**The internet is a big deal yall:**

The use of the Internet has become integrated into daily life as a means of accessing information, interacting with others, and tending to required tasks. The International Telecommunication Union reports that over half the world is online, and 70% of 15-24 year olds are on the internet (ICT report). Further, the Nielson Total Audience report from 2016 indicates that Americans spend nearly 11 hours a day in media consumption (Nielson report). Researchers discovered that online data collection can be advantageous over laboratory and paper data collection, as it often cheaper and more efficient (Illieva, Baron, and Healey 2002; Schuldt and Totten

1994, Reips chapter). Internet questionnaires first appeared in the early 90s when HTML scripting code integrated form elements, and the first experiments appeared soon after (Musch & Reips, 2000, Reips 1997). The first experimental lab on the internet was the Web Experimental Psychology Lab formed by Reips (<http://www.wexlab.eu)>, and the use of the Internet to collect data has since grown rapidly. What started with email and HTML forms has since moved to whole communities of available participants including websites like Amazon’s Mechanical Turk, and Qualtric’s Participant Panels. Participants of all types and forms are easily accessible for somewhat little to no cost.

**But are the data the same in person versus online**

Our ability to collect data on the Internet has inevitably lead to the question of measurement invariance between in person and online data collection methods. Invariance implies that different forms, data collection procedures, or even target demographics produce comparable sets of responses, which is a desirable characteristic to ensure a minimal number of confounding variables.

Same:

Different:

* Work by Dillman and his group (Dillman & Bowker, 2001; Dillman, Smyth, & Christian, 2008; Smyth, Dillman, Christian, & Stern, 2006), among others, has shown that many Web surveys are plagued by problems of usability, display, coverage, sampling, nonresponse, or technology.

**Everything has really been about person versus online but we should go deeper into the research design question**

With the development of advanced online survey platforms, such as Qualtrics and Survey Monkey, researchers have the potential to control potentially confounding research design issues. Randomization has been a hallmark of good research practice, as the order or presentation of stimuli can be a noise variable in a study with multiple measures. Thus, researchers have often randomized scales by rotating the order of presentation in paper format or simply clicking the randomization button for online studies. This practice has counterbalanced out any order effects of going from one scale to the next. However, within scale structure has remained constant, maintaining the order of presentation for particular items on a scale. These items are still stimuli within a larger construct, therefore, potentially, they have the ability to influence the items that later on the survey, which we will call *item reactivity*. For example, a question about *being prepared for death* or *thoughts about suicide* might change the responses to further questions, especially if previous questions did not alert participants to be prepare for that subject matter.

Scale development typically starts with an underlying latent variable that a researcher wishes to examine through measured items or questions. Question design is a well-studied area that indicates that measurement is best achieved through questions that are direct, positively worded, and understandable to the subject. CITE suggests that researchers design a multitude of items to investigate and invite subject matter experts to examine these questions. After suggested edits from these experts, a large sample of participant data is collected. While item response theory is gaining traction, classical test theory has dominated this area through the use of exploratory and confirmatory factor analysis (EFA, CFA). EFA elucidates several facets of how the measured items represent the latent trait through factor loadings and overall model fit. Factor loadings represent the correlation between each item and the overall latent variable, where a researcher wishes to find items that are strongly related to the latent trait. Items that are not related to the latent trait, usually with factor loadings below .300 (Preacher) are discarded. Model fit is examined when simple structure has been achieved (i.e. appropriate factor loadings for each item), and these fit indices inform if the items and factor structure model fit the data well. Well-designed scales include items that are highly related to their latent trait and have excellent fit indices. Scale development additionally includes the examination of other measures of reliability (alpha) and construct validity (relation to other phenomena) but the focus of the scale shifts to subscale or total scores. Published scales are then distributed for use in the form that is published, as item order is often emphasized through important notes about reverse scoring and creating subscale scores.

**What exactly do the items do to each other though?**

These scale development procedures focus on items, and EFA/CFA statistically try to mimic variance-covariance structure by creating models of the data with the same variance-covariance matrix. If we imagine that stimuli in a classic experimental design can influence the outcome of a study because of their order, then certainly the stimuli on a scale (i.e., the items) can influence the pattern of responses for items. This area of study is relatively unexplored, as easy randomization has only recently become available for researchers. In our study, we contribute the literature on in person versus online testing by exploring how paper and pencil formats compare to non-randomized online surveys. Because this literature is mixed, we provide multiple sources of evidence (*p*-values, effect sizes, Bayes Factors, and tests of equivalence) to determine if differences found are not only statistically significant, but also practically significant. Second, we expand to item randomization for online based surveys, examining the impact on item loadings to their latent variable, variance-covariance structure, item means, and total scores again providing evidence of difference/non-difference from multiple statistical sources. Finally, we examine these scenarios with a unique set of scales that have both dichotomous True/False and traditional Likert formats to explore how the answer response options might impact any differences found between in person, online, and randomized methodologies.

**This is in the method section but also needs to jive with the intro and talk about BF and TOST because most audiences won’t know them:**

## Hypothesis and Data-Analytic Plan

Each hypothesis was tested using three dependent measures. The variance-covariance matrix for each type of delivery was estimated and compared to each other by using root mean squared error (RMSE; CITE). RMSE estimates the difference between covariance matrices and is often used in structural equation modeling to determine if models have good fit to the data. A criterion of < .06 for good fit, .06-.08 for acceptable fit, and > .10 for bad fit was used (Hu & Bentler, 1999?). This analysis was used to determine if the change in delivery changed the structure of the item relationships to each other. RMSE values were calculated using the monomvn package in R (Gramacy, CITE).

Next, item averages were calculated across all participants for each item. These 20 items were then compared in a matched dependent t-test (i.e. item 1 to item 1, item 2 to item 2) to determine if delivery changed the mean of the item. While covariance structure elucidates the varying relations between items, we may still find that item averages are pushed one direction or another by a change in delivery and still maintain the same correlation between items. If this test was significant, we examined the individual items across participants for large effect sizes, as large sample sizes would create significant t-test follow ups. Last, the total scores for each participant were compared across delivery type using an independent t-test. Item analyses allow a focus on specific items that may show changes, while total scores allow us to investigate if changes in delivery alter the overall score that is used in other analyses. In both the item and total score analyses, d values and Bayes Factors are provided to examine the size of effects for interpretation, instead of p-values that are biased by sample size.

Hypothesis 1. Paper forms were compared computerized non-random forms to examine the method of delivery on the relationships between items, item means, and total scores. We expected to find XX consistent with previous research by XX.

Hypothesis 2. Computer forms were then analyzed by randomized and nonrandomized groups to examine the impact of randomization on covariance structure, item means, and total scores. We expected to find that these forms would vary across covariance structure and item means, which would indicate differences in reactivity to questions (i.e. item 4 always has item 3 as a precursor on a nonrandom form, while item 4 may have a different set of answers when prefaced with other questions). We examined total scores; however, it was unclear if these values would change. A difference in item means may result in changes in total scores, but may also result in no change if some item means decrease, while others increase.

Use these links later for citations.

<http://www.nielsen.com/us/en/insights/reports/2016/the-total-audience-report-q1-2016.html>

<http://www.uni-konstanz.de/iscience/reips/pubs/papers/chapters/Reips2012APAwith_edits.pdf>