**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 is 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.

According to Deutskens, Ruyter, and Wetzels (2006), mail surveys and online surveys produce nearly identical results with regards to response characteristics. The authors suggest the number and length of responses were also similar, as well as the accuracy of the data collected online versus by mail (Deutskens, Ruyter, & Wetzels, 2006). Only minor differences arise between online surveys and mail in surveys when it comes to participant honesty and suggestions. For example, participants who responded to surveys online provided more suggestions, lengthier answers, and greater information about competitors in the field that they may prefer (Deutskens, Ruyter, & Wetzels, 2006). The hypothesis as to why individuals may be more honest online than in person is that the individual may feel more anonymity and less social desirability effects due to the nature the online world, therefore less concerned about responding in a socially polite way. A trend found by Fang, Wen, and Prybutok (2012) shows individuals are more likely to respond to surveys online with extreme scores, rather than mid-range responses on Likert scales due to the lessened social desirability factor. There may be slight cultural differences in responses online. For example, collectivistic cultures showed greater tendency toward mid-range responses on Likert scales via in-person and online due to placing greater value on how they are socially perceived; however, the trend is still the same as scores are more extreme online versus in person or by mail (Fang, Wen, & Prybutok, 2012).

Although 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, other studies have found internet data to be reliable and almost preferable as it produces a varied demographic response compared to the traditional sample of introduction to psychology college students while also maintaining data equivalence (Lewis, Watson, & White, 2009).

**Limitations to Web-Based Survey Designs**

It is important to consider accessibility of different populations when developing web-based surveys, among other limitations. Nua Internet surveys estimate that only around 3% of the world’s population has access to internet, 60% of which is accounted for by Canadian and American citizens (Dillman & Bowker, 2001). Additionally, For December of 1998, it was estimated that around 40% of American households owned computers. Among these households, however, only 20% had access to the internet (do we have more updated information to put here possibly?). Thus, it is not prudent to assume that our web-based sampling methods are adequately representative of the desired population. Coverage becomes less of a problem when examining the desired population in a given context, as some populations all have the ability to access the internet (e.g. college students) (Dillman & Bowker, 2001). An additional limitation that warrants caution and consideration in developing web-based studies concerns the response rate for the given study. In a somewhat-recent meta-analysis, Cook, Heath, & Thompson (2000) compiled participant response data from 56 web-based surveys in 39 studies. They found the response rate to be relatively low at 34.6% (de Leeuw & Hox, year). While web-based surveys are both convenient and cost-effective, it is necessary to consider the limitations of these particular surveys and how they could potentially confound data.

Web-based surveys also provide a different means of response, one that is typed instead of spoken or written (De Leeuw, 1992, 1998). Tourangeau, Rips, & Rasinki (2000) claim that these responding methods differ in the amount of privacy and burden that they offer the participant. Perhaps participants are less focused working from their computer at home. Based on this information, one might expect the responses of a web-based survey to differ from those of an in-person survey. It is important to consider the implications of these different responses, as they have the potential to confound research hypotheses and designs. It is necessary to know whether or not they do in fact affect item means and descriptive statistics in the same sample size.

**Controlling Web-Based Surveys**

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 through randomization. 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 web-based studies. This practice has counterbalanced out any order effects of going from one scale to the next. However, while scale structure has remained constant, these items are still stimuli within a larger construct. Therefore, these construct-related items 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.

The question is no longer whether web-based surveys are reliable sources of data collection; the theory now is in need of a shift to whether or not item-randomization in survey data collection creates differences…. (add in stuff about item randomization here… if it exists; or maybe talk about why it could make a difference/what our paper offers)

**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.

**Evidentiary Value**

Talk about p value conundrum and why we should use multiple forms of evidence and that we are going to do that.