J Bus Psychol
DOI 10.1007/s10869-015-9404-y

Evaluating Item-Sort Task Methods: The Presentation of a New Statistical Significance Formula and Methodological Best Practices

Matt C. Howard[1] · **Robert C. Melloy**[1]

Abstract
Purpose An item-sort task is a common method to reduce over-representative item lists during the scale-creation process. The current article delineates the limitations and misapplications of the accepted statistical significance formula for item-sort tasks and proposes a new statistical significance formula with greater utility across a wider range of item-sort tasks.
Design First, a simulation study compares the two formulas in an array of conditions that vary on sample size and number of assignment choices. Second, an empirical study compares the results of three separate item-sort tasks across the two formulas for statistical significance.
Findings In the empirical study, the proposed formula produces more correct retention decisions than the existing formula across all three item-sort tasks. In the simulation study, the proposed formula is more appropriate than the existing formula under most conditions. The two formulas function identically in item-sort tasks with only two assignment choices.
Implications Researchers could obtain erroneous results when misapplying the existing item-sort task statistical significance formula to cases with more than two assignment choices. The proposed formula corrects this limitation, ultimately providing accurate results more often than the existing formula. Applying the proposed formula could help future research and practice throughout the scale development process.
Originality Despite widespread use, few attempts have been made to improve scale-creation pretest methods, particularly item-sort tasks. The current study demonstrates that even conventional statistical methods are susceptible to misuse and misapplication, and future research could benefit from the reexamination of other common methods.

Keywords Item-sort tasks · Measurement · Scale creation · Psychometrics

Introduction

When constructing a new scale, researchers often create an over-representative item list in order to gage all aspects of their construct of interest (Hinkin 1995, 1998). Once this list has been established, it must be subsequently reduced (Stanton et al. 2002). Many previous studies do so through administering their items and analyzing the resulting factor structure through exploratory factor analysis (EFA) and/or confirmatory factor analysis (CFA; Dahling et al. 2009; Leach et al. 2008; Spector 1992; Stanton et al. 2002); however, this is not possible for all researchers for several important reasons.

First, many scholars do not have access to the required resources to produce sufficient power for these statistical techniques. Previous authors have recommended minimum sample sizes between 100 and 500, which are dependent upon an array of factors (i.e., communalities, item-to-construct ratios, etc.), to obtain accurate EFA and CFA results (Comrey and Lee 1992; MacCallum et al. 1999). Researchers are consistently bound by a finite number of participants, causing samples of this size to be unavailable. This problem is exacerbated if repeated administrations of the same scale are required, especially if time is also

✉ Matt C. Howard
mch248@psu.edu

[1] Industrial/Organizational Psychology, The Pennsylvania State University, 142 Moore Building, University Park, PA 16802, USA

Published online: 28 March 2015

limited. Second, if the initial item pool is large, it is often difficult to obtain a sample large enough to produce accurate EFA estimations. With a 2:1 participant to item ratio, there is only a 10 % chance of obtaining the correct factor structure for a scale; even with an ideal 20:1 participant to item ratio, there is still only a 70 % chance (Costello and Osbourne 2005). Third, CFA requires firm theoretical predictions regarding items' underlying factor structure. Given that over-representative item lists are exploratory in nature, the assumed theoretical factor structure may not be replicated during the item-reduction phase of the scale development process. The possible disparity between theoretical predictions and empirical results provides difficulties for using CFA as an item-reduction technique. Fourth, these methods of reducing items cannot discover problematic wording. Consequently, items that are confusing, double-barreled, or leading may not be removed, and their biases could remain in the final measure (Hardy and Ford 2014; Hinkin 1995, 1998). These four factors are very problematic for researchers. Therefore, alternative methods to reduce initial scale item pools may be more appropriate.

A variety of qualitative pretest methods exist, such as practice verbal administrations (Aday and Cornelius 2011; Krosnick 1999; Presser and Blair 1994), cognitive interviewing (Conrad and Blair 2004; DeMaio and Landreth 2004; Willis 2005), expert discussion panels (Czaja 1998; Krosnick 1999; Olson 2010), and other proposed methods (Collins 2003; Groves et al. 2013; Presser et al. 2004; Schaeffer and Presser 2003). While these methods are useful in discovering wording-related issues, they cannot discover items that fail to gage the intended construct (for a review, see Presser et al. 2004). Alternatively, several quantitative pretest methods have been created. Many of these have limited use, require sample sizes substantially larger than common pretest methods, and are unable to detect wording-related issues (Hinkin and Tracey 1999; Holden and Jackson 1979). Fortunately, a pretest method exists which can discover wording-related issues, distinguishes items that may not gage the intended construct, requires a sample size comparable to other pretest methods, and has been used in a large amount of scale development studies. This pretest method is called an item-sort task.

Anderson and Gerbing (1991) were the first to propose the use of an item-sort task, drawing from other similar pretest methods (Hemphill and Westie 1950; Lawshe 1975; Rovinelli and Hambleton 1977). To perform an item-sort task, researchers present participants with potential scale items along with a list of constructs, and the participants are meant to assign each item to the construct they believe the item measures. Items with frequent assignments to the posited construct are retained, whereas the others are discarded. Anderson and Gerbing (1991) claimed that, although this method could not give definitive statistical evidence of a scale's underlying factor structure, it could still provide support for each item's substantive validity. Substantive validity is "the extent to which that measure is judged to be reflective of, or theoretically linked to, some construct of interest" (Anderson and Gerbing 1991, p. 732). If an item is shown to have high substantive validity, then it should be retained for further analyses.

Since their original article, studies utilizing item-sort tasks have almost exclusively applied Anderson and Gerbing's (1991) formula (Ferris et al. 2008; Fiore et al. 2013; Lawrence et al. 2007; Linderbaum and Levy 2010; Michel et al. 2014). Currently, Google Scholar reports that Anderson and Gerbing's (1991) original publication has been cited 652 times (November 11, 2014), and many reviews of measure creation methods endorse item-sort tasks using Anderson and Gerbing's (1991) formula and proposed methods (Hardy and Ford 2014; Hinkin 1995, 1998). Further, some authors solely cite these reviews when performing item-sort tasks, perhaps underestimating the true impact of Anderson and Gerbing's (1991) contribution (Hardesty and Bearden 2004). Despite the acceptance of their formula as a conventional pretest tool, certain unintended consequences have emerged through the use of this test of statistical significance. Particularly, Anderson and Gerbing's (1991) formula for an item-sort task's statistical significance does not accept items at a 95 % level of statistical significance when there are more than two assignment choices. Therefore, the current article addresses and challenges the assumption that their formula is always appropriate when testing the results of item-sort tasks.

To begin, we first present a general overview of Anderson and Gerbing's (1991) method and formula and explain how it tests for statistical significance. Second, we note the limitations that restrict the use of Anderson and Gerbing's (1991) original formula and argue that it is being misapplied when testing for statistical significance in item-sort tasks with more than two assignment choices. Then, we propose an alternative formula that can be applied to item-sort tasks comparing two *or more* assignment choices. We apply the proposed formula to both simulation and empirical studies to compare the results to those obtained when using Anderson and Gerbing's (1991) original formula. Finally, we discuss the important statistical and methodological implications of the proposed formula that justify its use to test item-sort tasks.

Anderson and Gerbing's (1991) Methods and Formulas

In their proposed method for item-sort tasks, Anderson and Gerbing (1991) suggest that each item from a researcher's over-representative item list should be presented to a

number of respondents. In contrast to the large sample size requirements of factor analytic procedures, item-sort tasks only require 20 or fewer participants to produce stable substantive-validity estimates (Anderson and Gerbing 1991). For each item, respondents are asked to choose which from a list of constructs the item best represents. Whether using a single alternative or multiple, the choice of constructs should include the posited construct as well as theoretically similar constructs. After the respondents have assigned each item to a construct, a statistical significance test is performed on each item to determine whether an item was assigned to the posited construct more so than an acceptable level of chance. In addition to testing for an item's substantive validity, an item-sort task also allows for respondents to provide qualitative feedback on each item's wording if the researcher includes a free-response blank after each item. This free-response blank allows respondents to identify items that are confusing, double-barreled, or leading. Although this particular method has such advantages, some concerns exist.

After administering an over-representative item list to a set of respondents, Anderson and Gerbing (1991) suggest that researchers should employ a specific formula to determine whether a particular item's results are statistically significant. However, in their seminal paper, the authors developed their formula and illustrated its use with only two constructs—the posited construct and a single alternative. Subsequent researchers have since applied this original formula to item-sort tasks that included longer construct lists, likely recognizing that their created items may inadvertently gage an array of possible alternatives. This misapplication is problematic because although Anderson and Gerbing's (1991) formula is appropriate for the two-construct case they described, it was not intended to be applied to situations with more than two constructs.

To elaborate, Anderson and Gerbing (1991) formula assumes that there is a 0.5 probability that a respondent assigns the item to a posited construct, and a 0.5 probability that the item is assigned to the alternative construct. All choices other than these two are discounted, and it is assumed that they have a zero probability of being assigned. This leaves the following:

$$H_0 : P(a) \leq 0.5$$

$$H_1 : P(a) > 0.5$$

where $P(a)$ denotes the probability that an item is assigned to its posited construct. So, the null hypothesis is retained when the probability that an item is assigned to its posited construct is equal to or less than 0.5, and it is rejected when the probability that an item is assigned to its posited construct is more than 0.5. If the null is rejected, then the item is considered to be representative of the posited construct. Following this assumption, Anderson and Gerbing (1991) give a series of formulas to test for statistical significance.

First, they propose the substantive-validity coefficient, c_{sv}. This coefficient indicates "the extent to which respondents assign an item to its posited construct more than to any other construct" (Anderson and Gerbing 1991, p. 734). It is defined as follows:

$$c_{sv} = \frac{n_c - n_o}{N}$$

where n_c represents the number of respondents assigning a measure to its posited construct, n_o represents the highest number of assignments of the item to any other construct in the set, and N represents the total number of respondents. Then, to receive a critical value for c_{sv}, assuming a 0.05 level of significance, they suggest the following:

$$P(n_c \geq m) < 0.05$$

where n_c is defined as before, and m represents that critical number of assignments. In this equation, m is determined through summing the binomial probabilities (0.5 probability) of a certain number of responses occurring, starting with the maximum possible amount and decreasing, where the sum of these probabilities is still <0.05. For example, if $N = 15$ and r is the number of assignments to the posited construct, the binominal probability for $r = 15$ is 0.0000, for $r = 14$ is 0.0000, for $r = 13$ is 0.0032, for $r = 12$ is 0.0138, and for $r = 11$ is 0.0416. The summation of these is 0.0592. Therefore, the cumulative probability that 11 out of 15 people assign the posited construct exceeds 0.05 (probability $r \geq 11 = 0.0592$), and the cumulative probability that 12 out of 15 people assign the posited construct is below 0.05 (probability $r \geq 12 = 0.0176$). This leaves m, the critical number of assignments, as 12 when there are 15 respondents. These calculations can easily be calculated on several openly available, online binominal distribution probability calculators.

With m obtained, the critical value of c_{sv}, denoted as $\bar{c}_{sv}$, is determined as follows:

$$\bar{c}_{sv} = \frac{m - (N - m)}{N} = \frac{2m}{N} - 1$$

where m and N are defined as before. Using the previous example ($m = 12$ and $N = 15$), $\bar{c}_{sv}$, would be 0.6. To determine whether a particular result is significant, c_{sv} should then be compared to $\bar{c}_{sv}$. If an item's c_{sv} value is equal or greater than the $\bar{c}_{sv}$, then it should be retained for further analysis. This concludes Anderson and Gerbing's (1991) equations; however, the following example may clarify it further.

an item-sort task is taken that was given to fifteen respondents. For each item, there are four possible assignments (i.e., constructs). A particular item was assigned

eleven times to the posited construct, three times to alternative construct A, one time to alternative construct B, and zero times to alternative construct C. The c_{sv} value would be found as follows:

$$c_{sv} = \frac{11 - 3}{15} = 0.5\bar{3}$$

It should be noted that construct A was used for n_o, and the alternative constructs B and C were not used. Then, c_{sv} $(0.5\bar{3})$ should be compared to the calculated $\bar{c}_{sv}$ value of 0.6. In this example, the item would not be significant. At this point, the item would likely be discarded from further use. While this formula has been used in a large number of studies (Ferris et al. 2008; Fiore et al. 2013; Lawrence et al. 2007; Linderbaum and Levy 2010; Michel et al. 2014), there are concerns toward its use.

Limitations to Applying Anderson and Gerbing's (1991) Formula

The same item-sort task scenario is imagined as before; however, now the item was assigned eleven times to the posited construct, two times to construct A, two times to alternative construct B, and zero times to alternative construct C. Now, c_{sv} is $\frac{(11-2)}{15} = 0.6$, which meets the critical value and would be considered statistically significant. In this example, the item would be considered to be representative of the posited construct, and the item would be used in further studies. But should it? Both examples have the same number of individuals responding that the item measures the intended construct and an identical number of responses indicating that it taps into alternative constructs. Each example's item appears to be equally representative of the posited construct, as displayed in Fig. 1. Despite both examples' similarities, only one of the examples deems the item usable for future studies.

These disparate results are a primary concern when applying Anderson and Gerbing's (1991) formula to situations in which more than two constructs may be selected. Although the posited construct had the same number of responses in each example, the distribution of the item to alternative constructs causes the item to become significant. This resulted in a larger number of responses being discounted, although they indicate that various other constructs are being measured within the item. Although this contamination is not due to a single alternative construct but rather a group of constructs, it can still bias participant responses to items. The application of Anderson and Gerbing's (1991) formula to discriminate from more than one alternative construct is a shortcoming of current item-sort task methodology. Researchers applying this method are assuming equal probability that an individual assigns the item to the posited construct and that just one other construct in a set of constructs receives assignments. This leaves all other alternatives with a zero probability. Obviously, with respondents assigning the item to other constructs, their probability is non-zero. The use of Anderson and Gerbing's (1991) formula to cases with greater than two constructs does not account for these other alternatives.

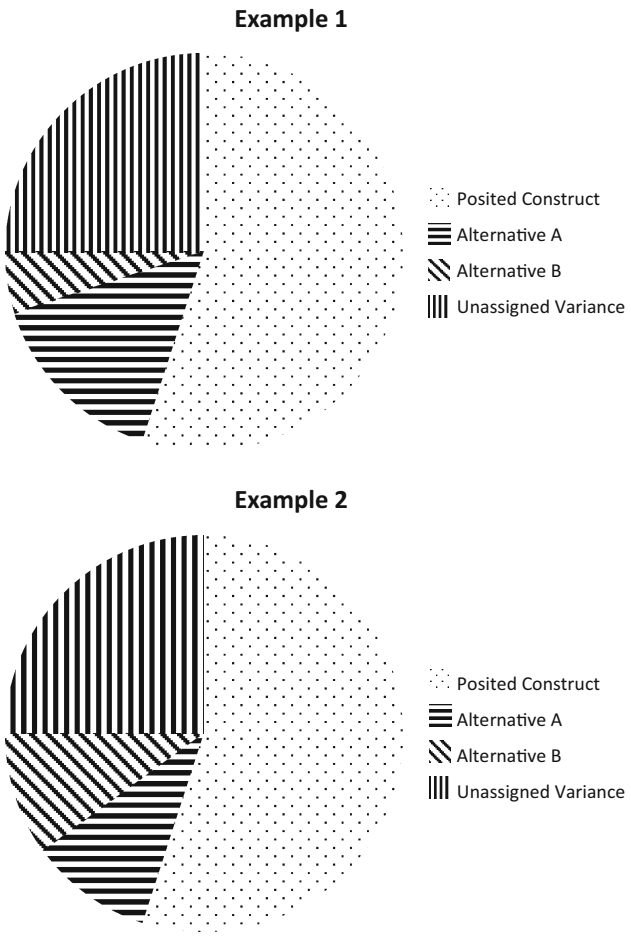

Fig. 1 Visual representation of item-sort task. Using Anderson and Gerbing's (1991) formula, Example 1 is not statistically significant while Example 2 is statistically significant. Using the proposed formula, both examples are not statistically significant

The exclusion of alternative constructs also creates some problematic statistical conundrums, given certain conditions. The following example is taken: an item-sort task with $k = 7$, where k is the number of constructs. For a particular item, a 40 % chance to assign the item to the posited construct is assumed and a 10 % chance for each alternative construct. With these conditions, the item should not be found significant given the null hypothesis that Anderson and Gerbing (1991) present; however, using Anderson and Gerbing's (1991) formula, this item can be statistically significant. With the probabilities given and a sample size of 40, the item will most likely receive 16 assignments to the posited construct and four assignments

to each other construct. Skipping the arithmetic, this will result in a c_{sv} of 0.3, which matches the $\bar{c}_{sv}$ of 0.3. With these conditions, the item is determined to be statistically significant under the most probable conditions (let alone with a 0.95 confidence level), although the predefined conditions mandate that it should not be statistically significant. Therefore, the alternative constructs' probabilities should not be assumed to be zero, as the assumption results in undesirable statistical properties and leaves Anderson and Gerbing's (1991) formula overly reliant on the distribution of assignments to alternative constructs.

Proposed Formula

Given the examples above, it seems that Anderson and Gerbing's (1991) formula can be too lenient in certain situations considering the desired a priori assumptions. These problematic results are partly due to the fact that the formula is misapplied to situations where there are more than two constructs. That is, constructs aside from the both primary and one alternative construct are not included within the formula for statistical significance. Fortunately, a slight modification to Anderson and Gerbing's (1991) original formula may provide a solution. Assume that there is a 0.5 probability that an individual assigns the item to the posited construct, and a 0.5 probability that *any* other construct is assigned. While this may seem to be a small change, it is important because it accounts for the probability of *all* choices being chosen, instead of only accounting for the probability of one alternative choice. A small note should be made, however. The choice of using a 0.5 probability is completely arbitrary. This number was chosen because of its previous use, but there is no reason that an alternative probability cannot be chosen. If a researcher wishes to be more conservative with their item assignments, then they could raise their pre-determined probabilities of assigning the posited construct to an item. However, as with all other statistical tests, choices regarding hypotheses should be chosen a priori. With the chosen probability of 0.5, this leaves the following equation:

$$H_0 : P(a) \leq 0.5$$

$$H_1 : P(a) > 0.5$$

where $P(a)$ is the same as before. To receive a critical value, assuming a 0.05 level of significance, the following equation is used:

$$P(n_c \geq m) < 0.05$$

where n_c and m are defined as before. In this equation, n_c is the number of respondents assigning an item to its posited construct, and m is determined through summing the binomial probabilities (0.5 probability) of a certain number of responses occurring, starting with the maximum possible amount and decreasing, where the sum of these probabilities is still <0.05. The proposed formula takes into account the probability of the item being assigned to all other constructs, and no further computations are needed. So, the determined m value should be compared to n_c to determine whether a set of responses is statistically significant. The following example may illustrate more clearly.

The first item-sort task example is taken as before. This item-sort task was given to 15 respondents. For a particular item on this item-sort task, there are four possible responses. The item was assigned eleven times to the posited construct, while it was assigned three times alternative construct A, one time to alternative construct B, and zero times to alternative construct C. Skipping the arithmetic, $m = 12$ for this example. This number should be compared to the number of responses the posited construct received and found to be non-significant. The same is true no matter the distribution of responses to the alternative constructs, unlike Anderson and Gerbing's (1991) formula.

Also, the proposed formula does not fall victim to the same statistical concerns that Anderson and Gerbing's (1991) formula encounters in the other examples. Once again, an item-sort task with $k = 7$ is taken, where k is defined as before. For an item, a 40 % chance to assign the posited construct and a 10 % chance for each alternative construct are assumed. With these conditions, the item should not be found significant given the null hypothesis. With a sample size of 40, the item will most likely receive 16 assignments to the posited construct and one or zero assignment to each other construct. Skipping the arithmetic, $m = 26$ which is more than the amount of assignments received, and the item is not considered statistically significant. Therefore, because the proposed formula includes the contribution of all alternative constructs, the correct result is found given the propositions.

Furthermore, the proposed formula provides improved results aside from this single example; in fact, the proposed formula provides improved results in several cases. Figure 2a, b demonstrates instances in which the two formulas provide disparate results, once again stemming from the inclusion of alternative constructs into the proposed formula's calculation of statistical significance. These figures present results across various categories, sample sizes, and assignments to the posited construct. As apparent from the table, divergences occur in several instances due to the comparison of one alternative construct or all alternative constructs. Given these numerous disparate results, one should heavily consider which formula to use.

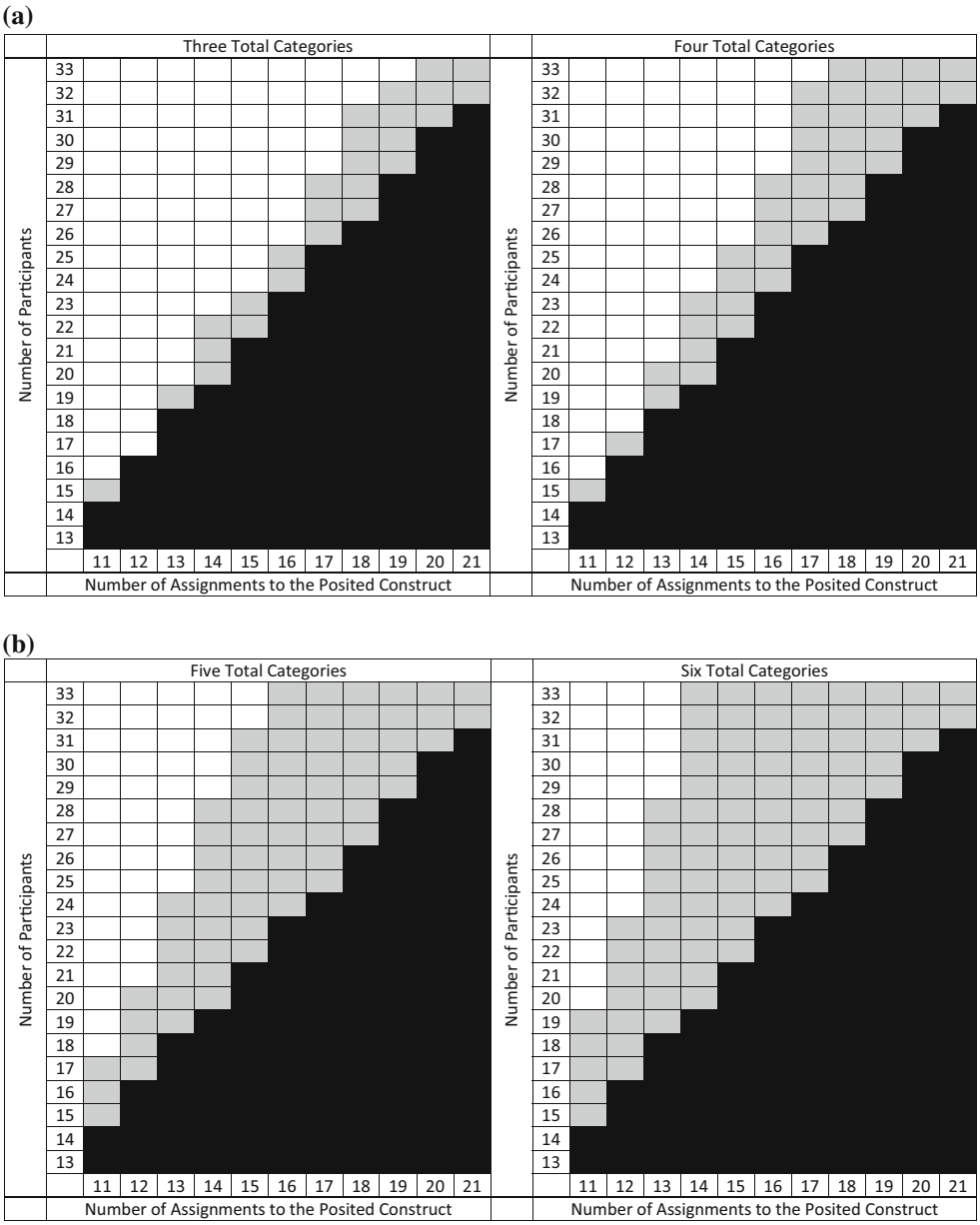

Fig. 2 **a** Visual representation of disparate results when using Anderson and Gerbing's (1991) formula and the proposed formula. **b** Visual representation of disparate results when using Anderson and Gerbing's (1991) formula and the proposed formula Cells in these tables represent Anderson and Gerbing's (1991) and the proposed formula's statistical test results, when the posited construct is assumed to have a 0.5 likelihood of being assigned. For all analysis, the number of assignments to the posited construct is listed on the *x* axis, and all other assignments are equally distributed across the alternative constructs. Cells that are black indicate instances that are always statistically significant using both formulas. Cells that are white indicate instances that are never statistically significant with both formulas. Cells that are gray are statistically significant with Anderson and Gerbing's (1991) formula but not statistically significant with the proposed formula

When using the proposed formula, all alternative response choices are accounted for. As previously mentioned, Anderson and Gerbing's (1991) formula only determines whether an item is assigned to the posited construct more so than a single alternative, as it was intended to compare item-sort task results with only two response choices. The proposed formula measures whether the item was assigned to the posited construct more so than *all* other alternatives, which leads to a greater chance of detecting items that are not representative of the posited construct. Additionally, the posited construct is held to a higher standard than the other constructs using the proposed formula, because it is simultaneously compared to all response choices. This lowers the amount of contamination and is stricter than Anderson and Gerbing's (1991) original formula. With these limitations overcome, the proposed formula seems to be more appropriate for item-sort tasks including multiple alternative construct choices; however, despite theoretical improvements in the proposed formula, simulation and experimental tests can demonstrate whether the proposed formula is appropriate in practice.

Testing the Formula

To test the differences between the two formulas, simulations were performed comparing the formulas across various sample sizes and number of total constructs. Then, an

empirical test of the two formulas was performed to compare their differences in a real-world scenario.

Simulation

To test the two formulas across a variety of scenarios, multiple simulations were performed. A simulation methodology was chosen because it allows for all aspects of the item-sort task to be chosen beforehand, and results could be understood given certain a priori conditions. To perform the simulations, three primary factors were altered across simulations. The first factor was the number of participants, and each simulation either included 10, 20, 30, 40, or 50 participants.

The second factor was the number of total constructs, and each simulation included 2, 3, 4, 5, 6, or 7 total constructs. The probability of assigning an item to the posited construct was 0.5, and the remaining 0.5 probability was equally distributed across the alternative constructs. This probably was chosen because it aligns with the null hypothesis of the two formulas (H_1: $P(a) > 0.5$). Also, with the chosen probability in alignment with the formulas' assumptions, a simulation should accept 5 % of the items to demonstrate a 95 % level of significance. In reality, a scale developer would likely not wish to eliminate 95 % of created items within an over-representative item list; however, the purpose of the current simulation is to demonstrate the efficacy of the formulas to specifically remove problematic items. The simulated distribution itself represents a collection of poor items, as it was generated from the null hypothesis (H_1: $P(a) > 0.5$), and an effective formula would remove most of these items.

The third factor is whether Anderson and Gerbing's (1991) or the proposed formula is applied. The combination of three simulation attributes (sample size, total constructs, formula applied) with five, six, and two categories resulted in a total of 60 conditions. For each condition, 1000 item-sort tasks were performed, and the number of retention decisions was recorded. This process was repeated 1000 times for each condition, and the number of retention decisions was averaged within each condition. For example, a condition may include 10 participants, 3 total categories, and apply the proposed formula. For the simulation, the item has a 0.5 chanced of being assigned to the posited construct, and each of the other two constructs receives a 0.25 probability. Then, 1000 item-sort tasks are performed with these a priori conditions, and the number of retention decisions using the prosed formula is recorded. This process is repeated a total of 1000 times, and the number of retention decisions for each simulation is averaged. The final number is the average number of items that are considered statistically significant out of 1000, given the item-sort task conditions. This process provides robust evidence on the number of items that are retained across various conditions and for each formula.

The results of the 60 total simulations are provided in Table 1. As apparent from this table, Anderson and Gerbing's (1991) formula is much more lenient than the proposed formula. When the sample and category size is small, the two formulas are fairly similar in their results and both are fairly close to accepting items at a 95 % level of significance. However, when the sample and category size increases, the retention decisions of the two formulas diverge. Anderson and Gerbing's (1991) formula's retention decisions gradually increase, whereas the proposed formula's number of retention decisions remains generally constant. Once the formulas reach the final simulation condition (50 participants with 7 total categories), Anderson and Gerbing's (1991) formula accepts the vast majority of item decisions (84 %). Overall, the proposed formula, on average, deviated 1.9 % from a 95 % level of significance, whereas Anderson and Gerbing's (1991) formula, on average, deviated 29.8 % from a 95 % level of significance. This is certainly an undesirable property of the formula, as it is meant to reduce over-representative item lists but instead accepts items at below a 95 % level of significance.

Four primary inferences can be taken from these simulations. First, the two formulas provide identical results when the total number of categories is two. As mentioned above, Anderson and Gerbing (1991) developed their formula for two construct item-sort tasks, and it is statistically correct in this situation. The simulation results for two possible assignment categories reinforce the correctness of Anderson and Gerbing's (1991) formula for this situation and provide support for the applicability of the proposed formula in this situation, too.

Second, Anderson and Gerbing's (1991) formula seems overly reliant on the number of alternative constructs and deems too many items statistically significant when the sample and number of categories increase. Given the a priori conditions (0.5 probability of posited construct), the item should only be retained approximately 5 % of the time to meet a 95 % level of significance within the simulation, and this is clearly not the case.

Third, the level of significance for Anderson and Gerbing's (1991) formula drastically changes along with the item-sort task conditions. This is undesirable, as statistical significance tests should adhere to a consistent level of significance. Fortunately, the proposed formula demonstrated a consistent level of significance, although some variation can be seen.

Fourth, the proposed formula may be slightly too conservative. This is due to the cutoffs based upon the sample size. For certain sample sizes, meeting the cutoff, which is statistically based upon the chosen m value, occurs less

Table 1 Statistical simulation results comparing Anderson and Gerbing's (1991) formula and the proposed formula

Sample size	Number of categories	*m* (associated statistical probability assuming H_1: $P(a) > 0.5$)	Number accepted with Anderson and Gerbing's Formula	Number accepted with proposed formula
10	2	9 (01 %)	10.888	10.888
10	3	9 (01 %)	21.885	10.749
10	4	9 (01 %)	25.249	10.737
10	5	9 (01 %)	27.390	10.811
10	6	9 (01 %)	28.253	10.781
10	7	9 (01 %)	29.141	10.625
20	2	15 (02 %)	20.667	20.667
20	3	15 (02 %)	56.981	20.554
20	4	15 (02 %)	101.529	20.620
20	5	15 (02 %)	124.915	20.672
20	6	15 (02 %)	143.539	20.635
20	7	15 (02 %)	162.264	20.896
30	2	20 (05 %)	49.426	49.426
30	3	20 (05 %)	230.602	49.602
30	4	20 (05 %)	362.188	49.326
30	5	20 (05 %)	470.237	49.607
30	6	20 (05 %)	540.159	49.363
30	7	20 (05 %)	595.262	49.159
40	2	26 (04 %)	40.170	40.170
40	3	26 (04 %)	194.236	40.163
40	4	26 (04 %)	382.463	40.379
40	5	26 (04 %)	512.630	40.041
40	6	26 (04 %)	599.720	40.570
40	7	26 (04 %)	662.474	40.406
50	2	32 (03 %)	32.426	32.426
50	3	32 (03 %)	299.366	32.675
50	4	32 (03 %)	553.615	32.388
50	5	32 (03 %)	704.795	32.341
50	6	32 (03 %)	791.231	32.732
50	7	32 (03 %)	844.619	32.318

To determine the number of items accepted with each formula when the null hypothesis is true, 1000 item-sort tasks were simulated for each condition, and the number of retention decisions was recorded. This process was repeated 1000 times, and the number of retention decisions was averaged. Therefore, the first row indicates that, given the null being true, an item-sort task with 10 participants assigning items to two possible constructs would provide significant results 10.888 out of 1000 items, when this process was averaged after 1000 times

often than a 5 % probability, and the next lowest cutoff is above a 5 % probability. For example, with a sample size of 20, the appropriate m value is 15. The statistical probability of meeting this m value given a 0.5 probability of the posited constructs being assigned is only 2 %. Alternatively, the statistical probability of meeting a cutoff of 14 is approximately 6 % and above a 5 % probability. This forces the m value of 15 to be used, resulting in a smaller number of items being accepted. While it is an undesirable property that the cutoff does not always align with a 95 % level of significance, it is inevitable due to the small sample sizes used within item-sort tasks. Taken together, these results indicate that the proposed formula is itself an improvement upon Anderson and Gerbing's (1991) formula, as their formula is biased by the sample size and number of conditions. While these results provide valuable information, an empirical test can analyze the results in a real-world scenario.

Empirical Study

To test whether the proposed formula to test an over-representative item-sort task's statistical significance is more appropriate, a replication of Anderson and Gerbing's

(1991) original experiment is performed. In their original experiment, the authors performed an item-sort task on an item list that had already been reduced through exploratory factor analysis. Then, they compared the item-sort task results to the exploratory factor analysis results to determine whether item-sort tasks were an appropriate method to reduce over-representative item lists. In the current study, this is repeated. The results of an item-sort task using both, the newly created and Anderson and Gerbing's (1991), formulas will be compared to exploratory factor analysis results. The results will provide evidence for which formula can better predict exploratory factor analysis results, and are better indicators of substantive validity.

Method

To test the two formulas, an existing over-representative item list reduced through factor analysis was used. In Dahling et al. (2009) article, the authors present a multiple-study process to create the Machiavellian Personality Scale. In their first study, the authors reduced an over-representative item list through exploratory factor analysis. In the reporting of this study, the exact wording and factor loadings of each item are presented. This comprehensive reporting allows the current study to compare item-sort task results to the original factor analysis. Therefore, an item-sort task using Dahling et al. (2009) collection of 45 items was conducted with two groups of participants.

Data from three groups of participants were collected, each serving as separate item-sort task respondents. Three groups were recruited comprised of 24, 24, and 22 students, respectively, recruited from the undergraduate psychology participant pool of a large northeastern university. They were given a small amount of course credit for their participation. These samples were chosen with the intentions of being as similar to Anderson and Gerbing's (1991) original item-sort task sample as possible. Because these participants did not respond based on their personal characteristics, no demographic information was recorded.

When creating the initial item list for the Machiavellian Personality Scale, Dahling et al. (2009) created items which were intended to gage four categories: Desire for Control, Desire for Power, Distrust of Others, and Amoral Manipulation. Participants were given the definition of each of these constructs. These definitions were adapted from Dahling et al. (2009) article and were slightly simplified to be understandable to an undergraduate sample. Then, participants were instructed on how to perform an item-sort task, but no other information or training was given. When answering each item, their answer choices were these four constructs and a "None of the Above" option.

Additionally, because item-sort tasks use small samples, their responses can be skewed even by the insufficient motivation of a single participant. For this reason, three insufficient motivation checks were added. The first was a set of three items mixed within the item-sort task, directing participants to give a particular response. The second was an item that read, "Do you believe that your data should be used for analysis," to which participants could respond, "Yes" or "No." The third was another item that read, "Using the slider below, how much effort did you put into completing this survey," and a slider ranging from 0 to 100 was presented. Participants that failed two-of-the-three motivation checks (incorrectly answering motivation check one, answering "no" to motivation check two, or responding below 50 to motivation check three) were removed from data analysis. The previously reported sample sizes for the three groups (24, 24, and 22) reflect those who passed two-of-the-three motivation checks.

Results

To test whether the two formulas provided a substantial change in item retention decisions, a paired sample t test was performed. For the data, two columns were created. The first column was the retention decision given Anderson and Gerbing's (1991) formula, and the second column was the retention decision given the proposed formula. Each row was an item presented to the participants in the separate sample. For example, the first column of the first row was the retention decision (1 = retained and 0 = discarded) for the first item within the first sample using Anderson and Gerbing's (1991) formula, whereas the second column of the first row was the retention decision for the first item within the first sample using the proposed formula. Then, each of the following rows contained the retention decisions for all the following items within each sample. This format allowed the two formulas to be compared across the item-sort tasks performed for both samples. The results demonstrate that a significant difference exists depending on the formula used ($t(134) = 4.97$; $p < 0.0001$), and that further analysis comparing the two formulas is appropriate.

Next, to determine which formula is more suitable, results using both formulas were compared to the previous item retention decisions determined through an exploratory factor analysis (Dahling et al. 2009). Responses from the three groups separately analyzed by both formulas are presented in Table 2. With all three groups combined, the proposed formula resulted in 81 % correct item decisions, whereas Anderson and Gerbing's (1991) formula resulted in 76 % correct item decisions. To be more specific with the overall combined results, the proposed formula demonstrated 85 % correct item removal decisions, whereas

Anderson and Gerbing's (1991) formula resulted in 90 % correct item removal decisions; the proposed formula demonstrated 74 % correct item retention decisions, whereas Anderson and Gerbing's (1991) formula resulted in 59 % correct item retention decisions. Therefore, the proposed formula is more accurate in its item decisions.

When analyzing the pattern of correct and incorrect retention decisions, the proposed formula is more conservative with item retention decisions which results in greater accuracy. This also resulted in the proposed formula having fewer false positives than Anderson and Gerbing's (1991) formula, also known as Type I error. The tradeoff for this reduction to Type I error, however, is an increase in Type II error when using the proposed formula. That is, the proposed formula was more likely to reject items that were retained through the EFA. As a long line of research has noted, a reduction to Type I error at the cost of Type II error is often preferred (Aguinis et al. 2010; Devlin and Jacobs 2010). Taking this into consideration, the proposed formula's reduction to Type I error makes it more suitable than Anderson and Gerbing's (1991) formula despite its increase to Type II error.

Discussion

Item-sort tasks are regularly used to reduce an over-representative list of items created early in the scale development process. The results of the item-sort task produce a list of items that constitute an initial scale for the construct of interest. For this reason, the results of the item-sort task are extremely important for measurement creation. Unfortunately, the extant formula for item-sort task statistical significance has certain concerns. The goal of the current study was to analyze Anderson and Gerbing's (1991) formula for item-sort task statistical significance, as well as propose a new formula that mitigates potential concerns.

As mentioned above, Anderson and Gerbing (1991) had developed and illustrated their formula in the case of two constructs, but scholars since have almost exclusively used item-sort tasks with multiple alternative construct categories for participants to assign items. Thus, the problem is not with the formula per se, but arises when this formula is misapplied to cases with more than two constructs. Therefore, while these errors were not present in Anderson and Gerbing's (1991) original intentions for item-sort tasks, subsequent authors' adaptations of item-sort task methods produce problematic and unanticipated results that give rise to notable concern. By accounting for both cases with two or more constructs, our proposed formula is more applicable and accurate than Anderson and Gerbing's formula (1991), and its errors are more preferred. A simulation study and empirical study demonstrated that the proposed formula is a statistical improvement upon Anderson and Gerbing's (1991) formula and is applicable to a far greater set of circumstances. Our findings indicate that the proposed formula should be used for future item-sort tasks.

Table 2 Item-sort task results using Anderson and Gerbing's (1991) formula and the proposed formula

Anderson and Gerbing's Formula			Proposed formula		
Item decision	Result	# of items	Item decision	Result	# of items
Group one					
Delete	Correct	20	Delete	Correct	26
	Incorrect	3		Incorrect	7
Retain	Correct	11	Retain	Correct	8
	Incorrect	11		Incorrect	4
Group two					
Delete	Correct	23	Delete	Correct	27
	Incorrect	2		Incorrect	4
Retain	Correct	13	Retain	Correct	11
	Incorrect	7		Incorrect	3
Group three					
Delete	Correct	22	Delete	Correct	26
	Incorrect	2		Incorrect	3
Retain	Correct	13	Retain	Correct	12
	Incorrect	8		Incorrect	4

$n = 24$ (Group 1), 24 (Group 2), and 22 (Group 3)

The table provides the total number of correct and incorrect item retention and deletion decisions provided by each formula across three separate item-sort task samples

In addition to being more statistically accurate, the proposed formula also avoids some methodological pitfalls. When applying Anderson and Gerbing's (1991) formula, the alternative constructs cannot be overly repetitive. This may not be desirable in practice if one is seeking to differentiate certain items from similar dimensions of a construct. For example, if a researcher performs an item-sort task to create a procedural justice scale, using distributive justice and interactional justice as alternative constructs may split responses between the two, but it is necessary to distinguish items from these constructs. When using Anderson and Gerbing's (1991) formula for cases with more than two constructs, this splitting will cause many of the items to be discounted, as only the most reported category is used in the statistical analysis. While this is a potential problem if one chooses to use Anderson and Gerbing's (1991) formula, it does not affect our proposed formula. This is because assignments to each alternative construct are mathematically included within the proposed formula, and the proposed formula compares the posited construct against all other constructs. It does not matter whether a single alternative construct is selected often or responses are split across several alternative constructs, as they are all cumulatively compared to the posited construct. Therefore, the proposed formula does not discount any alternative responses and so accounts for assignments to multiple constructs.

Applying Item-Sort Tasks

Notes should be made about applying item-sort tasks within research and practice. Regardless of the statistical formula used, great consideration should still be given to the alternative constructs presented in an item-sort task. Item-sort tasks can only determine whether an item is assigned to the posited construct more so than the presented constructs, not more so than all other possible constructs. Thus, the influence of an unlisted construct will remain unknown.

Unfortunately, despite the discovery of better statistics and methods, there is no guarantee that researchers will follow them. When reporting study results, many researchers do not give adequate explanation in regards to their item-sort tasks. At best, some authors include separate phases to their studies to describe their item-sort tasks (Ferris et al. 2008). At worst, some only leave a sentence or a footnote (Bansal 2005; Fassnaght and Ibrahim 2006; Gopalakrishnan and Bierly 2001). In many studies, researchers also fail to use the correct critical value to determine statistical significance. Some studies include arbitrary critical values without any explanation of how they reached their value (Bauer et al. 2001; Geyskens and Steenkamp 2000; Menor and Roth 2007; Tinsley 1998), and others never mention their critical values at all (Anderson et al. 1994; Bansal 2005; Hibbard et al. 2001; Mathwick et al. 2001; Ulaga and Eggert 2006). The repeated publication of studies that do not use appropriate cutoff values is problematic. In light of this, a table is included which illustrates the proper critical values for item-sort tasks of varying sample sizes using Anderson and Gerbing's formula (1991; Table 3). Researchers can use this table to determine their critical value without having to calculate it themselves. Another table is also included for the correct critical values when using the proposed formula (Table 4).

It may be reasonable to alter a priori cutoffs based upon item-sort task purposes, but only after careful consideration has been devoted to this decision. Previous researchers have solely relied on the 0.5 probability of assigning the item to the posited construct based upon Anderson and Gerbing (1991)'s original formula. It was noted that this assumption is completely arbitrary. Logically, this assumption also appears rather liberal. Given that items are meant to be direct reflections of the construct of interest, it seems unlikely that any created item is only assigned to the posited construct at a 0.5 probability, and requiring items to be significantly different than a 0.5 probability may be too lenient for certain situations. It may be more realistic to expect a 0.6 or 0.75 probability. Of course, certain factors can influence this probability, such as the alternative constructs or respondents. Therefore, if the alternative constructs are notability different than the construct of interest or the respondents are experiences subject matter experts, it may be appropriate to alter the assumption of assignment at a 0.5 probability and apply a stricter cutoff. Future research should seek to determine the adequacy of the conventional cutoff and under what circumstances an alternative cutoff should be implemented.

The results of the current manuscript may be of particular importance to practitioners. When creating a scale, researchers at universities sometimes have access to an undergraduate student participant pool. This allows for a large sample size for measure pretesting, although the validity of undergraduate student samples is often debated. However, organizational practitioners less often have the luxury of a large pretest participant pool. For this reason, practitioners may be more reliant on small sample sizes for measurement pretesting. The current article can certainly aid practitioners who must rely on these small pretest sample sizes, as the proposed formula results in more accurate item-sort task decisions. Also, the use of subject matter experts may eliminate the need of other qualitative pretest methods, and item-sort tasks can be completed with as few as five participants. These factors further reduce the required pretest sample size, proving to be even more beneficial to practitioners.

Table 3 Critical values for varying sample sizes for item-sort tasks using Anderson and Gerbing's (1991) formula

Sample size (n)	Critical value ($\bar{c}_{sv}$)	Sample size (n)	Critical value ($\bar{c}_{sv}$)
5	1.000	23	0.391
6	1.000	24	0.417
7	1.000	25	0.440
8	0.750	26	0.385
9	0.778	27	0.407
10	0.800	28	0.357
11	0.636	29	0.379
12	0.667	30	0.333
13	0.538	31	0.355
14	0.571	32	0.375
15	0.600	33	0.333
16	0.500	34	0.353
17	0.529	35	0.314
18	0.444	36	0.333
19	0.474	37	0.297
20	0.500	38	0.316
21	0.429	39	0.333
22	0.456	40	0.300

All critical values rounded to three decimal places

Table 4 Critical values for varying sample sizes for item-sort tasks using the proposed formula

Sample size (n)	Critical value (m)	Sample size (n)	Critical value (m)
5	5	23	16
6	6	24	17
7	7	25	18
8	7	26	18
9	8	27	19
10	9	28	19
11	9	29	20
12	10	30	20
13	10	31	21
14	11	32	22
15	12	33	22
16	12	34	23
17	13	35	23
18	13	36	24
19	14	37	24
20	15	38	25
21	15	39	26
22	16	40	26

Item-Sort Task Methodological Comments

Several concluding notes should be made regarding item-sort tasks. The current article touched-upon the difference between Type I and Type II error for item-sort tasks, but this issue should be reiterated. When creating a measure, a researcher or practitioner hopes that their measure gages all aspects of the construct of interest, known as content validity, while avoiding the inclusion of irrelevant constructs, known as construct contamination. Within item-sort tasks, committing Type I error results in the inclusion of items that gage irrelevant constructs. Including these items within a final measure would distort results and obfuscate the true nature of the construct of interest. This is particularly problematic for

researchers, as subsequent scholarship based upon these measures would be misleading. These concerns can be alleviated through additional measurement testing, such as an EFA, but they are still important to consider. Conversely, committing Type II error results in the removal of items that gage the construct of interest. Removing these items would require the creation of additional items, as well as potentially prompting additional pretesting. While inconvenient for researchers, this may pose particular problems for practitioners. The creation of additional items would result in additional working hours for practitioners, costing valuable time, and money; however if a practitioner is unable to perform additional pretesting after the item-sort task, then it may be advisable to err on removing too many items to ensure the validity of the measure before actual use. For these reasons, researchers and practitioners should carefully consider the costs of Type I and II error within their item-sort tasks.

Also, researchers almost exclusively use sample sizes of 20 when performing item-sort tasks. Although it may be possible that they conveniently have samples of 20 for their studies, it may also be possible that they are exactly imitating the sample used in Anderson and Gerbing (1991). For item-sort tasks, smaller sample sizes can be acceptable. Using either formula, sample sizes as small as five can yield statistically significant results, but only if all respondents unanimously assign an item to the posited construct. As with all other statistical tests larger sample sizes are better but, when necessary, sample sizes smaller than 20 can be sufficient.

It was noted that a free-response blank after each item allows respondents to qualitatively indicate which items are confusing, double-barreled, or leading. Most likely any respondent, including an undergraduate student pool participant, can indicate if any item is confusing; however, it may be unrealistic to expect these individuals to indicate whether an item is double-barreled or leading. For this reason, when using undergraduate student pool participants, it may be appropriate to additionally apply other qualitative pretest methods to detect for these wording issues, such as practice verbal administrations (Aday and Cornelius 2011; Krosnick 1999; Presser and Blair 1994) or cognitive interviewing (Conrad and Blair 2004; DeMaio and Landreth 2004; Willis 2005). Alternatively, subject matter experts, such as relevant practitioners and researchers, are likely able to detect these wording issues, and the application of qualitative pretest methods may not be as necessary.

Conclusion

In this article, an overview of item-sort tasks was provided. Although item-sort tasks have been consistently used for decades, the misuse of common methods and statistics are problematic. This review will hopefully clarify the proper method to use when performing an item-sort task. In addition, an alternative formula was provided to test for an item-sort task's statistical significance. This proposed formula has more acceptable assumptions and avoids certain limitations of Anderson and Gerbing's (1991) formula. Also, the proposed formula was shown to have more satisfying results in practice. Finally, important considerations were noted that should be taken into account for future item-sort tasks. Therefore, the current article will hopefully guide future researchers in all aspects of item-sort tasks and improve upon the current methods.

Acknowledgments We would like to thank Matt Crayne for his comments on a previous version of this manuscript.

References

Aday, L. A., & Cornelius, L. J. (2011). *Designing and conducting health surveys: A comprehensive guide*. Hoboken, NJ: Wiley.

Aguinis, H., Werner, S., Abbott, J., Angert, C., Park, J. H., & Kohlhausen, D. (2010). Customer-centric science: Reporting significant research results with rigor, relevance, and practical impact in mind. *Organizational Research Methods, 13*, 515–539. doi:10.1177/1094428109333339.

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*, 732–740. doi:10.1037/0021-9010.76.5.732.

Anderson, J. C., Håkansson, H., & Johanson, J. (1994). Dyadic business relationships within a business network context. *Journal of Marketing, 58*, 1–15.

Bansal, P. (2005). Evolving sustainably: A longitudinal study of corporate sustainable development. *Strategic Management Journal, 26*, 197–218.

Bauer, T. N., Truxillo, D. M., Sanchez, R. J., Craig, J. M., Ferrara, P., & Campion, M. A. (2001). Applicant reactions to selection: Development of the selection procedural justice scale (SPJS). *Personnel Psychology, 54*, 387–419.

Collins, D. (2003). Pretesting survey instruments: An overview of cognitive methods. *Quality of Life Research, 12*, 229–238.

Comrey, A. L., & Lee, H. B. (1992). *A first course in factor analysis*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Conrad, F. G., & Blair, J. (2004). Aspects of data quality in cognitive interviews: The case of verbal reports. In S. Presser, J. M. Rothgeb, M. P. Couper, J. T. Lessler, E. Martin, J. Martin, & E. Singer (Eds.), *Methods for testing and evaluating survey questionnaires* (pp. 67–88). New York: Wiley.

Costello, A. B., & Osbourne, J. (2005). Best practices in exploratory factor analysis: Four recommendations for getting the most from your analysis. *Practical Assessment, Research & Evaluation, 10*, 1–9.

Czaja, R. (1998). Questionnaire pretesting comes of age. *Marketing Bulletin-Department of Marketing Massey University, 9*, 52–66.

Dahling, J. J., Whitaker, B. G., & Levy, P. E. (2009). The development and validation of a new Machiavellianism scale. *Journal of Management, 35*, 219–257.

DeMaio, T. J., & Landreth, A. (2004). Do different cognitive interview techniques produce different results? In S. Presser, J. M. Rothgeb, M. P. Couper, J. T. Lessler, E. Martin, J. Martin, & E. Singer (Eds.), *Methods for testing and evaluating survey questionnaires*. NJ: Wiley, Hoboken. doi:10.1002/0471654728.ch5.

Devlin, A., & Jacobs, M. (2010). Antitrust error. *William and Mary Law Review, 52*, 75–132.

Fassnaght, M., & Ibrahim, K. (2006). Quality of electronic services: Conceptualizing and testing a hierarchical model. *Journal of Service Research, 9*, 19–37.

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*, 1348–1366. doi:10.1037/a0012743.

Fiore, A. M., Niehm, L. S., Hurst, J. L., Son, J., & Sadachar, A. (2013). Entrepreneurial marketing: Scale validation with small, independently-owned businesses. *Journal of Marketing Development and Competitiveness, 7*, 63–86.

Geyskens, I., & Steenkamp, J. M. (2000). Economic and social satisfaction: Measurement and relevance to marketing channel celationships. *Journal of Retailing, 76*, 11–32.

Google Scholar (2014). Articles citing "Predicting the performance of measures in a confirmatory factor analysis with a pretest assessment of their substantive validities." http://scholar.google.com/scholar?cites=6941685027650661312&as_sdt=5,39&sciodt=0,39&hl=en.

Gopalakrishnan, S., & Bierly, P. (2001). Analyzing innovation adoption using a knowledge-based approach. *Journal of Engineering and Technology management, 18*(2), 107–130.

Groves, R. M., Fowler, F. J, Jr, Couper, M. P., Lepkowski, J. M., Singer, E., & Tourangeau, R. (2013). *Survey methodology*. Hoboken, NJ: Wiley.

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*, 98–107.

Hardy, B., & Ford, L. (2014). It's not me, it's you: Miscomprehension in surveys. *Organizational Research Methods, 17*, 138–162.

Hemphill, J. K., & Westie, C. M. (1950). The measurement of group dimensions. *The Journal of Psychology, 29*, 325–342.

Hibbard, J. D., Kumar, N., & Stern, L. W. (2001). Examining the impact of destructive acts in marketing channel relationships. *Journal of Marketing Research, 38*, 45–61.

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

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

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

Holden, R. R., & Jackson, D. N. (1979). Item subtlety and face validity in personality assessment. *Journal of Consulting and Clinical Psychology, 47*, 459.

Krosnick, J. A. (1999). Survey research. *Annual Review of Psychology, 50*, 537–567.

Lawrence, S. A., Gardner, J., & Callan, V. J. (2007). The support appraisal for work stressors inventory: Construction and initial validation. *Journal of Vocational Behavior, 70*, 172–204.

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

Leach, C. W., van Zomeren, M., Zebel, S., Vliek, M. L. W., Pennekamp, S. F., Doosje, B., et al. (2008). Group-level self-definition and self-investment: A hierarchical (multicomponent) model of in-group identification. *Journal of Personality and Social Psychology, 95*, 144–165. doi:10.1037/0022-3514.95.1.144.

Linderbaum, B. A., & Levy, P. E. (2010). The development and validation of the feedback orientation scale (FOS). *Journal of Management, 36*, 1372–1405.

MacCallum, R. C., Widaman, K. F., Zhang, S., & Hong, S. (1999). Sample size in factor analysis. *Psychological Methods, 4*, 84–99.

Mathwick, C., Malhotra, N., & Ridgon, E. (2001). Experiential value: Conceptualization, measurement and application in the catalog and internet shopping environment. *Journal of Retailing, 77*, 39–56.

Menor, L. J., & Roth, A. V. (2007). New service development competence in retail banking: Construct development and measurement validation. *Journal of Operations Management, 25*(4), 825–846.

Michel, J. S., Pace, V. L., Edun, A., Sawhney, E., & Thomas, J. (2014). Development and validation of an explicit aggressive beliefs and attitudes scale. *Journal of Personality Assessment, 96*, 327–338.

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

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

Presser, S., Couper, M. P., Lessler, J. T., Martin, E., Martin, J., Rothgeb, J. M., & Singer, E. (2004). Methods for testing and evaluating survey questions. *Public Opinion Quarterly, 68*, 109–130. doi:10.1093/poq/nfh008.

Rovinelli, R. J., & Hambleton, R. K. (1977). On the use of content specialists in the assessment of criterion-referenced test item validity. *Dutch Journal of Educational Research, 2*, 49–60.

Schaeffer, N. C., & Presser, S. (2003). The science of asking questions. *Annual Review of Sociology, 29*, 65–88.

Spector, P. E. (1992). *Summed rating scale construction*. Newbury Park, CA: Sage Publications.

Stanton, J. M., Sinar, E. F., Balzer, W. K., & Smith, P. C. (2002). Issues and strategies for reducing the length of self-report scales. *Personnel Psychology, 55*, 167–194.

Tinsley, C. (1998). Models of conflict resolution in Japanese, German, and American cultures. *Journal of Applied Psychology, 83*, 316–323. doi:10.1037/0021-9010.83.2.316.

Ulaga, W., & Eggert, A. (2006). Value-based differentiation in business relationships: Gaining and sustaining key supplier status. *Journal of Marketing, 70*, 119–136.

Willis, G. B. (2005). *Cognitive interviewing: A tool for improving questionnaire design*. Thousand Oaks, CA: Sage.