When Two Factors Don’t Reflect Two Constructs: How Item Characteristics Can Produce Artifactual Factors

Paul E. Spector
University of South Florida

Paul T. Van Katwyk
Personnel Decisions International and University of South Florida

Michael T. Brannick
University of South Florida

Peter Y. Chen
Ohio University

Factor analyses of scales that contain items written in opposite directions sometimes show two factors, each of which contains items written in only one direction. Such item direction factors have been found in scales of affect and personality that have been used in organizational research. We discuss how patterns of subject responses to items that vary in direction and extremity can produce an artifactual two factor structure in the absence of multiple constructs. Response patterns are demonstrated in Study 1 with job satisfaction data gathered from employed subjects. The production of two factors is illustrated in Study 2 with simulated data based on item response characteristic equations.

Many scales that assess variables of interest to organizational researchers (e.g., employee personality) contain items written to reflect both ends of the continuum of the construct of interest. Factor analyses of such scales sometimes suggest the existence of two factors, each containing items written in one direction. When assessing a scale that contains oppositely worded items that form separate factors, it is important to ascertain whether the scale measures one or two constructs. The seeming independence of oppositely worded items can be caused, not by underlying constructs, but by the way in which people respond to items.

In this paper, we first review literature showing that item direction factors within scales is widespread. Next, we discuss how people’s patterns of responses to oppositely worded items can produce two-factor structures. The existence of these patterns is demonstrated in Study 1 with job satisfaction data from a sample...
of employed subjects. The production of factors based on response patterns to different types of items is shown with simulated data in Study 2. Finally, we present guidelines for determining when two-factor structures are likely artifactual.

**Item Type Factors In Existing Scales**

Factors produced by items phrased in opposite directions have been noted for several scales in the organizational domain. Such factors have been found in Hackman and Oldham's (1975) Job Diagnostic Survey (Idaszak & Drasgow, 1987), Meyer and Allen's (1984) Affective and Continuance Commitment Scales (Magazine, Williams & Williams, in press), Rizzo, House and Lirtzman's (1970) role ambiguity and conflict scales (McGee, Ferguson & Seers, 1989), and Spector's (1988) Work Locus of Control Scale (Spector, 1992). In all cases these factors were interpreted to be artifacts of wording direction. However, there has been limited work done on the underlying mechanisms that produce artifactual factors.

Several possible mechanisms have been suggested for artifactual factors, including lack of ability to understand negatively worded items (Cordery & Sevastos, 1993) and carelessness in reading items (Schmitt & Stults, 1985). In addition Schriesheim and his colleagues noted that negatively worded items are often less reliable and valid than positively worded items (Schriesheim, Eisenbach & Hill, 1991; Schriesheim & Hill, 1981), possibly due to the greater difficulty for people to interpret negatively worded items. All of these researchers have recommended that care be taken in mixing negatively and positively worded items in the same questionnaire.

We discuss in this paper how the distribution of item responses can produce direction factors in some scales. This mechanism is not necessarily responsible for item direction factors in all of the scales noted. For example, it seems quite likely that the Schmitt and Stults (1985) carelessness response explanation is correct for the JDS. However, not all scales produce negatively worded items by negating positively worded items as does the JDS. Our mechanism does not rely on item negation (adding "not" to a positively worded item, e.g., "I do NOT like my job"). It is most likely to occur with scales that assess affect or personality because of the tendency to use oppositely worded items that in some cases are rarely endorsed by subjects.

In the affect domain, item direction factors have often been interpreted as reflecting multiple constructs. For example, items that reflect negative versus positive emotions have been interpreted as assessing the independent constructs of negative and positive affectivity (Brief & Roberson, 1989; Burke, Brief & George, 1993; George, 1991; Watson, Clark & Carey, 1988; Watson & Tellegen, 1988). However, Gotlib and Meyer (1986) argued that the two-factor structure of the Multiple Affect Adjective Checklist was likely caused by response sets. Scheier and Carver's (1985) dispositional optimism scale (the LOT) has been interpreted as containing two independent dimensions—optimism and pessimism (Marshall, Wortman, Kusulas, Hervig & Vickers, 1992). Pilotte and Gable (1990)
suggested that the two factors they found with their scale of computer anxiety might reflect two constructs.

Item direction factors have been found in several scales, but there are many scales that do not produce this effect. For example, the phenomenon has not been noted for job satisfaction scales. The Job In General Scale has been found to be clearly unidimensional (Ironson, Smith, Brannick, Gibson & Paul, 1989). The nine facet Job Satisfaction Survey did not form item direction factors even though approximately half the items were phrased in each direction (Spector, 1988). With scales that produce the item direction factors, there are several possible mechanisms that might cause them.

Although the item direction factors in scales could be produced by independent constructs, it seems more likely that they are artifacts. An examination of the literature on the psychometric properties of item responses suggests that the way in which people respond to items that vary in direction can produce item direction factors even when the underlying construct is unidimensional. Nunnally and Bernstein (1994) noted that pseudofactors can emerge when items with different directions or distributional properties are factor analyzed. Van Schuur and Kiers (1994) argued that factor analysis is inappropriate for many scales because of how people respond to them. We next discuss how such responses can produce factors based on direction of item wording.

How Responses To Items With Different Characteristics Can Produce Artifactual Factors

In computing total scores with scales that contain oppositely worded items, negative items are reverse scored so that strong disagreement is given the same number as strong agreement with positive items. Therefore, the scoring of disagreement with the first item type will be equivalent to agreement with the second. The assumption underlying such a scoring method is that individuals who disagree strongly with an item on one side of the continuum will usually agree with items on the other side. This should result in a similar magnitude of correlations between oppositely and similarly worded items. If similarly worded items intercorrelate around .40, then oppositely worded items should intercorrelate around -.40. The uniformity in correlation magnitude across a set of items that assess the same construct should produce a single factor. If a subset of items based on wording direction intercorrelate more highly with one another than with oppositely worded items, a two-factor structure will emerge even in a set of items that assesses a single construct.

Although we assume that reverse scoring oppositely worded items makes the item scores equivalent, this is not necessarily the case. As pointed out by Thurstone (1928) in his classic paper on attitude scale construction, a person will tend to agree with items that are close to his or her attitude and disagree with items that are far from it in either direction. This idea that people agree with items that are close to their level on the trait and disagree with all others is termed the ideal point (Cliff, Collins, Zatkin, Gallipeau & McCormick, 1988) or the unfolding principle (Andrich, 1988). This implies that there is a curvilinear, rather than linear, relation.
between the extent of agreement with an item and the person's standing on the construct of interest. Given a particular item, people whose standing is either higher or lower than the item will be likely to disagree with it.

Figure 1 illustrates how the ideal point idea works. The line represents the underlying trait continuum from extreme negative on the left to extreme positive on the right. The letters below the line represent the standing of four items on the continuum. Numbers above the line represent the standing of four people on the continuum. In each case, individuals are likely to agree with the item that is closest to their position on the continuum (e.g., person 1 will be most likely to agree with item a). A person who has a moderate position, (e.g., #2) may disagree with the extreme negative item (a) even though it is in the direction of his or her standing. It is also possible that person 1 will disagree with item b, even though it is in the negative direction. If a scale contains only extreme items that are at positions a and d, some people might disagree with all items because the items are too far from the people on the continuum of interest.

For example, take a simple case in which a sample of people are asked to respond to the following two items, taken from the Positive and Negative Affect Scale (PANAS, Watson, Clark & Tellegen, 1988). For simplicity, a two-choice yes/no format is used:

**Figure 1.** An Illustration of the Ideal Point Principle
Do you feel enthusiastic?  yes  no
Do you feel distressed? yes  no

For those people who are either enthusiastic or distressed, it is expected that they would say yes to one item and no to the other. Yet, what if a person is in a more neutral mood and is neither enthusiastic nor distressed? Might he or she not say no to both items since neither reflects his or her feeling? For example, we could expect that a person who is feeling calm and perhaps somewhat sleepy would say no to feeling enthusiastic and also not say yes to feeling distressed. Yet the failure to say that they feel enthusiastic does not necessarily imply a negative feeling, and, conversely, the failure to feel distressed does not necessarily imply a strong positive feeling.

If the response format for the affect items given above is changed to agreement/disagreement with five to seven response choices, the same point holds. An individual who disagrees with negatively worded items that reflect strong negative affect will not necessarily agree, even slightly, with positively worded items that reflect strong positive affect. Although some of the individuals who agree with one type of item will consistently disagree with the opposite type, others may disagree slightly with both types of items. Responses to oppositely worded items are not always mirror images of one another.

The response patterns described above produce a two-factor structure in the following way. Presume that a sample of respondents vary in their affective dispositions. Further, presume that we have an affectivity scale with half the items written to reflect strong positive affectivity and half to reflect strong negative affectivity. If we have an internally consistent set of items, then within each type of item (negatively or positively worded), respondents will be consistent in their responses, regardless of their underlying true affective state. Respondents who are highly negative or highly positive will agree with one set of items and disagree with the other. Respondents who are mildly positive, mildly negative, or neutral may disagree with both sets of items. Note that we have not mentioned that some respondents might agree with both types of items, which would be an acquiescence response set. Although some people might have this tendency, research on acquiescence has found it to be relatively rare (e.g., Diener & Iran-Nejad, 1986). Furthermore, disagreement with items that are too extreme in both directions is not because of a response set, although uniform disagreement can be caused by one.

The consistency in response within item direction type will produce strong correlations among these items. Most people who agree (disagree) with one will agree (disagree) with the others, meaning that one could accurately predict responses to one item from responses to another. The lack of consistency across item types on the part of some people will attenuate correlations between the two types of items. This is because a person who disagrees with one item might agree or disagree with an item that is oppositely worded. There will be an inconsistent association between responses to the two items, which will cause a lower correlation across item types than within item types.
If all respondents can agree with one type and disagree with the other, or if all respondents disagree with all items, the two factors will not appear. In the former case, almost all respondents would agree with one type of item and disagree with the other, so that the positively and negatively worded items would be highly positively correlated after reverse scoring. In the latter case, almost all subjects would disagree with all items so that all items would be highly positively correlated.

Therefore, the appearance of two factors depends upon the joint distributions of both items and people on the underlying measurement continuum. If the items are extreme relative to where most people are (e.g., like items a and d in Figure 1), and the people are spread throughout the continuum, there will be a mix of people who are consistent and inconsistent in agreeing with items on opposite ends of the continuum. Correlations between items at opposite ends of the continuum will be attenuated, while correlations between items at the same end will be relatively strong. As a result, artifactual factors would appear based on the response patterns to these items. Specifically, these factors merely reflect the greater magnitude of correlation within versus across item type based on the distributions of responses to items rather than multiple constructs.

In order to demonstrate that this phenomenon actually occurs, we provide two forms of evidence. First, Study 1 shows that subjects exhibit the patterns of responses that have been discussed here. Item direction and extremity affect the patterns of responses to items. Second, Study 2 shows with simulated data that artifactual factor structures can emerge in the absence of multiple constructs.

**Study 1: Demonstrating Hypothesized Response Patterns**

Our explanation of how two factors can be produced artfully is based on the assumption that people only endorse (agree with) items that are close to their true level of the construct, and will disagree with items that are far away in either direction. To provide evidence that this is in fact what people do, we created four sets of items concerning job satisfaction which varied in direction (negative or positive) and degree (extreme or moderate). Job satisfaction was chosen as the construct because existing scales have not shown the item direction effect, as we discussed previously. By manipulating item extremity, we demonstrate that two factors can be produced. We assume that the majority of individuals are moderate in their feelings about their jobs, and are either somewhat satisfied or dissatisfied, as opposed to extremely satisfied or dissatisfied. Subjects who were divided into three groups (dissatisfied, neutral, and satisfied) completed all four groups of items. If our supposition is correct, we should find that individuals agree more strongly on average with moderate items that reflect their direction of feeling than they do with extreme items. This is because the moderate items are closer than extreme items to the feelings of the majority of subjects.

An initial pool of 60 items, half written in the positive and half in the negative direction, was generated. Half of each type was intended to be moderate and half extreme in wording direction. The moderate items were similar to items that are contained in available job satisfaction instruments, and the Job in General
Scale (Ironson, Smith, Brannick, Gibson & Paul, 1989) was used to help generate these items. The extreme items were written to reflect a stronger feeling than is typical of satisfaction items. We reasoned that most respondents would find the extreme items difficult to endorse.

Table 1. Positive and Negative Satisfaction Items

<table>
<thead>
<tr>
<th>Extreme Negative Satisfaction Items</th>
<th>Moderate Positive Satisfaction Items</th>
</tr>
</thead>
<tbody>
<tr>
<td>I hate my job.</td>
<td>I enjoy my job.</td>
</tr>
<tr>
<td>I can't stand to go to work.</td>
<td>I have interesting things to do at work.</td>
</tr>
<tr>
<td>I loathe my job.</td>
<td>My job is rewarding.</td>
</tr>
<tr>
<td>I have a horrible job.</td>
<td>My job is pleasant.</td>
</tr>
<tr>
<td>I often feel that I can't stand to go to work for even one more day.</td>
<td>My job is rewarding.</td>
</tr>
<tr>
<td>I have a rotten job.</td>
<td>My job is pleasant.</td>
</tr>
<tr>
<td>I wouldn't wish my job on my worst enemy.</td>
<td>I have a decent job.</td>
</tr>
<tr>
<td>I can't imagine a worse job than mine.</td>
<td>I have a meaningful job.</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Extreme Positive Satisfaction Items</th>
<th>Moderate Negative Satisfaction Items</th>
</tr>
</thead>
<tbody>
<tr>
<td>I love my job.</td>
<td>My job is unpleasant.</td>
</tr>
<tr>
<td>My job is the best thing that ever happened to me.</td>
<td>I would be happier with a different job.</td>
</tr>
<tr>
<td>My job is excellent.</td>
<td>My job is dull.</td>
</tr>
<tr>
<td>I have the best job in the world.</td>
<td>I don't enjoy going to work.</td>
</tr>
<tr>
<td>I can't think of a better job than mine.</td>
<td>I sometimes wish I had a better job.</td>
</tr>
<tr>
<td>My job is the most enjoyable part of my life.</td>
<td>My job could be better.</td>
</tr>
<tr>
<td>My job makes me feel glad to be alive.</td>
<td>I don't really like my job very much.</td>
</tr>
<tr>
<td>I am in love with my job.</td>
<td>Moderate Positive Satisfaction Items</td>
</tr>
<tr>
<td></td>
<td>I enjoy my job.</td>
</tr>
<tr>
<td></td>
<td>I have interesting things to do at work.</td>
</tr>
<tr>
<td></td>
<td>My job is rewarding.</td>
</tr>
<tr>
<td></td>
<td>My job is pleasant.</td>
</tr>
<tr>
<td></td>
<td>My job is enjoyable</td>
</tr>
<tr>
<td></td>
<td>I am happy with my job.</td>
</tr>
<tr>
<td></td>
<td>I have a decent job.</td>
</tr>
<tr>
<td></td>
<td>I have a meaningful job.</td>
</tr>
</tbody>
</table>
Initially we generated items of all four types and compiled the best 60 items into a list. Two of the authors (Paul E. Spector and Paul T. van Katwyk) independently sorted the items into the four categories of extremely negative, moderately negative, moderately positive, and extremely positive. They disagreed on five items which were replaced. They then independently sorted the new items and agreed completely on their categories. The list of 60 items was administered to eight industrial/organizational psychology doctoral students who were asked to sort them into the four categories. We kept eight of each type of item which were most consistently placed into their category. For 24 items, there was perfect agreement, for seven items seven of eight judges agreed, and for one item six of eight judges agreed. There was perfect agreement for all 16 of the extreme items. The 32 items can be seen in Table 1.

The items were placed into a questionnaire format with six response choices ranging from “Strongly disagree” to “Strongly agree.” An additional question asked subjects to indicate their overall feeling about the job using a scale ranging from 0 (hate the job as much as possible) to 10 (love the job as much as possible). We divided the scale into approximate thirds according to the possible range of the scale, and subjects were categorized as dissatisfied (scored 0 to 3), neutral (scored 4 to 6), and satisfied (scored 7 to 10).

The questionnaires were administered to 138 employed undergraduate students at the University of South Florida. This university is an urban school with mainly commuter students. A sizable proportion of students are above traditional age, as reflected in the mean graduation age of 26. The majority of the student population at this university works at least part time, with many of the students working full time and attending school part time. The number of subjects falling into the dissatisfied, neutral, and satisfied groups were 16, 45, and 77, respectively.

Item analyses were conducted for the 16 extreme and 16 moderate items separately, as well as for all 32 items combined. The 16 extreme items had an internal consistency reliability (coefficient alpha) of .93 with item remainder correlations of .38 to .83. The 16 moderate items had a coefficient alpha of .94 with item remainder correlations of .49 to .85. All 32 items had an internal consistency reliability of .97 with item remainder correlations ranging from .43 to .87. Two of the 32 items would have slightly raised the coefficient alpha, if removed (by .0001). One was an extreme negative item and the other was a moderate positive item. These results clearly show that combined or separately, the two types of items produced internally consistent scales.

For each subject, the mean response across the eight items of each type was computed. This resulted in four scores per subject, one for each item type. Means were then computed across subjects in each satisfaction group for each item type, resulting in 12 means. These means represented responses to each of the four types of items by each satisfaction group.

Figure 2 is a plot of mean agreement (disagreement) as a function of item type, ranging from extreme negative to extreme positive for each satisfaction group separately. The horizontal line across the center of the graph at a value of 3.5 represents the midpoint of the scale. The shapes of the three curves (one for
each satisfaction group) were an inverted U. For the dissatisfied group, the highest agreement was for the moderately negative items, which was the only type of item for which there was a mean above 3.5 (the midpoint of the scale). As can be seen, this group disagreed with all other item types, including the extreme negative. Similarly, the satisfied group mean was in the agree range (over 3.5) only for the moderate positive items. The neutral group means fell between the means of the other two groups for all items. They tended to be close to 3.5 for the moderate items, and were well below 3.5 into the disagree range for the extreme items.

The moderate items tended to behave as one would expect items to behave. The satisfied group agreed with the positive and disagreed with the negative items, while the dissatisfied group agreed with the negative items and disagreed with the positive. The neutral group fell in the middle, close to the middle of the scale on both types of items. For the extreme items, means for all three groups were to the disagree side of 3.5. Even the dissatisfied people tended to disagree with extreme negative items, and even the satisfied people tended to disagree with extreme positive items. What is not shown in the figure is that 60 of the 77 satisfied subjects disagreed with the extreme positive items, and, likewise, 10 of the 16 dissatisfied subjects disagreed with the extreme negative items. Clearly, people will disagree with items that are more extreme than their standing on the trait of interest.

![Figure 2](image-url)
Figure 3 shows item response curves for the four types of items, using the same means as shown in Figure 2. In this figure, item agreement is a function of people's overall satisfaction score rather than item type. This figure shows the lack of symmetry between the two extreme item types. The curves for all four item types were relatively linear, with positive slopes for positive items and negative slopes for negative items. With moderate items, there is a symmetry around the neutral point (about 3.5 on the scale from 1 to 6). Corresponding agreement and disagreement is about evenly far from 3.5 for each group (see points a+ vs. a− for the dissatisfied group, and b+ vs. b− for the satisfied group). This means that disagreement with a negative item is equivalent to agreement with a positive item, which makes it reasonable to reverse score items in one direction. For extreme items, the curves were completely below the middle point of 3.5 on the agreement continuum, but the slopes were similar to those of the moderate items. The major difference in the curves for the two types of items is that both extreme items are shifted down the vertical axis. That is, neither of the extreme item types are frequently endorsed. This means that the reverse scoring of negatively worded extreme items will not produce an equivalent score to the extreme positive items for most respondents. This is exactly the pattern of responses that would theoretically be expected with items that are extreme relative to the underlying scores of a sample of subjects.

Figure 3. Item Characteristic Curves for Four Types of Job Satisfaction Items
Study 2: Item Characteristics Produce Factors

Study 1 showed that extreme items can produce patterns of responses whereby individuals might disagree with items written at opposite ends of a continuum. However, we also must show that such response patterns can produce artifactual factors. We could do this by showing a factor analysis of the Study 1 data or some other data set taken from real subjects. However, with real subjects we cannot be absolutely certain what the structure of the underlying constructs might be. One could argue that it is the dual nature of job satisfaction that produced the factors rather than an artifact. To eliminate this possibility we conducted a computer simulation where data were generated to represent a single underlying continuum.

To demonstrate how item characteristics (direction and extremity) can produce factor structures, we used equations to model subject behavior. Item functions were used to represent items that were either extreme or moderate at both ends of an underlying measurement continuum. Data were generated to represent subject responses, including both the "true" score component and measurement error. Data from 1600 simulated subjects were factor analyzed using both confirmatory and exploratory methods to show the effects of item characteristics on a factor structure for both extreme and moderate items.

The item characteristic functions (equations) were created to relate the degree of agreement with an item to the underlying trait score of an individual. The agreement continuum ranged from 1 (strongest disagreement) to 6 (strongest agreement), with neutrality falling between 3 and 4 at 3.5. The trait continuum ranged from 1 (strongest negative) to 6 (strongest positive) for the true score of an individual. Each of the four types of items, extreme negative, moderate negative, moderate positive, and extreme positive was represented by a function that related the extent of agreement with an item to an individual's true score on the underlying trait.

For the moderate items, the functions were straight lines with either negative or positive slopes, depending upon the direction of the item. For a moderate positive item, a person who is high on the trait should strongly agree and a person who is low on the trait should strongly disagree. There was perfect correspondence between item score and trait score. The function equated scores on both continua, as in:

Agreement = Trait.

For a moderate negative item, the relation between the trait and agreement was reversed. A person who is high on the trait should strongly disagree with the item, and a person who is low on the trait should strongly agree with the item. The function is:

Agreement = 7 - Trait.

This resulted in an agreement score of 1 when the true score was 6, 2 when the true score was 5, 3 when the true score was 4, 4, when the true score was 3, 5
when the true score was 2, and 6 when the true score was 1. Both functions for the moderate items are illustrated in Figure 4.

Extreme items had somewhat more complex functions than the moderate items. With extreme items in either direction, most individuals will disagree. Only the most extreme individuals will agree with items that are written in a direction that matches their trait level (negative or positive). This results in functions that are linear in part of their range and flat in the remaining part of their range. Functions for an extreme negative and an extreme positive item are illustrated in Figure 5. For the positive item, individuals with trait scores ranging from 1 to 4 had agreement scores of 1. The function relating agreement to trait was linear for individuals with trait scores greater than 3. Agreement scores of 2 corresponded to trait scores of 5, and agreement scores of 3 corresponded with trait scores of 6. Thus, only the most extreme subjects on the trait would agree with the item. This pattern matches that of the extreme items in Study 1. The extreme item functions used in the simulation were the following for the negative and positive items, respectively:

\[
\begin{align*}
\text{Agree} &= \text{Trait} - 3 \\
\text{Agree} &= 4 - \text{Trait}
\end{align*}
\]
where values of Agree less than 1 were made equal to 1.

Each subject’s response in the simulation had both a trait and random error component. The trait component was derived from the item functions. The random error component was drawn from a normal distribution with a mean of 0 and standard deviation of 1, and was computed with the random number generator function (RANNOR) from the SAS language. The distribution of trait scores across cases was symmetrical, with 100 cases having values of 1 and 6, 200 cases having values of 2 and 5, and 500 cases having values of 3 and 4. Each generated agreement score for each case was rounded to its nearest integer value between 1 and 6.

A computer program written in the SAS language was used to generate data for 1600 cases on eight items, four written in each direction. One simulation modeled extreme items and the other modeled moderate items. To test for the existence of a one- or two-factor structure for each simulation (extreme or moderate items), confirmatory factor analyses (CFAs) were conducted with LISREL 6 (Joreskog & Sorbom, 1984). Each model contained two factors, one containing all negatively worded items and the other containing all positively worded items. By fixing or freeing the estimated correlation between the two factors, as estimated in the phi matrix, it was possible to conduct a nested models test of whether the data were best represented by a one- or two-factor structure.
For the extreme items, it was hypothesized that a two-factor structure would underlie the data. The two factors are not expected to be orthogonal, but might correlate to a moderate degree. For the moderate items, a one-factor structure was expected to best fit the data. In both cases the same three models were tested. First, a model was tested in which the correlation between the two factors was estimated (the oblique factor model). Next, a model was tested in which the correlation between the two factors was constrained to be $-1.0$ (the one-factor model). Finally, a model was tested in which the correlation between the two factors was constrained to be 0 (the orthogonal factor model). For the extreme items, the analysis should indicate that the oblique or orthogonal models would fit the data. For the moderate items, the one-factor model should fit the data, and with the oblique factor model, the correlation between the two factors should be approximately $-1.0$, indicating that they reduced to a single factor. Two factors that are perfectly correlated are equivalent to a single factor.

For the extreme items, the oblique two-factor model fit the data quite well ($\chi^2(19) = 6.9, p = .995, \text{GFI} = .999, \text{AGFI} = .998, \text{RMSR} = .009$) with an estimated correlation between the two factors of $-0.357$. The orthogonal factor model did not fit as well, although the goodness of fit statistics were well above the .90 criterion for good fit ($\chi^2(20) = 129.21, p = .000, \text{GFI} = .981, \text{AGFI} = .957, \text{RMSR} = .106$). The chi square test for nested model comparison suggested that the oblique model fit significantly better than the orthogonal ($\chi^2(1) = 122.31, p = .000$). The single factor model was forced by fixing the correlation between the two factors to $-1.0$. This one factor model fit quite poorly ($\chi^2(20) = 1085.55, p = .000, \text{GFI} = .792, \text{AGFI} = .533, \text{RMSR} = .124$). The nested models test comparing this model to the oblique was significant ($\chi^2(1) = 1078.65, p = .000$), indicating a significantly better fit of the oblique two-factor model over the one-factor model.

For the moderate items, the oblique factor model fit the data very well ($\chi^2(19) = 23.17, p = .207, \text{GFI} = .996, \text{AGFI} = .992, \text{RMSR} = .009$), with an estimated correlation between the two factors of $-0.995$. For comparison, a single factor model was run with the two factors correlated $-1.0$. The fit was almost identical to the oblique factor model ($\chi^2(20) = 24.22, p = .233, \text{GFI} = .996, \text{AGFI} = .991, \text{RMSR} = .009$). The nested model comparison test for the oblique factor model versus the single factor model was nonsignificant ($\chi^2(1) = 1.05, p > .05$). These results suggest that the oblique factor model was essentially no different from the single factor model. An orthogonal two-factor model was forced on the moderate item data by fixing the correlation between the two factors to zero. The orthogonal two factor model fit very poorly ($\chi^2(20) = 1960.30, p = .000, \text{GFI} = .849, \text{AGFI} = .660, \text{RMSR} = .379$). Although it was not possible to compare the single factor to the orthogonal two-factor model, it was possible to compare the oblique to the orthogonal. This test was significant ($\chi^2(1) = 1937.13, p = .000$), indicating a significantly better fit of the essentially one-factor oblique model over the orthogonal two factor.

For each simulation, a principal axis factor analysis was computed with SAS Proc Factor. The maximum correlation of each item with other items was used as communality estimates, and oblique rotation (promax) was performed. The
purpose of this analysis was to show that an exploratory factor analysis would produce the expected number of factors. Extreme items were expected to produce two factors based on wording direction, and moderate items were expected to produce one-factor.

For the extreme items, the eigenvalues were 2.4213, 1.1526, .0499, .0450, .0321, .0245, -.0032, -.0118. The scree criterion suggested that there were two factors, and they accounted for 96% of the variance. Two-factors were rotated (see Table 2). The four negatively worded items had loadings ranging from -.02 to .01 on the first factor and loadings ranging from .61 to .67 on the second factor. Conversely, the positively worded items had loadings ranging from .62 to .67 on the first factor and loadings ranging from -.01 to .01 on the second factor. The two factors were correlated -.37. These results clearly indicate a two-factor structure consisting of the negatively and positively worded items, respectively. The moderately worded items produced eigenvalues 4.5847, .0787, .0478, .0342, .0282, .0004, -.0077, -.0188. The scree criterion clearly shows a single factor, which accounted for 97% of the variance. These results suggested that there was a very strong single factor among the moderately worded items.

**Discussion**

Our conclusion is that the relative independence of factors based on items with opposite wording directions may be artifactually produced by how people respond to items. Although computing separate subscale scores for items written in opposite directions along a seemingly unidimensional continuum might result in the assessment of different constructs, our simulation has shown that item characteristics can produce such structures.

As we illustrated above with job satisfaction items, the ideal point principle does apply when people respond to items that are extreme on the underlying continuum. In other words, many people disagree with items at both ends of a unidimensional continuum if those items are sufficiently far from their standing on that continuum. The result is that extreme items at opposite ends of a continuum have quite skewed distributions, with few people agreeing with them. As we discussed above, this produces low correlations between extreme items with

<table>
<thead>
<tr>
<th>Item wording direction</th>
<th>Factor 1</th>
<th>Factor 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Negative</td>
<td>-.02</td>
<td>.61</td>
</tr>
<tr>
<td>Negative</td>
<td>.01</td>
<td>.67</td>
</tr>
<tr>
<td>Negative</td>
<td>-.02</td>
<td>.61</td>
</tr>
<tr>
<td>Negative</td>
<td>.01</td>
<td>.66</td>
</tr>
<tr>
<td>Positive</td>
<td>.67</td>
<td>.01</td>
</tr>
<tr>
<td>Positive</td>
<td>.62</td>
<td>-.01</td>
</tr>
<tr>
<td>Positive</td>
<td>.64</td>
<td>.01</td>
</tr>
<tr>
<td>Positive</td>
<td>.62</td>
<td>-.01</td>
</tr>
</tbody>
</table>

Table 2. Rotated Factor Loadings For Extreme Items
opposite wording direction and high correlations between items with the same wording direction. The result when factor analyzed is a two factor structure with items of each wording direction loading on separate factors.

Our results provide an alternative explanation for the existence of two constructs underlying the two factors found in several domains. For example, affect scales that contain items reflecting both negative and positive emotions typically form factors based on direction of item wording. This consistent finding has led some researchers to conclude that negative and positive affect represent orthogonal constructs. Watson, Clark, and Carey (1988) concluded that:

> Although their names might suggest that they are opposite poles of the same dimension, Positive and Negative Affect are in fact highly distinctive dimensions that can be meaningfully represented as orthogonal (uncorrelated) factors (p. 347).

**How To Distinguish Artifact From Construct**

When two factors are detected in a set of oppositely worded items, several procedures can be used to determine if the factors are more likely artifactual or substantive. Since no single procedure is conclusive by itself, several should be applied to build a case for either artifact or substance. The mechanism that we have discussed tends to produce skewed distributions of item responses. The process simulated here also produces joint distributions whereby some respondents score low on both negatively and positively worded items and few if any score high on both. A check for both of these tendencies in the data can be suggestive of a possible artifact.

Factors that come from normally distributed item data that have perfectly homoscedastistic, linear joint distributions are unlikely to have arisen from the artifact. Finding data that do not conform to distributional assumptions for parametric statistics indicates the possibility of an artifact. Further tests should be conducted to determine if the artifact is the most likely explanation. Of course care should be taken in interpreting results of factor analyses when the underlying variables violate distributional assumptions, and some have argued that such analyses are inappropriate in these circumstances (e.g., van Schuur & Kiers, 1994).

Van Schuur and Kiers (1994) reviewed procedures that can be used to determine if a sample of item data would fit an unfolding model consistent with our explanation. Some of these procedures require that responses be converted to two choice formats (e.g., yes or no). Some also require the use of statistical procedures that are not contained in the major computer packages, such as SAS or SPSSX. Perhaps one of the simplest tests requires a scaled set of items that vary on the underlying construct of interest. For example, in Study 1 the items were placed in four categories from extreme dissatisfaction to extreme satisfaction. The items are sorted from low to high on the construct. A correlation matrix of the items is computed with the items in the sorted order. Those items that are closest together in their scale value should be most highly intercorrelated. This produces a simplex pattern in the matrix whereby the magnitude of correlations decreases with distance from the main diagonal (as you go down and toward the left). To apply this proce-
WHEN TWO FACTORS DON'T REFLECT TWO CONSTRUCTS

One would probably have to generate items that vary on the underlying construct of interest, since most scales do not have a mixture of item scale values. It can be helpful in showing that the artifact is a likely explanation, although other response models can also produce a simplex pattern (van Schuur & Kiers, 1994).

One should be cautious in making a case for two substantive factors when constructs are confounded with wording direction. Since the artifact of item wording direction can produce factors, one must provide strong evidence that the factors are substantive. This involves both ruling out the artifact using the procedures we discussed, as well as validating both substantive measures. Relying only on the correlations of the new scales with other scales is not sufficient because the artifact that produces factors can produce a pattern of relations with other scales. In using scales as criteria for validation, care must be taken that the content and wording direction of the criterion is not more like one of the new scales than the other. For this reason, much of the evidence cited in support of the viability of NA and PA is suspect.

Evidence cited in support of the two construct hypothesis for NA and PA has been the somewhat different pattern of correlations with criteria (Watson, Clark & Tellegen, 1988). One must look carefully at the nature of the criteria which elicit different correlations. Self-report scales that tend to correlate more strongly with NA than PA and which provide the majority of criterion related evidence for a dual construct hypothesis have content that reflects negative affect such as depression, stress, negative mood, or somatic symptoms (e.g., Watson, Clark & Carey, 1988; Watson, Clark & Tellegen, 1988). Thus, many of these scales that have been shown to correlate more strongly with NA than PA may simply have items that also reflect negative emotions and experiences.

Examples that provide support for an argument based on similar item content can be found by considering the items on the Beck Depression Inventory (BDI, Beck, Ward, Mendelson, Mock & Erbaugh, 1961), Hopkins Symptom Checklist (HSCL, Derogatis, Lipman, Rickels, Uhlenhuth & Covi, 1974) and the A-State scale (Spielberger, Gorsuch & Lushene, 1970), which have been found by Watson, Clark, and Tellegen (1988) to correlate more strongly with the NA than PA subscale of the PANAS. The BDI overlaps the PANAS NA scale by including items concerning guilt and irritability; the HSCL overlaps the PANAS NA scale by including items concerning irritability, guilt, and feeling nervous; and the A-State scale overlaps the PANAS NA scale by including items concerning being upset, nervous, and jittery. In addition, these scales contain words and phrases which reflect the same idea through somewhat different words, such as afraid, anxious, and nervous. Since these scales share content both conceptually and semantically, and reflect negative emotions and experiences, it is not surprising that they correlate. Interestingly, Marshall et al. (1992) claimed as evidence for the validity of the two factors of the LOT, differential correlations with measures of NA and PA. The sum of the optimism (positive) items correlated more strongly with PA than NA, and the sum of the pessimism (negative) items correlated more strongly with NA than PA.

Required are studies that go beyond scales and use alternative means of validation. With affect one might develop a series of differential hypotheses that NA and PA would predict different behaviors. One might also relate scores on the
scales to appropriate physiological measures. The establishment of separate nomological networks for each construct should serve as the basis for their acceptance. The factor analysis results suggest what might be rather than what is concerning the existence of constructs.

Conclusions

Although we have suggested that oppositely worded items can produce artefactual factor structures, we are not suggesting that they should always be avoided. It should be noted that the use of extreme items in bipolar scales of unidimensional constructs is often necessary despite the creation of two seemingly independent factors. Extreme items are necessary when one intends to distinguish individuals who are extreme on the construct from those who are moderate. In fact, it may be preferable to use extreme items in scales that have traditionally used primarily moderate items as it would reduce the possibility of basement and ceiling effects.

The idea that orthogonal factors comprised of oppositely worded items reflects a methodological artifact has been suggested for several popular scales in the organizational domain (e.g., Cordery & Sevastos, 1993; Idaszak & Drasgow, 1987; Magazine et al., McGee et al., 1989; Schmitt & Stults, 1985; Spector, 1992). However, in the study of affect, the interpretation of independent constructs has gained considerable prominence. We argue that the two factors found in affect scales are probably produced artefactually by how subjects respond to items, rather than by independent constructs. This artifact is likely at work when distributions of item responses are very skewed with the majority of subjects disagreeing. We question the suggestion that constructs such as NA and PA, or optimism and pessimism, are independent. To establish the viability of positive and negative item subscale scores as independent constructs will require stronger evidence than what has been advanced to date.

Acknowledgment: Requests for reprints should be addressed to Paul E. Spector, Department of Psychology, University of South Florida, Tampa, FL 33620. We thank Kevin Kelloway, Chuck Lance, and Nambury S. Raju for their helpful comments.

References


