Practical Assessment, Research & Evaluation

A peer-reviewed electronic journal.

Volume 23 Number 5, April 2018 ISSN 1531-7714

Scale Pretesting

Matt C. Howard, *University of South Alabama*

Scale pretests analyze the suitability of individual scale items for further analysis, whether through judging their face validity, wording concerns, and/or other aspects. The current article reviews scale pretests, separated by qualitative and quantitative methods, in order to identify the differences, similarities, and even existence of the various pretests. This review highlights the best practices and objectives of each pretest, resulting in a guide for the ideal applications of each method. This is followed by a discussion of eight questions that can direct future research and practice regarding scale pretests. These questions highlight aspects of scale pretests that are still largely unknown, thereby posing a barrier to their successful application.

Most guides for the scale development process suggest that researchers and practitioners should begin by generating an over-representative item list, which helps ensure adequate content coverage (Hinkin, 1995, 1998; MacKenzie et al., 2011; Meade & Craig, 2012). These guides typically suggest that the second step should be a reduction of this item list via exploratory (EFA) or confirmatory factor analysis (CFA) to minimize construct contamination. An increasing number of authors, however, have suggested that a distinct intermediate step should be taken between item development and EFA/CFA (Anderson & Gerbing, 1991; DeVellis, 2016; Hardesty & Bearden, 2004; MacKenzie et al., 2011). This intermediate step is the scale pretest.

Most often, scale pretests use a small number of participants (i.e., 5 to 30) to initially reduce the item list before reducing it further via EFA or CFA. As prior authors have suggested (DeVellis, 2016; Presser et al., 2004), scale pretests have arisen primarily for four reasons. First, many recommended sample sizes for EFA and CFA depend on the number of items, such as 10 participants for every item analyzed (Brown, 2015; Hinkin, 1995, 1998; Howard, 2016; Thompson, 2004). If the initial item list is large, it may be difficult – if not impossible – for some researchers to obtain a sufficient sample size, but an item-sort task can reduce the item list into a more manageable size for EFA or CFA. Second, even with a reduced item list, a sufficient sample size may still be unobtainable. In these instances, scale pretests have been used in place of EFA or CFA. Third, scales may need to be created for constructs that are not central to the research effort. In these cases, it may be unreasonable for a researcher or practitioner to undergo the full-scale development process, but scale pretests can provide some inferences regarding the ability of a developed scale to gauge its intended construct. Fourth, some pretest methods can ascertain aspects of items that cannot be identified through EFA or CFA (Presser et al., 2004).

Discussions of pretest methods are beginning to appear in broader reviews of the scale development process, but focused reviews of pretests are still scarce (Hunt et al., 1982; Howard & Melloy, 2016; Presser et al., 2004). As shown below, the dearth of pretest reviews results in the application of many different pretest methods, but authors rarely provide justification for applying their chosen method. Likewise, notable differences can be seen between applications of the same pretest method. This suggests that pretest methods are possibly being used in a haphazard manner, and researchers may be applying pretest methods that are not ideal for their research needs. Due to these concerns, we contend that researchers and practitioners may be unaware of the differences, best practices, and even existence of the various pretest methods. To address these concerns and prompt a more systematic application of scale pretests, we review the best practices of scale pretesting and identify eight questions to direct future research.

Scale Pretests

The goal of a scale pretest is to identify items that may be justifiably retained for further testing. The manner in which pretest methods achieve this goal differs, but it is often consistent with whether the pretest is quantitative or qualitative. Most often, quantitative pretests obtain a numerical measure of face validity, which is assumed to contribute to the overall construct validity of the eventual scale (DeVellis, 2016; Hardesty & Bearden, 2004; Howard & Melloy, 2016). Construct validity is "the degree to which a test measures what it claims, or purports, to be measuring" (Brown, 1996, p. 231). The construct validity of a scale can never be known, but it is supported by the cumulative results of the scale development process. Face validity is the extent that a scale or item is subjectively judged to represent its intended construct. A scale consisting of items with adequate face validity is often assumed to have adequate construct validity (although this is not always the case). For this reason, items that are judged to have sufficient face validity are retained for further analysis when using quantitative pretest methods.

On the other hand, qualitative pretest methods do not judge the validity of items as directly as quantitative pretest methods (Blair et al., 2013; Fowler, 2013; Presser et al., 2004). Instead, qualitative pretest methods primarily identify whether items have certain wording concerns, such as being double-barreled, leading, or confusing (Leech, 2002). Some qualitative pretest methods are able to identify items with face validity concerns, but these pretest methods do not provide a direct numerical indicator that can, for example, be used to rank the items by their face validity. For this reason, these qualitative pretest methods may be able to remove items with large face validity concerns, but they cannot be used to solely retain the items with the greatest face validity. Below, both quantitative and qualitative pretest methods are reviewed.

Quantitative Pretest Methods

Three quantitative pretest methods are reviewed: item-rating tasks, item-sort tasks, and Hinkin and Tracey's (1999) ANOVA method. These methods were chosen for their popularity and importance, but we also provide brief summaries of lesser-used quantitative pretest methods.

Item-Rating Task

Item-rating tasks and item-sort tasks are among the most popular quantitative pretest methods (Anderson & Gerbing, 1991; DeVellis, 2016; Hardesty & Bearden, 2004; Howard & Melloy, 2016; Hunt et al., 1982; Lawshe, 1975). Despite the popularity of the former, many authors do not call item-rating tasks as such, instead only calling the procedure a pretest or assessment. We label this method an item-rating task to avoid any confusion.

To perform an item-rating task, participants are given a definition of the focal construct. Then, they are provided each item and asked to evaluate the extent that the item represents the focal construct. As noted by Hardesty & Bearden (2004), a common response scale consists of "clearly representative," "somewhat representative," and "not representative," but authors may also use other response scales, such as "very good," "good," "fair," and "poor." No firm rules exist for the recommended sample size for item-rating tasks, but researchers typically use sample sizes ranging from 10 to 30 (Anderson & Gerbing, 1991; Goetz et al., 2013; Heene et al., 2014).

Once responses have been collected, three approaches are most popular to make item retention decisions. First, a sumscore can be calculated for each item. Each response choice is assigned a corresponding value (e.g., very good – 4, good – 3, fair – 2, poor – 1); responses are summed for each item; and the highest scoring items are retained. Second, items that receive a certain percentage of the highest (e.g., very good) or two highest responses are retained. Third, items that receive any of the lowest response (e.g., poor) are discarded. In one of the few studies on item-rating tasks, Hardesty and Bearden (2004) provided support that the first and second approaches provide the most accurate item-rating task results, as defined by the likelihood that the approach replicated the item retention decisions of the entire scale development process.

When these three approaches are applied, authors often use a numerical cutoff that will retain a certain number of items, rather than an a priori chosen number (Hardesty & Bearden, 2004; Howard & Melloy, 2016). For instance, a researcher may be interested in creating a reduced item list of 30 items. When using the sumscore approach, 11 items may have received a score of 24 or greater, 33 items may have received a score of 23 or greater, and 40 items may have received a score of 22 or greater. If this were the case, the researcher would likely use a sumscore cutoff of 23 in order to retain 33 items for subsequent analyses.

While item-rating tasks have been successfully used in ample prior studies, the method poses certain concerns. Item-rating tasks may be poor at identifying items that represent more than one construct. If an item represents the focal construct and an alternative construct equally well, most researchers would not want this item in their final scale; however, an item-rating task may identify this item as adequately representing the focal construct.

Further, using item retention cutoffs that will retain a certain number of items may be useful, but this method goes against the notion of statistical testing. That is, statistical decisions are almost always made through a priori guidelines with statistical justifications, such as p-values, confidence intervals, and effect size guidelines (Bosco et al., 2015; Cohen, 1992, 1994; Nakagawa & Cuthill, 2007). Without such justifications, it should be questioned whether this approach is a true statistical method. More importantly, it should be questioned whether this method provides accurate and statistically-supported results. For instance, an item with 80% of respondents reporting “very good” may not be significantly more representative than an item with 75% of respondents reporting “very good.” Also, using cutoffs to retain a certain number of items results in different values being used from study-to-study, even if the number of items and participants remains the same, which again draws into question the validity of this method.

Item-Sort Task

Fortunately, another method alleviates some of these concerns noted above: the item-sort task. To perform an item-sort task, participants are given a detailed definition of the focal construct(s) as well as several other theoretically similar constructs. Then, participants are instructed to indicate which construct that they believe each item best represents. The list of choices should include the focal construct(s), other theoretically similar constructs, and an “any other construct” option. Typically, sample sizes for item-sort tasks include between 20 and 30 participants, but Howard and Melloy (2016) show that sample sizes as small as five can be used.

Once responses have been collected, authors calculate the number of times that each item was considered representative of the focal construct (Anderson & Gerbing, 1991). Items with a sufficient number of assignments to the focal construct are considered representative of that construct and not others. Two approaches can be used to make item retention decisions. First, authors can choose a cutoff that would result in the desired number of items to be retained. Second, authors can use the cutoff values provided by Howard and Melloy (2016) that are based on traditional statistical significance testing. Using this latter approach, the results of item-sort tasks have a sound statistical justification and have been shown to replicate EFA results.

Further, no matter the approach, item-sort tasks can address the noted concern of item-rating tasks. Item-sort tasks are not only able to identify items that poorly represent the focal construct, but they are also able to identify items that may represent multiple constructs. If an item represents two constructs equally well, then this item would be expected to have only half of the participants to indicate that it represents the focal construct. Using the cutoff values provided by Howard and Melloy (2016), an item that is considered representative of the focal construct half of the time is not statistically significant no matter the sample size. Likewise, items that only partially represent other constructs can still be identified using item-sort tasks (Anderson & Gerbing, 1991; Howard & Melloy, 2016). Thus, item-sort tasks address the notable concerns of item-rating tasks, while still providing the benefits of this other method.

Hinkin and Tracey’s ANOVA Method

A third quantitative pretest is Hinkin and Tracey’s (1999) ANOVA method, which was intended to be an improvement beyond item-rating and item-sort tasks. Participants are given a detailed definition of the focal construct as well as several other theoretically-related constructs. Then, the participants are provided each item and asked to evaluate the extent that the item represents each of the construct choices. The typical response scale ranges from 1 (not at all) to 5 (completely). While no firm guideline exists for sample size requirements, Hinkin and Tracey (1999) used samples of 57 and 173, but they also noted that samples of 30 may be acceptable. Once responses have been collected, a one-way ANOVA is performed for each item, comparing the item's mean value for each category. If an item has a significantly greater value for a certain category, then it is considered representative of that construct and not others.

Hinkin and Tracey (1999) suggested that their method could gauge the extent that an item represents multiple constructs, which was an improvement beyond item-rating tasks. They also suggested that item-sort tasks do not rely on statistical testing or take into account, "the extent to which an item may correspond to a given dimension" (Hinkin & Tracey, 1999, p. 180). While their method achieves these goals, recent developments in item-sort tasks also satisfy these goals.

Despite the proposed benefits, Hinkin and Tracey's (1999) method is not as widespread as item-rating tasks or item-sort tasks. While the reason is unclear, some suggestions can be made. First, Hinkin and Tracey's (1999) sample sizes in the demonstration of their method were very large for scale pretests, and researchers may have been wary of their method's accuracy with samples smaller than their examples. Second, providing individual ratings for each item in regard to each possible construct is cognitively taxing for participants. Researchers may have felt that most participants would not be motivated or have the ability to provide accurate ratings. Third, researchers may have believed that Hinkin and Tracey's (1999) method was not a sufficient improvement beyond item-rating tasks and item-sort tasks, as the application of these two methods persisted after Hinkin and Tracey (1999). Fourth, this method is more involved than item-rating and item-sort tasks, and researchers may prefer the easier alternatives. Despite these possibilities, there seem to be no statistical concerns with Hinkin and Tracey's (1999) ANOVA method, and the method may still be able to provide insightful results regarding initial item lists.

Other Quantitative Methods

Most other quantitative pretest methods are variants of the item-rating task. For instance, researchers have asked participants to rate the importance or difficulty of items, rather than their ability to gauge the focal construct (Coste et al., 1997; Goetz et al., 2013; Smith et al., 2000). These studies typically use the same guidelines as standard item-rating tasks, but they are used when item relevance may not be the most important determinant to retaining items.

Schriesheim and colleagues (1993) also developed a pretest method. Participants are provided each item and asked to evaluate the extent that the item represents each of the construct choices. The data is then used to calculate a q-correlation matrix, and this matrix is subject to a principal components analysis. The item loadings can be used to determine whether an item is representative of a construct. Despite being more sophisticated, Schriesheim and colleagues' (1993) method has not seen as much use as item-rating and item-sort tasks. This may be because Hinkin and Tracey (1999) directly compared their method to Schriesheim and colleagues' (1993) method, and Hinkin and Tracey (1999) argued that their method was superior.

Lastly, other quantitative pretest methods have seen modest use and provide little beyond the methods detailed above (Blair et al., 2013; Clark & Watson, 1995; Fowler, 2013; Goetz et al., 2013; Hardesty & Bearden, 2004; Rea & Parker, 2014). We do not review these methods, and instead turn to another important category of scale pretests: qualitative methods.

Qualitative Methods

Three qualitative pretest methods are reviewed in the following: cognitive interviews, focus groups, and traditional interviews. These were also selected for their popularity and importance, but we also provide brief summaries of other qualitative pretest methods.

Cognitive Interviews

The origins of cognitive interviewing date back to between the 1940s and 1970s (Belson, 1981; Cantril & Fried, 1944), in which researchers applied variations of the method with little standardization in their approaches. It was not until the 1980s that researchers more strongly considered the utility and accuracy of the approach. This shift, paired with the creation of several federally-funded "cognitive laboratories," began a more systematic application of cognitive interviewing as a scale pretest method (see Presser et al. 2004 for a review).

To perform a cognitive interview, participants complete the over-representative item list, and information is collected regarding the process of answering each item. Most often, cognitive interviews involve verbal data collection (Beatty & Willis, 2007; Presser et al., 2004), which requires the researcher to be present. The recorded information is then used to evaluate whether the participant perceives the item as intended and/or whether the participant had difficulty understanding the item, both of which may be indicative of item quality (Beatty & Willis, 2007). In the words of Presser and colleagues (2004), a cognitive interview is, "essentially a dress rehearsal" (p. 110), but the nature of the "dress rehearsal" may differ in many regards.

When performing a cognitive interview, researchers may use think-alouds, probes, or a combination of both. A think-aloud is when a participant is asked to speak their thoughts while completing the items, which may uncover any item confusion. For an item intended to gauge conscientiousness, a participant may say, "The item reads, I am organized and a hard-working worker. Well, I am organized, but I am not a hard-worker. I guess that I will mark strongly disagree." This would indicate that the item has concerns. On the other hand, probes are prompts given to participants about the items. Beatty (2004) identified several types of probes, including re-orienting (asking for an answer), elaborating (asking for information), cognitive (asking for introspection), confirmatory (asking for confirmation), expansive (asking for elaboration), functional (asking for clarification), and feedback (providing information). Although each probe provides useful information, there seems to be no consensus regarding when to use them. Several authors have suggested, however, that trained interviewers are better at choosing the correct occasion than untrained interviewers (Beatty, 2004; Beatty & Willis, 2007; Presser et al., 2004).

Also, researchers may choose to apply concurrent or retrospective reporting. Proponents of concurrent reporting argue that participants may be unable to remember their thoughts about particular items after the fact, and only information about the overall item list may be accurate (Beatty & Willis, 2007; Willis, 2004). Alternatively, proponents of retrospective reporting argue that responding to prompts alters participants' thought processes while completing the survey, and the social interaction involved with prompting during administration may alter the response process (Beatty & Willis, 2007; Willis, 2004). It appears that more authors recommend the use of retrospective reporting, but it is always strongly recommended that researchers understand the benefits and detriments of each approach before performing a cognitive interview.

Researchers also need to determine how to analyze cognitive interview results. It is difficult to determine whether a participant interpreted an item correctly or whether they "missed the mark" altogether. Likewise, it is difficult to determine whether a participant struggled "too much," but it is up to the researcher to draw these lines. Resources exist to determine coding guidelines (Beatty, 2004; Willis, 2004), but no "hard and fast" rules exist.

Lastly, researchers must choose whether to use non-essential coding methods. Two of the most popular are behavior coding and response latency. Behavior coding involves coding the reports and/or behavior of participants and interviewers, such as whether an item was read incorrectly (Van der Zouwen & Smit, 2004). Items with many atypical behaviors should be removed. Response latency involves recording the time it takes to answer a question (Bassili & Scott, 1996; Draisma & Dijkstra, 2004). Items with longer latencies are believed to perform poorly, and they should be removed. Both methods need further research before they can be applied reliably (Beatty, 2004; Beatty & Willis, 2007; Presser et al., 2004; Willis, 2004).

While cognitive interviews are widely used, prior studies have discovered some concerns. DeMaio and Landreth (2004) showed that cognitive interviews vary greatly, and cognitive interviews performed by two separate organizations may produce very different results. Even when the same cognitive interviewing techniques are used, inter-rater agreement is often low (Conrad & Blair, 2004; DeMaio & Landreth, 2004; Presser & Blair, 1994). Likewise, little research has compared the utility of multiple qualitative pretest methods. Other less-cumbersome pretests may identify poor items at a similar, or even better, rate than cognitive interviewing.

Focus Groups

Focus groups are used to gather a wide range of experiences from several diverse participants. Often, focus groups are used during the item generation phase to produce items from multiple perspectives and ensure that the entire content domain of a construct is gauged (Brod et al., 2009; Sweeney & Soutar, 2001). The method can also be used immediately after the item generation phase to ensure that the items are free from wording concerns and represent the focal construct (DeVellis, 2016; Lynn, 1986; Kim et al., 1999). To perform a focus group, participants are gathered at a common location and provided the over-representative item list (Morgan, 1996). Then, they are asked to provide feedback regarding each item. They either provide the feedback as a group, individually, or a combination of both.

Focus groups may differ by the type of feedback elicited. Kim and colleagues (1999) performed a focus group that consisted of three phases: review the items for (1) grammatical accuracy and readability, (2) construct accuracy, (3) and construct deficiency. Many other authors have used focus groups that include a combination of these same phases, most commonly the first and second phases (Rosen et al., 2004; Yang et al., 2004). The second phase, gauging construct accuracy, requests participants to provide feedback about the ability of each item to gauge the focal construct, which largely forces them to judge the face validity of each item. While qualitative methods do not provide a numerical metric to rank items' face validity, focus groups still allow this aspect of validity to be included in item retention decisions.

Further, sample size suggestions vary, but all authors suggest that researchers should conduct focus groups until a saturation point is reached (Kim et al., 1999; Yang et al., 2004). That is, the focus groups fail to provide novel information. Brod and colleagues (2009) suggest creating a list of novel information generated after each focus group and to stop the data collection process when the list from a focus group is notably smaller. Brod and colleagues (2009) also note that this often occurs after three or four focus groups of four to six participants.

While focus groups have several benefits, they also pose unique concerns. Participants in a focus group may feel unable to provide certain feedback, or they may even have their perceptions changed by others' feedback (Brod et al., 2009; Greenbaum, 2000; Kitzinger, 1995). Prior authors have also supported that participants in focus groups may provide more extreme responses than they normally would otherwise (Brod et al., 2009; Morgan, 1996). Focus groups also require multiple participants to gather together in a common location, and it may be almost impossible to gather participants from certain populations. Thus, while focus groups can provide important information, they may be more difficult to perform than other pretest methods.

Traditional Interviews

While focus groups can provide information regarding a wide range of experiences, interviews are typically able to provide more in-depth information (Brod et al., 2009; Greenbaum, 2000; Kitzinger, 1995). Some authors have also suggested that participants are more willing to provide honest feedback in interviews compared to focus groups, as they may feel less pressure from others to provide certain responses (Morgan, 1996). When performing a traditional interview, participants read the item list and provide feedback on each item. Items that are consistently identified as concerning are removed. Thus, this method can provide similar information as focus groups without needing to gather participants together.

Like most other qualitative pretest methods, it is still unclear whether this design can provide accurate feedback, and little research has investigated the ability of traditional interviews to identify problematic items. Also, many researchers include traditional interviews to reduce item lists, but these researchers rarely report applications of this method as a full study (Ferris et al., 2008; Howard et al., 2016). Instead, it is usually presented as a single sentence or paragraph after the item generation phase. This insinuates that researchers may not perceive this approach as important to the scale development process. Nevertheless, traditional interviews may provide important information regarding the items, and this method should be applied and studied.

Other Qualitative Pretest Methods

Other qualitative pretest methods exist aside from cognitive interviewing, focus groups, and traditional interviews. These methods have seen little discussion, and much is still unknown regarding their validity. One of these methods is free response prompts, which are brief questions such as "Did you find this item confusing? If so, why?" Participants are provided the item list and asked to respond to the prompt after each item. Items with several participant responses are removed. A benefit of free response prompts is their ease to administer, and they can be included in an online survey. It is still unclear, however, whether participants can accurately provide feedback regarding each item without using more intensive methods, such as cognitive interviewing.

Also, some researchers have used qualitative participant observations to directly ensure the face validity of each item (Brod et al., 2009). In these instances, researchers observe the behaviors of target participants to ensure that each item represents an observed behavior. Most often, participant observations are performed when participants are unable to provide the intensive self-reports required in cognitive interviews, focus groups, traditional interviews, and other qualitative methods. Beyond these, few other qualitative methods can be seen in research.

Discussion

Several aspects of scale pretests should be apparent from the above review (Table 1). Most notably, (1) an array of pretest methods exist, (2) these pretests may achieve various goals, (3) much remains unknown about these pretests, (4) and more research is needed to understand their similarities, differences, benefits, and detriments. With this in mind, the following presents eight research questions to guide the future study and application of scale pretesting methods.

Future Research Questions

1) Which method provides the best results?

Researchers always want to apply the best method possible, and it is natural to want a single pretest method that is best across all situations. Unfortunately, current pretest methods cannot provide this solution. Each method has particular strengths and weaknesses, and they should be applied when the research situation is suitable. Thus, researchers should not ask "which method provides the best results?" but rather "when should each method be used?"

2) When should each method be used?

To determine which method to use, a researcher should first determine whether they are most concerned with (a) face validity or (b) wording issues and (somewhat) face validity. If the former is the primary concern, a quantitative pretest method should be applied. If the latter is the primary concern, then a qualitative pretest method should be applied.

If a quantitative pretest method is chosen, then the researcher also needs to determine whether they are interested in items' relationship with (a) the focal construct alone or (b) the focal construct and other constructs. If the researcher is only interested in the focal construct, then item-rating tasks are ideal. If the researcher is interested in the focal construct and other constructs, then they should use either an item-sort task or Hinkin and Tracey's (1999) ANOVA method. Because current research has not directly compared these two methods to determine which provides more accurate results, the researcher can choose whichever of these two methods that they prefer. It should be kept in mind, however, that research has yet to show that Hinkin and Tracey's (1999) ANOVA method performs well with sample sizes typical of pretests.

Table 1. Summary of Scale Pretest Method Attributes

	Item-Rating Task	Item-Sort Task	ANOVA Method	Cognitive Interviews	Focus Group	Interviews
1.) Identify items that gauge focal construct?	Yes	Yes	Yes	No	Somewhat[a]	Somewhat[a]
2.) Identify items that gauge multiple constructs?	No	Yes	Yes	No	Somewhat[a]	Somewhat[a]
3.) Identify items with wording concerns?	No	No	No	Yes	Yes	Yes
4.) Identify confusing items?	No	No	No	Yes	Yes	Yes
5.) Able to be administered via online survey?	Yes	Yes	Yes	No	No	No
6.) Typically use SMEs?	Yes	Yes	Yes	No	Yes	Yes
7.) Typically use group settings to collect data?	No	No	No	No	Yes	No
8.) Typical Sample Size?	10 - 30	5 - 30	30 - 150	3 - 6	3 - 4 Groups of 5 - 6 People	3 - 6

[a]Focus groups and interviews can obtain some indicators of face validity, but not in a manner that the items can rank-sorted on these attributes.

If a qualitative pretest method is chosen, the researcher needs to determine whether they are concerned with (a) wording issues alone (b) or wording issues and face validity. If wording issues are the primary concern, then cognitive interviewing is ideal. If both wording issues and face validity are concerns, then it should be determined whether the larger concern is (a) the breadth of responses (b) or the depth of responses. If the breadth of responses is the concern, then focus groups should be used. If the depth of responses is the concern, then traditional interviews should be used. Nevertheless, it should be kept in mind that cognitive interviewing has the most empirical support for its validity, although these prior results are mixed. To aid in future scale pretesting decisions, Figure 1 is included as a visual guide.

3) Which methods can be effectively used in conjunction?

Researchers almost always apply a single pretest method when developing measures. Applying two or more methods from the same category (quantitative or qualitative) could benefit the development of scales, as it could provide a triangulation of results (Jick, 1979; Morse, 1991). More importantly, applying two or more methods from different categories could identify attributes of items that could not be discovered with one category alone, and applying both a quantitative and qualitative pretest method could address the weaknesses of the other. Perhaps the pretest methods that would provide the most utility, in regard to difficulty to implement and information obtained, would be the application of any quantitative pretest method and free response blanks. While free response blanks are only sparsely used, they are among the very few qualitative pretest methods that can be administered through an online survey. When applying the discussed quantitative pretest methods, the free response blank can be placed after the numerical rating for each item. A visual demonstration of this is provided in the supplemental material, in which free response blanks are applied alongside an item-sort task.

Table 2. Summary of Eight Questions, Answers, and Directions for Future Research

Question	Answer	More research is needed on…
Which method provides the best results?	None, quantitative and qualitative methods have different goals.	Which methods with similar goals provides more accurate results. For instance, do item-sort tasks or the ANOVA method provide more accurate results?
When should each method be used?	In general, quantitative methods should be used when face validity is a concern, whereas qualitative methods should be used when wording issues (and perhaps face validity) are a concern. Further decisions vary on the context.	
Which methods can be effectively used in conjunction?	Using a qualitative and quantitative methods in conjunction appears to be ideal. Also, using methods that use general participants and SMEs together may be ideal.	Which methods can perform well together. For instance, should focus groups or traditional interviews be used with item-sort tasks?
Are SMEs required for certain methods?	Perhaps, but many methods that traditionally use SMEs may not need to do so.	Whether SMEs provide more accurate results than general participants.
What is the required sample size for these methods?	The bottom-range for moth methods has yet to be identified, but 30 should be sufficient for most methods.	Whether prior sample size recommendations are supported by empirical and statistical research.
Must scale pretesting methods always precede traditional psychometric evaluations?	Not always.	Which methods should can used with and without follow-up evaluations.
What are some concerns with existing pretest methods?	Identifying repetitive items, removing repetitive items, and considering other types of validity	The creation of new and modification of old pretest methods and to address these concerns.
What is the future of scale pretesting?	The application of scale pretests will continue to thrive, and the study of the methods themselves will increase.	Empirically testing the accuracy of existing pretest methods and the creation of new methods that address old concerns.

4) Are SMEs required for certain methods?

Before discussing which methods require SMEs, another question should be asked first:

What exactly are SMEs in the context of scale pretests? Typically, SMEs are those with relevant academic experience. For instance, a researcher creating a conscientiousness scale may use graduate students or graduates of Ph.D. programs in Psychology. Many authors have also used undergraduates but noted that these SMEs were current or prior students of a relevant course and/or research lab. It is interesting to note, however, that researchers less frequently use target populations as SMEs for scale pretests, although they are regularly considered SMEs for the item generation phase. This may be because these SMEs are believed to have relevant knowledge of the behaviors that may compose the criterion space for a construct, but they are unable to identify the exact boundaries of a construct. Like most other aspects of scale pretests, it is unclear whether this notion is actually true without supporting research.

Further, when using quantitative pretest methods, the decision to use general participants or SMEs is often unclear. For item-rating methods and item-sort tasks, authors almost always use SMEs; however, neither Anderson and Gerbing (1991) or Howard and Melloy (2016) used SMEs in their empirical studies on item-sort tasks, and little research has empirically shown that SMEs provide more accurate judgements. Further, Hinkin and Tracey (1999) used graduate and undergraduate students to test their ANOVA method, but these students were not specified to be in classes relevant to the item-lists. Thus, it is unclear whether any quantitative method explicitly requires SMEs. When using general participants, provided construct definitions need to be clear and comprehensive, as this information may be their only exposure to certain constructs.

Regarding qualitative pretest methods, cognitive interviews are almost always performed with general participants. If SMEs were used to complete the item list, their prior knowledge of the focal construct may alter their responses. Alternatively, focus groups and interviews may or may not require SMEs. Most research has used SMEs to identify wording issues and construct contamination, but some authors have used target populations relevant to the focal construct. For instance, people with health conditions have been used as SMEs when creating a scale for severity of symptoms (Mangione et al., 2001; Olson, 2010). Like quantitative pretest methods, research has yet to show that SMEs provide more accurate results than general participants.

5) What is the required sample size for these methods?

The recommended sample sizes for the various pretest methods are more direct than the decision to use SMEs. Typically, 10 to 30 participants are recommended for item-rating tasks and item-sort tasks, although Howard and Melloy (2016) showed that statistical significance can be calculated with sample sizes of five for item-sort tasks. Hinkin and Tracey (1999) suggest that sample sizes as small as 30 can be used for their ANOVA method, but their examples included samples larger than 150. For qualitative methods, prior researchers have suggested that three to six participants may provide accurate results for cognitive interviews and traditional interviews. For focus groups, Brod and colleagues (2009) suggested that three or four focus groups of four to six participants can provide accurate results. Aside from item-sort tasks (Howard & Melloy, 2016), however, prior research has not provided empirical or statistical evidence for these sample size cutoffs. Instead, these findings are largely based on conjecture and prior experience.

6) Must scale pretesting methods always precede traditional psychometric evaluations?

Scale pretests almost always precede traditional psychometric evaluations, such as EFA and CFA, and many researchers may believe that scale pretests are useless without such follow-up investigations. The origin of this belief may have arisen from prior empirical studies on the ability of quantitative pretest methods to predict the results of EFA and CFA (Anderson & Gerbing, 1991; Howard & Melloy, 2016), and suggestions that quantitative pretest methods are able to identify items that perform well in an EFA or CFA. This tradition should be reconsidered.

Of course, the entire scale development process has several steps, and each should be followed to ensure a psychometrically sound scale that is valid for gauging the focal construct (Hinkin 1995, 1998). Researchers are often unable to undergo the entire scale development process due to limited time and/or resources. In these instances, scale pretests can provide valuable information even in the absence of follow-up analyses. In other words, providing some reassurance that administered items are adequate is better than providing no such evidence. We strongly suggest that future researchers should apply these discussed methods in these instances, which is only seen sparingly in current research (Howard & Melloy, 2016; Olson, 2010), and they should apply both a qualitative and quantitative pretest method when doing so.

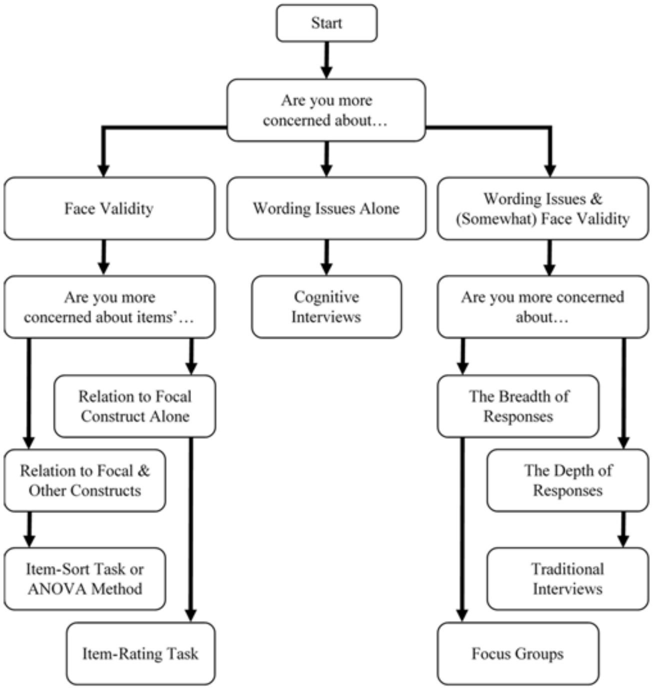
Start
Are you more concerned about…
Face Validity
Wording Issues Alone
Wording Issues & (Somewhat) Face Validity
Are you more concerned about items'…
Cognitive Interviews
Are you more concerned about…
Relation to Focal Construct Alone
The Breadth of Responses
Relation to Focal & Other Constructs
The Depth of Responses
Item-Sort Task or ANOVA Method
Traditional Interviews
Item-Rating Task
Focus Groups

Figure 1. Flowchart of Scale Pretest Applications

7) What are some concerns with existing pretest methods?

In general, quantitative pretest methods select items that are judged to be representative of the focal construct, and items that more accurately gauge the focal construct are more likely to be retained. Selecting the most accurate items may reduce content coverage, however, and only items that are closely-related synonyms may be retained. Similarly, qualitative pretest methods primarily select items that are free from wording concerns, but participants may also judge the face validity of each item during a focus group or traditional interview. It is again possible that participants may perceive certain items as being irrelevant that actually gauge important aspects of the focal construct, thereby reducing the content coverage of the item list. We suggest that researchers should apply methods and cutoffs that retain more items than needed when pretest methods are used with subsequent psychometric analysis. This would help ensure the content validity of the measure, and these items can also be further reduced in subsequent steps.

Further, overly repetitive items may pose other concerns aside from content validity issues. These items provide little information regarding the focal construct when included in the same scale, and they may also negatively influence model fit when performing a CFA (Brown, 2015; Fabrigar et al., 1999). Unfortunately, none of the discussed quantitative or qualitative methods are regularly used to identify repetitive items; however, focus groups and traditional interviews may achieve this objective – if a phase is added to specifically identify repetitive items. For this reason, it may be useful for researchers to more often apply focus groups and traditional interviews with these phases for their scale pretesting.

8) What is the future of scale pretesting?

Scale pretest methods provide valuable information, and researchers are increasingly recognizing their benefits. For these reasons, we believe that the application of scale pretests will continue, but three new directions will be seen. First, the application of pretest methods will continue in a more systematic manner. With the continued usage, authors will begin to recognize situations in which these methods are best applied, and more best practices will begin to emerge.

Second, more research will analyze the characteristics of scale pretests themselves. For instance, several pretests have similar objectives that are achieved in a similar manner, but it is unclear which of these pretests perform better. Likewise, future research should perform more detailed investigations into the manner that scale pretests retain items, such as whether quantitative methods actually have concerns with retaining repetitive items, and whether SMEs actually provide more accurate results for pretest methods. Similarly, future research should determine when the applications of these methods are most appropriate. While the current article suggested applying quantitative and qualitative pretest methods together, certain pretest methods may perform particularly well together. Certain methods may also perform poorly in the absence of subsequent psychometric evaluation, but these methods cannot be identified without further research. Together, these suggestions are only the beginning of further pretest investigations.

Third, discussions of pretest methods focus on their relation to face validity and ability to replicate EFA and

CFA results, but it is important to consider each type of validity together. Face validity is interlinked with content, convergent, discriminant, and other types of validity. We suggest that new pretest methods should analyze multiple aspects of validity together.

References

Anderson, J. C., & Gerbing, D. W. (1991). Predicting the performance of measures in a confirmatory factor analysis with a pretest assessment of their substantive validities. *Journal of Applied Psychology*, 76(5), 732-740.

Bassili, J. N., & Scott, B. S. (1996). Response latency as a signal to question problems in survey research. *Public Opinion Quarterly*, 60(3), 390-399.

Beatty, P. (2004). The dynamics of cognitive interviewing. In S. Presser, J. M. Rothger, M. P. Couper, J. T. Lessler, E. Martin, J. Martin, & E. Singer (Eds.), Methods for testing and evaluating survey questionnaires (pp. 45-66). Hoboken, NJ: John Wiley & Sons, Inc.

Beatty, P. C., & Willis, G. B. (2007). Research synthesis: The practice of cognitive interviewing. *Public Opinion Quarterly*, 71(2), 287-311.

Belson, W. A. (1981). The design and understanding of survey questions. Aldershot, UK: Gower.

Blair, J., Czaja, R. F., & Blair, E. A. (2013). Designing surveys: A guide to decisions and procedures. Tyne, UK: Sage.

Bosco, F. A., Aguinis, H., Singh, K., Field, J. G., & Pierce, C. A. (2015). Correlational effect size benchmarks. *Journal of Applied Psychology*, 100(2), 431-449.

Brod, M., Tesler, L., & Christensen, T. (2009). Qualitative research and content validity: developing best practices based on science and experience. *Quality of Life Research*, 18(9), 1263-1278.

Brown, T. A. (2015). Confirmatory factor analysis for applied research. New York, NY: Guilford Publications.

Brown, J. D. (1996). Testing in language programs. Upper Saddle River, NJ: Prentice Hall Regents.

Cantril, H. (1944). The meaning of questions. In Gauging public opinion (pp. 3-22). Princeton, NJ: Princeton University Press.

Clark, L. A., & Watson, D. (1995). Constructing validity: Basic issues in objective scale development. *Psychological Assessment*, 7(3), 309-319.

Cohen, J. (1992). A power primer. *Psychological Bulletin*, 112(1), 155-159.

Cohen, J. (1994). The earth is round (p < .05). *American Psychologist*, 49(12), 997-1003.

Conrad, F., & Blair, J. (2004). Data quality in cognitive interviews: The case of verbal reports. In S. Presser, J. M. Rothger, M. P. Couper, J. T. Lessler, E. Martin, J. Martin, & E. Singer (Eds.), Methods for testing and evaluating survey questionnaires (pp. 67-87). Hoboken, NJ: John Wiley & Sons, Inc.

Coste, J., Guillemin, F., Pouchot, J., & Fermanian, J. (1997). Methodological approaches to shortening composite measurement scales. *Journal of Clinical Epidemiology*, 50(3), 247-252.

DeMaio, T. J., & Landreth, A. (2004). Do different cognitive interview techniques produce different results? In S. Presser, J. M. Rothger, M. P. Couper, J. T. Lessler, E. Martin, J. Martin, & E. Singer (Eds.), Methods for testing and evaluating survey questionnaires (pp. 89-108). Hoboken, NJ: John Wiley & Sons, Inc.

DeVellis, R. (2016). Scale development: Theory and applications (Vol. 26). Tyne, UK: Sage.

Dietrich, H., & Ehrlenspiel, F. (2010). Cognitive interviewing: A qualitative tool for improving questionnaires in sport science. *Measurement in Physical Education and Exercise Science*, 14(1), 51-60.

Draisma, S., & Dijkstra, W. (2004). Response latency and (para) linguistic expressions as indicators of response error. In S. Presser, J. M. Rothger, M. P. Couper, J. T. Lessler, E. Martin, J. Martin, & E. Singer (Eds.), Methods for testing and evaluating survey questionnaires (pp. 131-147). Hoboken, NJ: John Wiley & Sons, Inc.

Fabrigar, L. R., Wegener, D. T., MacCallum, R. C., & Strahan, E. (1999). Evaluating the use of exploratory factor analysis in psychological research. *Psychological Methods*, 4(3), 272-299.

Ferris, D. L., Brown, D. J., Berry, J. W., & Lian, H. (2008). The development and validation of the Workplace Ostracism Scale. *Journal of Applied Psychology*, 93(6), 1348-1366.

Fowler Jr, F. J. (2013). Survey research methods. Tyne, UK: Sage.

Goetz, C., Coste, J., Lemetayer, F., Rat, A. C., Montel, S., Recchia, S., ... & Guillemin, F. (2013). Item reduction based on rigorous methodological guidelines is necessary to maintain validity when shortening composite measurement scales. *Journal of Clinical Epidemiology*, 66(7), 710-718.

Greenbaum, T. L. (2000). Moderating focus groups: A practical guide for group facilitation. Tyne, UK: Sage.

Hardesty, D. M., & Bearden, W. O. (2004). The use of expert judges in scale development: Implications for improving face validity of measures of unobservable constructs. *Journal of Business Research*, 57(2), 98-107.

Heene, M., Bollmann, S., & Bühner, M. (2014). Much ado about nothing, or much to do about something? Effects of scale shortening on criterion validity and mean differences. *Journal of Individual Differences*, 35(4), 245-249.

Hinkin, T. R. (1995). A review of scale development practices in the study of organizations. *Journal of Management*, 21(5), 967-988.

Hinkin, T. R. (1998). A brief tutorial on the development of measures for use in survey questionnaires. *Organizational Research Methods*, 1(1), 104-121.

Hinkin, T. R., & Tracey, J. B. (1999). An analysis of variance approach to content validation. *Organizational Research Methods*, 2(2), 175-186.

Howard, M. C. (2016). A review of exploratory factor analysis decisions and overview of current practices: What we are Doing and how can we improve? International *Journal of Human-Computer Interaction*, 32(1), 51-62.

Howard, M. C., & Melloy, R. C. (2016). Evaluating item-sort task methods: The presentation of a new statistical significance formula and methodological best practices. *Journal of Business and Psychology*, 31(1), 173-186.

Hunt, S. D., Sparkman Jr, R. D., & Wilcox, J. B. (1982). The pretest in survey research: Issues and preliminary findings. *Journal of Marketing Research*, 19(2), 269-273.

Jick, T. D. (1979). Mixing qualitative and quantitative methods: Triangulation in action. *Administrative Science Quarterly*, 24(4), 602-611.

Kim, B. S., Atkinson, D. R., & Yang, P. H. (1999). The Asian Values Scale: Development, factor analysis, validation, and reliability. *Journal of Counseling Psychology*, 46(3), 342.

Kitzinger, J. (1995). Qualitative research. Introducing focus groups. BMJ: *British Medical Journal*, 311(7000), 299.

Lawshe, C. H. (1975). A quantitative approach to content validity. *Personnel Psychology*, 28(4), 563-575.

Leech, B. L. (2002). Asking questions: techniques for semistructured interviews. *Political Science & Politics*, 35(4), 665-668.

Lynn, M. R. (1986). Determination and quantification of content validity. *Nursing Research*, 35(6), 382-386.

MacKenzie, S. B., Podsakoff, P. M., & Podsakoff, N. P. (2011). Construct measurement and validation procedures in MIS and behavioral research: Integrating new and existing techniques. *MIS Quarterly*, 35(2), 293-334.

Mangione, C. M., Lee, P. P., Gutierrez, P. R., Spritzer, K., Berry, S., & Hays, R. D. (2001). Development of the 25-list-item national eye institute visual function questionnaire. *Archives of Ophthalmology*, 119(7), 1050-1058.

Meade, A. W., & Craig, S. B. (2012). Identifying careless responses in survey data. *Psychological Methods*, 17(3), 437-455.

Morgan, D. L. (1996). Focus groups as qualitative research (Vol. 16). Tyne, UK: Sage.

Morse, J. M. (1991). Approaches to qualitative-quantitative methodological triangulation. *Nursing Research*, 40(2), 120-123.

Nakagawa, S., & Cuthill, I. C. (2007). Effect size, confidence interval and statistical significance: a practical guide for biologists. *Biological Reviews*, 82(4), 591-605.

Olson, K. (2010). An examination of questionnaire evaluation by expert reviewers. *Field Methods*, 22(4), 295-318.

Presser, S., & Blair, J. (1994). Survey pretesting: Do different methods produce different results. *Sociological Methodology*, 24(1), 73-104.

Presser, S., Couper, M., Lessler, J., Martin, E., Martin, J., Rothgeb, J., & Singer, E. (2004). Methods for testing and evaluating survey questions. *Public Opinion Quarterly*, 68(1), 109-130.

Rea, L. M., & Parker, R. A. (2014). Designing and conducting survey research: A comprehensive guide. Hoboken, NJ: John Wiley & Sons.

Rosen, R. C., Catania, J., Pollack, L., Althof, S., O'Leary, M., & Seftel, A. D. (2004). Male Sexual Health Questionnaire (MSHQ): Scale development and psychometric validation. *Urology*, 64(4), 777-782.

Schriesheim, C. A., Powers, K. J., Scandura, T. A., Gardiner, C. C., & Lankau, M. J. (1993). Improving construct measurement in management research: Comments and a quantitative approach for assessing the theoretical content adequacy of paper-and-pencil survey-type instruments. *Journal of Management*, 19(2), 385-417.

Smith, G. T., McCarthy, D. M., & Anderson, K. G. (2000). On the sins of short-form development. *Psychological Assessment*, 12(1), 102-111.

Sweeney, J. C., & Soutar, G. N. (2001). Consumer perceived value: The development of a multiple item scale. *Journal of Retailing*, 77(2), 203-220.

Thompson, B. (2004). Exploratory and confirmatory factor analysis: Understanding concepts and applications. Washington DC: American Psychological Association.

Van der Zouwen, J., & Smit, J. H. (2004). Evaluating survey questions by analyzing patterns of behavior codes and question–answer sequences: A diagnostic approach. In S. Presser, J. M. Rothger, M. P. Couper, J. T. Lessler, E. Martin, J. Martin, & E. Singer (Eds.), Methods for testing and evaluating survey questionnaires (pp. 109-130). Hoboken, NJ: John Wiley & Sons, Inc.

Willis, G. B. (2004). Cognitive interviewing: A tool for improving questionnaire design. Tyne, UK: Sage.

Yang, Z., Jun, M., & Peterson, R. T. (2004). Measuring customer perceived online service quality: scale development and managerial implications. International *Journal of Operations & Production Management*, 24(11), 1149-1174.

Supplemental Material – Item Sort Task with Free Response Blank Example

Instructions: In the following, you will be asked to indicate which construct that you believe several items represent from the options provided. For this reason, it is very important that you are familiar with the constructs of interest. Please read the following definitions to familiarize yourself with these constructs. Afterwards, using the options provided, please indicate the construct that you believe the following items represent. If you believe that the item does not represent any of the options provided, please mark "Other Construct."

Conscientiousness - A fundamental trait that influences whether people adhere to long-range goals, avoid acting impulsively, act carefully in their behaviors, desire performing well, and remain committed to social obligations.

Extraversion – A fundamental trait that influences whether people are outgoing, talkative, social, seek new sensations, and receive gratification outside of oneself.

Neuroticism – A fundamental trait that influences whether people are moody, experience negative emotions, and response more negatively to stressors.

Lastly, a final column is added that is labeled "Confusing / Wording Concerns." If you believe that the item is confusing or possesses any wording concerns, please write a brief note describing the concerns.

	Conscientiousness	Extraversion	Neuroticism	Other Construct	Confusing / Wording Concerns?
1.) I am talkative.					
2.) I am hard working.					
3.) I am emotionally stable.					
4.) I enjoy running.					
5.) I like to be orderly.					
…	…	…	…	…	…

Citation:

Howard, Matt C. (2018). Scale Pretesting. *Practical Assessment, Research & Evaluation*, 23(5). Available online: http://pareonline.net/getvn.asp?v=23&n=5

Corresponding Author

Matt C. Howard
Assistant Professor
Marketing and Quantitative Methods
Mitchell College of Business
University of South Alabama

email: mhoward [at] southalabama.edu