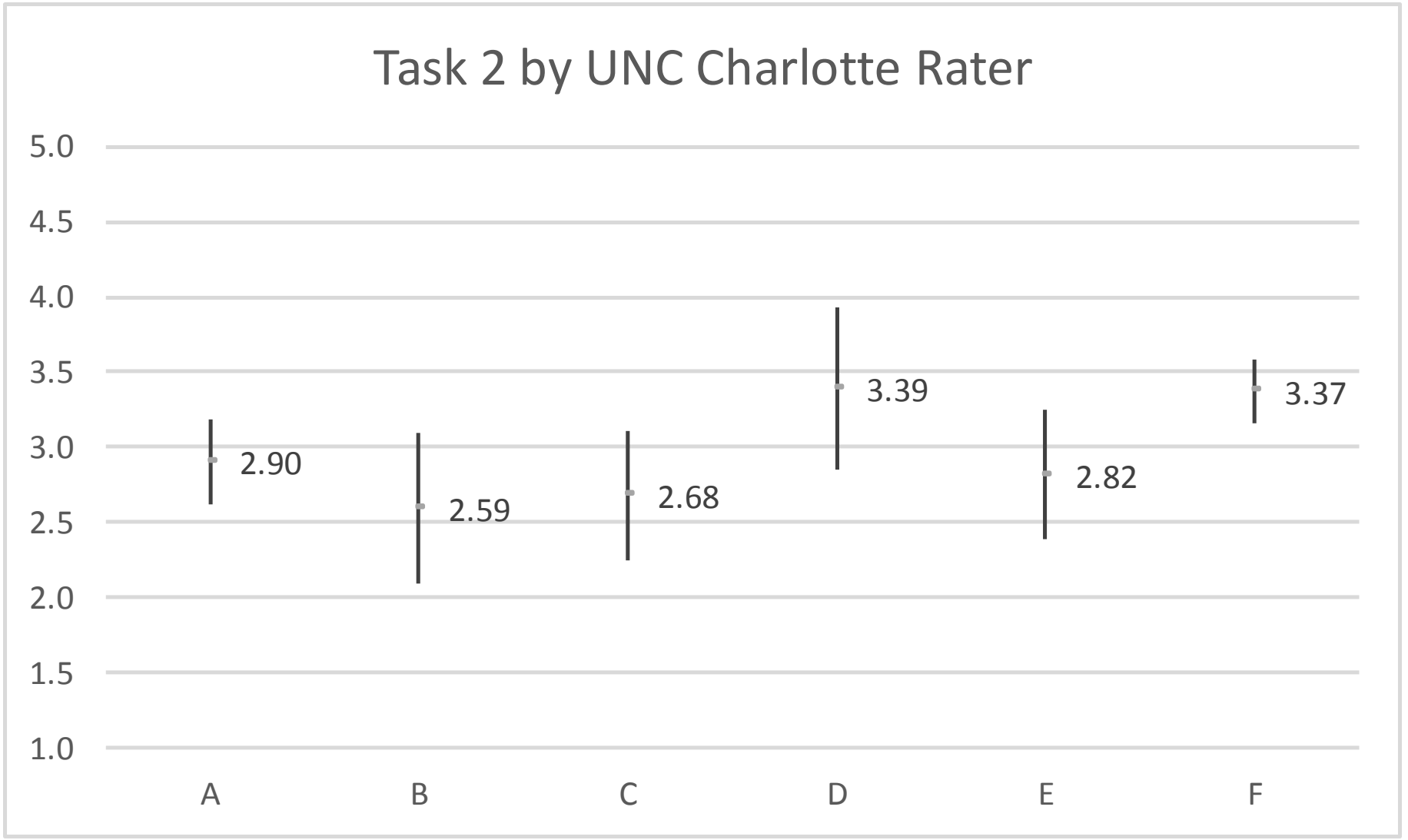
Data Sources

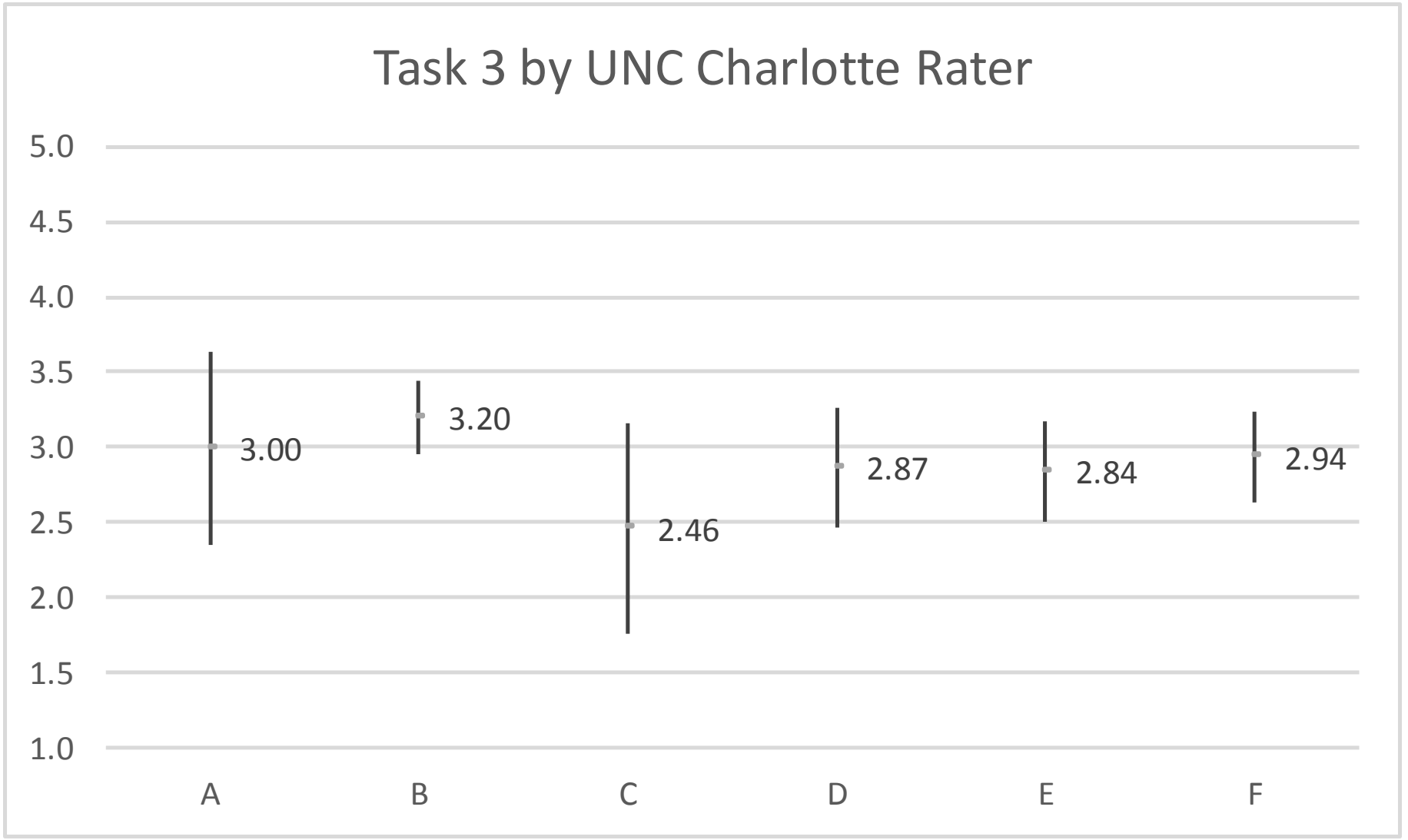
- Descriptive Data
  - Scores from Formative edTPA tasks scored by UNC Charlotte faculty
  - Scores from summative edTPA data (Pearson)
- Feedback
  - Survey data from ELED faculty

Examination of the edTPA Data

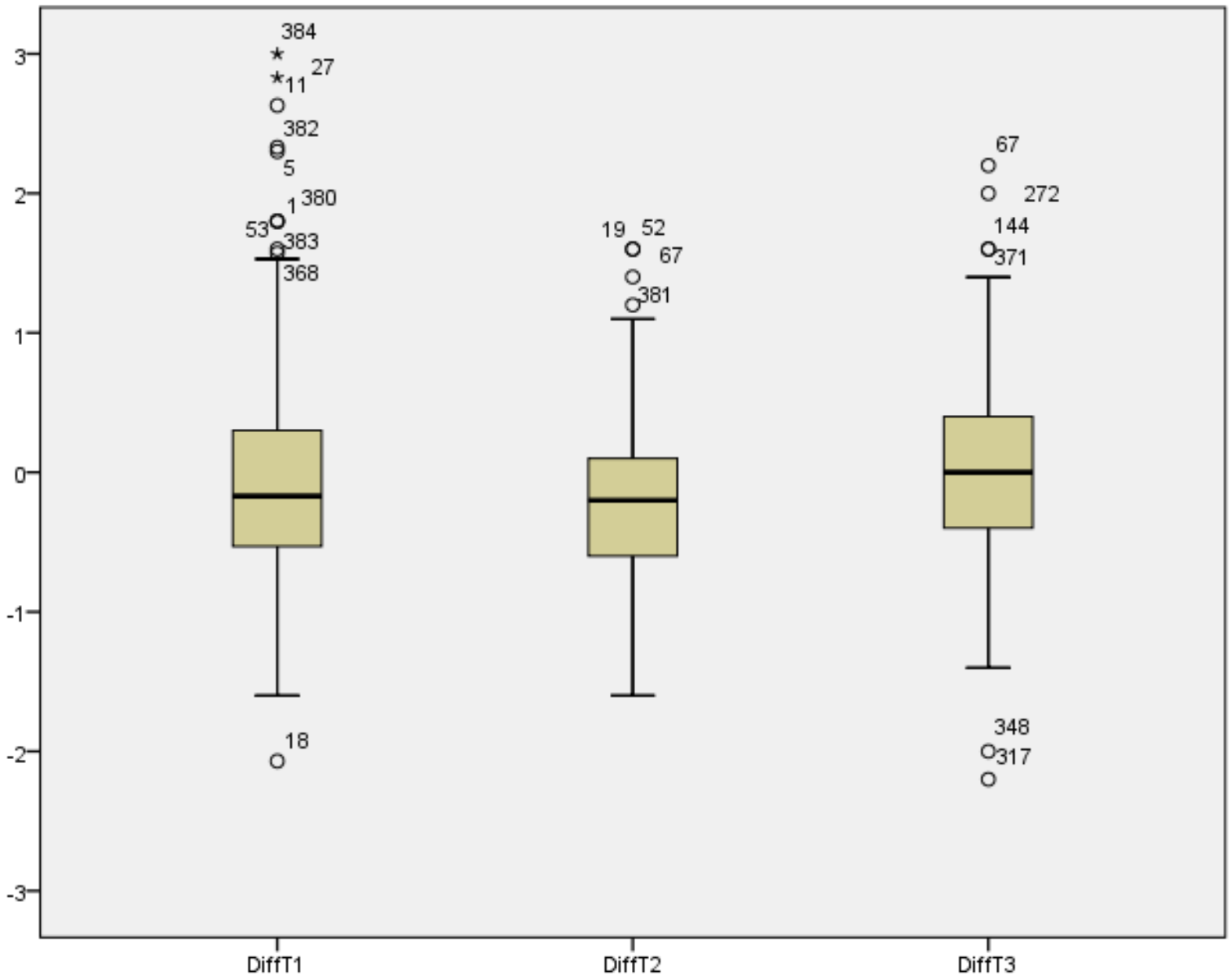
- Statistically significant differences between our raters in means and variances by task
- Low correlations between our scores and Pearson scores
- Variability between our raters in their agreement with Pearson scores
- Compared Pass and Fail Students on our Practice Scores
- Created models to predict scores based on demographics

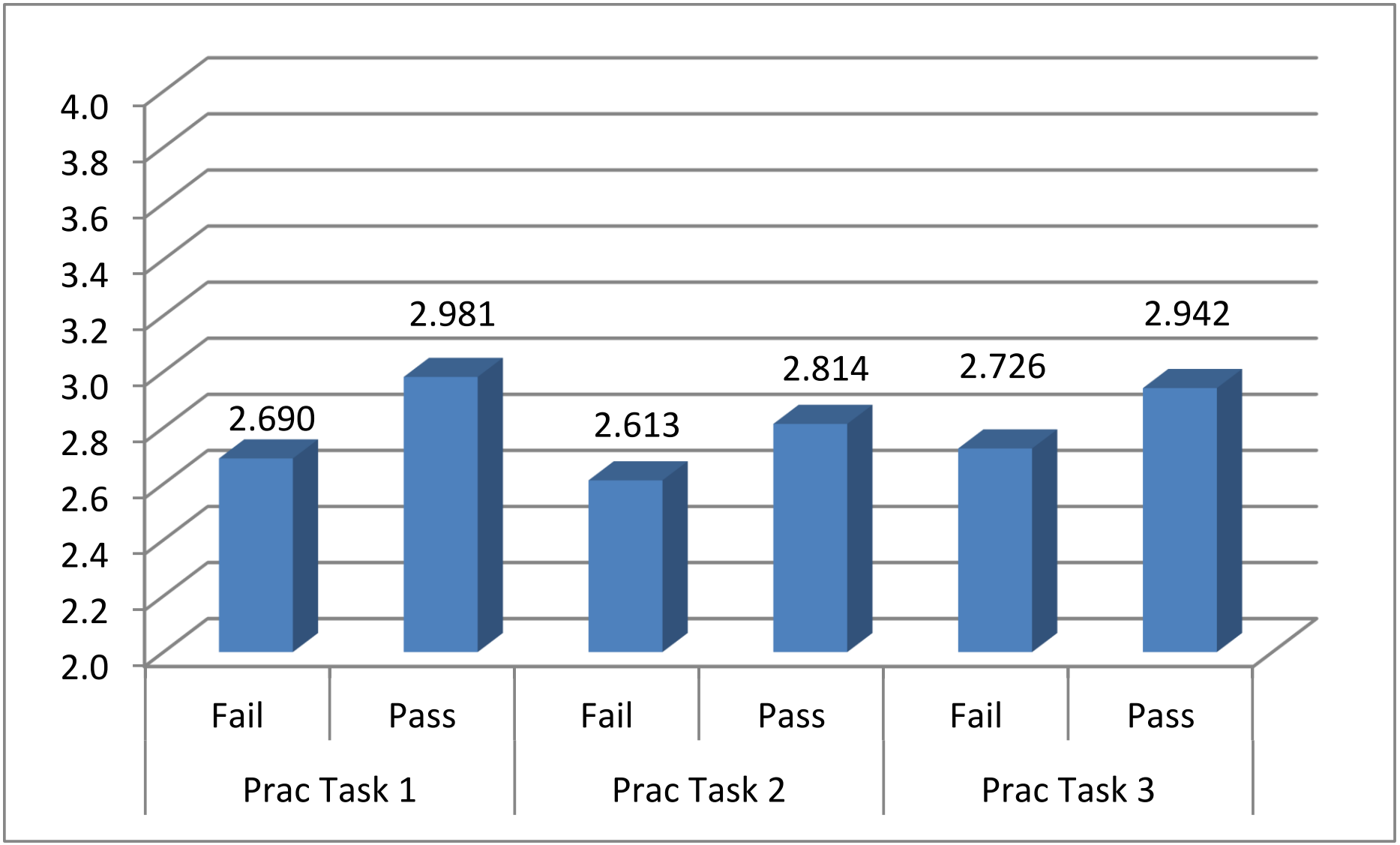

Task 1 by UNC Charlotte Rater
5.0
4.5
4.0
3.5
3.0
2.5
2.0
1.5
1.0
2.38
2.81
2.26
2.68
2.99
4.07
2.97
2.93
A
B
C
D
E
F
G
H

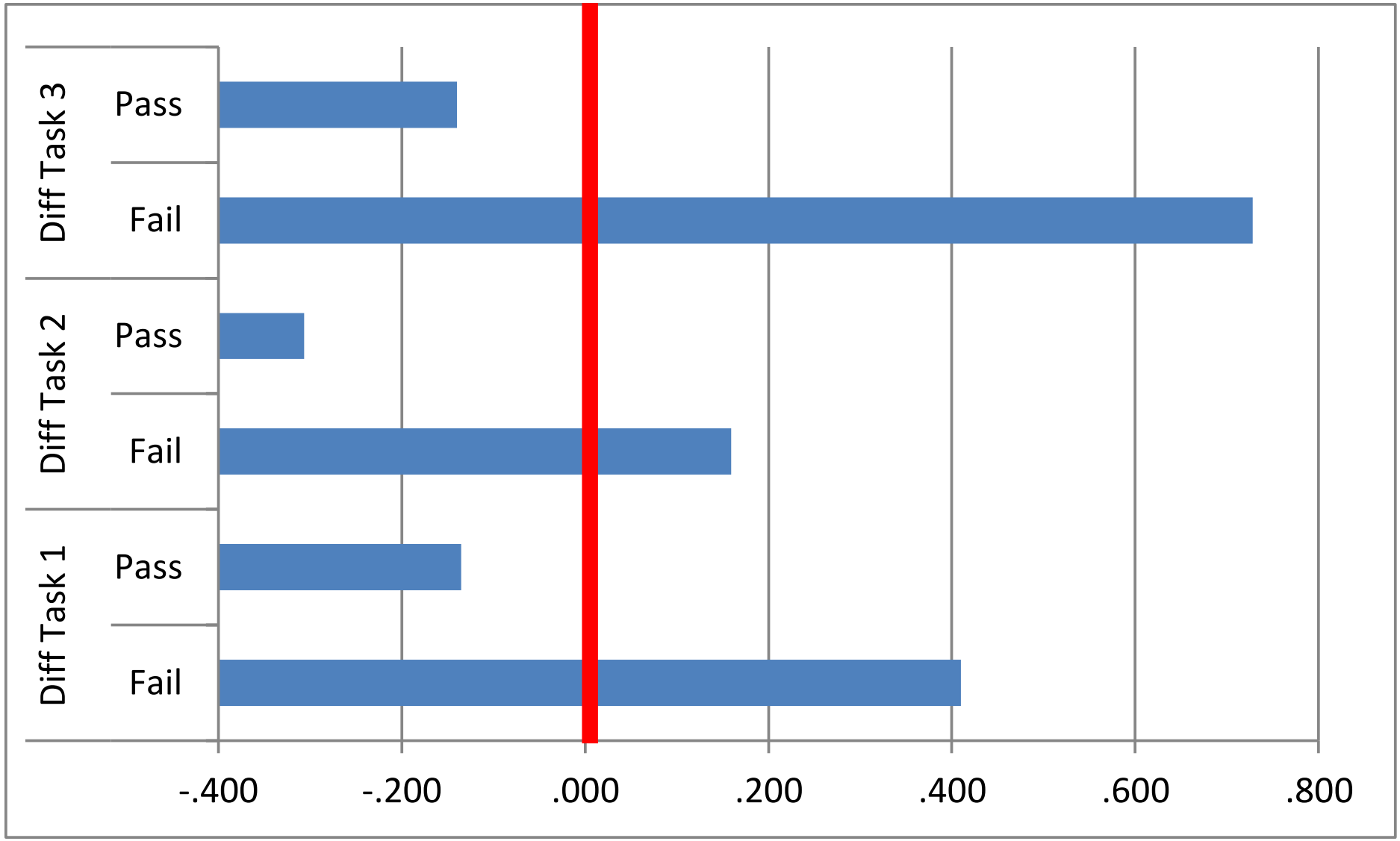

Task 2 by UNC Charlotte Rater
5.0
4.5
4.0
3.5
3.0
2.5
2.0
1.5
1.0
2.90
2.59
2.68
3.39
2.82
3.37
A
B
C
D
E
F


Task 3 by UNC Charlotte Rater
5.0
4.5
4.0
3.5
3.0
2.5
2.0
1.5
1.0
3.00
3.20
2.46
2.87
2.84
2.94
A
B
C
D
E
F

	Task 1	Task 2	Task 3
Pearson Total Score with UNCC Rater	.302	.138	.107
Pearson Task Score with UNCC Rater	.199	.225	.227
Lowest by UNCC Rater	.037	.125	.094
Highest by UNCC Rater	.629	.301	.430
Pearson Task Score with Pearson Total Score	.754	.700	.813
Difference Between Pearson and UNCC Rater			
Minimum	-2.070	-1.600	-2.200
25th	-0.530	-0.600	-0.400
50th	-0.130	-0.200	0.000
75th	0.330	0.200	0.400
Maximum	3.000	2.000	2.200


384
27
11
382
5
380
1
53
383
368
18
19
52
67
381
67
272
144
371
348
317
3
2
1
0
-1
-2
-3
DiffT1
DiffT2
DiffT3


4.0
3.8
3.6
3.4
3.2
3.0
2.8
2.6
2.4
2.2
2.0
2.690
2.981
2.613
2.814
2.726
2.942
Fail
Pass
Fail
Pass
Fail
Pass
Prac Task 1
Prac Task 2
Prac Task 3


Diff Task 3
Pass
Fail
Diff Task 2
Pass
Fail
Diff Task 1
Pass
Fail
-.400
-.200
.000
.200
.400
.600
.800

Predicting Pearson Scores - Task 1

Effect	*B*	*t*	*p*
Intercept	2.996	69.821	.000
Track	.002	.031	.975
Male	.051	.434	.665
Non-white	-.010	-.154	.878
Ages 23-28	-.037	-.589	.556
> 28	.020	.237	.813

Predicting UNCC Scores - Task 1

Effect	*B*	*t*	*p*
Intercept	2.875	59.570	.000
Track	.623	7.130	.000
Male	-.033	-.242	.809
Non-white	-.151	-1.948	.052
Ages 23-28	-.102	-1.412	.159
> 28	.154	1.519	.130

Predicting Pearson Scores - Task 2

Effect	*B*	*t*	*p*
Intercept	3.007	80.998	.000
Track	.010	.166	.868
Male	.094	.929	.353
Non-white	.000	.004	.996
Ages 23-28	-.029	-.530	.596
> 28	.014	.185	.853

Predicting UNCC Scores - Task 2

Effect	*B*	*t*	*p*
Intercept	2.649	66.185	.000
Track	.507	7.112	.000
Male	-.064	-.538	.591
Non-white	.046	.730	.466
Ages 23-28	.009	.140	.889
> 28	.040	.475	.635

Predicting Pearson Scores - Task 3

Effect	*B*	*t*	*p*
Intercept	2.936	58.646	.000
Track	-.053	-.628	.530
Male	-.062	-.450	.653
Non-white	-.024	-.319	.750
Ages 23-28	-.041	-.558	.577
> 28	-.037	-.366	.714

Predicting UNCC Scores - Task 3

Effect	*B*	*t*	*p*
Intercept	2.939	87.114	.000
Track	-.418	-6.141	.000
Male	-.020	-.195	.845
Non-white	-.016	-.283	.778
Ages 23-28	.077	1.537	.125
> 28	.040	.544	.587

Feedback from faculty to inform results – next steps

- Survey data

1. I teach a course that includes a formative assignment uploaded to Taskstream to record formative performance of candidates.

Yes No

2. I teach a course that provides formative feedback to candidates in regards to edTPA that is not uploaded to Taskstream.

Yes No

*If you answered No to **both** the questions above, thank you for your time. Your survey is complete. If you answered Yes to **either** question, please proceed to question 3.*

3. How often do you provide feedback to candidates on an assignment that addresses edTPA content?

One time when I score it
One time in draft form and one time when I score it
More than three times

4. What type of feedback do you provide to candidates on a final submission of an assignment that addresses edTPA content? I provide students:

An overall numeric score only (I don't use a rubric)
Only numeric scores following a rubric that is either the edTPA rubric or a modified version of the edTPA rubric
Numeric scores WITHOUT a rubric and few written comments of specific feedback
Numeric scores WITH a rubric (edTPA or modified) and few written comments of specific feedback
Numeric scores WITHOUT a rubric and paragraphs of specific feedback
Numeric scores WITH a rubric (edTPA or modified) and paragraphs of specific feedback

Answer questions 5 and 6 as applicable.

5. On average per student, it takes about how long to score the assignment that addresses edTPA content?

_____

6. On average per student, it takes about how long to score the formative edTPA task (include all the time for drafts)?

_____

8. Do you score the student's written communication skills as part of the grade for the assignment or formative task?

Yes No

7. How would you rate the following in regards to the formative (within coursework) feedback you provide a candidate?

1= not at all; 2= not usually; 3= sometimes; 4= usually; 5= always

I provide timely feedback.	1	2	3	4	5
My feedback is extensive— descriptive and detailed.	1	2	3	4	5
Students use and apply the feedback I provide in future work.	1	2	3	4	5
I meet with each student to review their performance.	1	2	3	4	5
I believe students find my feedback helpful.	1	2	3	4	5
I use edTPA resources to score the formative tasks.	1	2	3	4	5

Please list the resources you use: _____

Considerations in data examination

- Not a “gotcha” for faculty but informative about scoring practices (too strict, too variable, not variable)
- Common guidance for what is “quality” for feedback (e.g., in a formative task that can be time consuming to grade drafts, final products, meet with students about submissions, etc., how much is “enough?”)
- Identify effective supports for faculty (e.g., should we expect reliability without Task-alike conversations or opportunities to score common tasks?)

Faculty PD opportunity

- 1 ½ day common scoring opportunity
- Review criteria, reviewed common work sample
- Debriefed in groups
- Rescored a different sample after training
- Results indicate faculty were much better aligned
- Will analyze 2017-18 results next year
- Scoring common work sample will be built into the faculty PD schedule each year

So to wrap up …

Questions??

UNC CHARLOTTE
Cato College of Education

- Laura Hart
  - Director of Office of Assessment and Accreditation for COED
  - Laura.Hart@uncc.edu
- Teresa Petty
  - Associate Dean
  - tmpetty@uncc.edu