A Thesis
Submitted to the Faculty
of
Xavier University
in Partial Fulfillment of the
Requirements for the Degree of
Master of Arts
By
Kara Jones
January, 2011

Approved:

Karl W. Stuckenberg, Ph.D., ABPP
Chair, Department of Psychology

Mordell Mullins, Ph.D.
Thesis Chair
The Effect of Item Stem Length on Face Validity,
Perceptions of Fairness, and Involvement
<table>
<thead>
<tr>
<th>Role</th>
<th>Member</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chair</td>
<td>Morell Mullins, Ph.D.</td>
</tr>
<tr>
<td></td>
<td>Associate Professor of Psychology</td>
</tr>
<tr>
<td>Member</td>
<td>Mark Nagy, Ph.D.</td>
</tr>
<tr>
<td></td>
<td>Associate Professor of Psychology</td>
</tr>
<tr>
<td>Member</td>
<td>Eric Barrett, M.A.</td>
</tr>
<tr>
<td></td>
<td>Adjunct Professor of Psychology</td>
</tr>
</tbody>
</table>
Acknowledgements

I would like to thank my committee for all of their help and great ideas. I am especially grateful of my advisor Morrie, for all of the support and assistance that he has given me during this process. My time at Xavier was made successful and meaningful because of the relationships I have developed with my professors and my cohort. It means the world to me that I have been able to grow professionally in an environment surrounded by wonderful individuals.

Finally, I would like to thank my family for all of the love and encouragement that they have given me throughout my education. Nothing I have accomplished would have been possible without them.
Table of Contents

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Acknowledgements</td>
<td>i</td>
</tr>
<tr>
<td>Table of Contents</td>
<td>ii</td>
</tr>
<tr>
<td>List of Tables</td>
<td>iii</td>
</tr>
<tr>
<td>List of Appendices</td>
<td>iv</td>
</tr>
<tr>
<td>Abstract</td>
<td>v</td>
</tr>
<tr>
<td>Chapter</td>
<td></td>
</tr>
<tr>
<td>I.  Review of Literature</td>
<td>1</td>
</tr>
<tr>
<td>II. Rationale and Hypotheses</td>
<td>12</td>
</tr>
<tr>
<td>III. Methods</td>
<td>14</td>
</tr>
<tr>
<td>IV. Results</td>
<td>18</td>
</tr>
<tr>
<td>V. Discussion</td>
<td>21</td>
</tr>
<tr>
<td>VI. Summary</td>
<td>28</td>
</tr>
<tr>
<td>References</td>
<td>40</td>
</tr>
<tr>
<td>Appendices</td>
<td>45</td>
</tr>
</tbody>
</table>
List of Tables

<table>
<thead>
<tr>
<th>Table</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Descriptive Statistics, intercorrelations, and reliabilities of scales</td>
<td>20</td>
</tr>
</tbody>
</table>
List of Appendices

A. Personality Measure Form 1 ................................................................. 45
B. Personality Measure Form 2 ................................................................. 46
C. Face Validity Measure and Perceptions of Fairness Measure ................. 47
D. Involvement Measure ........................................................................ 48
E. Demographic Information .................................................................. 49
F. Hypothetical Situation ......................................................................... 50
G. Informed Consent Form ....................................................................... 51
H. IRB Approval Letter ........................................................................... 52
Abstract

The present study examined the effects of item stem length on face validity, perceptions of fairness, and involvement. A single word item stem length personality questionnaire was modified to have sentence length item stems. No significant differences were found between the two measures on the reaction variables. The measures were also found to be psychometrically similar in their reliabilities and factor structures, although the factors were dramatically different from what was anticipated. These findings could indicate that non-content changes, such as item stem length modifications, could have little effect on reactions and the psychometrics of rating scales. Further research should examine the reaction effects on other types of questionnaires as well as their factor structures.
Chapter I

Review of Literature

In Industrial-Organizational (I-O) psychology, and most other fields of psychology, questionnaires (sometimes discussed as rating scales or surveys) are important measurement tools in research and applied areas. The creation of questionnaires helps to get results that determine performance ratings, hiring decisions, personality dimensions, job analysis, and integrity assessments, to name a few. Although the content of such tools is important, the questionnaires themselves have the potential to influence outcomes due to the format and manipulations of the format (Borman, Buck, Hanson, Motowild, Stark, & Drasgow, 2001). It is important in I-O psychology to explore all the variables that affect the questionnaires used to make essential decisions. The more we know about how differences in questionnaires can affect their outcomes, the better we can attain accurate, meaningful results.

The decisions made in scale construction could have perceptual consequences that influence respondents. The number of response categories, for example, has been found to have effects on perceptions of variables such as fairness and face validity (Mullins, Polson, Lanch, & Kehoe, 2006). Having a scale that is face valid can lead to cooperation and motivation as well as decrease dissatisfaction and feelings of injustice of low scorers, among other things (Nevo, 1985). In addition, perceptions of fairness can lead to positive perceptions of an organization (Hausknecht, Day, & Thomas, 2004). Another construct, involvement, while not thoroughly used in Industrial-Organizational Psychology research, is potentially important because it can show
how motivated a respondent is to process information about a given rating scale or issue (Johnson & Eagly, 1989).

The current study will attempt to explore differences in face validity, perceptions of fairness, and involvement based on item stem length. To achieve this, two parallel scales have been created that have either single-item stem lengths, small phrase item stem lengths, or full sentence item stem lengths. For example, the different lengths could be, “Happy,” “A happy person,” or “I consider myself a happy person.” It is predicted that the item stem lengths will create different perceptions of the scales. The following literature review will examine scale construction, face validity, perceptions of fairness, and involvement as they deal with rating scales.

**Scale Construction**

Scale construction has multiple stages, each of which, depending on the scale, involves multiple decisions. The main decisions of scale construction include the type of scale, scale format, and item format. Past research has included comparisons between scale types such as behavioral expectations scales and graphic rating scales (Keaveny & McGann, 1975), absolute versus relative rating formats (Roch, Sternburgh, & Caputo, 2007), effects of negatively worded items (Barnette, 2000), and response categories (Mullins et al., 2006) to name a few. Dawis (1987) reviewed different options in scale construction, including item format. The format of an item consists of two parts, the item stem and the response. The various topics in and approaches to scale construction research will be considered in more detail next.

Research on scale types has examined the effects that various scales have on accuracy. Keaveny and McGann (1975) researched the occurrence of halo effects (a tendency to rate based on overall impressions rather than attending to distinct performance dimensions), leniency (a
tendency to give higher ratings than might be warranted, based on performance), and the amount of discriminant validity on behavioral expectation scales (BES) and graphic rating scales (GRS). A BES consists of a scale with behavioral anchors, or examples, that demonstrate behaviors a ratee would be expected to exhibit. A GRS has trait labels, which are a trait or statement to be evaluated on, followed by a numerical rating. College students rated professors on 13 categories with a 9-point scale for either the BES or GRS. Neither scale type was found to have less leniency than the other, however the BES did demonstrate a reduction in halo effects. The BES was also found to have more discriminant validity than the GRS. In a similar study by Tziner (1984), a behaviorally anchored rating scale (BARS), which is made up of behavioral anchors that are examples of work behaviors and represents an extension of the BES methodology, was compared against a GRS for differences in halo effects, leniency, and interrater reliability. Managers from an aircraft plant rated employees on five dimensions with a 4-point scale using both the BARS and the GRS. The results showed the BARS to be less susceptible to halo effects, leniency, and have higher interrater agreement. In circumstances where decisions are based off of rating scales, choosing the scale that creates the most accurate results can aid to make sure that processes and decisions are deemed fair and justified.

Scale format has been researched by comparing absolute versus relative formats. Absolute formats require raters to rate against an absolute standard, as with BARS and GRS, and relative formats are more subjective to the situation, such as forced distributions and rankings. Roch et al. (2007) studied perceptions of fairness on absolute and relative rating scales. The study found that absolute rating formats were perceived as more fair than relative formats, not all absolute and relative formats were viewed as equally fair, the type of rating format influenced perceptions of procedural justice, and finally, the influence of rating format on procedural justice
perceptions depended on the level of distributive justice. These findings showed that while an organization may feel that a certain format is beneficial for rating, it is important to take into consideration respondents' perceptions of the scale.

Negatively worded items have also been a subject of format research. Barnette (2000) examined item stem directions and item response pattern direction effects on internal consistency reliability. Negatively worded items are those that are phrased in the opposite semantic direction from the other items in the scale. For example, on a personality questionnaire where most items do not contain any negative words (e.g., “not”), an item that contains such a word would be considered negative. Item response pattern direction was varied by going from strongly disagree (SD) to strongly agree (SA) and vice versa on a 5-point Likert scale. The scales tested were structured in a 2x3 framework where stem type was either all directly worded or half directly and half negatively worded. Response options were all SD to SA, all SA to SD, or mixed (half SD to SA and half SA to SD). The highest reliability resulted from the item stems all in the same direction but with the response options mixed. The authors speculated that although this may be seen as confusing for the respondents, it could be useful for detecting acquiescence or response set bias.

A main topic of research for item manipulations is the number of response categories. Previous research has determined the minimum number of response categories is four and the maximum between five and seven in order to have an appropriate reliability (Lozano, Garcia-Cueto, & Muniz, 2008). More recent research was concerned with similar aspects such as validity and perceptions of the number of response categories. Perceptions based on the number of response categories have been researched on integrity and personality measures by Mullins et al. (2006). The concerns of this study included perceptions of fairness, face validity, and
performance on each of the measures. For the integrity measure, response categories were either 2- or 5-point scales, where on the personality measure, the response categories were 2-, 5-, or 9-point scales. The research found that for the integrity measure, the 5-point scale was perceived as more fair, face valid, and participants viewed themselves as having performed better. However, for the personality measure, no differences were observed based on the number of response points on the scale.

While there have been a reasonable number of studies on scale construction, the literature is largely scattered. The above-reviewed literature is representative of studies done from the different areas of scale construction. To the best of the present author's knowledge, no studies have examined the effect of item stem length on respondent perceptions. This line of research could prove to be important regarding the perceptions of fairness, face validity, and involvement. Just as having too few response categories has an effect on these perceptions, having too small or too large of a description may influence these perceptions as well.

**Face Validity and Perceptions of Fairness**

Face validity has been defined as "the extent to which applicants perceive the content of the selection procedure to be related to the content of the job" (Smither, Reilly, Millsap, Pearlman, & Stoffey, 1993, p. 54). Nevo (1985) argued that face validity can create cooperation and motivation in participants, attract candidates, lower dissatisfaction and feelings of injustice for low scorers, convince policymakers, employers, and administrators to implement a test, and improve relations with mass media and courts. By creating scales that are face valid there may be more acceptance than is present when a scale that is not face valid is used.

Face validity is normally controllable by the test constructor and may be an important determinant of other reactions to rating scales (Chan, Schmitt, DeShon, Clause, & Delbridge,
1997). A test constructor has the ability to create items that can be distinguished as pertaining to
the scale or items whose true purpose is harder to discern. The decision about level of desired
face validity can lead to significant differences in how respondents react to a scale. Having high
face validity can increase rater motivation and effort (Bornstein, 1996) which in turn can

The face validity of a scale may reflect its perceived job relatedness, and can therefore
have implications for perceptions of fairness (Ployhart & Harold, 2004). However, while they
may be related, they are distinct concepts that need to be assessed separately. Face validity
concerns the scale itself without influences from external sources. Perceptions of fairness can be
influenced by how the scale is created or how outcomes will be distributed due to the scale. This
distinction could lead to differences in how each of these are rated thus there should be different
measures to assess their extent.

Perceptions of fairness are important because a participant’s feelings toward a measure
can have an effect on their attitudes and behaviors (Hausknecht et al., 2004; Roch et al., 2007).
This effect on attitudes and behavior due to a measure can have implications for organizational
commitment, job satisfaction, performance, organizational citizenship behaviors, and questions
over an organization’s policies (Roch et al.). Applicants will hold more positive perceptions about
an organization if a scale is seen as procedurally fair and job relevant (Hausknecht, et al., 2004).
In addition, if an applicant is aware of the type of selection process and believes the process to be
unfair, the applicant may be dissuaded from applying (Gilliland, 1993).

Three types of fairness have been a central part of perceptual research. These include
procedural justice, distributive justice, and interactional justice (Fassina, Jones, & Uggerslev,
2008; Gilliland, 1993; Kim & Leung, 2007). Procedural justice is the perceived fairness of
procedures that are used to establish outcomes that could include promotions, raises, or rewards. Distributive justice is the fairness of how an organization distributes outcomes. Finally, interactional justice is the fairness of treatment from an organization. For the purposes of this study, procedural justice will be focal. Studying procedural justice in the context of respondent reactions is consistent with past research (Maertz, Bauer, Mosley, Posthuma, & Campion, 2004; Truxillo, Bauer, Campion, & Paronto, 2002).

Procedural information can be gathered directly by assessing whether normative principles of process fairness have been violated or by making judgments where there is uncertainty or unclear information about whether fairness is occurring (Blader, 2007). Gilliland (1993) proposed that fairness is determined by several procedural justice rules such as job relatedness, consistency, and opportunity to perform. The extent to which these rules are violated or satisfied combine to make an overall fairness evaluation.

Blader and Tyler (2003) suggested that in addition to the content of procedural justice evaluations, the source of the procedural experience can influence judgments. The central sources of procedural justice evaluations are the official rules and procedures of an organization (formal) and the specific group who manages the employees’ work lives (informal) such as supervisors and heads of departments. Formal sources tend to be constant across time, people, and situations, as well as slow to change. Informal sources are dynamic and unique in that they vary depending on the employee, supervisor, and the relationship between them. Blader and Tyler propose that the source of justice has a distinct influence on overall evaluations of procedural justice. This means that employees recognize the differences between the sources and view each source as a separate influence on the overall judgment of fairness in the organization. An employee who can isolate the source of fairness or unfairness can have a negative reaction to
one source and a positive reaction to another. As such, while an employee may not feel that his or her supervisor is procedurally fair, an assessment or rating scale given by the organization or the supervisor can still be seen as fair.

Past research on procedural justice has included studies on the presence of fairness information and its effects on procedural justice. Truxillo et al. (2002) conducted a longitudinal study on this effect by measuring procedural fairness at four times in the application process. The four measurements occurred before the first selection test which was a multiple choice test (T1), before the second (video) test (T2), after the second test (T3), and when applicants received their results (T4). Information on the job relatedness and feedback timeliness was given just before the video test. Truxillo et al. found that applicants who were given information about the video test rated it as more fair at the time of testing and rated the selection process to this point as more positive one month later when they received their results. Providing information on the selection process to applicants is an inexpensive and relatively easy way to increase perceptions of fairness. This study shows that with the addition of this type of information, applicants can feel that the selection process is more fair even after the outcome of the process is given.

Procedural justice and organizational outcomes have been studied by Kernan and Hanges (2002). Their study investigated the relationship of survivor (an employee who is still part of an organization after layoffs or downsizing) perceptions of fairness and organizational outcomes after a restructuring initiative. The organizational outcomes focused on were organizational commitment, job satisfaction, turnover intention, and management trust. Findings of the study showed that procedural fairness was strongly related to all four of the organizational outcomes. These results illustrate how central procedural justice perceptions can be in an organization, especially when there are major changes in the organizational environment.
Involvement

Involvement has typically been an important topic in advertising research (Andrews, Durvasula, & Akhter, 1990). Although there has been little crossover into organizational psychology, involvement could be important in the perceptions of and engagement with rating scales. Involvement can be defined as, “an internal state of arousal with intensity, direction, and persistence” (Andrews et al., p. 28), and may affect respondent motivation to answer scale items accurately. Intensity refers to the degree of arousal with respect to the goal related behaviors. Direction of involvement is the target of the involvement intensity level. Finally, persistence is the duration of the involvement intensity. This definition emphasizes the participant focusing on a stimulus or goal, which is valuable when considering how respondents react to a rating scale.

Topics previously studied in I-O research concerning involvement include employee involvement (Cox, Zagelmeyer, & Marchington, 2006; Fenton-O’Creevy, 1998) and job involvement (Brown, 1996; Probst 2000). Employee involvement entails how an employee is involved in an organization, for instance their participation in organizing and carrying out their work (Fenton-O’Creevy, 1998). Job involvement entails the extent to which an employee identifies with their job (Probst, 2000). These constructs concern the employee after a relationship with the organization has been formed. Using involvement in the manner presented in this study, as a focus on a stimulus or goal, concerns the employee independent of a relationship with the organization, such as an application process. This could expand the I-O research literature to include another dimension of involvement to be used in the context of assessments.

Three main types of involvement are value relevant involvement, outcome relevant involvement, and impression relevant involvement (Cho & Boster, 2005). Value relevant
involvement is the psychological state created by the activation of attitudes that are linked to important values. This type of involvement is associated with attitude extremity such that those high on value relevant involvement are difficult to persuade. For example, if a person has strong feelings about politics even a solid argument against their views will not likely change their thoughts. Outcome relevant involvement refers to a condition that has future consequences for the respondent. High outcome relevant involvement fosters evaluation of information and issue relevant thinking. For example, if a person is buying a car they will be more likely to listen to and evaluate information on cars that the individual may not have liked initially. Impression relevant involvement is the concern with public perception of the self. This type of involvement represents a motivation to behave in a manner that is adaptable to the expectations of others. For example, in a discussion a person that has certain feelings about the legal drinking age may not voice their opinions for fear that others may not approve.

Outcome relevant involvement could be important to rating scales because it is concerned with currently important goals and outcomes (Petty & Cacioppo, 1990). Determining how involved a respondent is with a rating scale could help create more engaging measures. By making a respondent more involved with a scale they may put in more effort and concentration which could help with performance or willingness to complete the scale.

A main topic of outcome relevant involvement research has been the effects of involvement on persuasion. Involvement has been argued to increase a message recipient’s motivation to process information about the given issue (Johnson & Eagly, 1989). As the intensity of processing increases, for example, the personal importance of the message, the quality of the argument accounts for more change in attitude (Petty & Cacioppo, 1990). In a meta-analysis conducted by Johnson and Eagly, the strength of an argument facilitated
persuasion such that outcome relevant involvement increased persuasion with strong arguments and inhibited persuasion with weak arguments. Therefore, messages that contain weak arguments seem to elicit unfavorable thoughts while messages containing strong arguments elicit favorable thoughts.

Personal importance is valuable in rating scales and can increase involvement due to potential consequences for the respondent. A rating scale that is personally important, such as one used for hiring, also needs to provide a strong “argument” for its use to generate favorable attitudes toward the scale and potentially the organization. Arguments are about the presentation of information, and it may be that the strength of an argument is partially reflected in the amount of perceived information. By adjusting the number of words in an item stem we can potentially change the level of perceived information. The increase in perceived information could therefore lead to increased involvement. If the items in a scale appear to provide insufficient information, less involvement could result.

Changes in item length could tell us if the size of the items is related to respondent perceptions of the scale as a whole. The length of an item could lead to perceptions of being too ambiguous or too detailed to be answerable in a confident manner. In addition, if a certain optimal length makes participants think and be more engaged in rating scales, this could lead to more accurate results.
Chapter II
Rationale and Hypotheses

Although each variable has different criteria, determining face validity, perceptions of fairness, and involvement all have ties to the amount of information available. Overall lack of information could lead to lower perceptions of a rating scale, while more information may lead to more positive perceptions.

A major part of perceptions of face validity is whether the contents of a scale are job-related (Smither et al., 1993). Single word item stem lengths have the least content so it may be harder for respondents to assess the face validity of a scale that has such short items. Full sentence item stem lengths hold the most content and are more so transparent than an item with less information. Therefore, the first hypothesis is as follows:

*Hypothesis 1:* Participants who receive the measure with single word item stem lengths will score the lower on face validity perceptions than those receiving the full sentence item stem lengths.

Uncertainty and lack of information can lead to negative evaluations of procedural fairness (Blader, 2007). Single word item stem lengths hold little information and could be seen as lacking explanation of what the scale is asking the respondent. Full sentence item stem lengths appear to give more information than single word and small phrase item stem lengths, so could reduce uncertainty and give a better explanation as to what the item is asking. Thus, the second hypothesis is:
Hypothesis 2: Participants who receive the measure with single word item stem lengths will score the lower on perceptions of fairness than those receiving the full sentence item stem lengths.

More involvement can increase a respondent’s motivation to process information (Johnson & Eagly, 1989). The amount of perceived information in an item stem can influence how involved respondents are in the content of a rating scale. If there is a lack of information, such as in a single word item stem length, there may be less involvement. By having more information, such as in a full sentence item stem length, there may be more involvement. For this reason, this study proposes the following:

Hypothesis 3: Participants who receive the measure with single word item stem lengths will score lower on the involvement scale than those receiving the full sentence item stem lengths.
Chapter III

Methods

Participants

A total of 122 participants, 63 in the single-word item stem version and 59 in the full sentence item stem version, took part in this study. Of the participants, 60.7% were female, 78.7% Caucasian, 6.6% African American, 4.9% Hispanic, 3.3% Asian, 2.4% Biracial, and 4.1% other. Only 8 participants reported not having a job in the last five years while 45.9% had 1-2 jobs, 35.4% had 3-4 jobs, and 13.1% had 5 or more jobs. Of those who had jobs in the past 5 years, 82 had worked in multiple industries and 22 had worked in an industry that was not food service, office work, manufacturing, or retail. Recruitment was conducted through the university’s participant pool and via direct recruitment from specific classes.

Measures

Personality measure. A Big Five personality measure was used as the scale to be rated. This measure, the Mini-Markers, is a 40-item adjective-based measure based on the work of Saucier (1994). Two forms of this measure were utilized. The first form used was the original, single-word item stem version (Saucier, 1994) (see Appendix A) and the second was adapted to create a full sentence form (see Appendix B). The response options were the original 9-point scale ranging from 1 (Very inaccurate) to 9 (Very accurate). Saucier found the single-word stem length form to produce internal consistency reliabilities of .76 (Extraversion), .79 (Conscientiousness), .62 (Agreeableness), .59 (Openness), and .63 (Emotional Stability). This and all other copyrighted measures are contained in a separate document. The present study
found the internal consistency reliabilities for the single word questionnaire to be; .86 (Extraversion), .83 (Conscientiousness), .83 (Agreeableness), .63 (Openness), and .76 (Emotional Stability). For the full sentence questionnaire reliabilities were found to be .86 (Extraversion), .84 (Conscientiousness), .86 (Agreeableness), .71 (Openness), and .70 (Emotional Stability).

Face validity measure. Face validity was measured using a 3-item scale adapted from Chan et al. (1998) (see Appendix C). The response options utilized were a 5-point Likert scale ranging from 1 (Strongly Disagree) to 5 (Strongly Agree). Cronbach’s alpha was reported to be .79 (Chan et al., 1998). An example of an item on this measure would be, “I can see how the focus of this survey is related to what is expected of employees.” In the present study Cronbach’s alpha was found to be .73.

Perceptions of fairness measure. Perceptions of fairness was measured using a 3-item scale adapted from Chan et al. (1998) (see Appendix C). The response set utilized was a 5-point Likert scale ranging from 1 (Strongly Disagree) to 5 (Strongly Agree). Cronbach’s alpha was reported to be .79 (Chan et al., 1998). An example of an item on this measure would be, “I feel that I was able to report my true beliefs.” In the present study Cronbach’s alpha was found to be .43. When the first item of this scale is removed, alpha was raised to .83. All analyses were conducted using the original version of this scale.

Involvement measure. Involvement was measured by a 5-item measure adapted from Cho and Boster (2005) (see Appendix D). Responses were made on a 7-point Likert scale ranging from 1 (Strongly Disagree) to 7 (Strongly Agree). Cronbach’s alpha was reported to be .82. Construct validity was demonstrated by correlating outcome relevant involvement with information seeking, r(281) = .39, p < .001, and also conducting a regression analysis with same
variables, $\beta = .23, p = .007$ (Cho & Boster). In the present study Cronbach’s alpha was found to be .73.

**Demographic measure.** Demographic information pertaining to age, gender, ethnicity, and a description of work experience was collected (see Appendix E). The description of work experience included whether the participant has held a job, if they have had three or more jobs in the past five years, the average length of service, and the type of work experience (e.g. retail, manufacturing, office work). This was collected by asking the participants in a questionnaire format. The demographic information collected was intended to test for any differences that may emerge based on non-study factors, although this was not anticipated. In addition, an open-ended question created for this study was included to gather qualitative information as to reasons why participants perceived the personality measure the way they did. The question was, “What factors were your reactions to the personality inventory based on?”

**Procedure**

The study utilized a two group between-subjects design. The groups had identical instructions and a hypothetical scenario (see Appendix F) as well as completed a consent form (see Appendix G) and a demographic information questionnaire. The scenario described the application process of a hypothetical employer. It described what type of employment the participant will be seeking and the duties involved in that job. Each of the measures had an identical opening paragraph detailing how respondents are to answer the items. For example, “Read each item and answer on a scale of 1-9 how much it describes you.” Data were collected through the SurveyMonkey online survey tool and separate survey URLs for each condition were used. Participants signed up for the study through the University Participant Pool. Two tear sheets, individual pages with precut tabs at the bottom, were at the sign up site and each had a
different URL on its tabs. Participants tore off the tab for one of the two tear sheets, thereby randomizing themselves.
Chapter IV

Results

In order to test the hypotheses, three independent samples t-tests were conducted to compare the single word questionnaire and the sentence-length questionnaire. The first t-test assessed the differences in perceptions of face validity based on item stem length (Hypothesis 1), the second assessed the differences in perceptions of fairness (Hypothesis 2), and the third assessed the differences in level of involvement (Hypothesis 3).

The test for Hypothesis 1 was non-significant, t(120) = -0.81, ns. The mean score of face validity for the single word rating scale was 3.60 (SD=0.77) and the mean score for the full sentence rating scale was 3.70 (SD=0.70). The test for Hypothesis 2 was non-significant, t(120) = -1.01, ns. The mean score of perceptions of fairness for the single word rating scale was 3.40 (SD=0.67) and the mean score for the full sentence rating scale was 3.53 (SD=0.67). The test for Hypothesis 3 was non-significant, t(120) = -0.58, ns. The mean score of involvement for the single word rating scale was 4.93 (SD=1.03) and the mean score for the full sentence rating scale was 5.04 (SD=0.95).

For the personality questionnaire, mean scores within each of the factors were similar. For the single word rating scale the mean for Openness was 6.84, Conscientiousness was 7.08, Extraversion was 6.20, Agreeableness was 7.48, and Emotional Stability was 6.03. For the full sentence rating scale the mean for Openness was 6.71, Conscientiousness was 7.20, Extraversion was 6.09, Agreeableness was 7.31, and Emotional Stability was 6.30.
No significant effects or interactions were detected for gender or ethnicity on any study variables. In addition, no significant test results were found when the first item of the face validity scale was removed. The intercorrelations of all study variables are included in Table 1.
Table 1

Descriptive Statistics, Intercorrelations, and Reliabilities of Scales

<table>
<thead>
<tr>
<th>Single Word</th>
<th>Mean</th>
<th>SD</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Face Validity</td>
<td>3.60</td>
<td>0.77</td>
<td>.73</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2. Perceptions of Fairness</td>
<td>3.41</td>
<td>0.68</td>
<td>.46**</td>
<td>.43</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3. Involvement</td>
<td>4.94</td>
<td>1.03</td>
<td>.19</td>
<td>.40**</td>
<td>.73</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4. Gender</td>
<td>1.54</td>
<td>0.50</td>
<td>.01</td>
<td>.10</td>
<td>.13</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5. Ethnicity</td>
<td>1.17</td>
<td>0.38</td>
<td>.06</td>
<td>.14</td>
<td>.18</td>
<td>.17</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6. Had Job</td>
<td>1.02</td>
<td>0.13</td>
<td>.07</td>
<td>-.14</td>
<td>.03</td>
<td>-.14</td>
<td>-.06</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7. Jobs In Last Five Years</td>
<td>2.73</td>
<td>1.58</td>
<td>.05</td>
<td>-.06</td>
<td>.05</td>
<td>-.08</td>
<td>.16</td>
<td>-.22</td>
<td></td>
<td></td>
</tr>
<tr>
<td>8. Length of Jobs</td>
<td>3.02</td>
<td>1.06</td>
<td>-.14</td>
<td>.04</td>
<td>-.02</td>
<td>.07</td>
<td>-.13</td>
<td>-.37**</td>
<td>.43**</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Full Sentence</th>
<th>Mean</th>
<th>SD</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Face Validity</td>
<td>3.71</td>
<td>0.70</td>
<td>.73</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2. Perceptions of Fairness</td>
<td>3.53</td>
<td>0.67</td>
<td>.30*</td>
<td>.43</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3. Involvement</td>
<td>5.04</td>
<td>0.95</td>
<td>.15</td>
<td>.12</td>
<td></td>
<td>.73</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4. Gender</td>
<td>1.68</td>
<td>0.47</td>
<td>.18</td>
<td>.02</td>
<td>.08</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5. Ethnicity</td>
<td>1.25</td>
<td>.44</td>
<td>.30**</td>
<td>-.09</td>
<td>.15</td>
<td>.07</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6. Had Job</td>
<td>1.10</td>
<td>0.31</td>
<td>.09</td>
<td>-.02</td>
<td>-.06</td>
<td>-.13</td>
<td>-.07</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7. Jobs In Last Five Years</td>
<td>2.51</td>
<td>1.79</td>
<td>-.32</td>
<td>.03</td>
<td>.03</td>
<td>.10</td>
<td>-.12</td>
<td>-.48**</td>
<td></td>
<td></td>
</tr>
<tr>
<td>8. Length of Jobs</td>
<td>2.53</td>
<td>1.31</td>
<td>.08</td>
<td>.08</td>
<td>.22</td>
<td>.00</td>
<td>.2</td>
<td>-.66**</td>
<td>.02</td>
<td></td>
</tr>
</tbody>
</table>

*p < .05, **p < .01

Note: Numbers in parenthesis are coefficient α reliabilities

Gender and ethnicity coded as dichotomous variables; for gender, 1 = males and 2 = females; for ethnicity, 1 = Caucasian and 2 = Non-Caucasian
Chapter V

Discussion

In the present study, item stem length did not significantly affect face validity, perceptions of fairness, or involvement. The means of the reaction variables were similar when comparing the single word rating scale against the full sentence rating scale and slightly trended towards higher scores for the full sentence scale, though the observed magnitudes of the effects coupled with the small sample size render any interpretation of the trends premature. The closeness of the means between the rating scales would seem to indicate that the length of item stems does not affect how people will react to it and could potentially be altered without changing participant reactions.

This finding suggests that there may be changes made to item stems that will not affect how participants perceive the scale, or at minimum, certain scales may be able to be changed in a way to not affect perceptions. Mullins et al. (2005) observed that modifying the response scale for an integrity test affected participant’s reactions but that parallel modifications to the response scale for a personality test did not. Other elements of rating scales may also be able to be changed without altering reactions.

Changing the length of the item stems seems to have resulted in little change to the psychometrics of the questionnaires. Both versions of the personality questionnaire exhibited strong reliabilities, showing reliabilities higher on each factor than Saucier (1994) reported. While there was not a significant difference between the two personality questionnaires, the presence of significant correlations varied slightly between the two scales. The single word
rating scale did not correlate significantly with any demographic variables while the full sentence rating scale had significant correlations between face validity and ethnicity as well as the number of jobs held in the last five years. Treating ethnicity as a dichotomous variable, Caucasian versus non-Caucasian, the results of an independent-samples t-test indicated that non-Caucasian participants found the rating scale to be more face valid than Caucasian participants, t(57) = -2.41, p < .05. While this may imply that there were ethnic differences, only 15 of the 59 participants in that condition were not Caucasian. Additionally, due to the number of variables, this finding may be the result of sampling error. Further research, with a larger and more diverse sample, may be needed to determine the actual impact of ethnicity on face validity for the different rating scales.

Face validity was negatively correlated with having a job in the last five years. Having fewer jobs could be an indicator of less experience or, at least, fewer opportunities to be in various work environments. Those who have had multiple jobs and potentially more varied work experiences could feel that a personality questionnaire is not an important part of job selection.

To further test psychometric similarities, an exploratory factor analysis of the personality questionnaires was conducted to determine whether the dimensionality of the measure was affected by the change in stem length. Principal components analysis with varimax rotation was used to analyze the current data. The same factor structure was found for each version of the questionnaire. However, instead of five factors being identified, in both cases there were 12. To further examine this issue, the original five factors (Openness, Conscientiousness, Extraversion, Agreeableness, Emotional Stability) were factor analyzed. For the single word scale, Openness had three factors, and the other four had two factors each. For the full sentence scale, Openness and Emotional Stability had three factors and the other three had two factors. This is dramatically
different from what would be expected for these measures based on prior research. The factors did not group together in a way that appears to represent the Five Factor model, indicating more research on these measures may be warranted. For the purposes of this study, though, the finding that the factor structure was roughly comparable across the two measures at least seems to indicate that no meaningful psychometric changes resulted from the study’s manipulation.

**Contributions**

The present study displays the impact of form changes to a rating scale. When the personality scale was altered through item stem lengths there were no effects observed on participant reactions. In addition to the lack of reaction differences, the tested psychometrics were very similar between the conditions. These findings could indicate that non-content changes, such as item stem length modifications, could have little effect on reactions and the psychometrics of rating scales.

Organizations that are creating or modifying rating scales could maintain perceptions of the scales while increasing or decreasing item stem length by reasonable amounts. If the item stem lengths are reduced, the overall length of the scale would be shorter. This could be useful if time is of concern, such as time limits on an application, where applicants and those administering the measures having less to read through. Also, if there are multiple parts of an application, long item stems could create a lapse of attention or increase fatigue. As such, having sections that can be shortened without harming the measure’s psychometric properties would be potentially beneficial. If changes were made to an existing scale, data would still be able to be compared and used for current reporting. The visual presentation of a scale, for example physical space of paper or websites, could be adjusted by altering the item stem length of a scale without losing perceptions of fairness, face validity, or involvement with the scale.
The psychometrics of a scale, when item stem lengths are altered, have been shown to be stable through the present study. Factor analyses displayed the ability of scales to maintain similar factor structures when adjusted, though more rigorous research utilizing structural equations modeling to explore the comparability of factor structures should be undertaken. This finding highlights how organizations could be confident that scales that are changed are preserving the traits of the original scale or that having more than one version of a scales that is adjusted through item stem length will be similar in the way that item are grouped together psychometrically, at least for this personality test.

Limitations and Directions for Future Research

A smaller than anticipated participant group was used for this study. To test to see if the use of smaller groups had influenced the non-significant findings, Cohen’s $d$ statistic was calculated to determine whether the effect sizes observed might have approached significance if a sample of the originally proposed magnitude had been utilized. Cohen’s $d$ statistic for face validity was found to be 0.14. For perceptions of fairness $d = .19$, and for involvement $d = .11$. The low effect sizes observed for this study show that the smaller number of participants likely had no meaningful effect on the significance of the findings. Having more participants take part in the present study would be unlikely to change the outcomes due to the effect sizes all being under .2, which tells us that the distributions of scores have a large amount of overlap. Not having the additional five participants to reach the desired level of power most likely did not have any negative consequences or result in findings being non-significant that would have been significant with a small number of additional participants.

Another potential limitation is the effect of changing the measure itself. By altering the measure to create different item stem lengths there may be information that is “lost” or added.
Actual validity of the measures is not within the scope of this study to assess, but validity may be lowered or raised depending on the length of the item stem, although the reliabilities in the present study for both versions of the scales were high. However, the factor analysis does present cause for concern being that it showed well above the number of factors anticipated. This could change how a participant answers the items and could affect the ratings given to the scale. At the same time, however, the finding that the factor structures appeared comparable across conditions would seem to indicate that any validity issues with the measure are independent of the manipulation. Although face validity is an explicit aspect of the present study, the effects of manipulations such as those used herein on more traditional psychometric indicators of validity is an area future research should explore. Tests should be created that are proved to be psychometrically valid with the same modifications to confirm the present study’s findings.

There is also the possibility that the participants disagree as to what constitutes “fairness”. Blader (2007) suggested that employees frequently disagree as to whether their boss or organization is procedurally fair. By participants, and possibly applicants, having different ideas as what is fair there could be moderators to how they rate the scales. Potential moderators could be the job that is being applied for. For example, a personality questionnaire may be deemed more fair for a sales position than a data entry position. Future research could examine rating scales that are widely used in application processes. In addition, a more in-depth look at perceptions of fairness in the context of selection may be worthwhile. Determining what applicants are concerned about when applying for a job, especially in pre-employment testing, could be beneficial for organizations that want to portray fair organizational procedures.

Participants could also have preconceived notions of scales that would affect their ratings. There is an association between general attitudes toward employment testing and
procedural justice (Van Vianen, Taris, Scholten, & Schinkel, 2004). For the present study personality test biases may not be of any concern, although it could be the focus of future studies. Having pre-set beliefs of testing could also have an effect on how participants will rate the scales. In this study a personality measure was used for the manipulated scale. If a participant believes that personality scales are not suitable for any application processes or other job functions then they may rate the scale more negatively than a participant with no pre-set biases. Having such a preconceived notion could lead a participant to disregard the situation in which the rating scale is given and instead base all reactions on their prior beliefs. If an individual believed that jobs are about relationships and interactions, a personality measure would seem as ideal test for job fit. However, if the individual is more concerned with ability or job tasks then a personality measure could be seen as of no use for selection.

In the single word item stem length condition, fairness and involvement were significantly correlated. Future research could investigate if this finding is replicable or due to sampling error.

The measure used as the rated scale could limit the generalizability of any findings. Results may not translate to other forms of scales such as integrity tests or cognitive ability tests as well as other personality tests. Integrity and cognitive ability tests can be more complex than personality measures. The complexity could lead to a change between what is being tested and the clarity of the items to the test taker if scales differed in item stem length. Also, different scales may not have the ability to be changed to one word stems. Future research should look into other forms of scales to determine if similar results could be obtained as well as if the same results can be found with item stems as full sentences versus a paragraph.
Conclusion

Rating scales may be able to have non-trivial modifications that will not have an effect on how individuals will react to certain measures, or on the psychometric properties of those measures. This finding could be impactful in that when creating rating scales focus can be put on psychometrics with little worry on how individuals will view scales. The lack of major differences amongst demographic variables also may imply that the present findings are generalizable.

Finally, individuals may be indifferent to various changes that do not change the overall content of a rating scale. Although the present study was a between subjects design, the closeness of the rating means suggest that the participants were not affected by the differences in item stem length. The present study provides a starting point for future research into participant reactions to modifications of rating scales where altering aspects of scales could prove to be of little consequence.
Chapter VI

Summary

In Industrial-Organizational (I-O) psychology, and most other fields of psychology, questionnaires (sometimes discussed as rating scales or surveys) are important measurement tools in research and applied areas. The creation of questionnaires helps to get results that determine performance ratings, hiring decisions, personality dimensions, job analysis, and integrity assessments, to name a few. Although the content of such tools is important, the questionnaires themselves have the potential to influence outcomes due to the format and manipulations of the format (Borman, Buck, Hanson, Motowildo, Stark, & Drasgow, 2001). It is important in I-O psychology to explore all the variables that affect the questionnaires used to make essential decisions. The more we know about how differences in questionnaires can affect their outcomes, the better we can attain accurate, meaningful results.

The decisions made while creating a scale format could have perceptual consequences that influence respondents. The number of response categories, for example, has been found to have effects on perceptions of variables such as fairness and face validity (Mullins, Polson, Lanch, & Kehoe, 2006).

Research on scale types has examined the effects that various scales have on accuracy. Keaveny and McGann (1975) researched the occurrence of halo effects (a tendency to rate based on overall impressions rather than attending to distinct performance dimensions), leniency (a tendency to give higher ratings than might be warranted, based on performance), and the amount of discriminant validity on behavioral expectation scales (BES) and graphic rating scales (GRS).
Negatively worded items have also been a subject of format research. Barnette (2000) examined item stem directions and item response pattern direction effects on internal consistency reliability.

While there have been a reasonable number of studies on scale construction, the literature is largely scattered. The above literature is representative of studies done from the different areas of scale construction. To the best of the present author’s knowledge, no studies have examined the effect of item stem length on respondent perceptions.

Having a scale that is face valid can lead to cooperation and motivation as well as decrease dissatisfaction and feelings of injustice of low scorers, among other things (Nevo, 1985). In addition, perceptions of fairness can lead to positive perceptions of an organization (Hausknecht, Day, & Thomas, 2004). Another construct, involvement, while not thoroughly used in Industrial-Organizational Psychology research, is potentially important because it can show how motivated a respondent is to process information about a given rating scale or issue (Johnson & Eagly, 1989).

This study will attempt to explore differences in face validity, perceptions of fairness, and involvement based on item stem length. To achieve this, two parallel scales have been created that have either single-item stem lengths or full sentence item stem lengths. For example, the different lengths could be, “Happy” and “I consider myself a happy person.”

**Face Validity and Perceptions of Fairness**

Face validity has been defined as “the extent to which applicants perceive the content of the selection procedure to be related to the content of the job” (Smither, Reilly, Millsap, Pearlman, & Stoffey, 1993, p. 54). Nevo (1985) argued that face validity can create cooperation and motivation in participants, attract candidates, lower dissatisfaction and feelings of injustice.
for low scorers, convince policymakers, employers, and administrators to implement a test, and improve relations with mass media and courts. By creating scales that are face valid there may be more acceptance than is present when a scale that is not face valid is used.

The face validity of a scale may reflect its perceived job relatedness, and can therefore have implications for perceptions of fairness (Ployhart & Harold, 2004). However, while they may be related, they are distinct concepts that need to be assessed separately. Face validity concerns the scale itself without influences from external sources. Perceptions of fairness can be influenced by how the scale is created or how outcomes will be distributed due to the scale. This distinction could lead to differences in how each of these are rated thus there should be different measures to assess their extent.

Perceptions of fairness are important because a participant’s feelings toward a measure can have an effect on their attitudes and behaviors (Hausknecht et al., 2004; Roch, Sternburgh, & Caputo, 2007). This effect on attitudes and behavior due to a measure can have implications for organizational commitment, job satisfaction, performance, organizational citizenship behaviors, and questions over an organization’s polices (Roch et al.). Applicants will hold more positive perceptions about an organization if a scale is seen as procedurally fair and job relevant (Hausknecht, et al., 2004). In addition, if an applicant is aware of the type of selection process and believes the process to be unfair, the applicant may be dissuaded from applying (Gilliand, 1993). Three types of fairness have been a central part of perceptual research. These include procedural justice, distributive justice, and interactional justice (Fassina, Jones, & Uggerslev, 2008; Gilliland, 1993; Kim & Leung, 2007). For the purposes of this study, procedural justice will be focal. Studying procedural justice in the context of respondent reactions is consistent with
past research (Maertz, Bauer, Mosley, Posthuma, & Campion, 2004; Truxillo, Bauer, Campion, & Paronto, 2002).

**Involvement**

Involvement has typically been an important topic in advertising research (Andrews, Durvasula, & Akhter, 1990). Although there has been little crossover into organizational psychology, involvement could be important in the perceptions of and engagement with rating scales. Three main types of involvement are value relevant involvement, outcome relevant involvement, and impression relevant involvement (Cho & Boster, 2005). Outcome relevant involvement refers to a condition that has future consequences for the respondent. High outcome relevant involvement fosters evaluation of information and issue relevant thinking. For example, if a person is buying a car they will be more likely to listen to and evaluate information on cars that the individual may not have liked initially. Outcome relevant involvement could be important to rating scales because it is concerned with currently important goals and outcomes (Petty & Cacioppo, 1990). Determining how involved a respondent is with a rating scale could help create more engaging measures. By making a respondent more involved with a scale they may put in more effort and concentration which could help with performance or willingness to complete the scale.

Changes in item length could tell us if the size of the items is related to respondent perceptions of the scale as a whole. The length of an item could lead to perceptions of being too ambiguous or too detailed to be answerable in a confident manner. In addition, if a certain optimal length makes participants think and be more engaged in rating scales, this could lead to more accurate results.
Hypotheses

Although each variable has different criteria, determining face validity, perceptions of fairness, and involvement all have ties to the amount of information available. Overall lack of information could lead to lower perceptions of a rating scale, while more information may lead to more positive perceptions.

_Hypothesis 1:_ Participants who receive the measure with single word item stem lengths will score the lower on face validity perceptions than those receiving the full sentence item stem lengths.

_Hypothesis 2:_ Participants who receive the measure with single word item stem lengths will score the lower on perceptions of fairness than those receiving the full sentence item stem lengths.

_Hypothesis 3:_ Participants who receive the measure with single word item stem lengths will score lower on the involvement scale than those receiving the full sentence item stem lengths.

Method

Participants

A total of 122 participants, 63 in the single-word item stem version and 59 in the full sentence item stem version, took part in this study. Recruitment was conducted through the university’s participant pool and via direct recruitment from specific classes.

Measures

**Personality measure.** A Big Five personality measure was used as the scale to be rated. This measure, the Mini-Markers, is a 40-item adjective-based measure based on the work of Saucier (1994). Two forms of this measure were utilized. The first form used was the original,
single-word item stem version (Saucier, 1994) and the second was adapted to create a full sentence form. The response options were the original 9-point scale ranging from 1 (*Very inaccurate*) to 9 (*Very accurate*). Saucier found the single-word stem length form to produce internal consistency reliabilities of .76 (Extraversion), .79 (Conscientiousness), .62 (Agreeableness), .59 (Openness), and .63 (Emotional Stability). This and all other copyrighted measures are contained in a separate document. The present study found the internal consistency reliabilities for the single word questionnaire to be; .86 (Extraversion), .83 (Conscientiousness), .83 (Agreeableness), .63 (Openness), and .76 (Emotional Stability). For the full sentence questionnaire reliabilities were found to be .86 (Extraversion), .84 (Conscientiousness), .86 (Agreeableness), .71 (Openness), and .70 (Emotional Stability).

*Face validity measure.* Face validity was measured using a 3-item scale adapted from Chan et al. (1998). The response options utilized were a 5-point Likert scale ranging from 1 (*Strongly Disagree*) to 5 (*Strongly Agree*). Cronbach’s alpha was reported to be .79 (Chan et al., 1998). An example of an item on this measure would be, “I can see how the focus of this survey is related to what is expected of employees.” In the present study Cronbach’s alpha was found to be .73.

*Perceptions of fairness measure.* Perceptions of fairness was measured using a 3-item scale adapted from Chan et al. (1998). The response set utilized was a 5-point Likert scale ranging from 1 (*Strongly Disagree*) to 5 (*Strongly Agree*). Cronbach’s alpha was reported to be .79 (Chan et al., 1998). An example of an item on this measure would be, “I feel that I was able to report my true beliefs.” In the present study Cronbach’s alpha was found to be .43. When the first item of this scale is removed, alpha was raised to .83. All analyses were conducted using the original version of this scale.
Involvement measure. Involvement was measured by a 5-item measure adapted from Cho and Boster (2005). Responses were made on a 7-point Likert scale ranging from 1 (Strongly Disagree) to 7 (Strongly Agree). Cronbach’s alpha was reported to be .82. Construct validity was demonstrated by correlating outcome relevant involvement with information seeking, $r(281) = .39, p < .001$, and also conducting a regression analysis with same variables, $\beta = .23, p = .007$ (Cho & Boster, 2005). In the present study Cronbach’s alpha was found to be .73.

Demographic measure. Demographic information pertaining to age, gender, ethnicity, and a description of work experience was collected. The description of work experience included whether the participant has held a job, if they have had three or more jobs in the past five years, the average length of service, and the type of work experience (e.g. retail, manufacturing, office work). In addition, an open-ended question created for this study was included to gather qualitative information as to reasons why participants perceived the personality measure the way they did. The question was, “What factors were your reactions to the personality inventory based on?”

Procedure

The study utilized a two group between-subjects design. The groups had identical instructions and a hypothetical scenario as well as completed a consent form and a demographic information questionnaire. The scenario described the application process of a hypothetical employer. It described what type of employment the participant will be seeking and the duties involved in that job. Each of the measures had an identical opening paragraph detailing how respondents are to answer the items. For example, “Read each item and answer on a scale of 1-9 how much it describes you.” Data was collected through SurveyMonkey.
Results

In order to test the hypotheses, three independent samples t-tests were conducted to compare the single word questionnaire and the sentence-length questionnaire. The first t-test assessed the differences in perceptions of face validity based on item stem length (Hypothesis 1), the second assessed the differences in perceptions of fairness (Hypothesis 2), and the third assessed the differences in level of involvement (Hypothesis 3).

The test for Hypothesis 1 was non-significant, $t(120) = -0.81$, ns. The mean score of face validity for the single word rating scale was 3.60 (SD=0.77) and the mean score for the full sentence rating scale was 3.70 (SD=0.70). The test for Hypothesis 2 was non-significant, $t(120) = -1.01$, ns. The mean score of perceptions of fairness for the single word rating scale was 3.40 (SD=0.67) and the mean score for the full sentence rating scale was 3.53 (SD=0.67). The test for Hypothesis 3 was non-significant, $t(120) = -0.58$, ns. The mean score of involvement for the single word rating scale was 4.93 (SD=1.03) and the mean score for the full sentence rating scale was 5.04 (SD=0.95).

For the personality questionnaire, mean scores within each of the factors were similar. For the single word rating scale the mean for Openness was 6.84, Conscientiousness was 7.08, Extraversion was 6.20, Agreeableness was 7.48, and Emotional Stability was 6.03. For the full sentence rating scale the mean for Openness was 6.71, Conscientiousness was 7.20, Extraversion was 6.09, Agreeableness was 7.31, and Emotional Stability was 6.30.

No significant effects or interactions were detected for gender or ethnicity on any study variables. In addition, no significant test results were found when the first item of the face validity scale was removed.
Discussion

In the present study, item stem length did not significantly affect face validity, perceptions of fairness, or involvement. The means of the reaction variables were similar when comparing the single word rating scale against the full sentence rating scale and slightly trended towards higher scores for the full sentence scale, though the observed magnitudes of the effects coupled with the small sample size render any interpretation of the trends premature. The closeness of the means between the rating scales would seem to indicate that the length of item stems does not affect how people will react to it and could potentially be altered without changing participant reactions.

This finding suggests that there may be changes made to item stems that will not affect how participants perceive the scale, or at minimum, certain scales may be able to be changed in a way to not affect perceptions. Mullins et al. (2005) observed that modifying the response scale for an integrity test affected participant’s reactions but that parallel modifications to the response scale for a personality test did not. Other elements of rating scales may also be able to be changed without altering reactions.

Changing the length of the item stems seems to have resulted in little change to the psychometrics of the questionnaires. Both versions of the personality questionnaire exhibited strong reliabilities, showing reliabilities higher on each factor than Saucier (1994) reported. While there were no significant differences between the questionnaires, the presence of significant correlations varied slightly between the two scales. The single word rating scale did not correlate significantly with any demographic variables while the full sentence rating scale had significant correlations between face validity and ethnicity as well as the number of jobs held in the last five years. While this may imply that there were ethnic differences, only 15 of the
59 participants in that condition were not Caucasian. Further research, with a larger and more diverse sample, may be needed to determine the actual impact of ethnicity on face validity for the different rating scales. Face validity was negatively correlated with having a job in the last five years. Those who have had multiple jobs and potentially more varied work experiences could feel that a personality questionnaire is not an important part of job selection.

To further test psychometric similarities, an exploratory factor analysis of the personality questionnaires was conducted to determine whether the dimensionality of the measure was affected by the change in stem length. The same factor structure was found for each version of the questionnaire. However, instead of five factors being identified, there were 12. To further examine this issue, the original five factors (Openness, Conscientiousness, Extraversion, Agreeableness, Emotional Stability) were factor analyzed. The factors did not group together in a way that appears to represent the Five Factor model, indicating more research on these measures may be warranted. For the purposes of this study, though, the finding that the factor structure was roughly comparable across the two measures at least seems to indicate that no meaningful psychometric changes resulted from the study’s manipulation.

**Contributions**

The present study displays the impact of form changes to a rating scale. When the personality scale was altered through item stem lengths there were no seen effects on participant reactions. In addition to the lack of reaction differences, the tested psychometrics were very similar between the conditions. These findings could indicate that non-content changes, such as item stem length modifications, could have little effect on reactions and the psychometrics of rating scales.
Organizations that are creating or modifying rating scales could maintain perceptions of the scales while decreasing length. If the item stem lengths are reduced, the overall length of the scale would be shorter. This could be useful if time is of concern, if there are multiple parts of an application that could create a lapse of attention, or the visual presentation of a scale, for example physical space of paper or websites, could be adjusted by altering the item stem length of a scale without losing perceptions of fairness, face validity, or involvement with the scale.

The psychometrics of a scale, when item stem lengths are altered, have been shown to be stable through the present study. Also, factor analyses have displayed the ability of scales to maintain similar structures when adjusted. This finding highlights how organizations could be confident that scales that are changed are preserving the traits of the original scale, at least for personality tests.

Limitations & Directions for Future Research

A smaller than anticipated participant group was used for this study. To test to see if the use of smaller groups had influenced the non-significant findings, Cohen's $d$ statistic was used. Cohen's $d$ statistic for face validity was found to be 0.14. For perceptions of fairness $d=.19$, and for involvement $d=.11$. The low effect sizes observed for this study show that the smaller number of participants likely had no meaningful effect on the significance of the findings.

A potential limitation is the effect of changing the measure itself. By altering the measure to create different item stem lengths there may be information that is "lost" or added. Actual validity of the measures is not within the scope of this study to assess, but validity may be lowered or raised depending on the length of the item stem, although the reliabilities in the present study for both versions of the scales were high. However, the factor analysis does present cause for concern being that it showed well above the number of factors anticipated.
There is the also the possibility that the participants disagree as to what is fair. Blader (2007) suggested that employees frequently disagree as to whether their boss or organization is procedurally fair. Participants could also have preconceived notions of scales that would affect their ratings. Having pre-set beliefs of testing or particular types of assessments could also have an effect on how participants will rate the scales. For the present study personality test biases may not be of any concern, although it could be the focus of future studies. Also, the measure used as the rated scale could limit the generalizability of any findings. Results may not translate to other forms of scales such as integrity tests or cognitive ability tests, as well as other personality tests. Integrity and cognitive ability tests can be more complex that personality tests.

**Conclusion**

Rating scales may be able to have non-trivial modifications that will not have an effect on how individuals will react to certain measures, or on the psychometric properties of those measures. This finding could be impactful in that when creating rating scales focus can be put on psychometrics with little worry on how individuals will view scales. The lack of major differences amongst demographic variables also may imply that the present findings are generalizable.

Finally, individuals may be indifferent to various changes that do not change the overall content of a rating scale. Although the present study was a between subjects design, the closeness of the rating means suggest that the participants were not affected by the differences in item stem length. The present study provides a starting point for future research into participant reactions to modifications of rating scales where altering aspects of scales could prove to be of little consequence.
References


Appendix A

Personality Measure Form 1

The items below ask you to report how accurately each word describes you. It is important to respond in a way that describes you now, not how you would like to be.

Please use the following scale to describe how much the trait describes you:

<table>
<thead>
<tr>
<th>Inaccurate</th>
<th>?</th>
<th>Accurate</th>
</tr>
</thead>
<tbody>
<tr>
<td>Extremely</td>
<td>Very</td>
<td>Moderately</td>
</tr>
<tr>
<td>1</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>Bashful</td>
<td></td>
<td>Moody</td>
</tr>
<tr>
<td>Bold</td>
<td></td>
<td>Organized</td>
</tr>
<tr>
<td>Careless</td>
<td></td>
<td>Philosophical</td>
</tr>
<tr>
<td>Cold</td>
<td></td>
<td>Practical</td>
</tr>
<tr>
<td>Complex</td>
<td></td>
<td>Quiet</td>
</tr>
<tr>
<td>Cooperative</td>
<td></td>
<td>Relaxed</td>
</tr>
<tr>
<td>Creative</td>
<td></td>
<td>Rude</td>
</tr>
<tr>
<td>Deep</td>
<td></td>
<td>Shy</td>
</tr>
<tr>
<td>Disorganized</td>
<td></td>
<td>Sloppy</td>
</tr>
<tr>
<td>Efficient</td>
<td></td>
<td>Sympathetic</td>
</tr>
<tr>
<td>Energetic</td>
<td></td>
<td>Systematic</td>
</tr>
<tr>
<td>Envious</td>
<td></td>
<td>Talkative</td>
</tr>
<tr>
<td>Extraverted</td>
<td></td>
<td>Temperamental</td>
</tr>
<tr>
<td>Fretful</td>
<td></td>
<td>Touchy</td>
</tr>
<tr>
<td>Harsh</td>
<td></td>
<td>Uncreative</td>
</tr>
<tr>
<td>Imaginative</td>
<td></td>
<td>Unenvious</td>
</tr>
<tr>
<td>Inefficient</td>
<td></td>
<td>Unintellectual</td>
</tr>
<tr>
<td>Intellectual</td>
<td></td>
<td>Unsympathetic</td>
</tr>
<tr>
<td>Jealous</td>
<td></td>
<td>Warm</td>
</tr>
<tr>
<td>Kind</td>
<td></td>
<td>Withdrawn</td>
</tr>
</tbody>
</table>
Appendix B

Personality Measure Form 2

The items below ask you to report how accurately each sentence describes you. It is important to respond in a way that describes you now, not how you would like to be. Please use the following scale to describe how much the trait describes you:

<table>
<thead>
<tr>
<th>Inaccurate</th>
<th>?</th>
<th>Accurate</th>
</tr>
</thead>
<tbody>
<tr>
<td>Extremely</td>
<td>Very</td>
<td>Moderately</td>
</tr>
<tr>
<td>1</td>
<td>2</td>
<td>3</td>
</tr>
</tbody>
</table>

I consider myself a bashful person. ______ | I consider myself a moody person. ______
The word “bold” describes me well. ______ | The word “organized” describes me well. ______
I am typically careless in situations. ______ | I could describe myself as philosophical. ______
I could describe myself as cold. ______ | I am typically practical in situations. ______
Being called “complex” is accurate for me. ______ | Being called “quiet” is accurate for me. ______
I consider myself a cooperative person. ______ | I consider myself a relaxed person. ______
The word “creative” describes me well. ______ | The word “rude” describes me well. ______
I could describe myself as deep. ______ | I am typically shy in situations. ______
Being called “disorganized” is accurate for me. ______ | I could describe myself as sloppy. ______
I am typically efficient in situations. ______ | Being called “sympathetic” is accurate for me. ______
I consider myself an energetic person. ______ | I consider myself a systematic person. ______
The word “envious” describes me well. ______ | The word “talkative” describes me well. ______
I could describe myself as Extraverted. ______ | I am typically temperamental in situations. ______
I am typically fretful in situations. ______ | I could describe myself as touchy. ______
Being called “harsh” is accurate for me. ______ | Being called “uncreative” is accurate for me. ______
I consider myself an imaginative person. ______ | I consider myself an unenvious person. ______
The word “inefficient” describes me well. ______ | I could describe myself as unintellectual. ______
Being called “intellectual” is accurate for me. ______ | Being called “unsympathetic” is accurate for me. ______
I could describe myself as jealous. ______ | The word “warm” describes me well. ______
I am typically kind in situations. ______ | I am typically withdrawn in situations. ______
Appendix C

Face Validity Measure and Perceptions of Fairness Measure*

Consider the personality questionnaire when responding to the next set of questions.

*Please indicate the degree to which you agree with the following statements using the following scale:

5 = Strongly Agree
4 = Mostly Agree
3 = Neither Agree nor Disagree
2 = Mostly Disagree
1 = Strongly Disagree

1. I can see how the focus of the survey related to what is expected of employees. ___

2. The questions of the survey are related to requirements of employees. ___

3. The survey does not seem to be related to work activities. ___

4. I feel that I was able to report by true beliefs. ___

5. It would be fair to use this survey to select employees. ___

6. I would believe that I received fair treatment if the survey were used as a part of my application process. ___

*Face validity measured by items 1-3. Perceptions of fairness measured by items 4-6.
Appendix D

Involvement Measure

Consider the personality questionnaire when responding to the next set of questions.

Please indicate the degree to which you agree with the following statements using the following scale:

7 = Strongly Agree
6 = Moderately Agree
5 = Slightly Agree
4 = Neither Agree nor Disagree
3 = Slightly Disagree
2 = Moderately Disagree
1 = Strongly Disagree

1. The questionnaire that was presented has little impact on me getting a job.  ____

2. It is easy for me to think of ways this questionnaire influences an employer’s perception of me.  ____

3. Using the right questionnaire affects the application process.  ____

4. My getting this job has little to do with the questionnaire given.  ____

5. All in all, the effect of this scale on my application would be little.  ____
Appendix E

Demographic Information

Please answer the following:

1. Age: _____
2. Gender:  Male
   Female
3. Race:
4. Work Experience:
   Have you held a job in the past or do you currently have a job? Y/N
   If you answered yes to the previous question, how many jobs have you
   had in
   the past five years? _____
   On average, how long have you stayed at your jobs?
   ___ Less than 6 months
   ___ 6 months to 1 year
   ___ 1 year to 2 years
   ___ More than 2 years
   What type(s) of organization have you held a position in? (Check all that apply)
   ___ Retail
   ___ Manufacturing
   ___ Office Work
   ___ Food Service
   ___ Other
5. What factors were your reactions to the personality inventory based on?

____________________________________________________________________________
____________________________________________________________________________
____________________________________________________________________________
Appendix F

Hypothetical Situation

Please consider the following scenario while completing the survey:

Suppose that you are seeking employment as a manager at a retail store. As a manager you would be responsible for supervising employees, devising more efficient operations, maintaining a budget, and inventory control.

The questionnaire you are being asked to complete is being used as a part of your application process.
Appendix G

Informed Consent Form

You are being given the opportunity to volunteer to participate in a project conducted through Xavier University.

The purpose is to examine perceptions of rating scales used for application processes. This information will be assessed via an online survey that should take no longer than 15 minutes to complete. Data will be collected confidentially and therefore answers cannot be linked to any participants. At the conclusion of the study, you will be redirected to a separate survey where you can input your name and instructor in order to receive credit. Data collected from the two surveys cannot be linked in any fashion to identify your responses. You are free to not participate or to discontinue participation at any time without adverse effects. There are no known risks associated with your participation in this study.

If you have any questions at any time (during or after the study), you may contact Kara Jones at (734) 812-4894 or her advisor, Dr. Morrie Mullins, at (513) 745-3170. Questions about your rights as a research subject should be directed to the Xavier University’s Institutional Review Board at (513) 745-2870.

I understand that my participation in this study is voluntary. I understand that if I agree to this study, I may receive course credit or extra credit, at my instructor’s discretion.

I understand that by clicking the ‘Next’ button below, I am consenting to participate in this study.
Appendix H

IRB Approval Letter

March 29, 2010

Ms. Kara Jones
7473 Valley View Place, Apt. 302
Cincinnati, OH 45244

Re: Protocol #0649 "The Effect of Item Stem Cell Length on Face Validity, Perceptions of Fairness & Involvement"

Dear Ms. Jones:

The IRB has completed the review of your study proposal and finds that it meets the criteria for the Exempt from Review category under Federal Regulation 45 CFR 46. Your protocol is approved as exempt research, and therefore requires no further oversight by the IRB.

If you wish to modify your study, it will be necessary to obtain IRB approval prior to implementing the modification. If any adverse events occur, please notify the IRB immediately.

Please contact our office if you have any questions.

Sincerely,

Kathleen J. Hart, Ph.D., ABPP
Interim Chair, Institutional Review Board
Xavier University

KH/dm

c: Dr. Morell Mullins, Advisor