**Faith**

1. Hinkin (1998) describes a deductive or an inductive method to develop scales. In order to deductively create a scale, there must be a theoretical foundation that provides enough information to generate the initial set of items. Whereas, an inductive approach is appropriate when the conceptual basis for a construct may not result in easily identifiable dimensions for which item can be generated. Is it always clear which method to use? If not, what is preferred or look upon more favorably?   
  
2. Podsakoff, MacKenzie, Lee, & Podsakoff describe multiple statistical remedies throughout their paper. It was quite an exhaustive list, but they were always advantages and disadvantages to their methods. Is there a statistical remedy that is most commonly used within our field or does it truly depend on the variables or circumstance?   
  
3. Two papers that we read mentioned reverse coded items. Podsakoff et al. mentions that research has shown that reverse-coded items may produce artifactual response factors consisting of negatively worded items that may disappear after the reverse-coded items are rewritten in a positive way. They continue on to state that negatively worded items may be a source of method bias. Do you agree with this statement?

**Angela**

1. Podsakoff et al. discusses the issue of item-embeddedness in surveys, specifically that neutral statements can be accidentally warped if located near positive or negative statements. But how can we avoid that? Positive and negative statements seem to me a necessity.

2. Hinkin mentioned several times that there are "no hard or fast rules" in designing measures for surveys. But, maybe there should be a list of specific rules for all researchers to abide by. What do you think?

3. We often reject self-ratings in favor of other-ratings. However, Schwarz & Oyserman mentioned actor-observer differences that can influence the reporting of others' behavior. The authors state that when rating the behavior of others, people tend to rely on assumptions of the target individual's dispositions, which isn't necessarily accurate. Which would you choose, self- or other-ratings?

**Ashley**

1. According to Hinkin (1998), attitudes toward reverse-scored items are inconsistent among researchers. For example, Price and Mueller (1986) believed that the use of reverse-scored items would help reduce response set bias. However, Harrison and McLaughlin (1991) found that the use of a few of reverse-scored items randomly interspersed within a measure may be a detrimental effect on psychometric properties of a measure. What do you think of reverse-scored items then? How, do you think, should we decide if it is appropriate to use such items under a given circumstance?

2. In Hinkin (1998), the author cited Bagozzi et al. that confirmatory factor analysis in construct validation overcomes the weakness of MTMM; the author also predicted that MTMM may have eventually got replaced by this technique. So how is MTMM doing now? Has it been replaced or is it being gradually replaced by confirmatory factor analysis? If it’s still doing well, do you believe that it will be replaced by the new technique someday?

3. One of the techniques for controlling common method biases, according to Podsakoff et al. (2003) was improving scale items. However, they also mentioned that trying to vary the scale anchors and formats and to avoid the use of bipolar scale value might have resulted in validity decrease. So what do you think is the best way to balance this kind of tradeoff between common method biases control and scale validity? Is there any way that we can maximize the control, and meanwhile sacrifice as little validity as possible?

**Chelsea**

1. In step 3, "initial item reduction" of measurement development in Hinkin's paper, the author only talked about EFA and reliability analysis. Before that, shouldn't there be a step for analyzing the discrimination of every item? (By taking out the highest 23% and lowest 23% of the responses in an item and calculating the mean difference of the two.

2. In this week's reading, both Hinkin and Schwartz et al. talked about procedure and criteria in measurement development. Among these readings, we've frequently come across words such as "no strict guideline" or "no hard rule" when describing the criteria for including, deleting items or deciding on measurement structure. What potential problems can a not having "strict criteria" bring about

3. For calculating reliability, Hinkin still recommended using Cronbach's alpha. Referring to Edwards' advice in using advanced reliability analysis methods that was mentioned in his 2003 paper, "Construct Validation in Organizational Behavior Research", one of our previous week's reading, how would the process of measurement development be changed if we followed Edwards' advice?

**Noam**

1. The authors claim to have limited the possible types of deviant behavior to improve reliability, and that behaviors that were eliminated were "very rare". Yet authors discarded such a classic as "calling in sick when you are not", which had a pretty high participation rate, and has been shown in the past to be a huge economical problem on a national scale. Anyway, the question is what "box of tools" do we have for selection criteria for items to go into such a scale, and how can we better encompass such a construct as "deviant behavior" in the workplace?

2. Maybe it's just me, but I didn't understand how (if at all) Colquitt supported his hypotheses when it comes to the structural model. For example, just by glancing at the data it seems leader evaluation is related to all forms of justice. Why not address this? And how should we analyze and present structural models in general?

3. Based in the Hinkin paper, one may think any construct could be represented using 4 to 6 items. Does the nature of the construct and its complexity not matter at all?

**Lauren**

1. Bennett & Robinson write that "It should be noted that the process of validating a construct is never complete; no measure can ever be said to be validated in any final sense." Under what circumstances should a measure be edited? What happens if it's changed many times - does it become something different than what the authors originally intended?

2. If a measure currently in use has more than 4-6 items per construct, should the number of items be reduced if possible? What is the benefit of reducing the number of items versus keeping it as is?

3. Podsakoff et al. mention positive and negative affectivity as potential source of common method bias. If positive or negative affectivity is something dispositional, how does it result in responses that are less "true" if the responses are indeed an accurate reflection of how the person feels?

**Colin**

1. With such high praise for event history calendars, do you think the method has any applicability for I/O research? Maybe it can help with better recall of CWB and OCB?  
  
2. Hinkin's 1998 item generation rules cover the development of new theoretical measures. I note that personality inventories often have no theoretical origin (16PF) or simply take and modify items from other inventories (MMPI). Is item generation arbitrary as long as you have a large enough pool to narrow down during factor analysis?  
  
3. Schwarz and Oyserman (2001) described the cognitive steps by which participants respond to items. Could it be that variation due to cross-cultural differences (at least those that have different languages) is due to different language and phrasing of the translated questionnaires?

**Jonathan**

1. Event history calendars (Schwartz) seems to have a lot going for them, I wonder if we can somehow apply them to 1) I/O research when it comes to behaviors, 2) apply it to attitudes as a means to get more variance? How would you apply this method to your field of study? I think this is lots of potential for vocational interest since it has behavioral elements in it!

2. Hinkin’s article spends some time on the number of items to be fitted into a scale and yet more time on both inductive and deductive processes. With so much effort required the creation of a decent scale, it is a wonder to me how some of the paper published in our top journals get away with “home grown” measures consisting of 1 or 2 items. What should editors / reviewers do to ameliorate this? Then again with the time taken to develop such scales, is there forgiveness to be found if the research had to opportunity to study an interesting phenomena but didn’t have the time to develop those scales?

3. For a number of the articles on scale development, I can’t help but notice these lean heavily on how to construct and deal with LIKERT scales. There are other scaling measurement models out there, such as Thurstone scaling. Do you think, given that Thurstone scaling is supposed to a better measure of attitudes that it would benefit our fields A LOT more if we did have that kind of scaling emphasized instead? What do you think is holding us back?

**Jing**

1. Hinkin mentions that “domain sampling theory states that it is not possible to measure the complete domain of interest, but that it is important that the sample of items drawn from potential items adequately represents the construct under examination”. Is it important to have a complete measure of the domain of interest? How should we judge the adequateness of the items?

2. Hinkin talks about developing items in inductive ways. How could we guarantee that items developed in this way tap different dimensions of the theoretical construct of interest?

3. After reading Podsakoff et al, how do you think should we deal with common method biases in our research?

**Bertha**

1. Both Hinkin and Schwartz & Oyserman mention there is no "hard rules" or "strict guidelines" for creating a new measurement structure (this explains why so many bad measures exist in our field!).  If we decided that 5 rules MUST be followed when creating a new measurement what would they be?   
  
2. The behaviors in the Bennett and Robinson workplace deviance scale tend to have a very low base rate. When validating their scale, one of the samples had a participation rate well over 50%.  The authors explained that the participants in this sample were young and younger inexperienced employees may engage in more deviant behaviors than older employees. What other factors could cause such a high participation rate (factors mentioned in Schwartz & Oyserman)? Also, if the sample was not representative of the population, how might it have affected the creation of this measure?  
  
3. Podsakoff et al recommend using a cover story to psychologically separate the predictor variable from the criterion variable measures.  If participants can tell that the researcher wants them to see them as separate, how would this affect their responses?

**Pete**

1. In their paper on the development of a measure for workplace deviance, Bennett and Robinson make a note that during their item selection process, 30 of their 58 items were dropped because they had very low variances. That this might be a very high number is another problem, but more interestingly, what might this say about (and how might researchers get around) items that are particularly heinous (and reported infrequently)?

2. Podsakoff et al. mention the more common types of method biases and where they might be found in the research process. As a class of both I/O, social, and personality psychologists, do you notice biases that apply more or less to your domain of studies? Do you feel there are method biases particular to your set of studies that the authors might have missed?

3. One of the more common types of method biases talked about in this weeks readings seems is the anchors used on Likert scales, and how these anchors can be ambiguous especially with respect to reporting the frequency of past behaviors. Does this pose a significant challenge for meta-analyses, if researchers attempted to combine scales that are implying temporally different time scales?

**Bennett**

1. Hinkin (198) lays out two methods of item development: inductive and deductive. There are upsides and drawbacks to each, and sometimes one is more appropriate than another. But is it really so clear-cut? Is it possible that we can, or perhaps often do, intermingle these approaches?

2. The steps laid out in todays readings serve as very helpful advice for the creation of future surveys, but what about ones that already exist and have been along for a long time? Usually people just sort of assume that they are valid because they are so long-lasting, but could the opposite actually be true? In fact, it seems like many of the things the authors advise against (e.g. reverse coding, unless done very carefully) were probably state-of-the-art when these measures were developed. Can we somehow subject these older measures to modern standards, or should we generate new ones?

3. My namesake & Robinson (2000) mention that the process of validating a scale is never complete, and that further work should be done to validate and refine their scale. Do you think this idea translates into practice, or do you think that people are often just glad that a scale exists, is "validated" (perhaps in a single study not as well carried out as this one) and will just use it without thinking twice? A lot of the readings we did this week present validation as a step-by-step process, at the end of which, presumably, you are done (although Hinkin does show some iteration in the figure)? Is this the right way to think about validation? More importantly, how do you think people typically think about it?

**Rachel**

1. Surveys are often modified to better fit the population being studied (e.g. questions about work outcomes are changed to refer to grades when administered to a student vs. a working adult sample). Are modifications (even minor ones) of validation scales justified? Do they require data to support validity or is this a waste of resources?

2. Schwarz and Oysterman (2001) discuss the implications of different frequency scales on respondents' answers. If you were designing a new measure in which participants were asked to recall how many times they performed a certain behavior, how would you go about determining the proper time frame (e.g. within the last week, within the last month, etc.)?

3. I once spent an entire work day debating with my supervisor whether or not "my organization provides quality customer service" and "my organization provides \*high\* quality customer service" were the same question. Do you think any of Schwarz and Oysterman's or Hinkin et al.'s scale development strategies would adequately determine whether participants interpret "quality" as inherently high? If so, which strategies would best determine this information.