Integration of Data Science and Computing into Introductory Statistics

Department of Mathematics & Statistics
North Carolina A&T State University

2025 Summer Research Symposium on
Infusing and Integrating Data Science and AI in
Research and STEM Education

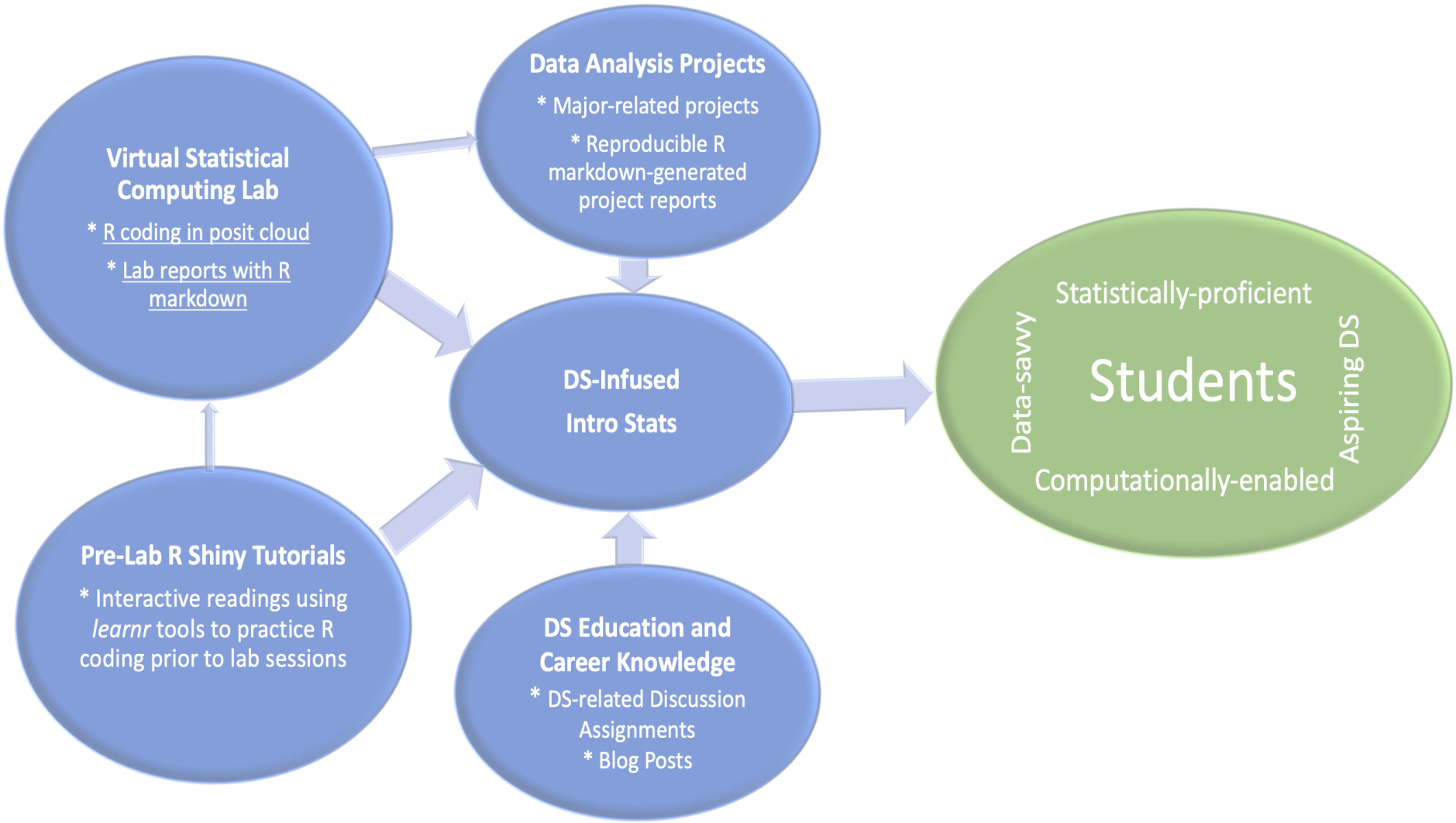
Why introduce DS/computing in Intro Stats?

- Help all students develop "computational thinking" skills.
- Intro Stats can help us attract and prepare a large diverse pool of UGs for DS education/careers:
  - At NCA&T, Intro Stats is an Algebra-based 3.00 credits course
  - **Large**: 7 sections each semester (~45 students in each section)
  - **Diverse**: serves STEM (~46%) and non-STEM (~54%) majors
- In Spring 2019, a survey of NCA&T's Intro Stats students ($n = 181$) found that a vast majority are unaware of DS opportunities:
  - Only **33.15%** of students surveyed had heard about DS,
  - Of those, only **27.12%** knew NCA&T offers DS courses.

Guiding Literature

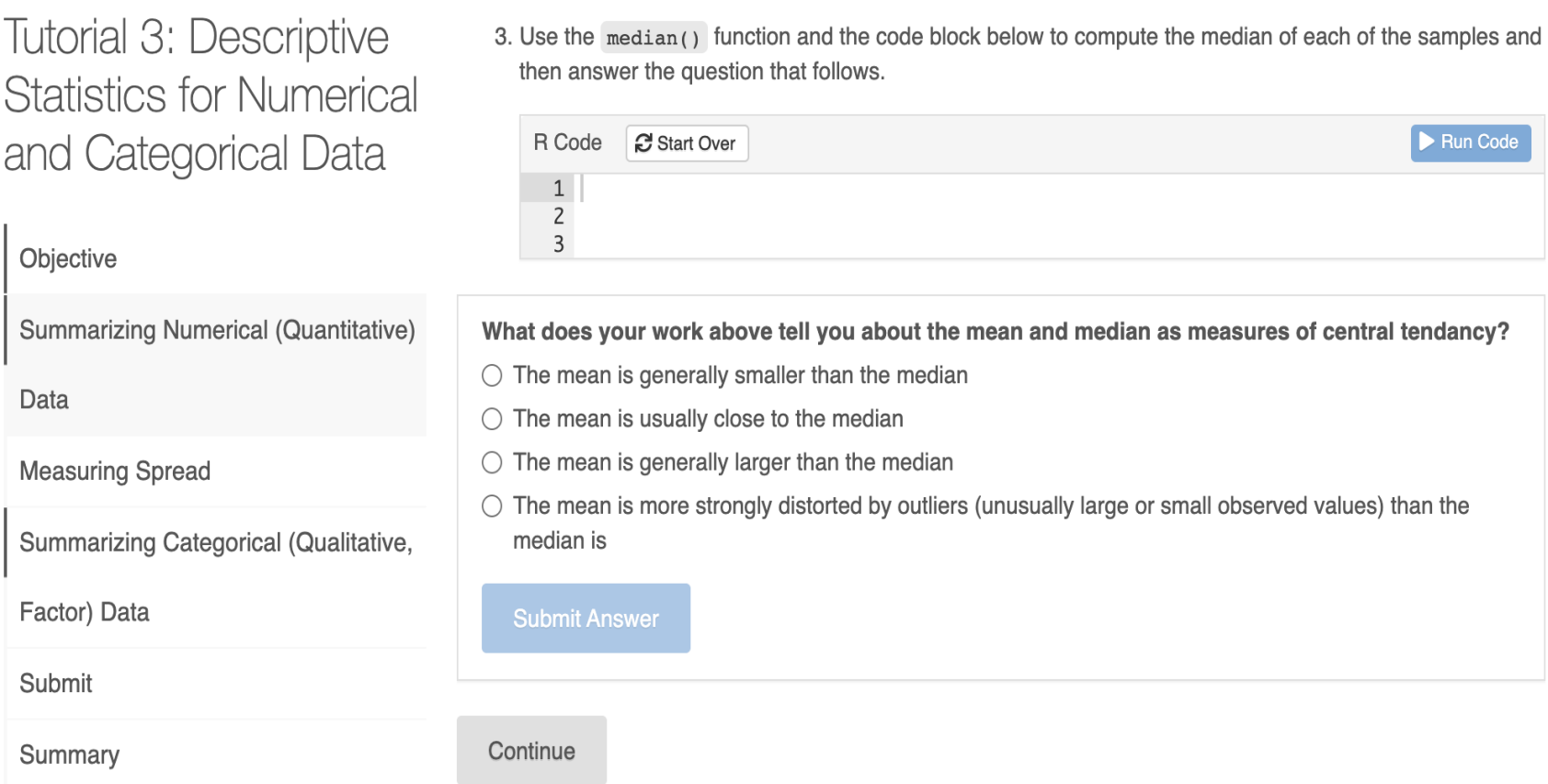
- The Intro Stats course should
  - introduce students to the entire data analysis cycle rather than pieces of it (**Cobb, 2015**),
  - expose students to multivariable thinking (**GAISE #1**),
  - leverage the use of technology for exploring concepts with simulations (**GAISE #2**),
  - help students learn statistics actively while analyzing real data using technology (**GAISE #3, 4 & 5**),
  - train students to think structurally with data and become data-savvy (**Horton et al., 2015**), and
  - expose students, early and frequently, to the elements of the DS workflow and the data scientist's toolbox (**Horton et al., 2015**)
  - See the Special Issue of the *JSDSE* on "**Integrating computing in the statistics and data science curriculum**" (**Horton & Hardin, 2021**).

DS/Computationally-Infused Intro Stats: Phase I-FA22



- **Implementation**: 2 treatment sections and 2 control sections

- Interactive Shiny Pre-Lab Tutorial (using the *learnr* package)



- Computing Lab Description (Static)

Getting started
Analysis
R as a big calculator
Adding a new variable to the data frame
Departure delays
Departure delays by month

You can also obtain numerical summaries for these flights:

```
lax_flights %>%
  summarise(mean_dd   = mean(dep_delay),
            median_dd = median(dep_delay),
            n         = n())
```

Note that in the `summarise` function you created a list of three different numerical summaries that you were interested in. The names of these elements are user defined, like `mean_dd`, `median_dd`, `n`, and you can customize these names as you like (just don't use spaces in your names). Calculating these summary statistics also requires that you know the function calls. Note that `n()` reports the sample size.

Summary statistics: Some useful function calls for summary statistics for a single numerical variable are as follows:

- `mean()` - The arithmetic mean is found by adding the numbers and dividing the sum by the number of numbers in the list
- `median()` - The middle number in a sorted, ascending or descending, list of numbers and can be more descriptive of that data set than the mean.
- `sd()` - The measure of the amount of variation or dispersion of a set of values.
- `var()` - the expectation of the squared deviation of a random variable from its population mean or sample mean.
- `IQR()` - the interquartile range is a measure of statistical dispersion, which is the spread of the data. The IQR may also be called the midspread, middle 50%.
- `min()` - The smallest value in the data set.
- `max()` - The largest value in the data set.

Note that each of these functions takes a single vector as an argument and returns a single value.

You can also filter based on multiple criteria. Suppose you are interested in flights headed to San Francisco (SFO) in February:

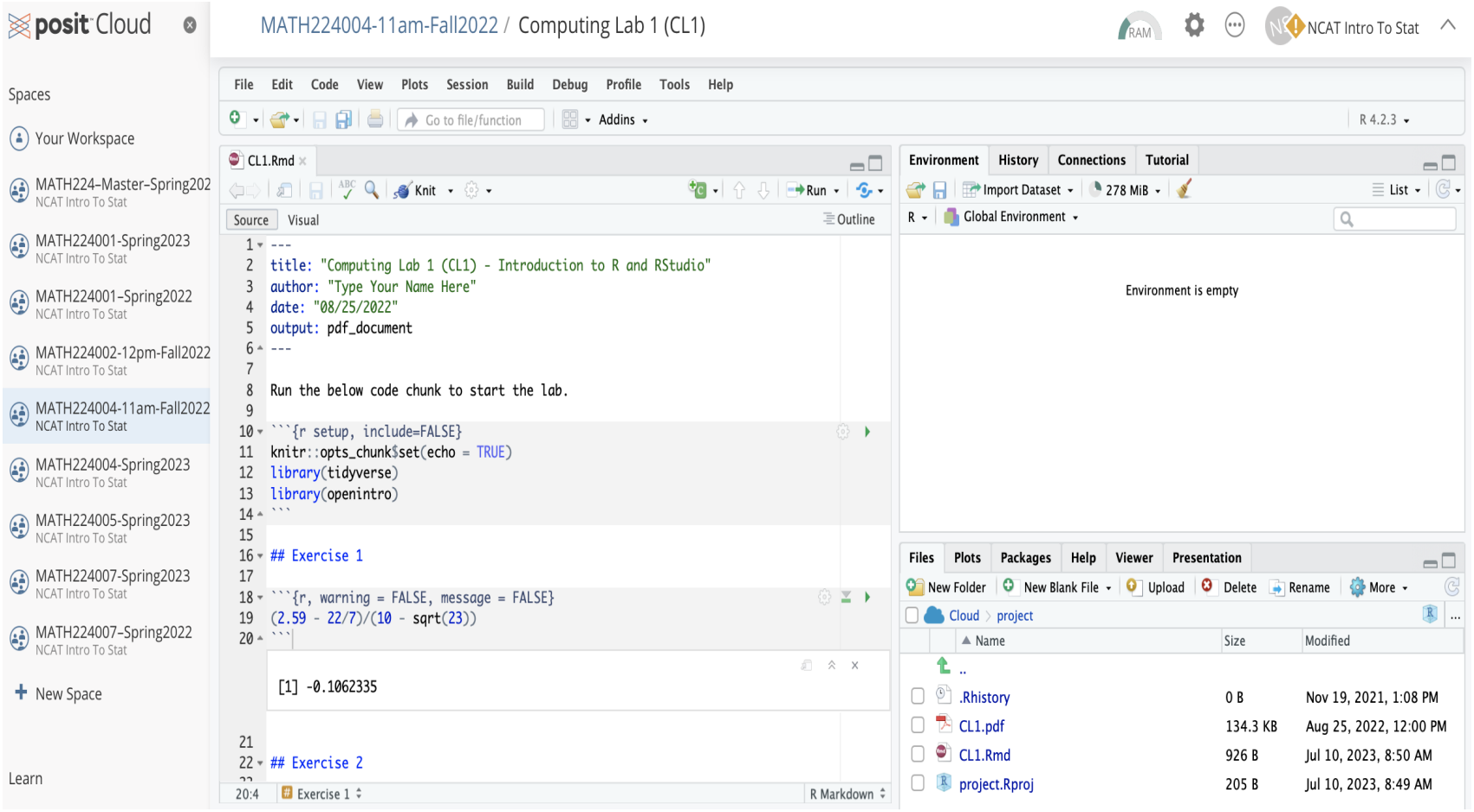
```
sfo_feb_flights <- nycflights %>%
  filter(dest == "SFO", month == 2)
```

Note that you can separate the conditions using commas if you want flights that are both headed to SFO **and** in February. If you are interested in either flights headed to SFO **or** in February, you can use the `|` instead of the comma.

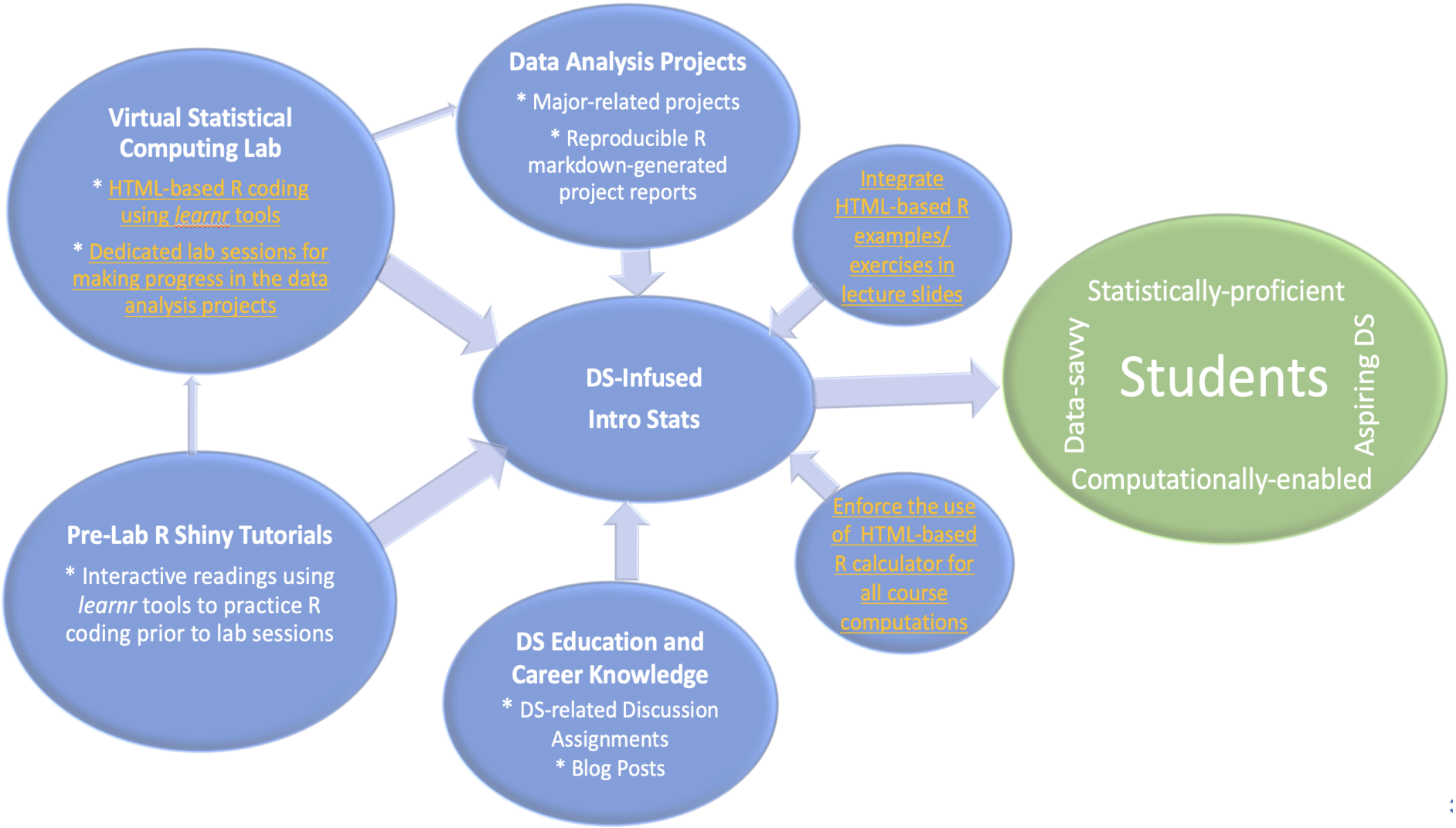
Exercise 2 Create a new data frame that includes flights headed to SFO in February, and save this data frame as `sfo_feb_flights`. How many flights meet these criteria?

Exercise 3 Describe the distribution of the **arrival** delays of these flights using a histogram and appropriate summary statistics. **Hint:** The summary statistics you use should depend on the shape of the distribution.

- Computing Lab R Markdown Template



DS/Computationally-Infused Intro Stats: Phase II-SP23



- **Implementation**: 4 treatment sections and 2 control sections
- **Research Design**: Traditional Pre/Post test

- Interactive Computing Lab (using the *learnr* package)

Exploratory Data Analysis Part I

Start Over

Recall that the five number summary includes the min, first quantile (Q1), median, third quantile (Q3), and max. Using the `mpg` dataset, we can compute the five number summary of the vehicle's highway mileage `hwy` as follows.

R Code | Start Over | Run Code

```
mpg %>%
  summarize(Min = min(hwy),
            Q1 = quantile(hwy, 0.25),
            Median = median(hwy),
            Q3 = quantile(hwy, 0.75),
            Max = max(hwy)
            )
```

Notice how the `quantile()` function is used to obtain quantiles by setting the proportion of data below the quantile (i.e., 0.25 or 0.75)

4. Use the code chunk below to calculate the measures of center (mean and median) for the vehicle's city mileage `cty`.

R Code | Start Over | Run Code | Submit Answer

```

```

5. Use the code chunk below to calculate the variation measures (standard deviation and interquartile range) for the vehicle's city mileage `cty`.

R Code | Start Over | Run Code | Submit Answer

```

```

- Slides with Interactive Coding

Examples

Example 1. Calculate the mean of a sample with five observations: 5, 3, 8, 5, 6.

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{5+3+8+5+6}{5} = \frac{27}{5} = 5.4$$

Using R, we can calculate the mean using the `mean()` command. Notice that we need to put the values in a vector using the `c()` function which stands for *concatenate*.

R Code | Start Over | Run Code

```
1 mean(c(5,3,8,5,6))
2
3
```

12/87

Discussions

1. If the data set has 5 observations, with $\bar{x} = 5.4$, find $\sum_{i=1}^{5} x_i$.
2. Continue discussion in 1, if add one more observation 10, will the mean $\bar{x}$ increase or decrease? What is the new $\bar{x}$?
3. Compare data sets 5, 3, 8, 5, 6 and 5, 3, 80, 5, 6, which one has the higher mean?

R Code | Start Over | Run Code

```
1
2
3
```

- Interactive R Calculator

Using R as a calculator

R can be used as an calculator as we already saw in the tutorial. So let's get a refresher on this.

Let's say we want to calculate $\frac{36}{29(15-9)}$. Then we would do the following:

R Code | Start Over | Run Code

```
36 / (29 * (15 - 9))


```

R also has built-in constants such as pi and mathematical functions such as e and log.

Let's find the radius of a circle with radius 4. Then using R we can get the area and the circumference.

R Code | Start Over | Run Code

```
radius = 4

area = pi * radius^2

circumference = 2 * pi * radius

c("Area" = area, "Circumference" = circumference)
```

We can also use `R` to calculate probabilities under the normal distribution. The following code returns the probability that a normal variable with mean 25 and standard deviation 15 is less than 50.

R Code | Start Over | Run Code

```
pnorm(q = 50, mean = 25, sd = 15)


```

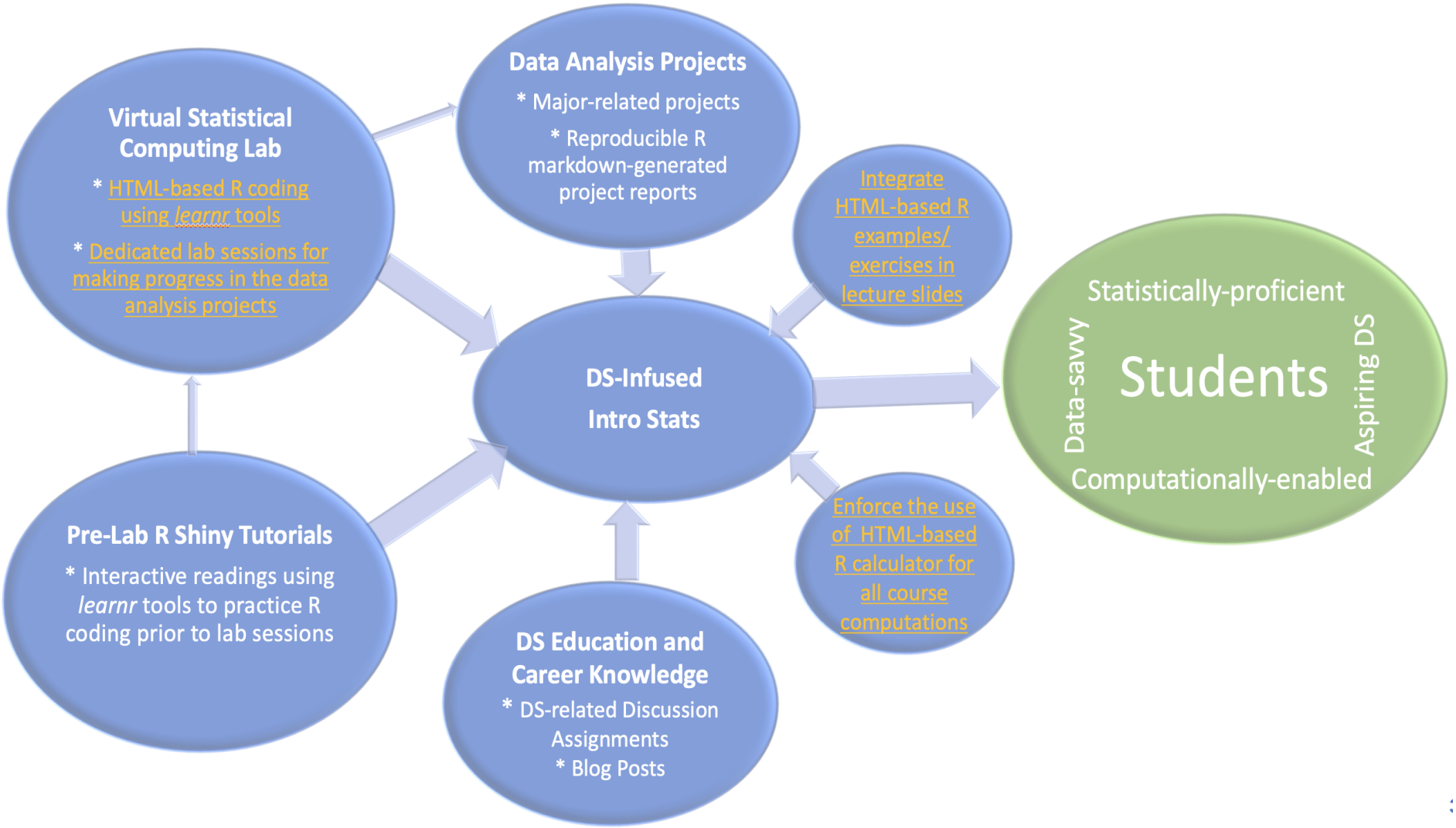
As you work on your homework assignments, feel free to use the below code chunks to perfom your calculations.

R Code | Start Over | Run Code

```


```

DS/Computationally-Infused Intro Stats:FA23/SP24



- **Implementation**: 10 treatment sections and 2 control sections
- **Research Design**: Retrospective Pre/Post test

Evaluating the DS-Infused Intro Stats Design

- **DS awareness, readiness & aspirations**
  - Students completed a DS awareness, readiness, and aspirations survey in Qualtrics
  - Pre-survey during 1st week of semester; Retrospective pre/post-survey at the end of semester
  - The survey was created in-house and validated through a series of exploratory factor analyses using pilot data from SP22
- **Statistical learning gains**
  - Students completed a revised version of the CAOS (Comprehensive Assessment of Outcomes in Statistics) scale (e.g., Tintle et al., 2018)
  - Pre/post-test approach

Survey Items

- DS Awareness questions [Yes/No]

Year	Awareness Questions
Pre/Post Survey	Have you ever taken a course in Data Science?
	Are you aware that NCA&T offers courses in Data Science?
	Are you aware that NCA&T offers an undergraduate certificate in Data Science?
	Are you aware that NCA&T offers degrees with concentration in Data Science?
	About how much is the average annual salary of data scientists?

Survey Items

- DS Readiness (confidence) questions [Strongly Agree to Strongly Disagree]

Year	Readiness Questions
Pre/Post Survey	I can summarize data sets with summary statistics and graphics using the RStudio software.
	I can perform basic statistical inference using the RStudio software.
	I can perform basic statistical modeling (linear and/or logistics regression).
	I can create reproducible data analysis reports using the R Markdown software.
	I am adequately prepared to apply statistical and data-analytical techniques and/or tools to study a given topic.
Retro-Pre Survey	**Prior to starting this class**, I could summarize data sets with summary statistics and graphics using the RStudio software.
	Prior to starting this class, I could perform basic statistical inference using the RStudio software.
	Prior to starting this class, I could perform basic statistical modeling (linear and/or logistics regression).
	Prior to starting this class, I could create reproducible data analysis reports using the R Markdown software.
	Prior to starting this class, I was adequately prepared to apply statistical and data-analytical techniques and/or tools to study a given topic.

Survey Items

- DS Aspiration questions [Yes/Unsure/No]

Year	Readiness Questions
Pre/Post Survey	Do you plan to **take Data Science course(s)** during your undergraduate program or during your graduate study (if you plan to do graduate studies)?
	Do you plan to **complete a certificate in Data Science** during your undergraduate program or during your graduate study (if you plan to do graduate studies)?
	Do you plan to **complete a minor in Data Science** during your undergraduate program or during your graduate study (if you plan to do graduate studies)?
	Do you plan to **complete a degree in Data Science** during your undergraduate program or during your graduate study (if you plan to do graduate studies)?
Retro-Pre Survey	**Prior to starting this class**, did you plan to **take Data Science course(s)** during your undergraduate program or during your graduate study (if you plan to do graduate studies)?
	Prior to starting this class, did you plan to **complete a certificate in Data Science** during your undergraduate program or during your graduate study (if you plan to do graduate studies)?
	Prior to starting this class, did you plan to **complete a minor in Data Science** during your undergraduate program or during your graduate study (if you plan to do graduate studies)?
	Prior to starting this class, did you plan to **complete a degree in Data Science** during your undergraduate program or during your graduate study (if you plan to do graduate studies)?

Year 1 (FA22-SP23) Key Results

- Sample distribution by demographic and academic profile factors

		N (%) / Mean (SD)		
Variable	**Design**	**Traditional**	**Design I**	**Design II**
Total N (%)		113 (38.2)	71 (24.0)	112 (37.8)
Pretest	Yes	87 (77.0)	70 (98.6)	109 (97.3)
Posttest	Yes	94 (83.2)	58 (81.7)	102 (91.1)
Presurvey	Yes	107 (94.7)	67 (94.4)	106 (94.6)
Postsurvey	Yes	86 (76.1)	52 (73.2)	78 (69.6)
Gender	Male	38 (33.6)	26 (37.1)	45 (40.2)
PELL	Yes	79 (69.9)	61 (87.1)	92 (82.1)
Rural	Yes	15 (13.3)	16 (22.9)	19 (17)
Residency	Out-of-State	42 (37.2)	22 (31.4)	37 (33)
STEM	Yes	76 (67.3)	49 (70.0)	68 (60.7)
Pre-course GPA	≥3.0	53 (52.5)	35 (55.6)	70 (62.5)
Grade	A	26 (23.0)	15 (21.1)	21 (18.8)
	B	27 (23.9)	12 (16.9)	39 (34.8)
	C	30 (26.5)	23 (32.4)	38 (33.9)
AP Stats	Yes	33 (30.8)	20 (29.9)	27 (25.5)
Attendance (%)		82.1 (15.4)	78.1 (20.9)	82.9 (15.9)

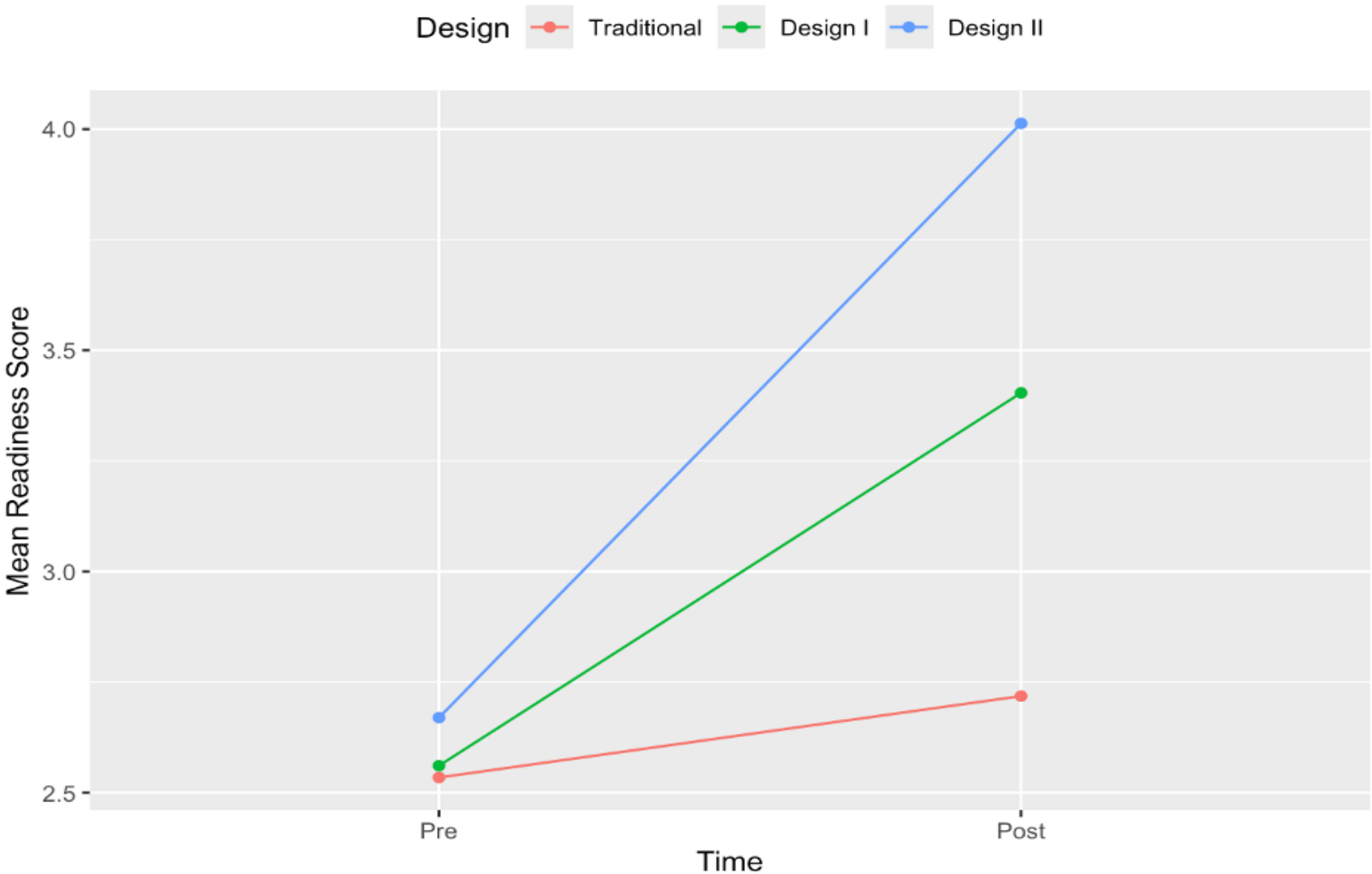
Year 1 (FA22-SP23) Key Results

- DS awareness by course design

		N (%) of Yes		
Question	**Design**	**Traditional**	**Design I**	**Design II**
Ever heard about DS?	Presurvey	56 (52.34)	34 (50.75)	56 (52.83)
	Postsurvey	58 (**70.73**)	43 (**87.76**)	59 (**76.62**)
	Ever heard about DS = Yes			
Know NCA&T has DS Courses?	Presurvey	26 (46.43)	18 (52.94)	33 (58.93)
	Postsurvey	35 (**60.34**)	33 (**76.74**)	53 (**89.83**)
Know NCA&T has UG DS Certificate?	Presurvey	11 (19.64)	11 (32.35)	24 (42.86)
	Postsurvey	29 (**50.00**)	30 (**69.77**)	46 (**77.97**)
Know NCA&T has DS Concentration?	Presurvey	15 (26.79)	9 (26.47)	29 (51.79)
	Postsurvey	28 (**48.28**)	31 (**72.09**)	48 (**81.36**)
Know salary range for Data Scientists?	Presurvey	30 (53.57)	14 (46.67)	36 (66.67)
	Postsurvey	30 (**52.63**)	26 (**68.42**)	39 (**70.91**)

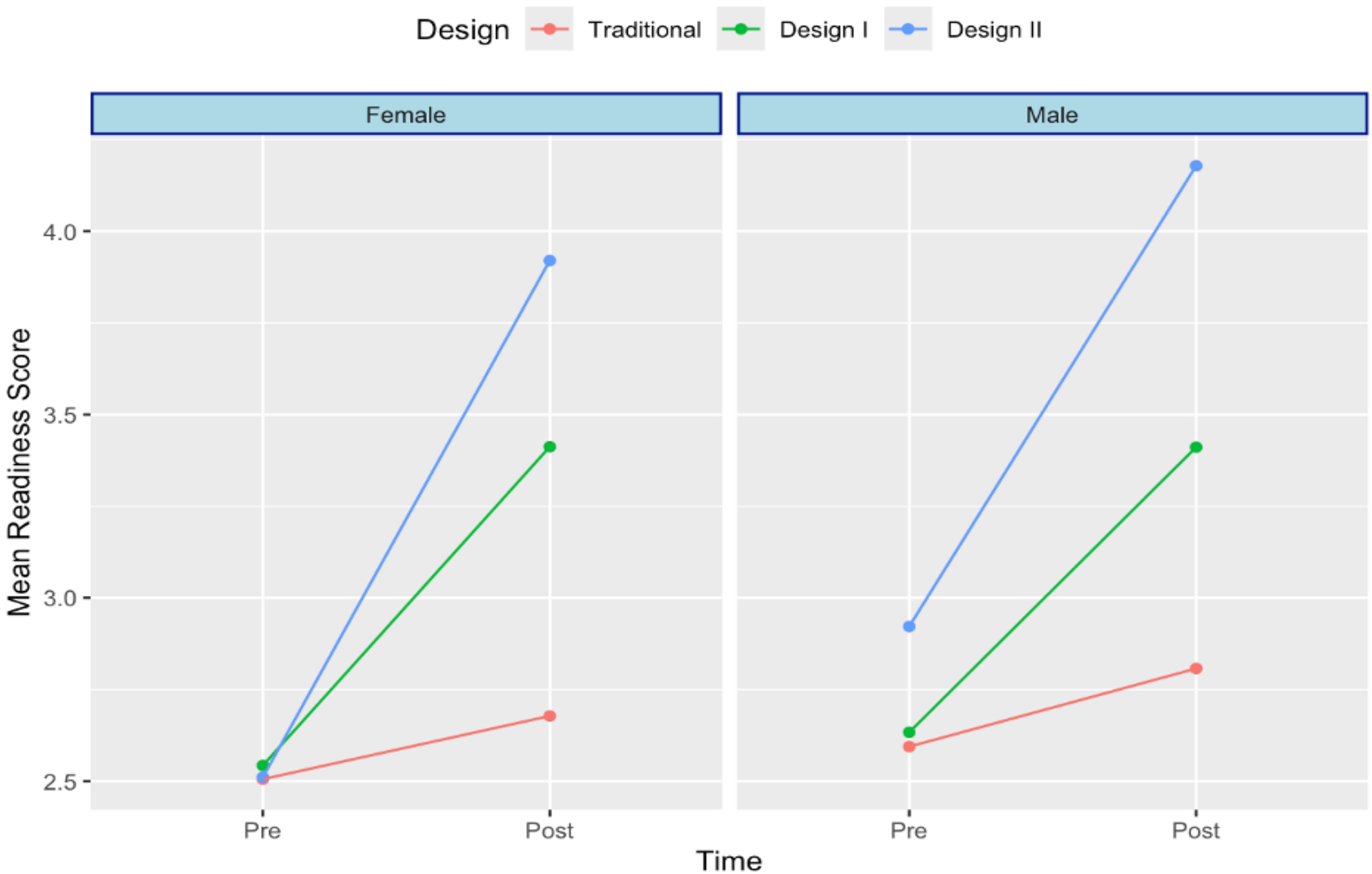
Year 1 (FA22-SP23) Key Results

- DS Readiness Score



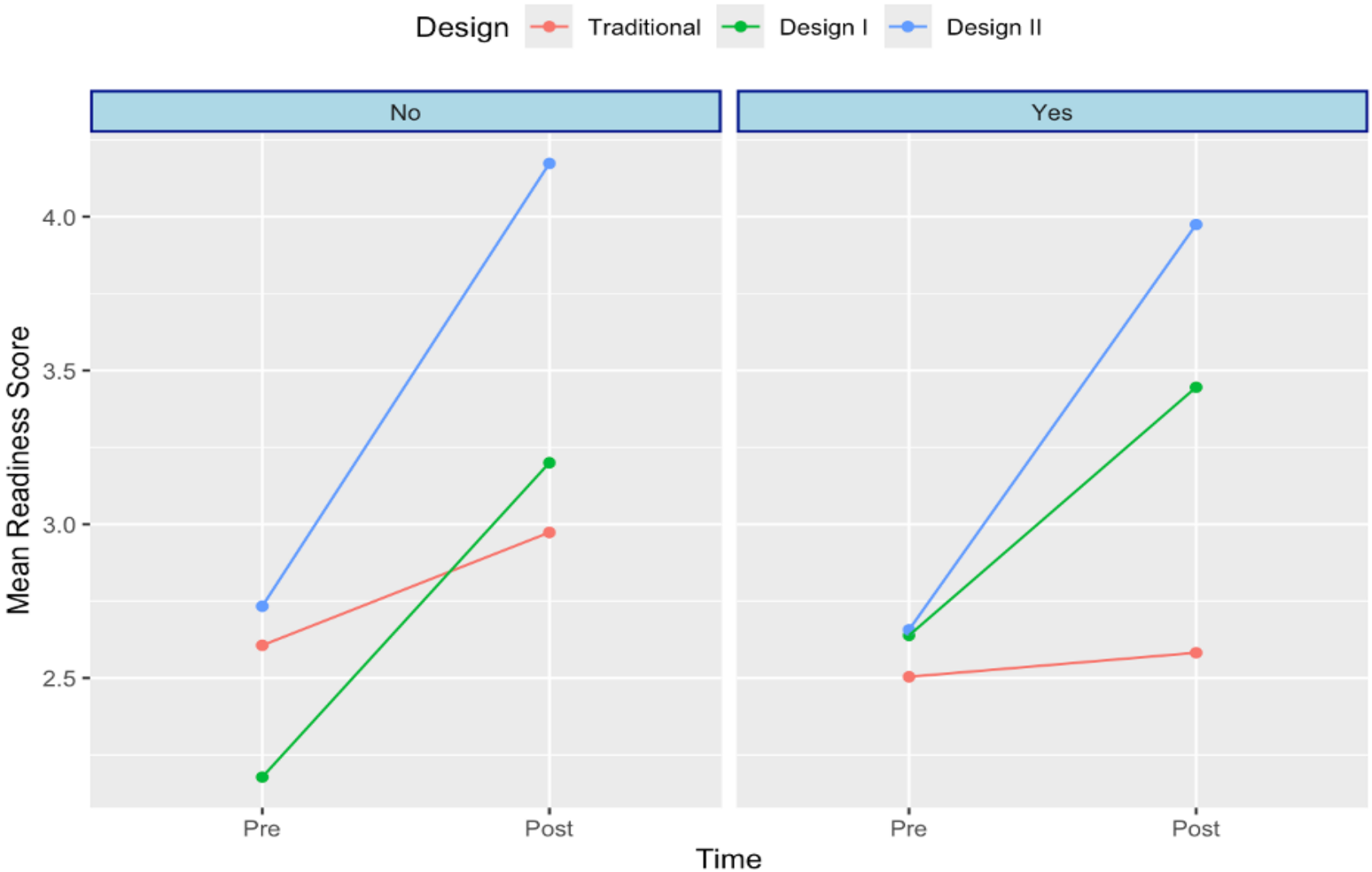
Year 1 (FA22-SP23) Key Results

- DS Readiness Score by Gender



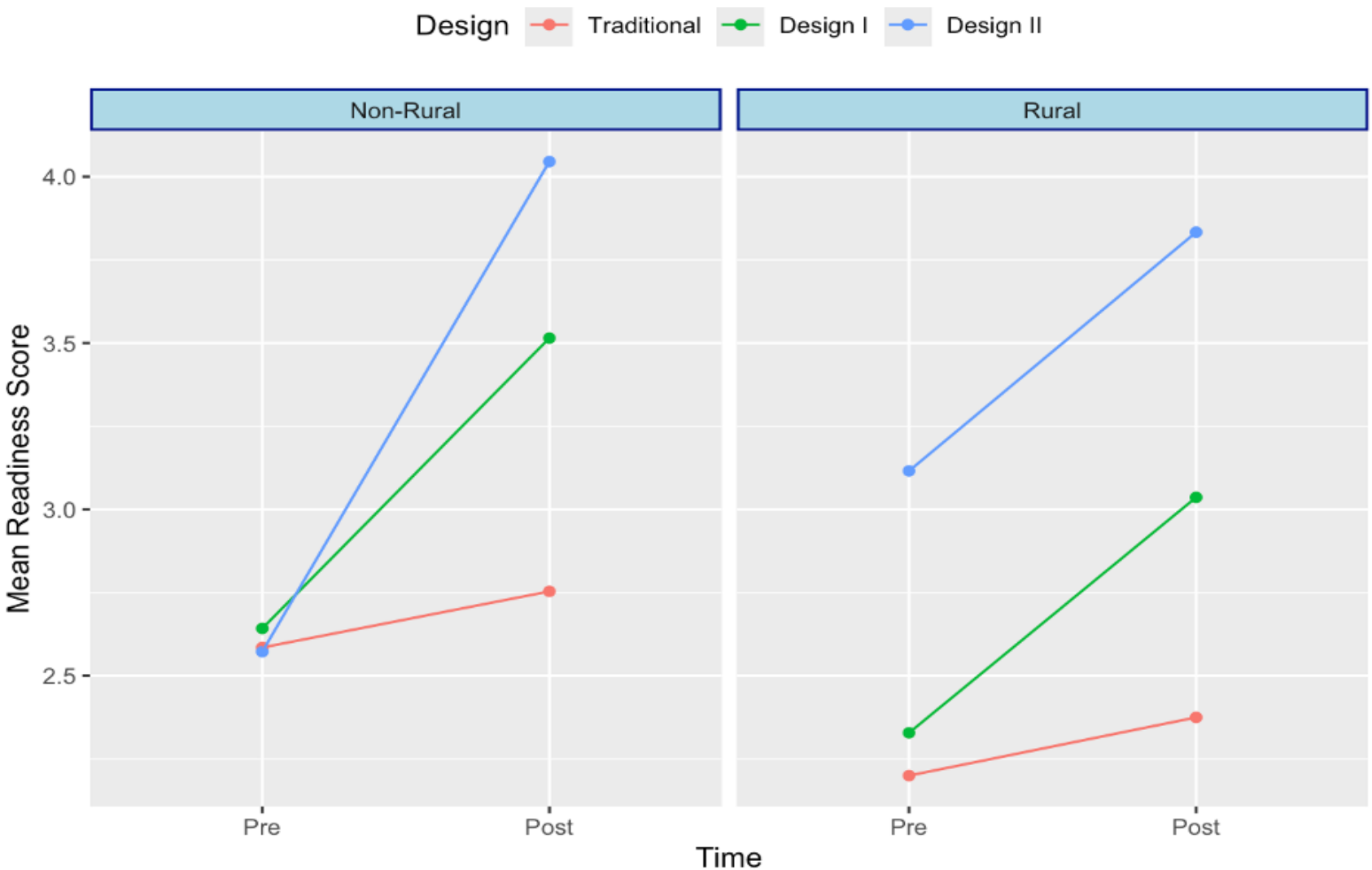
Year 1 (FA22-SP23) Key Results

- DS Readiness Score by PELL



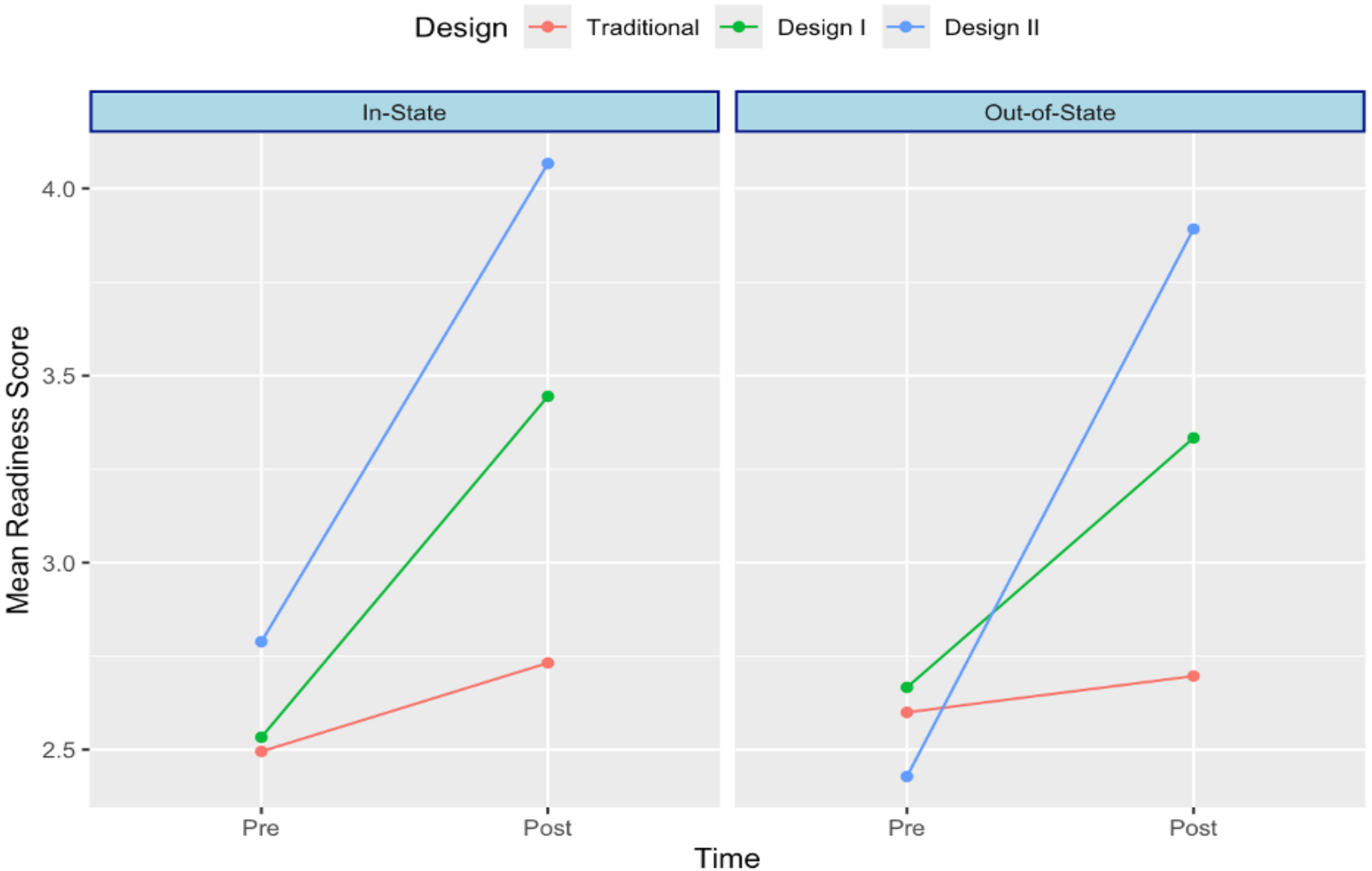
Year 1 (FA22-SP23) Key Results

- DS Readiness Score by Rural



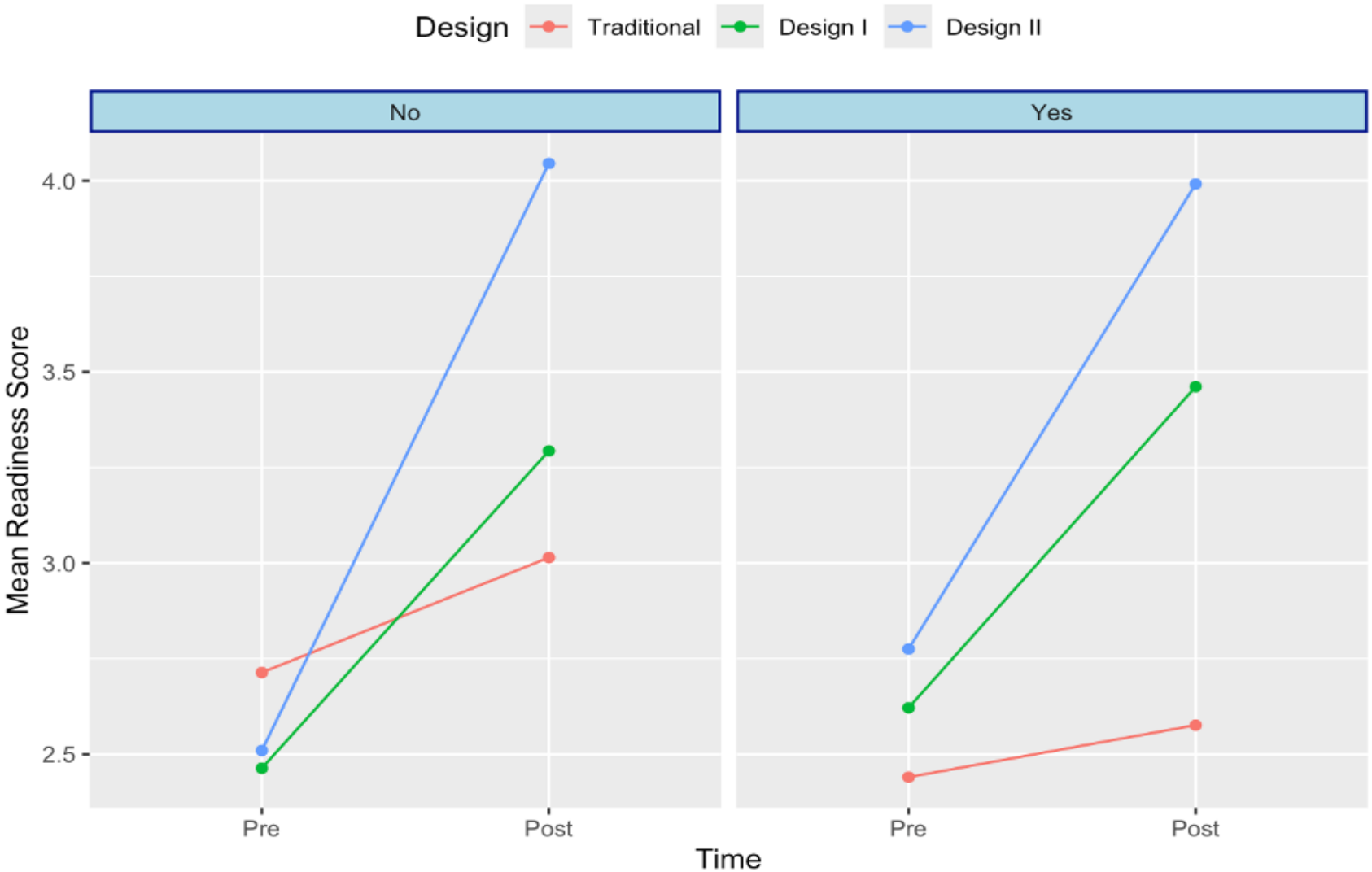
Year 1 (FA22-SP23) Key Results

- DS Readiness Score by Residency



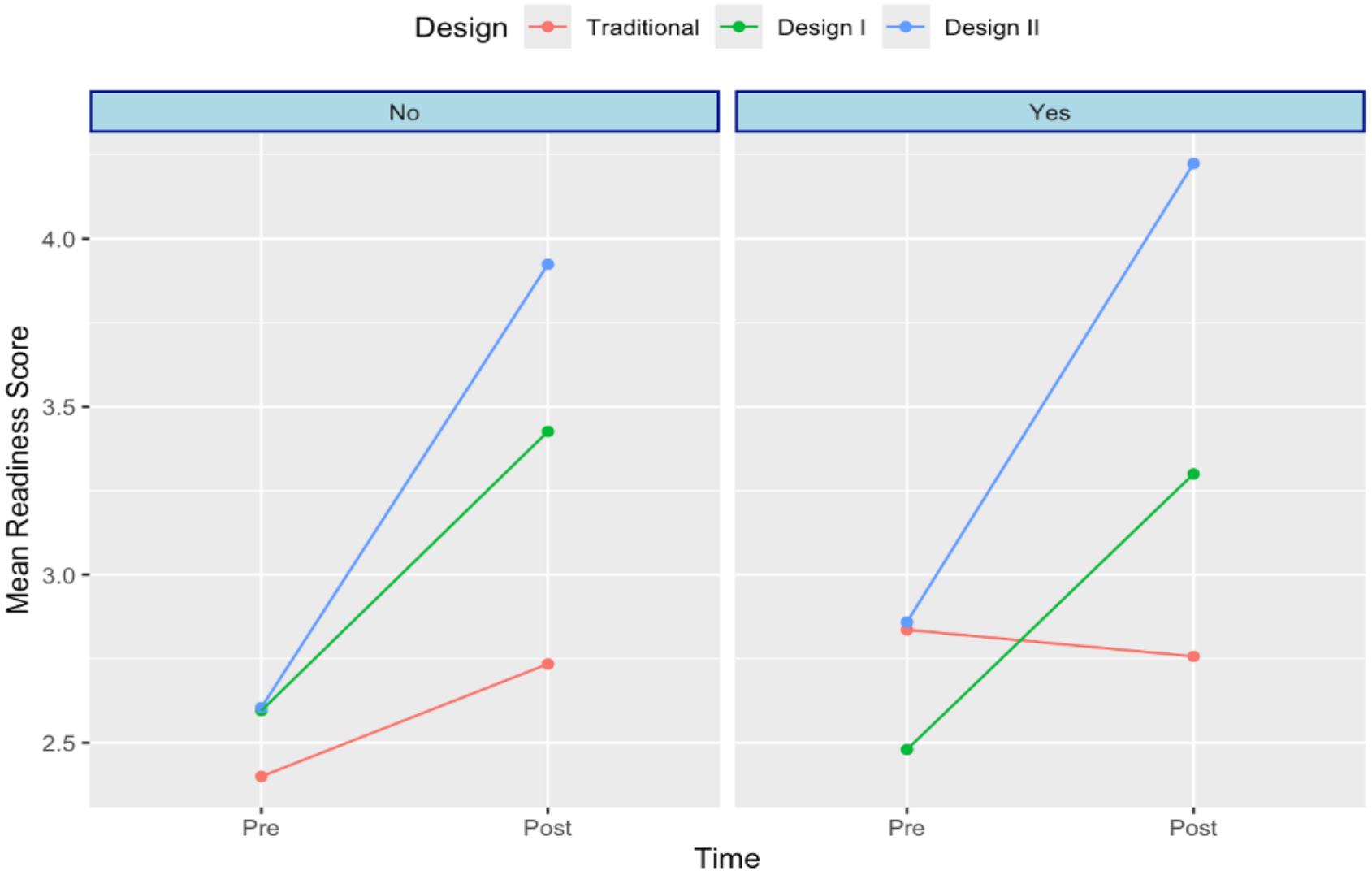
Year 1 (FA22-SP23) Key Results

- DS Readiness Score by STEM



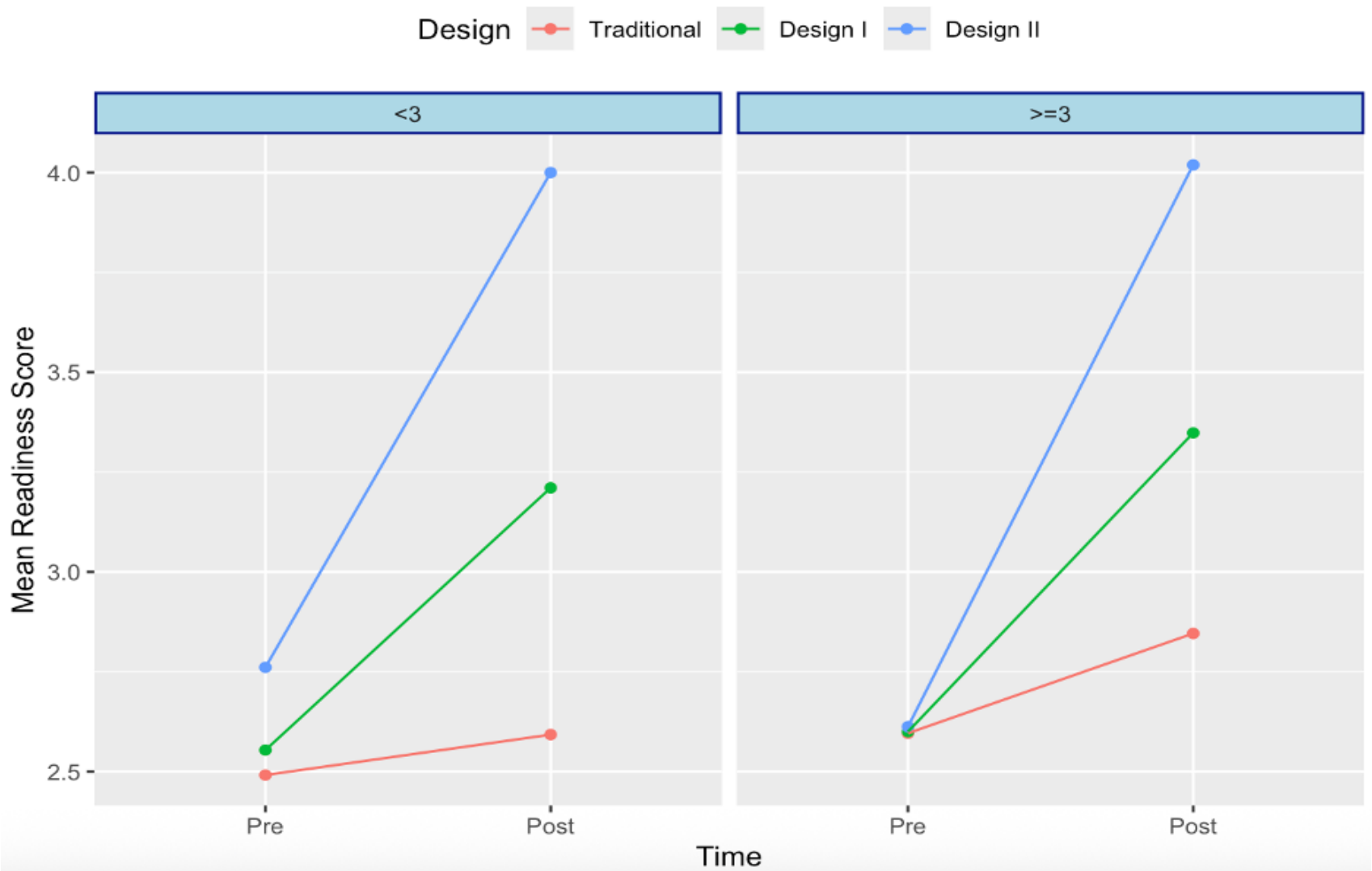
Year 1 (FA22-SP23) Key Results

- DS Readiness Score by AP Statistics



Year 1 (FA22-SP23) Key Results

- DS Readiness Score by GPA



Year 1 (FA22-SP23) Key Results

- Results of (partially overlapping) significance tests for DS readiness (confidence) by course design

Design	Difference Estimate	Test Statistics	P-value
Traditional	0.18	1.35	0.0901
Design I	**0.84**	4.68	**<0.0001**
Design II	**1.34**	9.54	**<0.0001**

Year 1 (FA22-SP23) Key Results

- Results of multivariable linear regression for the DS readiness post-test score on course design and students' characteristics

Term		Estimate	SE	95% LCL	95% UCL	P-value
Intercept		1.84	0.458	0.92	2.76	0.0002
Pretest Score		**0.20**	0.075	0.05	0.35	**0.0108**
Design	Design I	**0.72**	0.193	0.33	1.12	**0.0005**
	Design II	**1.28**	0.148	0.99	1.57	**<0.0001**
SEX	Male	0.12	0.149	-0.17	0.42	0.4161
PELL	Yes	-0.04	0.185	-0.41	0.33	0.8203
RURAL	Yes	-0.36	0.206	-0.77	0.05	0.0848
RESIDENCY	Out-of-State	-0.03	0.149	-0.33	0.26	0.8277
STEM	Yes	-0.12	0.149	-0.42	0.18	0.4346
AP STAT	Yes	-0.10	0.152	-0.40	0.21	0.5304
Pre-Course GPA	≥3.0	-<0.01	0.159	-0.32	0.31	0.2958
Course Grade	A	0.26	0.248	-0.23	0.75	0.2958
	B	0.04	0.226	-0.41	0.49	0.8589
	C	0.15	0.211	-0.27	0.58	0.4741
Attendance (%)		<0.01	0.005	-0.006	0.015	0.3677
R^2 = 31.95%		**Adjusted R^2 = 28.56%**			**P-value = <0.0001**	

Year 1 (FA22-SP23) Key Results

- DS **Aspirations** among students who heard about DS by course design

		N (%) of Yes		
Question	**Design**	**Traditional**	**Design I**	**Design II**
Plan to take DS courses?	Presurvey	9 (16.07)	8 (23.53)	9 (16.07)
	Postsurvey	8 (**13.79**)	4 (**9.30**)	5 (**8.62**)
Plan to complete a UG DS Certificate?	Presurvey	2 (3.57)	3 (8.82)	2 (3.57)
	Postsurvey	3 (**5.17**)	3 (**6.98**)	1 (**1.72**)
Plan to complete a UG DS minor?	Presurvey	1 (1.79)	0 (0.00)	0 (0.00)
	Postsurvey	1 (**1.72**)	3 (**6.98**)	0 (**0.00**)
Plan to complete a UG DS Degree?	Presurvey	1 (1.79)	0 (0.00)	1 (1.79)
	Postsurvey	2 (**3.45**)	2 (**4.65**)	0 (**0.00**)

Year 1 (FA22-SP23) Key Results

- Results of (partially overlapping) significance tests for performance on the CAOS test by course design

Design	Difference Estimate	Test Statistic	P-value
Traditional	6.07	2.77	0.0037
Design I	6.73	3.66	0.0003
Design II	**9.31**	4.65	**<0.0001**

Year 1 (FA22-SP23) Key Results

- Results of multivariable linear regression for the CAOS post-test score on course design and students' characteristics

Term		Estimate	SE	95% LCL	95% UCL	P-value
Intercept		32.14	8.763	14.85	49.43	0.0003
Pretest Score		0.18	0.089	0.007	0.36	**0.0421**
Design	Design I	-0.31	2.836	-5.90	5.29	0.9142
	Design II	-0.69	2.457	-5.53	4.16	0.7804
SEX	Male	2.17	2.194	-4.77	3.78	0.3242
PELL	Yes	1.15	2.827	-4.43	6.73	0.6844
RURAL	Yes	-5.91	3.049	-11.93	0.11	0.0542
RESIDENCY	Out-of-State	-6.73	2.443	-11.55	-1.91	**0.0065**
STEM	Yes	-0.49	2.167	-4.77	3.78	0.8204
AP STAT	Yes	2.26	2.366	-2.41	6.93	0.3402
Pre-Course GPA	≥3.0	4.95	2.433	0.15	9.75	**0.0433**
Course Grade	A	7.63	3.973	-0.22	15.47	0.0563
	B	8.93	3.330	2.36	15.50	**0.0080**
	C	4.29	3.170	-1.96	10.55	0.1773
Attendance (%)		<0.01	0.086	-0.17	0.17	0.9996
R^2 = 22.39%		**Adjusted R^2 = 16.35%**			**P-value = 0.0032**	

Year 2 (FA23-SP24) Key Results

- Sample distribution by demographic and academic profile factors

		N (%) / Mean (SD)	
Variable	**Design**	**Traditional**	**Treatment**
Total N (%)		74 (17.3)	354 (82.7)
Pretest	Yes	71 (95.9)	349 (98.6)
Posttest	Yes	71 (95.9)	328 (92.7)
Presurvey	Yes	66 (89.2)	300 (84.7)
RetroPre/Postsurvey	Yes	48 (64.9)	287 (81.1)
Gender	Male	19 (25.7)	100 (28.5)
PELL	Yes	36 (48.6)	177 (50.4)
Rural	Yes	15 (20.3)	52 (14.7)
Residency	Out-of-State	30 (40.5)	157 (44.7)
STEM	Yes	47 (63.5)	211 (60.1)
AP Stats	Yes	18 (27.3)	81 (26.8)
Pre-course GPA	≥3.0	57 (79.2)	224 (65.9)
Grade	A	21 (28.4)	105 (29.7)
	B	29 (39.2)	130 (36.7)
	C	17 (23)	77 (21.8)
Attendance (%)		80.1 (19.6)	84.2 (18.4)

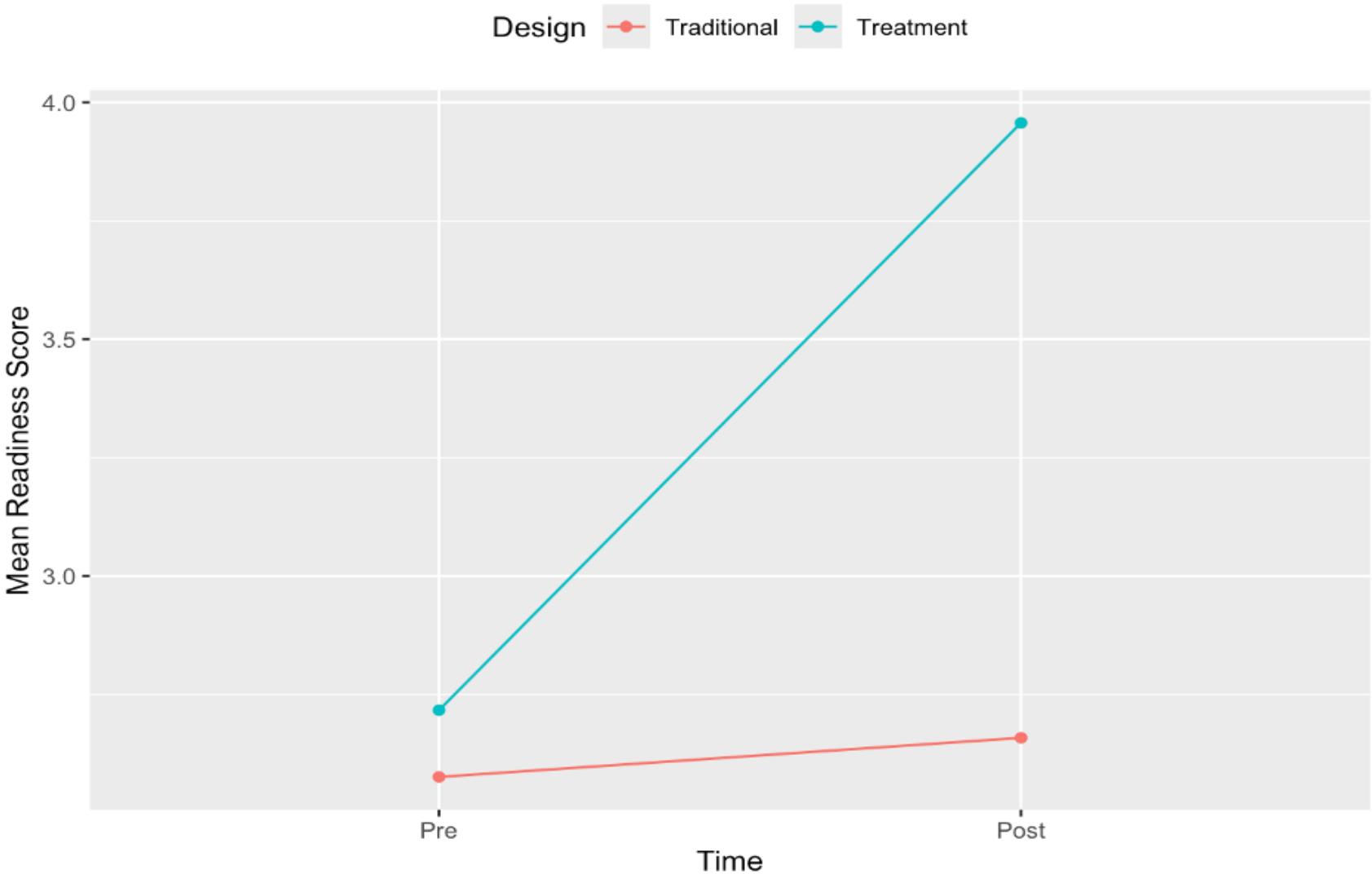
Year 2 (FA23-SP24) Key Results

- DS **Awareness** by course design

		N (%) of Yes	
Question	**Design**	**Traditional**	**Treatment**
Ever heard about DS?	Presurvey	33 (50.00)	178 (59.14)
	Postsurvey	30 (**66.67**)	224 (**85.17**)
	Ever heard about DS = Yes		
Know NCA&T has DS Courses?	Presurvey	19 (57.58)	97 (54.80)
	Postsurvey	19 (**63.33**)	178 (**79.46**)
Know NCA&T has UG DS Certificate?	Presurvey	8 (24.24)	71 (40.11)
	Postsurvey	10 (**33.33**)	161 (**71.88**)
Know NCA&T has DS Concentration?	Presurvey	13 (39.39)	71 (40.11)
	Postsurvey	15 (**50.00**)	152 (**67.86**)
Know salary range for Data Scientists?	Presurvey	20 (60.61)	118 (71.08)
	Postsurvey	14 (**56.00**)	143 (**68.75**)

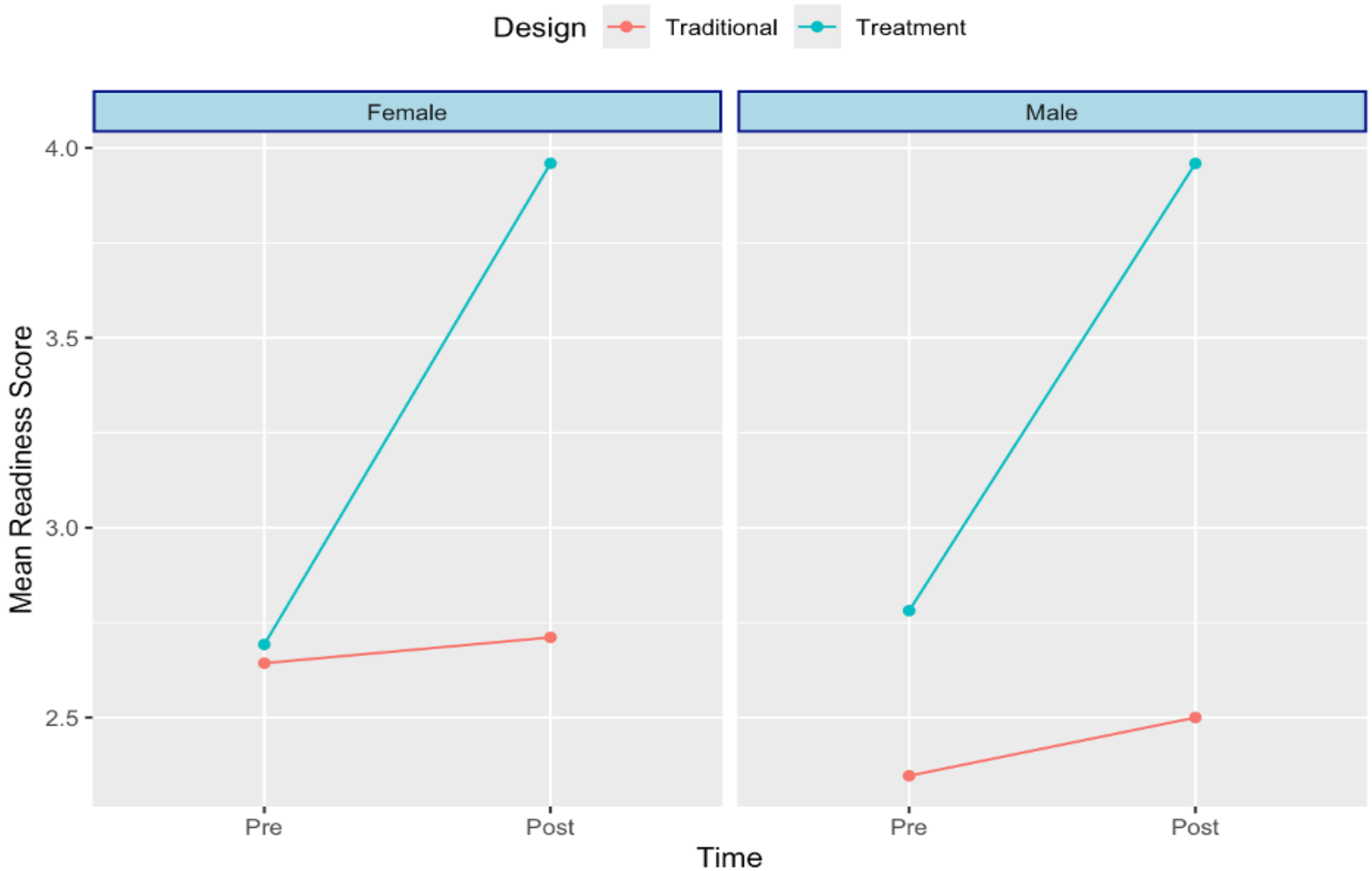
Year 2 (FA23-SP24) Key Results

- DS Readiness Score



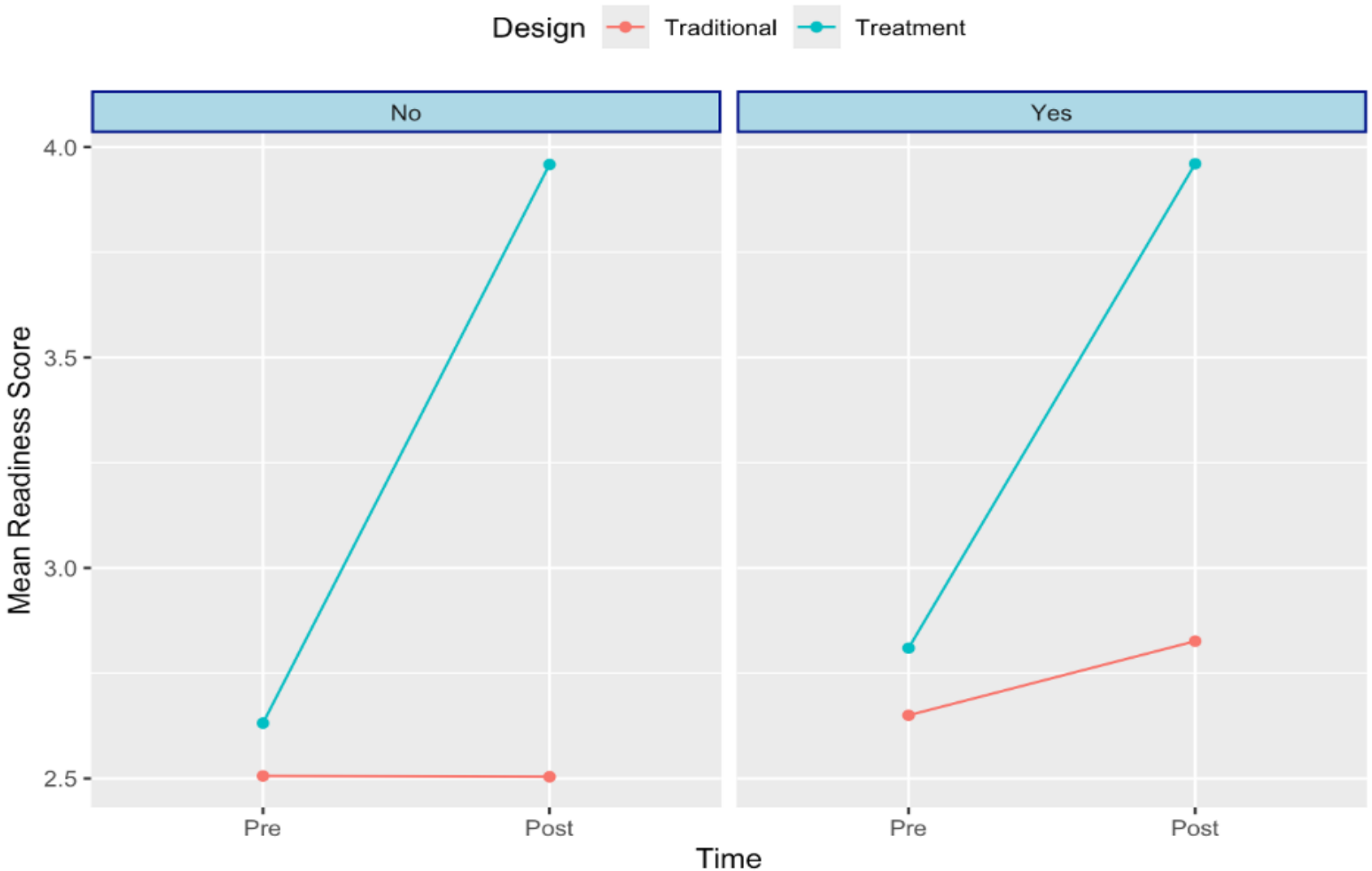
Year 2 (FA23-SP24) Key Results

- DS Readiness Score by Gender



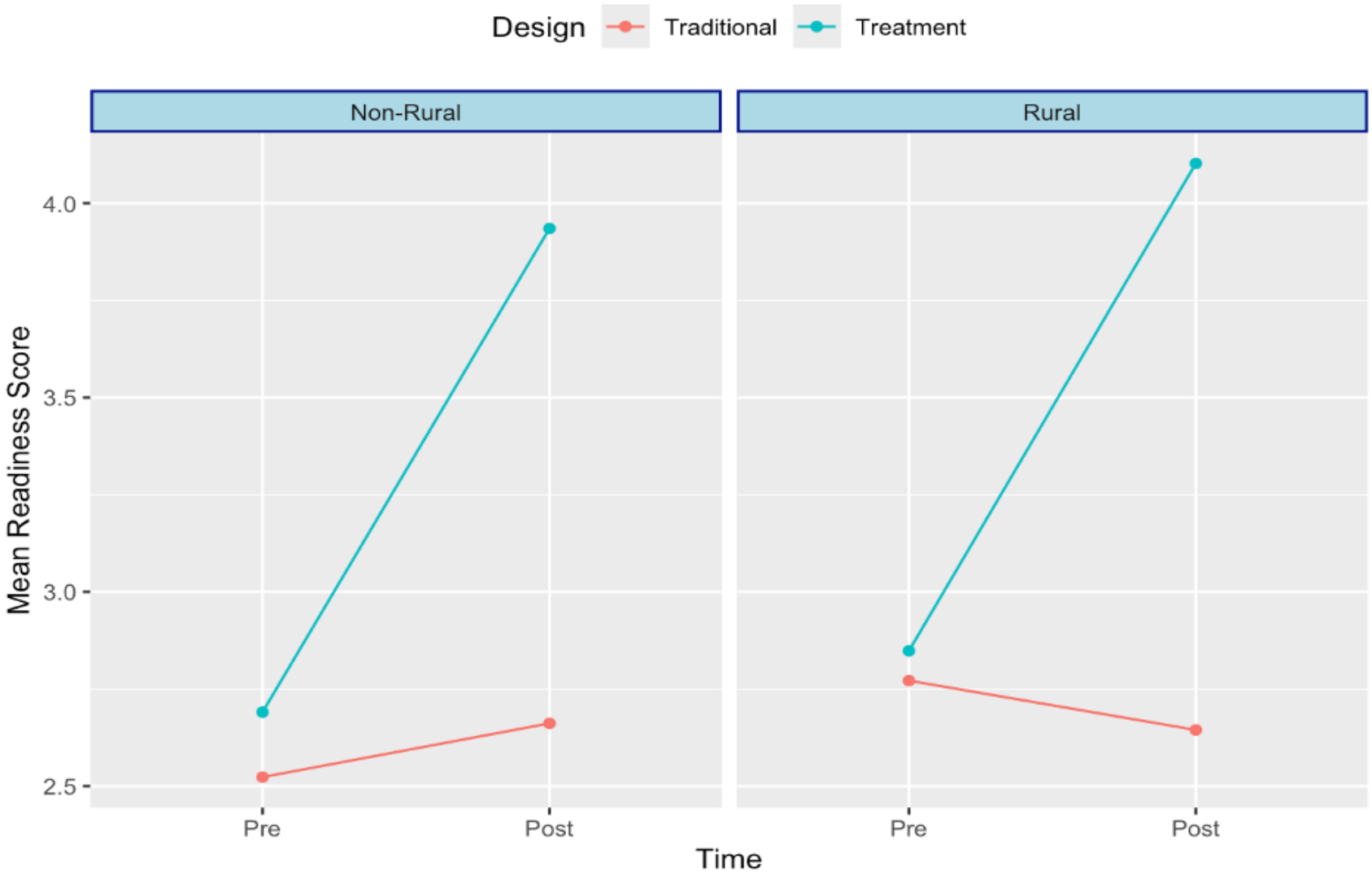
Year 2 (FA23-SP24) Key Results

- DS Readiness Score by PELL



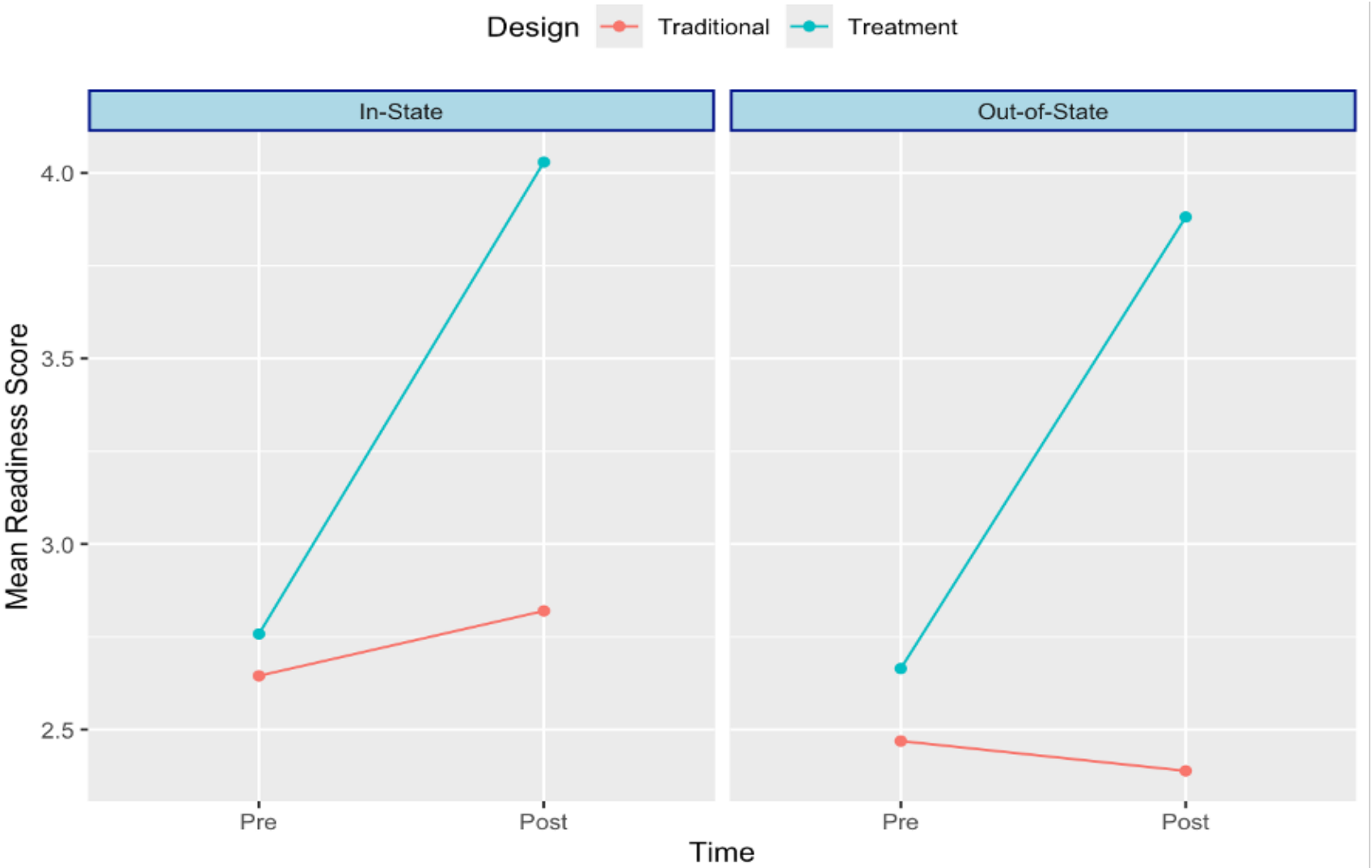
Year 2 (FA23-SP24) Key Results

- DS Readiness Score by Rural



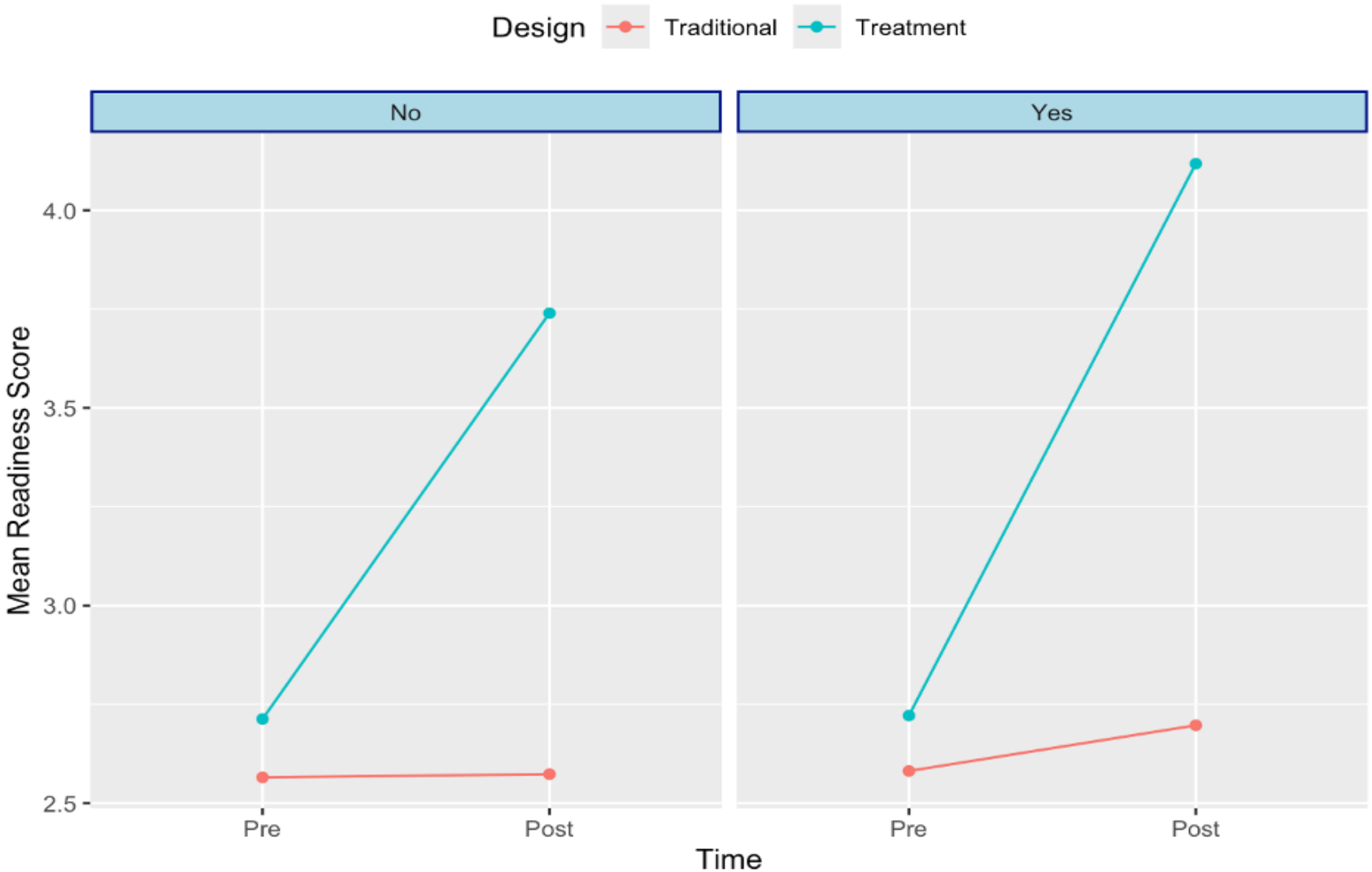
Year 2 (FA23-SP24) Key Results

- DS Readiness Score by Residency



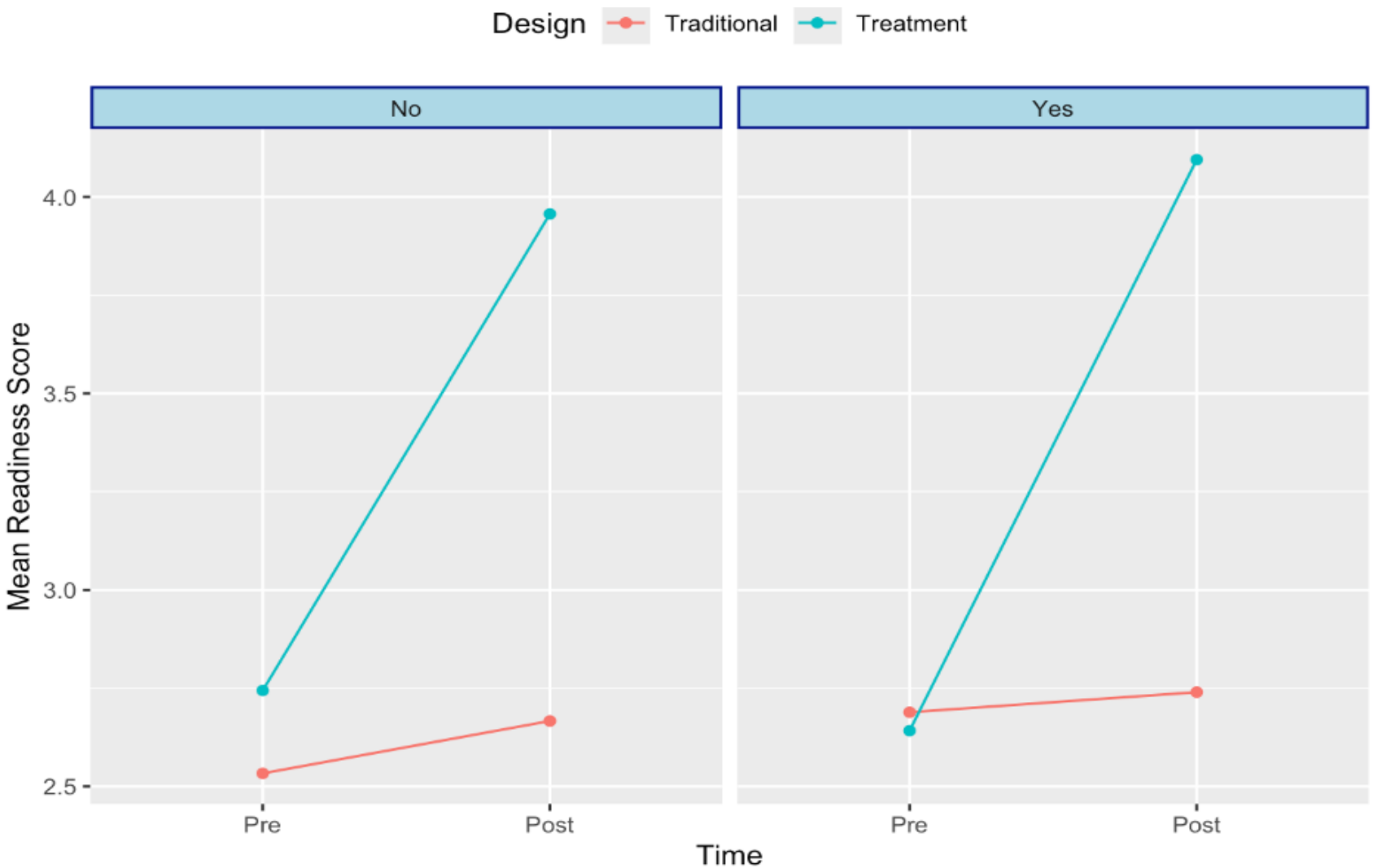
Year 2 (FA23-SP24) Key Results

- DS Readiness Score by STEM



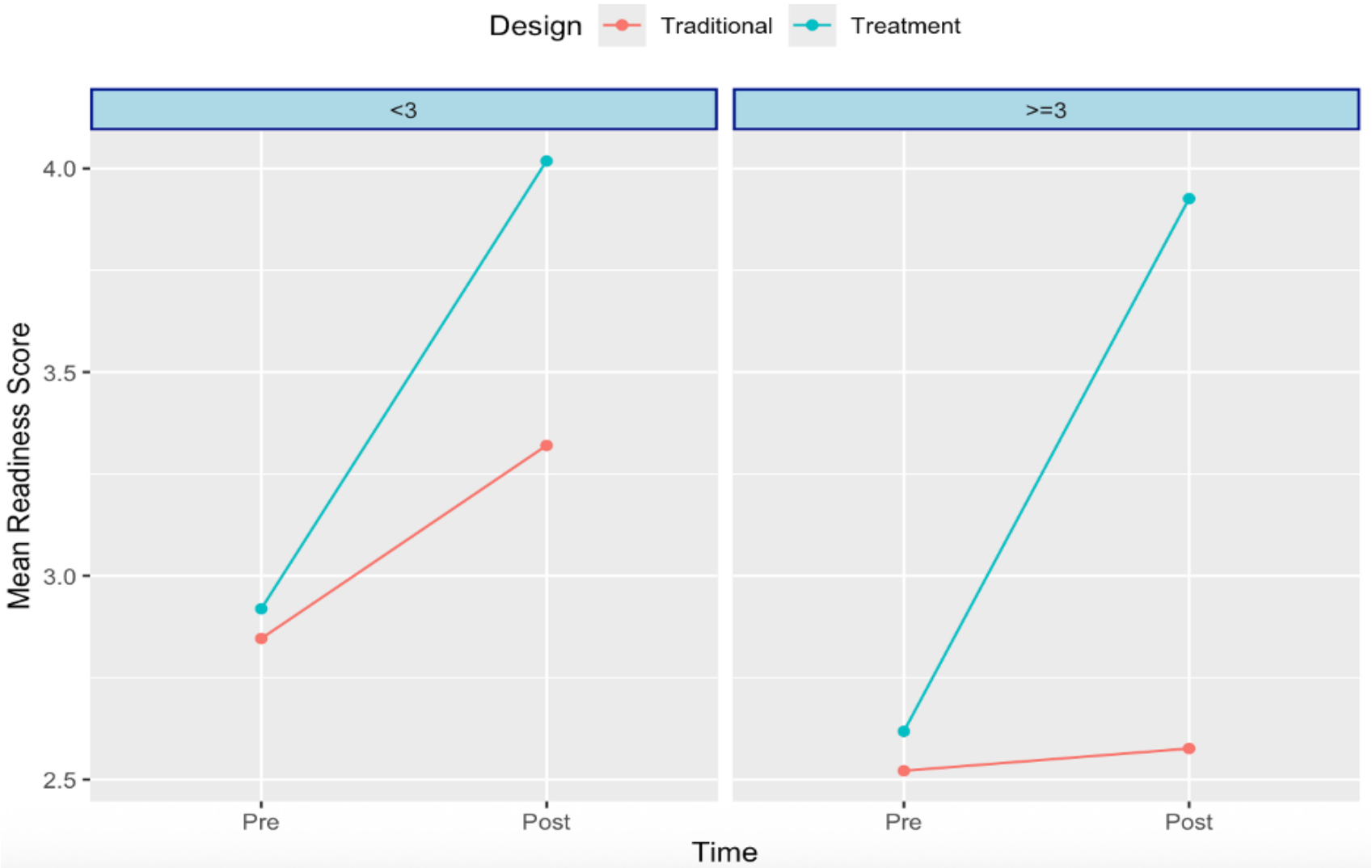
Year 2 (FA23-SP24) Key Results

- DS Readiness Score by AP Statistics



Year 2 (FA23-SP24) Key Results

- DS Readiness Score by GPA



Year 2 (FA23-SP24) Key Results

- Results of (partially overlapping) significance tests for DS readiness (confidence) by course design

Design	Difference Estimate	Test Statistic	P-value
Regular Pre/Post Design			
Traditional	0.08	0.53	0.2986
Treatment	**1.24**	15.35	**<0.0001**
Retrospective Pre/Post Design			
Traditional	**0.45**	3.91	**0.0001**
Treatment	**1.81**	22.45	**<0.0001**

Year 2 (FA23-SP24) Key Results

- Results of multivariable linear regression for the DS readiness post-test score (regular pre/post design)

Term		Estimate	SE	95% LCL	95% UCL	P-value
Intercept		1.32	0.329	0.67	1.97	<0.0001
Pretest Score		**0.23**	0.045	0.14	0.32	**<0.0001**
Design	Treatment	**1.21**	0.168	0.87	1.55	**<0.0001**
SEX	Male	-0.13	0.128	-0.39	0.12	0.2938
PELL	Yes	0.01	0.111	-0.21	0.23	0.9080
RURAL	Yes	-0.02	0.165	-0.35	0.31	0.8962
RESIDENCY	Out-of-State	-0.10	0.110	-0.32	0.12	0.3559
STEM	Yes	**0.38**	0.109	0.16	0.59	**<0.0001**
AP STAT	Yes	0.10	0.155	-0.22	0.41	0.5419
Pre-Course GPA	≥3.0	-0.22	0.152	-0.53	0.08	0.1466
Course Grade	A	**0.84**	0.294	0.24	1.45	**0.0084**
	B	**0.53**	0.259	0.008	1.07	**0.0469**
	C	0.59	0.286	-0.009	1.180	0.0532
Attendance (%)		<0.01	0.003	-0.004	0.009	0.4792
R^2 = 30.41%		**Adjusted R^2 = 28.21%**			**P-value = <0.0001**	

Year 2 (FA23-SP24) Key Results

- Results of multivariable linear regression for the DS readiness post-test score (retrospective pre/post design)

Term		Estimate	SE	95% LCL	95% UCL	P-value
Intercept		0.93	0.387	0.16	1.69	**0.0175**
Retro-Pretest Score		**0.27**	0.055	0.16	0.38	**<0.0001**
Design	Treatment	**1.38**	0.174	1.03	1.72	**<0.0001**
SEX	Male	-0.23	0.1431	-0.51	0.05	0.1100
PELL	Yes	-0.03	0.124	-0.27	0.21	0.8076
RURAL	Yes	0.06	0.182	-0.30	0.42	0.7363
RESIDENCY	Out-of-State	-0.13	0.13	-0.39	0.13	0.3135
STEM	Yes	**0.39**	0.125	0.14	0.63	**0.0022**
AP STAT	Yes	0.09	0.139	-0.19	0.36	0.5298
Pre-Course GPA	≥3.0	-0.24	0.156	-0.55	0.07	0.1262
Course Grade	A	**1.15**	0.267	0.63	1.68	**<0.0001**
	B	**0.81**	0.250	0.32	1.30	**0.0014**
	C	**0.77**	0.263	0.25	1.29	**0.0037**
Attendance (%)		<0.01	0.004	-0.004	0.01	0.3160
R^2 = 32.62%		**Adjusted R^2 = 29.14%**			**P-value = <0.0001**	

Year 2 (FA23-SP24) Key Results

- DS aspirations among students who heard about DS by course design

		N (%) of Yes	
Question	**Design**	**Traditional**	**Treatment**
Plan to take DS courses?	Presurvey	7 (21.21)	21 (11.86)
	Retro-Presurvey	1 (**3.33**)	14 (**6.28**)
	Postsurvey	0 (**0.00**)	19 (**8.52**)
Plan to complete a UG DS Certificate?	Presurvey	2 (6.06)	11 (6.21)
	Retro-Presurvey	3 (**10.00**)	2 (**0.90**)
	Postsurvey	0 (**0.00**)	14 (**6.28**)
Plan to complete a UG DS minor?	Presurvey	0 (0.00)	2 (1.13)
	Retro-Presurvey	0 (**0.00**)	2 (**0.45**)
	Postsurvey	0 (**0.00**)	4 (**1.79**)
Plan to complete a UG DS Degree?	Presurvey	0 (0.00)	1 (0.56)
	Retro-Presurvey	0 (**0.00**)	1 (**0.45**)
	Postsurvey	0 (**0.00**)	7 (**3.14**)

Year 2 (FA23-SP24) Key Results

- Results of (partially overlapping) significance tests for performance on the CAOS test by course design

Design	Difference Estimate	Test Statistic	P-value
Traditional	9.11	3.09	0.0018
Treatment	**12.36**	10.96	**<0.0001**

Year 2 (FA23-SP24) Key Results

- Results of multivariable linear regression for the CAOS post-test score on course design and students' characteristics

Term		Estimate	SE	95% LCL	95% UCL	P-value
Intercept		39.66	6.577	26.72	52.60	<0.0001
Pretest Score		**0.37**	0.078	0.22	0.52	**<0.0001**
Design	Treatment	2.59	2.454	-2.24	7.42	0.2491
SEX	Male	**-4.32**	2.209	-8.78	-0.08	**0.0480**
PELL	Yes	-1.26	1.916	-5.04	2.51	0.2306
RURAL	Yes	**-8.05**	2.674	-8.05	2.67	**0.0028**
RESIDENCY	Out-of-State	-3.14	2.083	-7.24	0.96	0.1332
STEM	Yes	**6.10**	1.961	2.25	9.96	**0.0020**
AP STAT	Yes	3.14	2.140	-1.07	7.35	0.1436
Pre-Course GPA	≥3.0	-0.08	2.521	-5.04	4.88	0.9761
Course Grade	A	-0.97	4.236	-9.31	7.36	0.8187
	B	0.09	3.956	-7.69	7.87	0.9817
	C	-2.04	4.015	-9.95	5.86	0.6109
Attendance (%)		-0.04	0.061	-0.16	0.08	0.4862
R^2 = 32.62%		**Adjusted R^2 = 12.03%**			**P-value = <0.0001**	

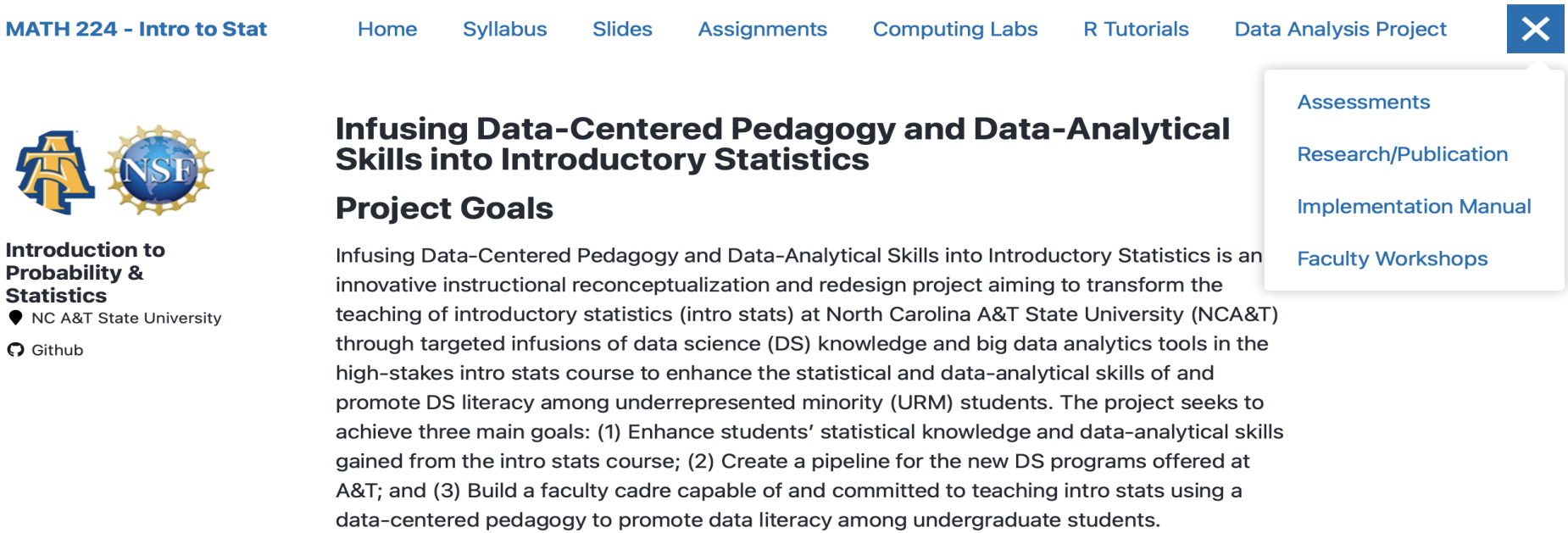
Main Findings

Integration of DS tools/knowledge into Intro Stats was associated with

- substantial gains in students' levels of DS awareness
- significant gains in students' levels of readiness for DS
- modest gains in students' aspirations of DS
  - only when using the retrospective pre/post design
- substantial statistical learning gains
  - higher than the national average of 11 points
  - Treatment learning gains $= 12.36$

Resources for Teaching a DS-Infused Intro Stats Course

- Project's Website on GitHub: https://introtostatncat.github.io



- This work is supported by NSF Grant #HRD2106945

- Project Team: Sayed Mostafa; Tamer Elbayoumi; Seongtae Kim; Mingxiang Chen; Guoqing Tang
- Implementation Team: Introductory Statistics Instructors and Graduate Assistants

References

- Cobb, G. (2015). Mere Renovation is Too Little Too Late: We Need to Rethink our Undergraduate Curriculum from the Ground Up. *The American Statistician*, 69, 266-282.
- Horton, N.J., Baumer, B.S. and Wickham, H. (2015). Setting the stage for data science: integration of data management skills in introductory and second courses in statistics. *CHANCE*, 28(2):40-50.
- Horton, N.J. and Hardin, J.S. (2021). Integrating computing in the statistics and data science curriculum: Creative structures, novel skills and habits, and ways to teach computational thinking. *Journal of Statistics and Data Science Education*, 29:sup1 S1-S3.
- Tintle, N., Clar, J., Fischer, K., Chance, B., Cobb, G., Roy, S., Swanson, T. and Vanderstoep, J. (2018). Assessing the Association Between Precourse Metrics of Student Preparation and Student Performance in Introductory Statistics: Results from Early Data on Simulation-Based Inference vs. Nonsimulation-Based Inference. *Journal of Statistics Education*, 26(2), 103-109.
- Woodard, V. and Lee, H. (2021). How students use statistical computing in problem solving. *Journal of Statistics and Data Science Education* 29(1), 1– 18.