Spring 2016

The effect of anchoring vignettes on factor structures: Student effort as an example

Carolyn A. Miesen
James Madison University

The Effect of Anchoring Vignettes on Factor Structures: Student Effort as an Example

Carolyn Angelica Miesen

A thesis submitted to the Graduate Faculty of

JAMES MADISON UNIVERSITY

In

Partial Fulfillment of the Requirements

for the degree of

Master of Arts

Department of Graduate Psychology

May 2016

FACULTY COMMITTEE:

Committee Chair: John Hathcoat

Committee Members/ Readers:

Deborah Bandalos

Sara Finney

Acknowledgments

I would first like to thank my thesis advisor, John Hathcoat. The door to his office was always open whenever I ran into trouble or had a question about my research or writing. He consistently allowed this paper to be my own work, but steered me in the right direction whenever I needed it. I would also like to thank the members of my thesis committee, Deborah Bandalos and Sara Finney. Without their passionate participation and input, my thesis could not have been successfully completed.

I would like to extend my very profound gratitude to all those who have supported and encouraged me throughout the master's program and through the process of researching and writing this thesis. I must thank all of the faculty at the Center for Assessment Research Studies for how positively their enthusiasm for both research and teaching as positively influenced my life. My peers in both the masters and doctoral program should also know how grateful I am for making the programs' culture so supportive and friendly. Specifically, I would like to thank my cohort, Thai Ong, for being a friend and cheering me on along the way. And of course, I must express my appreciation for Wayne Miesen, Joey Tucciarone, Catharine Burkholder, and Benny Schwartz for always believing in me.

This accomplishment would not have been possible without all of you. From the bottom of my heart, thank you.

Carolyn

Table of Contents

List of Tables

List of Figures

Abstract

Anchoring vignettes are used as a methodological technique for removing differential interpretation of response categories (DIRC) from scores on subjective self-report measures (King, Murray, Slomon, & Tandon, 2004). This technique requires participants to read one or more short scenarios, or vignettes, designed to represent various levels of a construct. Vignette ratings are used as an indication of DIRC, which is a source of differential item functioning (DIF). Prior research primarily used indirect methods for evaluating vignette quality. In response, the present set of studies proposes using invariance testing as a more direct evaluation of how the use of anchoring vignettes impacts the presence of DIRC. The effort subscale from the Student Opinion Scale (SOS) is used to demonstrate this set of procedures. It is also argued that DIRC will manifest as non-uniform DIF given that corrections using anchoring vignettes should impact the rank order of cases.

In these studies, 819 participants were randomly assigned to either a control group (n = 478) or a group that received vignettes (n = 341) prior to responding to the SOS. Invariance testing was completed in two studies. The first study examined the factor structure between the control group and the vignette group before adjusting scores using the vignettes to determine what effect reading the vignettes may have had on the factor structure. The second study examined the invariance between the control group and the vignette group after score adjustment to determine what effect adjusting scores using the vignettes may have had on the factor structure. Results for the first study supported strict factorial invariance (configural, metric, and scalar invariance, and residuals) and equivalent latent means, which suggests that just viewing the vignettes had insubstantial

impact on the factor structure of the SOS effort subscale. Results for the second study also supported strict factorial invariance, but there was a substantial difference in the group's latent means. This result suggests that DIRC was not removed from the sample, however using anchoring vignettes to adjust scores resulted in systematically lower observed scores after adjustment. Implications for measuring effort along with general conclusions about using invariance testing to evaluate anchoring vignettes is also provided.

CHAPTER ONE

Introduction

Differential Interpretations of Response Categories

Consider a hypothetical scenario where two participants, Benny and Catharine, were asked about the extent to which they agreed with the statement, "I am in good physical health." Benny had never been hospitalized for physical health issues. He indicated that he strongly agreed with this item. Conversely, Catharine had been hospitalized multiple times since she was a child for recurring health-related issues which she still suffered from but to a lesser extent than when she was a child. Since Catharine had not recently been hospitalized, she also chose to strongly agree with this item. Clearly this item should not be used as an indication of their "objective" physical health since both participants received the same score despite the fact that Benny is in better physical health than Catharine. In this example, each participant endorsed the same response option for this item because they used different standards to evaluate their level of health, which influenced the comparability of their scores.

An assumption of using self-report measures is that participants with equal levels of a construct are just as likely to place themselves within a response category. Researchers expect participants to use response categories in a similar way when endorsing an item. However, this assumption may be violated in practice when participants interpret response categories differently. Differential interpretations of response categories (DIRC) may occur when participants use different standards to decide how they will respond to an item. This was the case with Benny and Catharine. Both participants had a very different idea of what it meant to strongly endorse an item

about their physical health because they used their personal experience to inform how they interpreted the response categories. As a result, their observed scores lacked comparability due to DIRC.

DIRC as a Source of Differential Item Functioning

Differential item functioning (DIF) is defined as "an unexpected difference among groups of examinees who are supposed to be comparable with respect to an attribute measured by the item and the test on which it appears" (Dorans & Holland, 1993, *p*. 37). Said another way, DIF occurs when two populations with an equal level of a latent construct have different probabilities of endorsing an item. DIF is often conceptualized as a result of different interpretations of the item stem, however DIF may also be the result of DIRC. For instance, the literature on anchoring vignettes often refers to DIRC as a source of DIF that occurs when the interpretation of the item stem is invariant across groups, but the interpretation of the response categories differs (King, Murray, Slomon, & Tandon, 2004). DIF due to DIRC occurs when individuals or populations with an equal level of a latent construct have different probabilities of endorsing an item due to their interpretations of the response categories.

Often researchers assume that DIRC does not exist in the sample. In other words, it is often assumed that when participants indicate a response to an item, the response category carries the same meaning for everyone who completed the measure. However, it may be inappropriate to make this assumption, especially considering individual differences that may affect interpretation of response categories, such as socioeconomic status (Van Doorslaer & Jones, 2003), age (Groot, 2000), and nationality (King et al., 2004).

For instance, Van Doorslaer and Jones (2003) examined DIRC present in the measurement of income-related health inequality. They referred to DIRC as 'state dependent reporting bias' (Kerkhofs and Lindeboom, 1995), 'scale of reference bias' (Groot, 2000) and 'response category cut-point shift' (Murray et al., 2001). However, they define it simply as when people have different "threshold levels" for placing themselves within a response category, despite having the same latent level of health. The authors used various methods, such as OLS, ordered probit, and interval regression models, to examine the impact of socio-economic factors on the estimation of threshold levels. They found that there was evidence of DIRC in self-reported general health.

Similarly, Goot (2000) examined DIRC present in the measurement of quality of life. Using an ordered probit model, he found that how people used the response categories on the quality of life measure changed with age. As such, even though younger people generally have better health than older people, the measure indicated the opposite. He suggested that this occurs due to increased "age-norming" or adaptability with age.

Another example of a variable that may impact DIRC is nationality. King et al. (2004) compared ratings of political efficacy across nationalities and found that support for the existence of DIRC. Specifically, they found that Chinese citizens had different threshold levels for rating political efficacy than Mexican citizens. This difference suggested that Chinese citizens used different standards for interpreting response categories than Mexican citizens.

Anchoring Vignettes

In response to this concern, anchoring vignettes were developed by King et al., (2004) as a methodological technique to control for DIRC. The technique requires

participants to read one or more short descriptions of scenarios or vignettes designed to represent various levels of a construct. With respect to Benny and Catharine, who were asked to report on their level of health, if a researcher wished to use anchoring vignettes to correct for DIRC, s/he could draft anchoring vignettes that represent various levels of healthiness. Additionally, how the participants respond to the anchoring vignettes would provide insight into how they interpret the response categories.

In drafting these vignettes, the researcher would first need to choose a construct that is unidimensional, meaning that the trait varies on a single continuum. It is important that the construct being measured is unidimensional because to be able to use the anchoring vignettes to adjust scores, participants must be able to rate the vignettes in a certain order. So in the case of healthiness, the researcher may narrow her/his focus to cardiovascular health. Then the researcher would need to operationalize cardiovascular healthiness at each level. Specifically, s/he would need to consider how many levels of cardiovascular healthiness vignettes should be drafted for and what behaviors or feelings are present at each level. In this example, levels of cardiovascular health could be operationalized by shortness for breath given various activities. For instance, a low level of cardiovascular health may be represented by a vignette of someone who experiences shortness of breath when performing everyday tasks that are not very demanding (i.e. walking a short distance). A high level of cardiovascular health, on the other hand, may be represented by a vignette of someone who experiences shortness of breath only after performing physically demanding tasks (i.e. running a long distance).

After the anchoring vignettes are drafted, they are presented to the participants, who rate the anchoring vignettes before they rate themselves. For example, Benny and

Catharine were asked to rate each vignette according to how much they agreed with the statement, "this person is healthy." With regard to how the participants rate the anchoring vignettes, DIRC can be defined as the differential rating of vignettes by participants. For instance, Figure 1 demonstrates a situation in which Benny and Catharine rate the vignettes differently. Benny rated the low vignette as 1 (*strongly disagree*) and the high vignette as 3 (*neutral*). Catharine, on the other hand, rated the low vignette as a 4 (agree) and the high vignette as a 5 (strongly agree). Assuming that the participants' interpreted the anchoring vignettes in the same way, the differential ratings of the anchoring vignettes indicate that DIRC is present.

Figure 2 demonstrates a situation where there is no DIRC present. In this situation, both Benny and Catharine rate the low vignette as a 4 (*agree*) and the high vignette as a 5 (*strongly agree*). Here it is clear that the participants used the response categories in the same way to rate the anchoring vignettes. So, no DIRC exists.

Then the participants complete the self-assessment by rating their own level of health on one or more health related items. The self-assessment is always rated on the same response scale as the anchoring vignettes. Once the data is collected from participants, the researcher then adjusts the participants' self-assessment scores such that the individual's scores are on a metric that places the self-assessment relative to how the individual rated the anchoring vignettes. There are two different approaches for adjusting scores: non-parametric and parametric. The non-parametric approach is computationally simpler than the parametric approach, and can be implemented without adopting additional assumptions. Since the current set of studies does not employ the parametric approach, only the non-parametric approach will be reviewed.

Scoring

Figure 3 shows how self-assessment scores are adjusted using anchoring vignettes with the non-parametric approach. Note that Benny rated the low health vignette with a 1 (*strongly disagree*) and the high health vignette with a 5 (*strongly agree*). Then he responded to the self-assessment item with a 2 (*disagree*). After the data was collected, a new metric was developed which placed the self-assessment score relative to how the anchoring vignettes were rated. Since there were two anchoring vignettes, the new metric shown in Figure 3 has five categories: 1 (self-assessment is lower than the low vignette), 2 (self-assessment is equal to the low vignette) 3 (self-assessment is higher than the low vignette and less than the high vignette), 4 (self-assessment is equal to the high vignette), 5 (self-assessment is higher than the high vignette). Since Benny rated himself (2) higher than how he rated the low health vignette (1) but lower than how he rated the high health vignette (5), his adjusted score becomes a 3.

The solution for the non-parametric approach described in Figure 3, as introduced by King et al. (2004), is mathematically conceptualized as follows:

$$C = \begin{cases} 1 \; if \; y < z_1 \\ 2 \; if \; y = z_1 \\ 3 \; if \; z_1 < y < z_2 \\ 4 \; if \; y = z_2 \\ \dots \\ 2J + 1 \; if \; y > z_J \end{cases}$$

C represents the value assigned to the participant after the self-assessment has been rescaled. *C* spans from 1 to 2J+1, which represents all possible categories as a function of the number of vignettes (J). The *y* represents the unadjusted self-assessment value, and z_1 to z_J represent the ratings of the various vignettes. This formula simply describes how the self-assessment scores are adjusted on a metric that places those scores relative to how

the vignettes were rated. In the case of Benny in Figure 3, the assigned value of *C* became 3 since he rated the low health vignette (z_1) as less than his self-assessment (y), which was rated as less than how he rated the high vignette (z_2).

When a participant rates each vignette with a unique value and in the expected order, as Benny had in Figure 3, the *C* value is described as scalar, meaning that it can be assigned a single value. In situations where the participant rated two vignettes as equal or rated them out of the expected order, the *C* value is described as interval, meaning that it cannot be assigned a single value. Figure 4 shows an example of an interval value of *C*. In this example, Catharine rated both the high and low anchoring vignettes with a 3 (*neutral*). Then she rated herself with a 3 (*neutral*) as well. Because she rated herself equivalent to both anchoring vignettes ($z_1 = y = z_2$), she can be described as being equal to both the low vignette and the high vignette. Therefore, when Catharine's score is adjusted, her assigned value of *C* spans from 2 (equal to low vignette) to 4 (equal to high vignette). Numerous approaches may be used to handle interval cases (Wand, King, & Lau, 2011). For example, one may simply omit interval cases given they are not very frequent in the sample. However, if interval cases are frequent in the sample, then this may indicate that the anchoring vignettes are problematic. This possibility is described in the following section. Since interval cases were minimal, this set of studies omitted interval cases.

Vignette Evaluation

Since the technique's first introduction, anchoring vignettes have been applied to a variety of contexts. They have been used when measuring constructs within educational assessment (Buckley, 2009; Programme for International Student Assessment, 2014),

health care (d'Uva, van Doorslaer, Lindeboom, & O'Donnell, 2008; Grol-Prokopczyk, Freese, & Hauser, 2011; Peracchi & Rossetti, 2012; Rice, Robone, & Smith 2011; Salomon, Tandon, & Murray, 2004), work disability (Kapteyn, Smith, & van Soest, 2007), political efficacy (King et al., 2004), and job satisfaction (Kristensen & Johansson, 2006). For instance, the study by King et al. was mentioned in a previous section. In their study, they compared ratings of political efficacy across nationalities and found that support for the existence of DIRC. Specifically, they administered five anchoring vignettes (translated to into appropriate languages) to participants in different countries. These anchoring vignettes were created to fall along a continuum of political efficacy. Then the participants completed a self-assessment on their perceived level of political efficacy. Before adjusting scores, they found that Chinese citizens tended to rate their political efficacy as higher than did Mexican citizens. These findings were counterintuitive since Mexican citizens would appear to have higher levels of political efficacy given recent political events. To support their expectation the authors observed that:

> The citizens of Mexico recently voted out of office the ruling PRI party in an election closely observed by the international community and widely declared to be free and fair… Despite the existence of limited forms of local democracy, nothing resembling this has occurred in China. (King, et al., 2004, p. 196)

After controlling for DIRC using the anchoring vignettes using a non-parametric approach, the researchers found that Mexican citizens' scores were relatively higher in political efficacy than the Chinese citizens, which was more consistent with their expectations.

In the instance of using anchoring vignettes for political efficacy, as well as many other contexts, the use of anchoring vignettes proved to be beneficial for making more

accurate inferences. However, before anchoring vignettes can be used to help make more appropriate inferences from data, there they must be evaluated. To evaluate how well the anchoring vignettes function, most of the current studies test whether the assumptions of anchoring vignettes have been met. Prior methods used to assess the assumptions of anchoring vignettes do not allow for direct observation of how the anchoring vignettes function. As such, these methods may be considered indirect investigations of the quality of the anchoring vignettes.

There are two main assumptions of anchoring vignettes: vignette equivalence and response consistency. Vignette equivalence is the assumption that the vignettes are interpreted in the same way by different participants apart from measurement error (King et al., 2004). The participants must perceive the vignettes as representing a point along a continuum of the same construct. For this reason, the construct being measured must be unidimensional; meaning each item measures a single trait. The assumption of vignette equivalence has been evaluated by examining how participants rate the anchoring vignettes. The expectation is that if the vignettes are interpreted in the same way then this perception will be reflected in the order that the participants rate the vignettes. To be clear, although the precise value of ratings of the vignettes can be different across participants, the order in which the vignettes are rated must be equivalent (e.g. $z_1 < z_2$). If there is a substantial number of cases where the vignettes were ranked out of order according to the rank they were created to represent or vignettes were rated as equal (e.g. high frequency of interval cases), then this would be an indication that participants did not perceive the vignettes as representing various points along a continuum.

For instance, Rice, Robone, and Smith (2011), demonstrated various means of examining vignette equivalence in the context of reporting health system responsiveness across different countries. In their study, they examined rating consistency, compared spearman rank order correlation coefficients (SROCC), and conducted a hierarchical ordered probit (HOPIT) model; each a method for testing vignette equivalence. In this context, consistency refers to the similarity of vignette ranking. In other words, the goal of examining order constancy is to establish that there is an average rank, which applies to both the overall sample and subgroups of the sample. In their study, Rice et al. (2011) compared average country rankings to an average global rank, and found that there was no substantial variation among countries in how vignettes were ranked. This result supports the notion that the vignettes were interpreted in the same manner across countries and implied that vignette equivalence had been met.

Response consistency is the assumption that participants use the same criteria to rate the vignettes as they do to rate themselves (King et al., 2004). This assumption would be violated if a participant held different standards for rating the vignettes than they did for rating the self-assessment. Numerous studies have attempted to evaluate response consistency by comparing how the scale with and without vignettes correlated to a relatively objective measure (d'Uva, Lindeboom, O'Donnell, & Doorslaer, 2011; King et al., 2004; van Soest, Delaney, Harmon, Kapteyn, & Smith, 2011). For instance, Van Soest et al. (2011) examined a scale meant to gauge problematic drinking behavior in students by comparing response scale differences between the scale (with and without vignettes) and an "objective" measure. They used a self-report of how many drinks the students consume (number count) as a relatively more objective measure of drinking

behavior. They found that the model which used the anchoring vignettes was closer to the relatively objective measure than the model without the anchoring vignettes.

In the case of Benny and Catharine, the researcher may compare the correlations between a self-report cardiovascular health questionnaire with and without using anchoring vignettes and the results of a spirometry assessment (a lung function test that measures how much air one can exhale). The idea is that if participants use the same criteria to rate the vignettes as they do to rate themselves, then the anchoring vignettes worked to control for DIRC. In other words, according to this procedure, if response consistency is met the correlation between the self-reported health ratings and the spirometry assessment should be smaller in magnitude than the same correlation after adjusting the scores using the vignettes.

However this strategy for assessing response consistency tends to be limited in practice. The first issue with this strategy is that the logic behind the procedure is somewhat circular. When participants use the same criteria to rate the vignettes as they do to rate themselves the correlation with the relatively more objective measure would be higher for adjusted scores than unadjusted scores. Hence, if the correlation between the self-assessment and the objective measure changed before and after adjustment, researchers assume that this result indicates that the participants must have used the same criteria to rate the vignettes as they had to rate themselves. This line of thought seems to assume that response consistency had already been met and it does not attempt to isolate the effect from alternative explanations. Another issue with testing the assumption of response consistency in this way is that there is often no viable objective measure for the

construct. It thus begs the question: "if there is a viable objective measure then why would the researcher continue to use a subjective self-report?"

In sum, DIRC leads to incomparable scores. Viewing how participants rate the anchoring vignettes provides insight into how participants interpret the response scale. Adjusting scores using anchoring vignettes allows for more comparable scores by placing the scores on a common metric. The goal of using anchoring vignettes is to control for DIRC as a source of DIF. So far in the literature, the evaluation of anchoring vignettes provides indirect methods for determining whether the use of anchoring vignettes to adjust scores functions in this way. In response, I propose that anchoring vignettes should also be directly evaluated by examining how the use of anchoring vignettes impacts the presence of DIRC in the sample.

Proposed Method for Evaluating Anchoring Vignettes with DIF Testing

The present paper argues for a two-step procedure for evaluating the impact of anchoring vignettes on the factor structure of item-level scores from a self-report measure. The two-step procedure assumes that individuals have been randomly assigned to either a control or experimental condition. The control group did not receive the anchoring vignettes whereas the experimental condition received the anchoring vignettes prior to answering self-assessment items. Each step is discussed in terms of study 1 and study 2.

This two-step procedure involves DIF testing using a multi-group confirmatory factor analysis framework. DIF is described as either non-uniform or uniform, both of which can be detected using multi-group confirmatory factor analysis. In this framework, item responses are viewed as a linear function of unobserved or latent variables. For

example, in a one factor model, item responses are a function of an intercept, weight given to an unobserved factor, and an error term. The intercept is the expected value of the observed score given a latent level of zero. The weight given to an unobserved factor represents the relationship between the observed item responses to the unobserved latent construct. In other words, this weight is an indication of how salient the item is to the latent factor. Finally, the error term represents the variance in the observed score that is not accounted for by the latent factor.

Non-uniform DIF at the item level, as shown in Figure 5, reflects an inequality of factor weights (and possibly intercepts as well) across the control group and vignette group, which suggests that the item has differential saliency to the latent variable across groups. In this case, finding non-uniform DIF between a control group and a vignette group would indicate that the use of anchoring vignettes made the saliency of the item to the latent factor different than when no anchoring vignettes were used. This result could occur if there was substantial DIRC in the control group (e.g. Figure 1) that was not present in the vignette group. How exactly this may occur will be discussed in a subsequent section.

Uniform DIF at the item level, as shown in Figure 6, reflects an equality of factor weights and an inequality in item intercepts across groups. Thus, uniform DIF indicates that while an item has equal saliency to the factor across groups, the observed item scores differ consistently across the latent continuum. In other words, the observed differences on the item reflect different amounts of the latent factor and systematic bias. When comparing a control group to a vignette group, uniform DIF is an indication that DIRC was not present in the control group or the vignette group, but that the vignette group's

adjusted scores were systematically different from the control group's scores. If each item contains uniform DIF, then this manifests at the scale level as no DIF, but with latent mean differences (Figure 7). Assuming that the control and vignette groups are actually equivalent, observed differences on the scale level reflect differences in item intercepts for all items across groups. This would occur if there was no DIRC (e.g. Figure 2), but through using anchoring vignette to adjust scores the participants' assigned scores were systematically transformed either up or down.

If, at the item level, each item does not contain either non-uniform or uniform DIF, and the groups are actually equivalent, then no DIF on the scale level with equivalent latent means, as shown in Figure 8, would be present. In this case, no DIF reflects an equality of both factor weights and intercepts across the groups. Assuming that the control and vignette groups are actually equivalent, using anchoring vignettes would not have made the factor structure for the vignette group any different from the control group. In this case, the anchoring vignettes would have no impact on the factor structure or score values. This could occur if participants rated anchoring vignettes in the same way and in the way that the researcher operationalized them. For instance, no DIF with equivalent latent means would occur if Benny and Catharine had rated the low vignette as 2 (*disagree, = low vignette*) and the high vignette as 4 (*agree, = high vignette*). There would be no DIF because the participants rated the vignettes in the same way. Additionally, the item intercepts would be equivalent because adjusted values for the vignette group would not differ from the self-assessment within the vignette group.

Now that the connection between DIF and the comparison between control and vignette groups has been established, on may consider how to implement the procedure

for DIF testing using a multi-group confirmatory factor analysis framework. An important consideration before completing the invariance testing is that the vignette-adjusted scale must have the same number of categories as the unadjusted scale. To this end, the researcher should have an idea of how the original response scale should map onto the vignette adjusted categories. For instance, in the health example the original response scale consisted of five categories: 1 (*strongly disagree*), 2 (*disagree*), 3 (*neutral*), 4 (*agree*), and 5 (*strongly agree*). Recall that the adjusted scale also consisted of five response categories: 1 (self-assessment is *lower than the low vignette*), 2 (*self-assessment is equal to the low vignette*) 3 (*self-assessment is higher than the low vignette and less than the high vignette*), 4 (*self-assessment is equal to the high vignette*), 5 (*self-assessment is higher than the high vignette*). Categories 2 and 4 imply equivalence to the low vignette and the high vignette. In order to ensure that the original response categories corresponds to the vignette adjusted categories, the low vignette should be drafted such that it operationalizes what it means to disagree with the items (2). The high vignette should also be drafted such that it operationalizes what it means to agree with the items (4). If the original and adjusted scales contain the same amount of categories, then invariance models can be assessed.

Invariance models

Multi-group confirmatory factor analyses of invariance involve examining a series of increasingly constrained, nested models. Typically, researchers will test configural, metric, and scalar invariance models. The first step, configural invariance, tests whether the basic structure of the confirmatory factor analysis model is equivalent across groups. For the current study, the two groups are the control group and the

vignette group. This model essentially tests whether the number and pattern of items defining the latent variable are the same across groups (Millsap & Meredith, 2007).

After configural invariance is supported, analyses proceed to testing the metric invariance model. Recall that non-uniform DIF occurs when items have differential saliency to the items across groups. The metric invariance model tests for non-uniform DIF by constraining unstandardized factor pattern coefficients to be equivalent between groups. If metric invariance is not supported then the groups differ in how strongly the items relate to the latent factor (Millsap & Meredith, 2007). Therefore, if this model fits worse than the configural model, then there is evidence of non-uniform DIF between groups. If the metric model does have adequate fit and does not fit worse than the configural model, then the scalar invariance model is assessed.

The scalar model tests uniform DIF by further constraining the item intercepts to be equivalent across groups. Recall that uniform DIF is present when items have equal saliency to the latent factor across groups, but the item intercepts differ. If scalar invariance is not supported, then uniform DIF is found in at least one item, but not all items. Uncovering that at least one, but not all items has non-equivalent intercepts across groups indicates that there is an issue with the item stem. Since the ratings of the anchoring vignettes are used to adjust all self-assessment items the adjustment of the items should result in a consistent transformation across all items. So, if adjusting one item results in a different intercept than the control group, then it would be expected that the other items also have different intercepts.

One may also test a model where error variances are equivalent across group, however this form of invariance is considered to be strict (Byrne, 1998; Cheung &

Rensvold, 1999). To test this model, the scalar model can be further constrained such that the error variance associated with each item is constrained to be equivalent across groups. Uncovering an item has non-equivalent error variance across groups indicates that using anchoring vignettes results in less variance not accounted for by the latent variable. One implication of this finding could be that the amount of measurement error had been reduced.

Given at least partial measurement invariance (i.e., scalar invariance is found for some items), latent means may be compared. Latent means can be modeled using structured means modeling (SMM) (see Hancock, 2003; Thompson & Green, 2006). This procedure is akin to a *t*-test in that it tests the statistical significance of group mean differences and can used to obtain effect size estimates. However, it has an advantage over observed mean difference testing because it is theoretically free of measurement error since the latent factors, instead of the observed variables, are modeled.

When modeling the latent means, one can find that either there is a difference is group means or that they are equivalent. Finding that there is a difference in the latent means suggests that uniform DIF was present across the groups on an item level for all of the items in the scale. As previously mentioned, this would occur if there was no DIRC (e.g. Figure 2), but through using anchoring vignette to adjust scores the participants' assigned scores were systematically transformed either up or down.

Finding that there the latent means are equivalent, on the other hand, reflects an equality of both factor weights and intercepts across the groups. This result indicates that using anchoring vignettes did not have an impact on the factor structure or the score

values. This could occur if participants rated anchoring vignettes in the same way and in the way that the researcher operationalized them.

The Expected Impact of Anchoring Vignettes on a Factor Structure

Study 1. The measurement invariance of the scale (i.e. configural, metric, scalar, error invariance) between the control and vignette groups before any score adjustment should be investigated because it is possible that simply viewing the vignettes changed how an individual interprets the response categories when rating themselves. In other words, just viewing the anchoring vignettes may have an impact on the factor structure of the scale. For instance, Buckley (2008) found evidence that self-assessments were influenced due to context effects introduced by reading anchoring vignettes.

The implication of Buckley's (2008) finding on the evaluation of anchoring vignettes using invariance testing is that any comparison between adjusted and unadjusted scores may be confounded by the fact that the participants viewed the anchoring vignettes. So, as to determine the degree to which this confound may impact the interpretation of further analyses, study 1 assesses the measurement invariance of the scale between the control group and the vignette group before self-report item scores are adjusted. Recall that the control and vignette groups were random samples from the same population, so only little difference due to sampling error in parameter estimates across the two groups is expected. Thus, if there were large differences in parameters then they could be attributed to exposure to the anchoring vignettes. It is unclear how a context effect of the vignettes would impact the factor structure. However knowing whether just viewing the anchoring vignettes impacted the factor structure would inform the interpretation of the results found in subsequent analyses.

Study 2. Study 2 assesses measurement invariance between the control group and the vignette group after vignettes were used to adjust the self-report item scores. If the item parameters across groups for study one were equivalent, then the invariance analyses in study two would examine whether the process of adjusting SOS scores using the anchoring vignettes affected the psychometric properties of the test. If the anchoring vignettes did control for DIRC as a source of DIF, then the configural model would be supported, but the metric model would not be supported. Metric non-invariance would indicate that non-uniform DIF between the groups occurred because there was DIRC present in the control group that was not present in the vignette group. In other words, metric non-invariance would be a detection of non-uniform DIF due to the presence of DIF in the control group that was controlled for in the vignette group.

If anchoring vignette did not control for DIRC then complete invariance (configural, metric, and scalar invariance) would be expected. Additionally, either latent means differences or equivalent mean differences could occur. Configural invariance would be expected even if using anchoring vignettes did not control for DIRC, because transforming scores using the anchoring vignettes should not result a number and pattern of factor loadings for the vignette group that differs from the control group. If each item in the scale contains uniform DIF across groups, then uniform DIF would be detected by a statistically significant latent mean difference. On the other hand, if no items contained uniform DIF across the groups, then no DIF would be detected by equivalent latent means.

In the following sections, I discuss the expectations for if anchoring vignettes control for DIRC and if they do not in more depth.

Expectation for configural invariance model. The configural model should be supported regardless of whether the anchoring vignettes function to remove DIRC. The use of anchoring vignettes should not affect the number or pattern of factor pattern coefficients, because exposure to the anchoring vignettes and the score adjustment should not change how the participant conceptualized the construct. Additionally, anchoring vignettes should only be applied to unidimensional constructs, because the vignettes must be perceived as falling along a single continuum. So, the model should be unidimensional with the same number of factor pattern coefficients across both the control and vignette groups. If configural invariance is not supported, then one should not use anchoring vignettes in that context because the assumption of unidimensionality was violated.

Non-uniform DIF as indication that DIRC was removed from sample. Given that the configural model is supported, one would next test the metric model. It is expected that the metric model would not be supported if the anchoring vignettes function to remove DIRC as a source of DIF. In other words, metric non-invariance would indicate that non-uniform DIF between the groups occurred because there was DIRC present in the control group that was not present in the vignette group. The presence of DIRC in the vignette group and absence of it in the control group would manifest in metric non-invariance because the use of anchoring vignettes to control for DIRC affects the rank order of cases.

For example, consider how Benny and Catharine used the response scales in Figure 9. In this scenario, both participants had very different standards for how to rate health. Notice that Benny rated the low health vignette as a 1 (*strongly disagree*) and the high health vignette as a 3 (*neutral*). On the other hand, Catharine rated the low and high

health vignettes as 4 (*agree*) and 5 (*strongly agree*) respectively. Both participants rated their own health as a 4 (*agree*), however because they had different interpretation of the response categories their adjusted scores changed their rank order. Benny's score became a 5 (*higher than high vignette*) and Catharine's score became a 2 (*equal to low vignette*). Benny and Catharine were rank ordered as equivalent before score adjustment, but when DIRC is controlled Benny was ranked as higher in health than Catharine. Figure 9 show an example with only one item, however the same logic would apply to multiple items. As long as the participants rated themselves in a similar matter across the items, the rank order would change in a similar way when their scores are adjusted.

This logic implies that using anchoring vignettes to adjust scores may impact covariance among items by affecting rank order among cases. Recall that if DIRC is present, then how scores are adjusted would vary across participants, because they rated the anchoring vignettes differently. Thus, scores across participants would be adjusted in different ways, which could impact the change in rank order of participants. If the covariances among items changed in magnitude because of the change in rank order, so would the estimate of factor pattern coefficients. Said differently, if adjusting scores using anchoring vignettes removed DIRC, then the item correlations would increase and thus the factor loadings could increase. So adjusting scores using anchoring vignette would result in stronger relationships between the items and thus between the items and the factors.

How the remove of DIRC in the vignette group can impact the factor pattern coefficient can be demonstrated through tracing rules. For example, consider a situation where a researcher has a covariance matrix for the control group consisting of several

items. Item 1 and item 2 have an observed correlation of .64 (the standardized solution is discussed for the sake of simplicity). Ideally, we want to create a model with parameter estimates that allows us to approximate the values within the original matrix. If item 1 and item 2 load on the same factor, and only that factor, then the completely standardized solution for the predicted correlation between two items can be obtained by the following equation (Brown, 2006):

$$COV(X1, X2) = \sigma_{2,1} = \lambda_{X11}\phi_{11}\lambda_{X21}$$

In the completely standardized solution (i.e., reproducing correlations), the factor variance (ϕ) will equal one. So the solution for the implied correlation between the two items is now the product of the item's factor pattern coefficients ($\lambda_{X11}\ and\ \lambda_{X21}$). Consequently, the correlation between item 1 and item 2 is reproduced if both pattern coefficients were equal to .80.

What would happen, however, if the researcher examined the covariance matrix for the vignette group after score adjustment? Recall that in the case of Benny and Catharine, their rank order before score adjustment was different from their rank order after score adjustment. Similarly, the rank order among cases in a control group would differ from a vignette group after score adjustment. In this hypothetical case, the correlation between item 1 and item 2 in the vignette group was different from the control group, which indicates that what parameter estimates are needed to reproduce their relationship will also need to change. For example, if the correlation increased to .81 then we could reproduce this correlation if both pattern coefficients increased to .90. Notice how the estimate of the items' factor pattern coefficients increased as the magnitude of the relationship between those items increased. Consequently, since vignettes should

impact the rank-order of cases, it is reasonable to anticipate that metric invariance will fail to hold when comparing a control group to a vignette group after score adjustment. Stated differently, non-uniform DIF occurs when DIRC is present in the control group and absent in the vignette group (due to score adjustment).

Uniform DIF as indication that DIRC was not present in either sample. Another possibility to consider is that the anchoring vignettes could have an impact on the factor structure of the scale without controlling for DIRC. Consider a scenario depicted in Figure 10, where Benny and Catharine had very similar criteria for how they rated the anchoring vignettes. Notice that in this example, the high vignette, which was designed to represent the 4 response category, was rated as a 5 (*strongly agree*) across participants. Benny and Catharine also rated the low vignette, which was designed to represent the 2 response category, as a 4 (*agree*). The participant's self-assessment scores, both 3 (*neutral*) were rescaled such that the value changed to a 1 (less than the low vignette), but the participants' rank order remained the same.

In this scenario, participants interpreted the response categories in the same way (no DIRC) and score adjustment resulted in a different value, but with the same relative rank order. As such, then comparing the control group to the vignette group, the only difference would be that using anchoring vignettes resulted in different item intercepts. So the factor structure between the two groups would be equivalent, but the item intercepts would differ across groups.

If all item intercepts were different in a consistent way across the control and vignette group, then this result would be reflected in a significant latent mean difference between the groups. Since the groups were randomly sampled from the same population,

there is no reason to believe that their actual latent means would differ substantially. Thus, finding that there is a difference in the latent means suggests that uniform DIF was present across the groups on an item level for all of the items in the scale. This result would occur if there was no DIRC (e.g. Figure 2), but through using anchoring vignette to adjust scores the participants' assigned scores were systematically transformed either up or down. As shown in Figure 2, on the item level, uniform DIF means that participants with the same latent level of the construct will systematically differ on their item score depending on their group. This systematic difference in scores would reflect the act that participants used the response categories for the anchoring vignettes in the same way, so scores were adjusted consistently either upward or downward.

If at least one item intercept, but not all, was different across groups then this result would be reflected in scalar non-invariance. However, scalar non-invariance would not be expected because the same vignette ratings are used in the vignette group to adjust scores. So scalar non-invariance would indicate an issue with at least one item stem.

No DIF as indication that DIRC not present in either sample. If, however, the metric model was supported when comparing the control and vignette group score, then this result would indicate that non-uniform DIF was not present in the sample. In this case, further invariance analyses (i.e. scalar invariance model, error invariance model, and latent means modeling) would be conducted. Complete invariance, as evidenced by adequate fitting configural, metric, and scalar models along with insignificant latent means differences, would indicate that anchoring vignettes had no impact on the factor structure of the scale. In this case, the adjusted score would be the same value as the self-assessment for most participants. This result is demonstrated in Figure 11, where Benny

and Catharine actually had the same criteria for rating the anchoring vignettes. Recall that the low vignette was designed to represent the 2 response category and that the high vignette was designed to represent the 4 response category. Now notice how both Benny and Catharine rated the low vignette as 2 and the high vignette as 4. As a result, their self-reported scores of 4 (*agree*) remained the same after adjustment.

Applied Example with Student Effort

To illustrate how anchoring vignettes can be evaluated using measurement invariance procedures, a self-report measure of student effort is used as an example. In the following sections, I demonstrate how to apply invariance testing to assess the impact of anchoring vignettes on the factor structure of the effort subscale of the student opinion scale (SOS; Sundre, 1999). The effort subscale consists of five self-report items that pertain to how much students tried on a previous assessment. The comparability of scores is a concern for any scale used to make comparisons across participants, however it is particularly a concern when studying diverse populations. As such, effort is an especially suitable construct to examine with anchoring vignettes given its use in low-stakes testing for higher education, a context in which the validity of score comparisons across participants has substantial consequences for the institution. Higher education is also a context projected to experience a substantial shift in student demographics (Hussar & Bailey, 2013). This change in student population suggests that the comparability of student effort scores may be considered a validity concern for using the SOS in higher educational assessment.

Use of SOS in Higher Education

With the increased demand for higher education institutions to demonstrate student learning, colleges and universities have begun to rely more heavily on low-stakes assessments. Low-stakes assessments, by definition, do not have consequences tied to individual performance. However, results of low-stakes assessments may be used to make consequential inferences about student learning. Various challenges have resulted from the increased use of low-stakes assessments in higher education (Liu, 2011).

A major challenge is addressing concerns about the validity of using test scores from low-stakes assessment to make inferences about student learning. Student effort may be a source of construct-irrelevant variance, which occurs when test scores partially reflect something other than the intended construct (Messick, 1989). Stated differently, effort is a concern for low-stakes assessment because, with no stakes tied to the student, some student may not put forth adequate effort on the assessment (Cole & Osterlind, 2008; Sundre, 1999). As a result, low-stakes assessments of student ability may be underestimated (Wise & DeMars, 2005).

Clearly, student effort has consequences on the appropriate interpretation of the assessment scores because the results will not accurately reflect student learning. As a way of controlling for student effort as a source of invalidity, the SOS is commonly used for measuring effort in low-stakes assessments. Thus, the presence of DIRC is problematic for using the SOS because in order to correctly control for student effort, scores must be comparable across individuals. However, similar to the health example mentioned above, the SOS is a subjective self-report measure that may be subject to DIRC, which limits the comparability of scores. For instance, consider a situation where two students report their level of effort on an assessment. Both students have the same

latent level of effort, however one student has a particularly high standard for what it would mean to consider themselves as having demonstrated high effort on the assessment, whereas the other student had a relatively lower standard. As a result, the students used the response categories differently, which resulted in scores that did not reflect how similar the students actually were.

This example demonstrates how the assumption that participants interpret the response categories in the same way may be violated in practice. As previously alluded to, incomparability of scores due to DIRC may be even more of a concern when studying diverse populations. This point leads to a second reason why student effort is a particularly suitable construct to examine with anchoring vignettes: higher education institutions are becoming more diverse.

Increased College Campus Diversity and Comparability of SOS Scores.

Over the last 30 years, there has been a marked increase in diversity on college campuses. For example, enrollment for Hispanic students, one of the largest growing student populations, has grown by more than twenty percent since 2010 (Hussar & Bailey, 2013). The National Center for Education Statistics (Hussar & Bailey, 2013) projects that between 2010 and 2021 enrollment will increase by forty-two percent for students who are Hispanic, twenty-five percent for students who are Black, twenty percent for students who are Asian/Pacific Islander, and one percent for students who are American Indian/Alaska Native. Furthermore, other student populations are also on the raise, such as international students, returning adult students, and students with disabilities (Kennedy & Ishler, 2008).

Clearly the student population is becoming less homogeneous. As Pascarella (2006) observes, the characteristics of the student population are changing such that it is becoming less so "made up of White undergraduates from middle or upper-middle class homes, ages 18 to 22, attending four-year institutions full time, living on campus, and having few if any family responsibilities" (*p.* 512). This increase in diversity among the student population may come with new challenges in collecting data. With a broader range of backgrounds it may be more likely that DIRC will be more prevalent when sampling this population. For instance, recall that individual differences may affect interpretation of response categories, such as socio-economic status (e.g. Van Doorslaer & Jones, 2003), age (e.g. Groot, 2000), and nationality (e.g. King et al., 2004). Hence, it may prove to be beneficial to use anchoring vignettes in the assessment of college students as to strategy to control for DIRC as a source of DIF.

The Current Studies

Purpose

The purpose of the current research was to demonstrate how anchoring vignettes can be evaluated by examining the impact of using anchoring vignettes on the factor structure of the SOS. To accomplish this, two studies were designed. Both studies involved examining measurement invariance between randomly assigned groups: a group that had not received anchoring vignettes (control) and a group that had (vignette). The first study was designed to examine whether the factor structure of the SOS effort subscale was impacted by simply responding to the anchoring vignettes. The second study sought to determine whether adjusting scores using anchoring vignettes resulted in non-uniform DIF, possibly due to the removal of DIRC, as compared to the control

group. For both studies, measurement invariance procedures within the multiple group confirmatory factor analyses framework were used to answer the following research questions.

It should be mentioned that methods of using invariance testing to evaluate anchoring vignettes are not restricted to the design used in this study. For instance, if a control group is not available then one may simply compare a single group before and after score adjustment. This design would be similar to the one presented in this study, however one would need to acknowledge the possibility that just being exposed to the anchoring vignettes may have affected the scale in addition to the score adjustment.

Research Questions

Research question 1. Does simply rating the anchoring vignettes (without adjustment to scores) impact the factor structure of the SOS effort subscale? This research question was addressed in study 1 by examining measurement invariance and latent mean difference between the control group and the vignette group before score adjustment. Recall that the control and vignette groups were random samples from the same population, so only little difference due to sampling error in parameter estimates across the two groups is expected. If the experimental manipulation (just viewing and responding to the anchoring vignette) did not have an effect, then complete invariance and equivalent latent means would be expected. Thus, if there were large differences in parameters then they could be attributed to the participants' exposure to the anchoring vignettes.

If measurement invariance (configural, metric, and scalar invariance) was supported then latent means would be modeled. If there was no DIF and no latent mean

difference between the groups then it would be concluded that the SOS effort items were not impacted by simply rating the anchoring vignettes. If however, measurement invariance was not supported or there was a latent mean difference, then it would be concluded that simply rating the anchoring vignettes had an impact on the factor structure of the scale or the total scores, which would have implications for interpreting the results of study 2.

Research question 2. Does using anchoring vignettes to adjust scores on the SOS effort subscale result in items that are more salient to the factor? This research question was addressed in study 2 by examining measurement invariance between the control group and the vignette group after score adjustment. The advantage of comparing a control group to the vignette group after adjustment rather than just comparing the vignette group before and after adjustment is that the control group represents a set of scores that would be similar to the vignette group's scores if they had not received anchoring vignettes. In this way, the comparison between the control group and the vignette group after adjustment is a more robust comparison since the control results are not confounded by exposure to the anchoring vignettes.

If adjusting scores resulted in more salient items then metric invariance would not be supported, which would indicate that non-uniform DIF, possibly due to DIRC, was present in the sample. If this was the case then scalar invariance could also be assessed to examine the impact of anchoring vignette further, but latent means would not be modeled. However, if metric invariance was supported then I would proceed with additional tests of invariance to examine whether different parameter estimates are also invariant across groups.

CHAPTER TWO

Literature Review of Student Effort

The Student Opinion Scale (SOS) is a widely used self-report measure of student effort. In fact, the SOS has been used to measure effort in at least 30 published studies, 25 unpublished studies, and 33 K-12 schools and universities (Sessoms & Finney, 2015). The SOS has also been used in at least 9 countries (e.g., Saudi Arabia, China, Canada: Abdelfattah, 2010; Ip, Lui, Chien, Lee, Lam, & Lee, 2012; Smith, Given, Julien, Ouellette, & DeLong, 2013). Within this section, I will first review the development of the SOS which is followed by an examination of existent validity evidence.

Development of the SOS

The SOS was adapted by Sundre (1999) from the Motivation Questionnaire (Wolf & Smith, 1993) to measure effort and importance. The development of this scale was informed by expectancy-value theory of achievement motivation theory (Pintrich and De Groot, 1990; Pintrich & Schunk, 2002), which postulates that effort is a function of the perceived importance of a task and the participant's expectancy for success. The SOS contains ten items equally divided between two subscales: effort and importance. Effort is defined as the amount of mental energy the participant expended on the test (e.g., "I gave my best effort on these tests") and importance is defined as how much value the participant placed on test performance (e.g., "Doing well on these tests was important to me"). Each item is rated using a 5-point Likert-type scale ranging from 1= strongly disagree to 5 = strongly agree.

Existing Validity Evidence for the Student Opinion Scale

Validity is defined as "…the degree to which evidence and theory support interpretations of test scores for propose use of tests" (American Educational Research Association, 2014, *p*.11). Many studies have examined validity evidence for the SOS as an indication of effort and importance. For instance, Thelk, Sundre, Horst, and Finney (2009) applied Benson's (1998) strong program of construct validation to examining SOS scores. This framework for construct validation includes three stages. First, the substantive stage involves defining the theoretical and empirical domains of the construct. Second, the structural stage involves examining the relations among the observed scores. This stage can be thought of as pertaining to examining structural validity evidence. Finally, the external stage examines hypothesized relations of the construct with other variables of theoretical interest. This stage can be thought of as pertaining to examining external validity evidence.

Substantive Validity

In their substantive stage, Thelk et al. (2009) discussed how the SOS items mapped onto the expectancy-value model of achievement motivation theory (Pintrich, 1989; Pintrich and De Groot, 1990). Their explanation of the scale's substantive validity mirrors that of Thelk and Sundre (2007) who argued that the development of the scale,

> …parallels current conceptions of motivation theory… in which an individual's willingness to put forth the effort to learn or to display learning would be contingent upon the individual's interest or the perceived importance of the task, as well as their disposition to put forth the necessary work to complete the task (*p*. 4).

In other words, the item writing process for the SOS was informed by expectancy-value model of achievement motivation theory. This model, as previously mentioned, states that effort is a function of the perceived importance of a task and the participant's

expectancy for success. However, it should be noted that the SOS items do not exactly map onto the expectancy-value model. The SOS was designed to represent only particular aspects of the model. The expectancy-value model defines perceived importance as how much value the participant places on the task. Expectancy for success is defined as the participant's perception of how likely s/he is to succeed on the task given her/his ability level. The SOS importance subscale items were written to reflect perceived importance by gauging how valuable the participant perceived the assessment to be. This subscale maps onto the theoretical model, however the effort subscale does not. The SOS effort subscale was designed to measure the mental energy the participant expended on the assessment, which theoretically is the result of expectancy and importance. So, expectancy, as defined by expectancy-value model, is not measured by the SOS.

Structural Validity

Thelk et al. (2009) examined the dimensionality of the SOS in a low-stakes testing context across independent groups of first-year and sophomore students from a four year university (two groups of entering freshmen and two groups of sophomores/juniors). For each group of students, they compared a one dimensional model (where all items load onto a single factor) against a two dimensional model (where the effort and importance items load onto separate but correlated factors). Across each sample a two factor structure represented the data significantly better than a one dimensional model. The effort and importance subscales were also moderately correlated across samples (r = 0.47 and 0.46 in first-year students; r = 0.50 and 0.49 in sophomore students), which suggests that the items represented unique, but related constructs. The

items had high internal consistency across the samples ($\alpha = 0.80$ to 0.89 for the importance subscale; $\alpha = 0.83$ to 0.87 for the effort subscale), which suggests that participants responded consistently to both subscales. Thelk et al. (2009) also found similar results when examining the factor structure of the SOS scores across participant gender and testing methods (i.e. computer-based and paper-and-pencil testing mediums).

Finally, it is noteworthy that effort subscale of the SOS was examined for longitudinal measurement invariance. Sessoms and Finney (2015) administered the SOS to college students at a four year university when they were incoming first-years. Those same students were then administered the SOS a year and a half later after they completed 45-70 credit hours (second-semester sophomores). They found complete invariance across time, which suggests that the subscale measured the same construct in the same way over time. They also found that the effort subscale items had high internal consistency at each time point ($\omega = .84$ for time 1; $\omega = .88$ for time 2).

External Validity Evidence

There are many examples of research investigating external validity evidence for the SOS (Kong, Wise, & Bhola, 2007; Sundre & Kitsantas, 2004; Sundre & Wise, 2003; Swerdzewski, Harmes, & Finney, 2011; Thelk et al., 2009; Wise, 2006; Wise & Kong, 2005). For example, in their external stage, Thelk et al. (2009) reviewed how the SOS has been shown to relate to response time effort, performance on assessments, and whether the assessments are high or low-stakes. Most studies have focused on how the SOS effort subscale related to these external variables.

Effort and response time effort. Response time effort (RTE; Wise & Kong, 2005) refers to the amount of time a participant spends responding to an assessment item.

RTE can be collected when the participant completes a computer-based assessment. RTE is theoretically related to self-reported effort because it can identify rapid-responding behavior, which provides a lower-bound effort threshold (DeMars, 2007). In other words, a participant with low effort is expected to engage in rapid-responding behavior, which involves not taking enough time to read each item and thus not providing a thoughtful response. Students who fail to rapidly respond however, have not necessarily tried since students may take a long time to respond to an item for numerous reasons that have nothing to do with effort. RTE is an appealing variable to compare to the SOS because it is not a self-report measure. In fact, it is usually collected without the student's awareness.

Wise and Kong (2005) examined whether participant's rating of their effort on the SOS was related to their RTE. They found that, indeed, these measures were moderately correlated ($r = 0.54$). They also examined whether SOS effort scores differed by level of RTE (i.e. low, medium, high RTE). Groups differed in their SOS effort scores, with low RTE showing low effort scores and high RTE showing high effort scores. These results were further replicated with varying procedures for operationalizing RTE grouping criteria (Kong, Wise, & Bhola, 2007).

Effort and performance. In a low-stakes testing context, effort is expected to have a positive correlation with performance on an achievement assessment because participants who try harder on an achievement assessment should be more likely to perform well than participants who do not try as hard. This same correlation may not be observed in a high-stakes testing context due to restriction of range since most students are likely motivated to perform well to avoid negative consequences of poor

performance. Indeed, past studies found that the SOS effort scores positively correlate with performance (Sundre & Wise, 2003; Thelk, 2006; Wise & Kong, 2005). For example, Thelk (2006) and Wise and Kong (2005) found moderate and positive correlations between the effort subscale and an achievement test ($r = 0.30$ and $r = 0.34$ respectively). These results suggest that highly motivated participants tend to perform better on an assessment than those who are unmotivated.

Discriminant evidence has been examined by investigating the correlation between SOS effort scores and proxies for general ability. The SOS effort scores do not strongly correlate with performance on general ability assessments, such as SAT scores (Steedle, 2014; Sundre & Wise, 2003; Swerdzewski et al., 2011; Wise & Kong, 2005). For instance, Wise and Kong (2005) found small correlations between effort and SAT scores ($r = 0.14$ for SAT verbal and $r = 0.01$ for SAT quantitative). These findings were expected given there are no theoretical reasons to expect students of different ability to systematically differ in the amount of effort expended on an achievement test.

Since effort scores relate to performance on achievement tests and not to general ability, the SOS is often used to control for effort in a low-stakes testing environment. Recall that in low-stakes testing effort may be considered a source of construct-irrelevant variance since students who fail to give effort have true score estimates that are systematically lower than their true score. Low SOS effort scores provides a means to identify, and potentially correct for, effort as a source of construct-irrelevant variance. Indeed, removing the scores of participants with low SOS effort scores results in an increased average score on the achievement assessment, reduced standard errors of measurement, and higher correlations between achievement and proxies for general

ability (Steedle, 2014; Sundre & Wise, 2003; Swerdzewski et al., 2011; Wise & Kong, 2005).

Known group differences: Effort and the stakes of an assessment.

Participants who complete an assessment in a low-stakes context tend to report lower SOS scores than those who complete the same assessment in a high-stakes context (Sundre & Kitsantas, 2004; Thelk et al., 2009). This relationship between effort and whether the assessment is low or high-stakes is expected given that effort is theoretically a product of the participant's perceived importance of a task and her/his expectancy for success. Specifically, the stakes of an assessment affect the participant's perceived importance of the task, because when the outcome of an assessment is perceived as important then the participant will expend more effort than when the outcome is not perceived as important.

In their external stage, Thelk et al. (2009) demonstrated how stakes impact achievement scores. They compared the psychometric properties of an achievement assessment between a group who had consequences tied to the assessment and a similar group that did not have consequences tied to the assessment. The group who had personal consequences tied to the assessment had a higher average score and less variability than the other group. This result indicates that as the stakes of the assessment decreases, participants display greater variability in how much effort they are willing to expend. As a result, there will be greater variability of performance scores in a low-stakes context than a high-stakes context.

The Student Opinion Scale and DIRC

Clearly, the SOS is a widely used scale for a reason. Users of the SOS can reasonably argue that the scale is well justified for their purpose given the literature has consistently demonstrated support for a two-factor structure and has examined multiple lines of external validity evidence. However, if the SOS is used to control for effort as a form of construct-irrelevant variance in low-stakes assessment it is important that the measurement of effort provides comparable scores between students. Comparability between scores is also needed for other research purposes, such as any situation where high stakes decisions or inferences are made based on comparing scores across individuals.

To date, existent validity evidence has yet to address the possibility that DIRC, as a source of DIF, is present in the measurement of student effort. Given that the presence of DIRC limits the comparability of scores it is conceivable that a failure to correct for this issue has also limited existent validity evidence. For example, DIRC may impact the magnitude of correlations among SOS items, correlations with external variables (d'Uva et al., 2011; King et al., 2004; van Soest, Delaney, Harmon, Kapteyn, & Smith, 2011), as well the identification of students who fail meet a minimal motivational threshold in low-stakes testing contexts.

Additionally and as previously mentioned, the SOS is commonly used in low-stakes higher educational assessment, which is experiencing substantial demographic shifts. The presence of DIRC within measures relying upon self-report may therefore be of growing concern since DIRC may be more prevalent in diverse populations. In response to this concern, two anchoring vignettes that were written for the effort subscale

of the SOS were examined by investigating their capacity to remove DIRC as a source of non-uniform DIF.

CHAPTER THREE

Methods

Data Collection Procedure

All first-year students were required to participate in the university's Assessment Day. During Assessment Day, two series of cognitive and non-cognitive assessments were administered to the students by trained proctors. Students were randomly assigned into one of two groups. Both groups received an identical series of assessments. The series consisted of two cognitive tests (72 items and 32 items) and a battery of non-cognitive items (83 items), which was followed by a measure of student effort. The only difference between the groups was that one group received anchoring vignettes prior to the effort measure (n = 341) and the other group did not receive anchoring vignettes prior to the effort measure (n = 478). The effort measures were both administered at the end of the assessment series and asked participants to rate their opinion about the extent to which they tried on all of the previous assessments.

Measures

Anchoring vignettes

Two anchoring vignettes were drafted to represent two levels of effort (see Appendix A). The behavioral description of each anchoring vignette was based on research on student effort. Specifically, research has shown that unmotivated students have a higher frequency of rapid responding (defined as responding to an item without sufficient time to read and comprehend the item) than motivated students (Wise, 2006). Effort levels were operationalized by the degree of guessing behavior displayed in each scenario. The high effort vignette displayed no guessing behavior by describing a student who carefully reads each item and provides a thoughtful answer. The low effort vignette

displayed some guessing behavior by carefully reading each item, but guessing on items that appear to be difficult. The low vignette was thought to represent a 2 (disagree) on the original response scale and the high vignette was thought to represent a 4 (agree).

The instructions for rating each vignette asked the participants to rate how much they agree that the person in each scenario had given her/his best effort. Each vignette was rated on a five point Likert agreement scale which ranged from 1 (*disagree strongly*) to 5 (*strongly agree*).

Student Opinion Scale

The Student Opinion Scale (SOS; Sundre, 1999) is frequently used as a self-report measure of student opinion about a previously administered assessment (see Appendix B). The SOS contains ten items and two subscales. Five items were written to measure effort, which was defined as the amount of mental energy expended on a test by the student. Five more items were written to measure importance, which was defined as how much value is placed on test performance by the student. Each item is scored on a five point Likert-type scale which ranges from 1 (*disagree strongly*) to 5 (*strongly agree*). Total scores can range from 5 to 25. For the current studies, only the student effort items was be examined. An example of an item written to measure effort is, "I engaged in good effort throughout these tests."

Participants

A total of 819 (including cases of missing or omitted data) incoming college first-year students at a mid-sized state university in the southern region of the United States participated in this study. Students were randomly assigned to two groups, one of which received the vignettes (n = 341) and the other did not receive the vignettes (n = 478).

After removing cases of missing data (i.e. cases that did not complete all SOS items and/or assessments), a sample of 703 were retained, with 301 students who received the anchoring vignettes and 402 students who did not receive the anchoring vignettes. The sample size for the vignette adjusted group decreased further to 274 when cases of interval scores were omitted. Table 1 shows the demographic information for both groups of retained students (including gender, ethnicity, and age). Both groups are predominately Caucasian, female, and of traditional college age.

Analytic Procedures

Data analysis occurred in three stages. First, descriptive statistics of the SOS scores were examined and interpreted. Measurement invariance models were then assessed between the control group and the vignette group before adjusting scores (study 1) and between the control group and the vignette group after adjusting scores (study 2). Lastly, given complete or partial measurement invariance was supported (configural, metric, scalar invariance), latent mean differences was examined for both studies.

Data screening

The effort scores for each group in both studies were screened for outliers and assessed for normality. Both univariate and multivariate outliers were detected in the control group, the vignette group before adjustment, and the vignette group after adjustment. Univariate outliers were detected by examining frequency distributions and plotting a histogram for each item. A nonparametric approach of identifying univariate outliers was also taken by comparing item scores to the first and third quartile. Multivariate outliers were detected by examining Mahalanobis distance, which measures how many standard deviations away a person's set of scores is from the sample means for

all variables correcting for intercorrelations. Invariance testing was completed both with and without univariate and multivariate outliers. Since the results and conclusions drawn from scores were not substantially different, those cases were retained.

The scores for the control group, the vignette group before adjustment, and the vignette group after adjustment were also assessed for univariate and multivariate normality. Univariate normality was assessed by examining skew and kurtosis of the effort items. Skewness greater than the absolute value of 2 and kurtosis greater than the absolute value of 7 would have been considered indications of non-normality. These cut-off values were chosen based on research, which demonstrated that larger values may bias standard errors and fit indexes (e.g. Finney & DeStefano, 2006). As shown in Table 2, no item values of skew or kurtosis from any of the groups exceeded these cut-off values. As such, univariate normality was supported even though there some were univariate outliers.

However, multivariate normality was not supported for any of the conditions. Mardia's coefficient, an indication of multivariate kurtosis was 33.39 for the control group, 23.07 for the vignette group before score adjustment, and 9.83 for the vignette group after score adjustment. There is no strict cut-off value for Mardia's coefficient. Bentler (1998), in a SEMNET post, commented on how to assess Mardia's coefficient: "Personally I wouldn't worry much if it is 3 or 4 or 5 in [absolute] value. But if it is more than that, perhaps 10, or 20, that's probably a pretty good indication that the data truly are not normal, and it's a pretty good guess that the normal statistics on the model will be very inaccurate." The Mardia's coefficients for the control group (33.39) and vignette group before score adjustment (23.07) clearly indicate that the data are truly non-normal.

Since Bentler's (1998) recommendation for evaluating Mardia's coefficients was intended as a rule of thumb rather than a strict cut-off, the Mardia's coefficients for the vignette group after score adjustment (9.83) is close enough to 10 that it can be considered an indication of non-normality.

Finally, multicollinearity was assessed by examining the correlation matrix for highly related items ($r > .80$). The SOS items were moderately correlated across groups. The largest correlations were .580 for the control group, .607 for the vignette group before adjustment, and .633 for the vignette group after adjustment, which indicates that there was no extreme multicollinearity present in the data.

Estimation of measurement invariance and latent means models

LISREL 8.80 (Jöreskog & Sörbom, 1996) was used to analyze covariance matrices from the unadjusted and vignette adjusted groups. Maximum likelihood (ML) estimation was used to produce fit indexes and estimate model parameters. ML estimation was chosen over other estimators (e.g., Generalized Least Squares, Weighted Least Squares) because it is more sensitive to model misspecification (see Olsson, Foss, Troye, & Howell, 2000). Since the data for all groups was not multivariate normal, the Satorra-Bentler correction to χ^2, fit indexes, and standard errors was used for all analyses (see Finney & DiStefano, 2006). The Satorra-Bentler correction to fit indexes corrects for non-normality in the $RMSEA_{SB}$, CFI_{SB}, and TLI_{SB}. The SRMR is not corrected for non-normality (Yu & Muthen, 2002).

Setting the metric of the latent variable

When performing a multiple-group CFA, it is best practice to set the metric of the latent variable by constraining a path from the factor to one of the indicators to a value of

1.00. This way of setting the metric is preferable over setting the latent variance to 1.00, because it does not assume that there is equal factor variances across groups (see Marsh, 1994). The item set to 1.00 is often referred to as the "referent variable." For the present study, item 1 ("I engaged in good effort throughout these tests") was constrained to a value of 1.00 for all models examined. This item was chosen to be the referent variable after it was found to be invariant across groups, which was determined by examining a series of models where item 1 was set to be invariant while each item in turn was used as the referent variable (Rensvold & Cheung, 2001).

Assessing model fit

Measurement invariance was examined by investigating a series of nested CFA models. The types of invariance examined were configural, metric, scalar, and equivalent error variances. Overall fit of the CFA models was assessed using chi-square (χ^2_{SB}) significance test of global fit, several indices of approximate fit, and correlation or mean residuals. The χ^2_{SB} significance test is a measure of the exact model-data fit, meaning it indicates how much the original covariance matrix deviates from the reproduced matrix. The χ^2_{SB} significance test is very sensitive to sample size in that the likelihood of obtaining a significant χ^2_{SB} increases as sample size increases (Hu & Bentler, 1998). Considering that large sample sizes are needed for assessing CFA models, indexes of approximate fit are often used to supplement the χ^2. These indexes are either absolute or incremental.

Both the standardized root mean square residual (SRMR) and the root mean square error of approximation ($RMSEA_{SB}$) are examples of absolute fit indexes. They both are indications of global model misfit between the observed and reproduced

covariance matrixes. However, researchers tend to use both indexes because while the SRMR is most sensitive to misspecified factor correlations (i.e. the relationship between factors), the $RMSEA_{SB}$ is most sensitive to misspecified factor pattern coefficients (i.e. the relationship between the factor and the item). Both SRMR and $RMSEA_{SB}$ values range from 0 to 1 with smaller values indicating better fit. Although the SRMR is not adjusted for non-normality, Yu and Muthen (2002) suggest a cut-off value of equal to or less than .07. When using Satorra-Bentler correction to fit indexes, the recommended cut-off value for the $RMSEA_{SB}$ is less than or equal to .05 (Yu & Muthen, 2002).

The comparative fit index (CFI_{SB}) and the Tucker-Lewis fit index (TLI_{SB}) are incremental fit indexes. They compare the fit of the reproduced model to a null model where all variable correlations are fixed to zero. The difference between the CFI_{SB} and the TLI_{SB} is that the CFI_{SB} uses the expected value of the χ^2 under the noncentral χ^2 distribution whereas the TLI_{SB} uses the expected values of the χ^2 under the central chi-square distribution. The CFI_{SB} ranges from 0 to 1 with larger values indicating better fit. The TLI_{SB}, on the other hand, is a nonnormed index so it does not necessarily range from 0 to 1, but larger values also indicate better fit. Both the CFI_{SB} and the TLI_{SB} are moderately sensitive to simple model misspecification & very sensitive to complex model misspecification (Hu and Bentler, 1998). The recommended cut-off value for both indexes is .95 or above (Yu & Muthen, 2002).

Finally, local model-data fit for each model was assessed by examining correlation residuals: the difference between the observed correlations and reproduced correlations. Correlation residuals greater than |.15| indicated that relationships among items were either greatly under or over estimated by the reproduced model (Kline, 2011).

For the scalar model, mean residuals, which are the difference between the observed item means and the reproduced item means for each group, were examined. There is no recommended cut-off value because the mean residuals are interpreted on the metric of the scale. As such, what counts as a large residual depends on the scale being used. For example, a residual of $|.50|$ is was this study considers a large residual for a five point scale.

Each nested model was compared using a corrected chi-square difference test ($\Delta\chi^2_{SB}$), difference in CFI_{SB} (ΔCFI_{SB}), and difference in correlation residuals. Specifically, a decrease in model fit between nested models as indicated by a statistically significant $\Delta\chi^2_{SB}$ and/or ΔCFI_{SB} values of > .01 would indicate worse model fit for the more constrained model (Cheung & Rensvold, 2002). In regards to computing Satorra-Bentler scaled $\Delta\chi^2_{SB}$, one cannot obtain it by computing the difference between Satorra-Bentler scaled χ^2_{SB} values as can be done with normal theory χ^2 values. Instead, Satorra-Bentler scaled $\Delta\chi^2_{SB}$ must be computed by using the formula introduced by Bryant and Satorra, (2012).

Given configural, metric, and scalar invariance, the latent mean differences between groups (κ) would be estimated using structured means modeling (SMM). The statistical significance of κ (at $p<.05$) and the effect size was used to make a final evaluation of whether there was a statistical and practical difference in the latent means. A difference in latent means across the control group and vignette group before adjustment would indicate that uniform DIF was present for all items. This result would indicate that just viewing and rating the vignettes systematically how participants rated themselves on the self-assessment. A difference in latent means across the control group

and vignette group after adjustment would indicate that the whole process of using the anchoring vignettes to adjust scores (including viewing and rating the vignettes) resulted in uniform DIF for all items. This result would suggest that there was a systematic difference between the control group's self-assessment scores and the vignette group's adjusted scores.

CHAPTER FOUR

Results

Study 1: Comparison between Control Group and Vignette Group before Adjustment

First, the vignette group's raw scores (before adjustment) were used for assessing measurement invariance between the control group and the vignette group. Recall that the only difference between the control group and the vignette group prior to being rescaled using the anchoring vignettes was that the experimental group was exposed to the anchoring vignettes whereas the control condition did not receive vignettes. Also, recall that the groups were randomly assigned from the same population so any group difference is attributable to just viewing and responding to the anchoring vignettes. The first set of invariance analyses tested the expectation that the SOS was equivalent between the control group and the vignette group prior to rescaling.

Descriptive Statistics

Descriptive statistics and inter-item correlations for the SOS items for the control group and the vignette group before score adjustment are presented in Table 3. The control and vignette groups had very similar inter-item correlations. The control group inter-item correlations ranged from .318 to .580. The vignette group inter-item correlations ranged from .329 to .607. Additionally, the two groups exhibited similar means across items. Control group item means ranged from 3.50 to 4.30 and the vignette group item means ranged from 3.43 to 4.33. This range of means indicates that participants tended to either feel neutral toward the item or agree with the item. Both groups had similar standard deviations across items. The standard deviations for the

control group ranged from 0.72 to 1.12 and standard deviations for the vignette group ranged from 0.69 to 1.11. Only item 3 across both groups had a standard deviation over 1, so there may be limited variability within the other items.

Given the similarity between the control and vignette groups on the inter-item correlations, item means, and item standard deviations, it would appear as though at least complete invariance (configural, metric, and scalar) with equivalent latent means was a likely result. Since the items inter-correlations were similar across groups, there would be no reason to believe that the factor pattern coefficients would differ. Additionally, since the item means and standard deviations are similar it would appear as though item intercepts and latent means would not differ.

For each item across both groups, the majority of participants endorsed either a 4 (*agree*) or a 5 (*strongly agree*), however for each item the entire scale was utilized. The small standard deviations and the fact that most people rated themselves highly implies that there may be a ceiling effect present in the data across both groups.

Configural Invariance

The first step for assessing configural invariance is to assess the unidimensional model effort for each group separately. As shown in Table 4, the unidimensional model for both groups obtained overall adequate global fit: χ^2_{SB} (5) = 24.67, $p < .001$, $RMSEA_{SB}$ = .10, SRMR = .052, TLI_{SB} = .90, CFI_{SB} = .95 for the control group; χ^2_{SB} (5) = 26.41. $p < .001$, $RMSEA_{SB}$ = .12, SRMR = .048, TLI_{SB} = .94, CFI_{SB} = .97 for the vignette group. The χ^2_{SB} values were both statistically significant, which was expected given the sample size. Additionally, the $RMSEA_{SB}$ values for the control (.10) and vignette (.12) groups was higher than the recommended cut off, however these values

likely reflected the simplicity of the model. The correlation residuals ranged from .000 to |.098| (see Table 6). Taking into account the local and global fit indexes, the unidimensional model fit the data for both groups.

The large standardized pattern coefficients for both groups, shown in Figure 12, demonstrate that the items functioned well as indicators of effort. Across the control and vignette groups, most items were associated with standardized pattern coefficients at or above .60 and .67, respectively. This result indicated that the factor explained at least 36% (control group) and 44% (vignette group) of each items variance, except for item 5 for control group (26%) and item 1 for vignette group (30%). The low performance for item 5 was consistent with previous studies (Thelk et al., 2009). Reliability, calculated using the CFA's unstandardized parameters (Raykov, 1997, 2004), was adequate (ω = .71 for control group and .79 for vignette group).

The second step for assessing configural invariance is to assess the unidimensional model effort across both groups concurrently. As expected given the fit of each group separately, the combined model fit of the data across both groups obtained adequate global fit: χ^2_{SB} (10) = 50.93, p < .001, $RMSEA_{SB}$ = .11, SRMR = .052, TLI_{SB} = .94, CFI_{SB} = .97 (see Table 5). Additionally, there was no area of localized misfit as all correlation residuals were smaller than an absolute value of .15. Given these results, configural invariance was supported and so this model was used as a baseline with which to compare the metric model.

Metric Invariance Model

In the metric invariance model, the unstandardized factor pattern coefficients were constrained to be equal across groups. As shown in Table 5, the metric model

obtained adequate global fit: χ^2_{SB} (14) = 55.69, $p < .001$, $RMSEA_{SB}$= 0.09, TLI_{SB} = 0.96, CFI_{SB} = 0.97. The correlation residuals for each group were also small (i.e., no correlation residuals over |.15|), which indicates good local fit. Those residuals, presented in Table 7, ranged from |0.006| to |0.119|. When compared to the configural model, the metric model did not fit statistically or practically worse than the configural model, $\Delta\chi^2_{SB}$ (4) = 7.660, $p = .105$, $\Delta CFI_{SB} < .001$. In other words, only reading and responding to the anchoring vignettes with no adjustment to scores did not significantly affect how strongly the items relate to the latent variable. Given metric invariance was supported, the metric model was used as a baseline model for testing scalar invariance.

Scalar Invariance Model

To examine scalar invariance, the metric model was constrained so that the magnitude of all item intercepts were the same across groups. As shown in Table 5, scalar invariance was supported in this sample. The scalar invariance model fit statistically and practically no worse than the metric model: $\Delta\chi^2_{SB}$ (4) = 7.047, $p = .133$, $\Delta CFI_{SB} = .01$. Additionally, the global fit indexes indicate adequate fit: χ^2_{SB} (18) = 78.72, $p < .001$, $RMSEA_{SB} = .01$, SRMR = .07, $TLI_{SB} = .96$, $CFI_{SB} = .96$. Additionally, there was good local fit as shown by the small mean residuals presented in Table 8, which ranged from |0.01| to |0.05|. The difference should be interpreted on a five point metric. Given that the scalar invariance model was supported, this model was used as a baseline to assess a model which held the error variances invariant.

Invariant Error Variance Model

Next, the scalar model was then further constrained by setting the error variances equivalent across groups. This model assesses whether the item's error variances are

equivalent across groups. Invariance in the error variances was supported in this sample. The model fit statistically and practically no worse than the scalar invariance model: $\Delta\chi^2_{SB}$ (5) = 7.422, p = .191, ΔCFI_{SB} = .01. Moreover, the invariant error variance model had adequate global fit: χ^2_{SB} (23) = 72.11, p < .001, $RMSEA_{SB}$ = .08, SRMR = .065, TLI_{SB} = .97, CFI_{SB} = .97. Most of the correlation residuals for each group were also small (i.e., correlation residuals under |.15|), which indicates overall good local fit. However, there was a localized area of misfit for the vignette group, where the expected relationship between items 5 and 4 was underestimated by the model. The correlation residuals are presented in Table 9 and ranged from |0.002| to |0.164|. Since the fit indexes were adequate in both groups and there was only one local area of misfit (a residual of -0.16), this model could still be considered to have overall adequate fit. This result supports the conclusion that reading and responding to anchoring vignettes did not substantially impact the SOS error variances. This result suggests that just viewing and responding to anchoring vignettes did not result in less variance not accounted for by the latent variable.

Latent Mean Difference Models

Since configural, metric, and scalar invariance were supported, the latent mean difference for the completely invariant model was estimated to evaluate if just viewing and responding to the anchoring vignettes resulted in a consistent difference in all items across groups. Said another way, the latent mean difference was examined to assess whether the vignette group's scores were systematically different from scores in the control group. Since the groups were randomly assigned to from the same population, it would be reasonable to attribute any difference between the groups as resulting from the manipulation: just viewing and responding to the anchoring vignettes. As such, a

statistically significant latent mean difference would suggest that just viewing and responding to the anchoring vignettes resulted in systematically different scores on the self-assessment. This result would appear as uniform DIF on the item level and if all items contain uniform DIF then the result would appear as latent mean difference on the scale level.

Finding a statistically significant latent mean difference between the control group and the vignette group before adjustment would have implications on the next study where the control group is compared to the vignette group after adjustment. The main implication would be that just viewing and responding to the anchoring vignettes impacted the participant's self-assessment, so any effect found after adjusting scores using the anchoring vignettes would be confounded with the effect of just being exposed to the vignettes. Of course, the effect of just being exposed to the vignettes would not be eliminated by finding that the control group and the vignette group before adjustment had equivalent latent means. However, finding that the latent means between these two groups was equivalent would suggest that just viewing and responding to the vignette had a minimal impact on scores. In this case, it would be reasonable to attribute differences in groups found in study two mainly to the process of score adjustment.

As shown in Table 10, the estimated latent mean difference was not statistically significant and the effect size for the group difference was less than .001, which indicated the latent effort mean for the vignette group before score adjustment was very similar to the control group's latent effort mean. Table 10 also shows the observed means for each group and the observed effect size, which demonstrates the similarity of the groups' observed effort means. Although the observed effect size is similar in magnitude to the

latent effect size, the latent effect size has the advantage of being theoretically free of measurement error (Hancock, 2003). The fact that the control group and the vignette group had similar latent and observed mean scores in effort was expected given the groups were randomly assigned from the same population.

In summary, the finding that the control group and the vignette group before adjustment were invariant and had equivalent latent means implies that there was no effect of just viewing and responding to the anchoring vignettes. In the next study, invariance between the control group and the vignette group after adjustment will be assessed to determine whether there was an effect of adjusting the scores using anchoring vignettes.

Study 2: Comparison between Control Group and Vignette Group after Adjustment

Next, the vignette group's scores were adjusted using the anchoring vignettes. Measurement invariance was then assessed between the control group and the vignette group (after adjustment). Given that the item parameters from the control group were equivalent to the vignette group's scores before adjustment, this next set of invariance analyses examined whether the process of adjusting SOS scores using the anchoring vignettes affected the psychometric properties of the test.

Descriptive Statistics

Descriptive statistics and inter-item correlations for the control group and the vignette group after score adjustment are presented in Table 11. The item standard deviations for the vignette group after score adjustment ranged from 0.76 to 0.98, which suggests that the variability of scores did not change substantially from score adjustment. Recall that the control group item means ranged from 3.50 to 4.30 and the standard

deviations ranged from 0.72 to 1.12. Also recall that before adjustment the item means for the vignette group ranged from 3.43 to 4.33. After adjustment the item means for the vignette group ranged from 2.80 to 3.53. Clearly, the item means for the vignette group after adjustment were much lower than the item means before adjustment, which were more similar to the control group. This shift in item means in the vignette group after adjustment suggests that adjusting scores may have systematically transformed scores downward.

It was expected that if DIRC were present in the control group, but not present in the vignette group after score adjustment then the inter-item correlations for the vignette group would be larger in magnitude then the inter-item correlations for the control group, hence the items would be more salient to the factor. Upon examining the item correlations in Table 11, it is apparent that the inter-item correlations for the vignette group are very similar to those in the control group. This lack of clear difference in the inter-item correlations suggests that the items in the vignette group are likely no more salient to the factor than the control group, which may indicate that metric invariance would be supported. In other words, it would not appear as though DIRC, if present in the control group, was corrected for in the vignette group by adjusting scores.

Table 12 shows the change in descriptive statistics and inter-item correlations within the vignette group from before adjustment to after adjustment. As previously discussed, the item means for the adjusted scores are lower than the unadjusted scores. Additionally, the distribution for the adjusted scores are more normal than the distribution for the unadjusted scores. The item standard deviations for the adjusted scores are very similar, but values for the skew and kurtosis changed slightly. The

kurtosis values for the unadjusted scores ranged from -0.46 to 3.45, whereas the kurtosis values for the adjusted scores ranged from -0.62 to 1.21. For all items expect item 3, the kurtosis values were closer to 0 for the adjusted scores. Also, the skew values for the unadjusted scores ranged from -0.47 to 1.45, whereas the skew values for the adjusted scores ranged from -0.99 to -0.26. For all items, the skew values were slightly lower for the adjusted scores.

The inter-item correlations in Table 12 demonstrate that while there was some change in rank order among cases due to adjusting scores, the impact was minimal. Consistently, the inter-item correlations for the adjusted scores was higher than the inter-item correlations for the unadjusted scores. However, most of the inter-item correlations for the adjusted scores were only slightly larger in magnitude than those for the unadjusted scores. For instance, the correlation between item 1 and item 2 was 0.509 for the vignette group before adjustment. This correlation increased to 0.633 after adjustment. While this change may be considered substantial, other correlations increased only slightly. The correlations between item 4 and item 5, for example, was 0.587 before adjustment and it increased only slightly to 0.591 after adjustment. This suggests that while adjusting scores may have accounted for some DIRC, it was not enough to substantially impact scores.

Another indication that the rank order among cases within the vignette group did not change substantially is the correlation between each unadjusted item with its corresponding adjusted item. For instance, if there was no change in rank order, then if the unadjusted item 1 was correlated with the adjusted item 1, the correlation would equal 1.00. If the correlation is moderate, then this result indicates that there was substantial

change in rank order among cases. So, the unadjusted items were correlated with their corresponding adjusted items to further examine how score adjustment impacted rank order among cases within the vignette group. Each correlation was high (item 1: $r = .765$, item 2: $r = .797$, item 3: $r = .822$, item 4: $r = .827$, item 5: $r = .814$). This result was expected given the previous invariance results and provides additional evidence that the rank order among cases did not change as a result of using anchoring vignettes to adjust scores.

Figure 13 through Figure 17 show the percent of participants who endorsed (or were placed into) each response category across the groups for each item, including the control group and the vignette group, both before and after adjustment. Notice how for most items, participants in the control group and the vignette group before adjustment tended to rate themselves with either a 4 (*agree*) or 5 (*strongly agree*). Then after the vignette group's scores were adjusted, most *C* values across items became either 3 (*more than low vignette and lower than high vignette*) or 4 (*equal to high vignette*).

The frequencies for how the anchoring vignettes were rated, presented in Table 13, shed light on this finding. Recall, that the low vignette was thought to represent a 2 (disagree) on the original response scale and the high vignette was though to represent a 4 (agree). If DIRC was present then one would expect that participants would be varied in how they endorsed the vignettes. For the low vignette, there was variety in how participants endorsed it. However, for the high vignette, there was very little variety. The majority (58.0%) of participants in the vignette group rated the low effort vignette with a 2 (*disagree*) and a large majority (79.6%) of participants rated the high effort vignette with a 5 (*strongly agree*). Given that before adjustment most participants rated

themselves as either a 4 or 5 *and* most participants also rated the high vignette with a 5, it makes sense that most participant's *C* values became either a 3 (*more than low vignette and less than high vignette*) or 4 (*equal to high vignette*). In other words, because most participants rated the high vignette as a 5, there was no way for them to rate themselves as higher than the high vignette. As a result, scores appear to have been systematically shifted downward. This shift downward suggests that there may be non-uniform DIF across all items between the control group and the vignette group after adjustment. As such, there may be a statistically significant latent mean difference.

Configural Invariance Model

First, the unidimensional model of effort was assessed for each group separately. Table 14 shows that adequate global fit was found for both groups: χ^2_{SB} (5) = 24.67, p < .001, $RMSEA_{SB}$ = .10, SRMR = .052, TLI_{SB} = .90, CFI_{SB} = .95 for the control group and χ^2_{SB} (5) = 21.58, p < .001, $RMSEA_{SB}$ = .11, SRMR = .037, TLI_{SB} = .96, CFI_{SB} = .98 for the vignette group. Despite significant χ^2_{SB} values and high $RMSEA_{SB}$ values, the adequacy of the other fit indexes was evidence that the unidimensional model obtained adequate global fit across groups. Also, all correlation residuals for each group were small (i.e., < |.15|), which indicates good local fit. Table 15 presents the correlation residuals, which ranged from .000 to |.104|.

Figure 18 shows that the items functioned well as indicators of effort across the control and vignette groups. Most items obtained a standardized pattern coefficients at or above .60 in the control group and .69 in the vignette group. The factor explained at least 46% (control group) and 48% (vignette group) of each items variance, except for item 5 for control group, for which the factor only explained 26% of variance in the item.

Figure 19 shows how the parameters changed within the vignette group from before adjustment to after adjustment. Most items obtained a standardized pattern coefficients at or above .67 and .60 respectively. The factor explained at least 44% (before adjustment) and 48% (after adjustment) of each items variance, except for item 1 for the vignette group before adjustment (30%). However, after adjustment the factor explained 67% of variance in item 1 for the vignette group. In sum, when comparing parameters within the vignette group before and after adjustment, it is apparent that most parameters are very similar with the exception of item 1. However, overall it would not appear as though the factor pattern coefficients did not change substantially as a result of adjustment.

Reliability, calculated using the CFA unstandardized parameters (Raykov, 1997, 2004), was adequate ($\omega = .71$ for control group and $\omega = .81$ for vignette group). Recall that reliability for the vignette group before adjustment was $\omega = .79$. So, the scale was similarly reliable before and after adjustment for the vignette group.

Next a unidimensional factor structure across both groups concurrently was modeled. This combined model obtained adequate global fit: $\chi^2_{SB}(10) = 46.74$, $p < .001$, $RMSEA_{SB} = .10$, $SRMR = .052$, $TLI_{SB} = .95$, $CFI_{SB} = .98$ (Table 16). Given these results, configural invariance was supported and so this model was used as a baseline with which to compare the metric model.

Metric Invariance Model

Table 16 shows that, when compared to the configural model, the metric model did not fit statistically or practically worse, $\Delta\chi^2_{SB}(4) = 4.547$, $p = .337$, $\Delta CFI_{SB} = .00$. Furthermore, the metric model obtained adequate global fit indexes: $\chi^2_{SB}(14) = 47.02$, p

< .001, $RMSEA_{SB}$ = .09, SRMR = .062, TLI_{SB} = .97, CFI_{SB} = .98. There were also no areas of localized misfit. All correlation residuals were small (i.e. < |0.15|). Those residuals can be viewed in Table 17 and ranged from |.012| to |.119|. Contrary to hypothesis, these results indicate that adjusting scores with the use of anchoring vignettes did not significantly affect how strongly the items relate to the latent variable. Given metric invariance was supported, the metric model was used as a baseline model for testing scalar invariance.

Scalar Invariance Model

The results for the fit of the scalar model were mixed, however it can reasonably be argued that scalar invariance was supported. As shown in Table 16, when compared to the metric model, the scalar model did fit statistically worse than the metric model, $\Delta\chi^2_{SB}$ (4) = 20.37, $p < .001$, ΔCFI_{SB} = .01, suggesting uniform DIF was present. However, the model obtained adequate global fit: χ^2_{SB} (14) = 68.66, $p < .001$, $RMSEA_{SB}$ = .09, SRMR = .083, TLI_{SB} = .97, CFI_{SB} = .98.

To diagnose misfit in the scalar model, the mean residuals were examined. Table 17 shows the observed means, expected means based on the scalar invariance model, and their difference for each item across the control group, the vignette group before adjustment, and the vignette group after adjustment. The expected mean, is simply the item mean one would expect if there was scalar invariance. So the difference mean residual represents the discrepancy between the observed mean and the mean that would be expected if there were scalar invariance. Notice that item 3 has the largest difference in both the control group (-0.09) and vignette group after adjustment (0.07). For the control group, the observed mean for item 3 was 3.50 and the expected mean was 3.59.

The observed mean for the vignette group after adjustment was 3.25 and the expected mean was 2.73. This result may suggest that this item contains more uniform DIF than the other items. If this were the case, then it may be the case that the process of adjusting scores using anchoring vignettes impacted this item differently than how it impacted the other items. This result would be unexpected given that the same ratings for the anchoring vignettes were used to adjust all items.

Recall that the difference should be interpreted on a five point metric. Thus, the mean for item 3 for each group can be said to have been either over or underestimated by less than a tenth of a point on a five point scale. This difference did not appear to be practically significant. To examine whether the scalar model fit practically worse than the metric model, latent means were modeled with and without the intercept of item 3 constrained. In other words, latent means were modeled for a completely invariant model (i.e., all items intercepts constrained) and a partially invariant model (i.e., all item intercepts constrained except for item 3 which was freely estimated). The effect sizes of both models were then compared with the expectation that if the scalar model did fit practically worse than the metric model, then the latent effect sizes should substantially differ.

When assessing scalar invariance models, the latent mean of the vignette group after adjustment was fixed to zero so that the latent mean of the control group reflected the latent mean difference between groups (Hancock, 1997; Steenkamp & Baumgartner, 1998; Thompson & Green, 2006). This procedure was done for both the completely invariant model and the partially invariant model. The resulting latent mean differences were used to calculate latent Cohen's *d* effect sizes. The effect size for the completely

invariant model was 1.37 and the effect size for the partially invariant model was 1.30, which was a minimal difference in effect size. So, it was concluded that the scalar model fit practically no worse than the metric model.

Invariant Error Variances Model

Next, the scalar model was further constrained by setting the error variances equivalent across groups to evaluate whether adjusting scores impacted the variance not explained by the factor. The results for the error invariances model were mixed in this sample. As shown in Table 16, the model fit statistically worse than the scalar invariance model: $\Delta\chi^2_{SB}$ (5) = 16.00, p = .007, ΔCFI_{SB} = .01. Notice that the ΔCFI_{SB} was acceptable, which indicates that the difference in fit between the scalar and error invariance model may not be practically significant.

Additionally, the invariant error variance model had adequate global fit: χ^2_{SB} (23) = 84.59, p < .001, $RMSEA_{SB}$ = .09, SRMR = .072, TLI_{SB} = .97, CFI_{SB} = .96. There was also evidence of overall local fit as most correlation residuals were small (i.e., < |.15|). There was one localized area of misfit for the vignette group, where the expected relationship between items 5 and 4 was underestimated by the model (r = -.19). The correlation residuals, which ranged from 0.00 to |0.19|, are presented in Table 19.

Since the fit indexes were adequate both groups and there was only one local area of misfit, this model could still be considered to have overall adequate fit. As such, it could reasonably be argued that the error invariance model was overall supported. This result supports the conclusion that using anchoring vignettes to adjust scores did not substantially impact the SOS error variances.

Latent Mean Difference Models

Since configural, metric, and scalar invariance were supported, attention was paid to interpreting the latent mean difference for the completely invariant model mentioned in a previous section. As shown in Table 20, the estimated latent mean difference was statistically significant ($\hat{\kappa} = 0.70$, $p < .001$). Additionally, the effect size for the group difference was 1.30, which indicated that the scores for the vignette group (after adjustment) were well over a standard deviation ($SD = 0.53$) lower than the control group's scores. Table 20 also shows the observed means for each group and the observed effect size. Although the observed effect size is similar in magnitude to the latent effect size, the latent effect size has the advantage of being theoretically free of measurement error (Hancock, 2003).

Recall that in the first study there was no latent mean difference between the control group and the vignette group before adjustment. As such, the latent mean difference in the second study reflects a systematic effect of rescaling that occurred across items rather than a true difference in the construct between the two groups. This systematic effect of rescaling can be observed when comparing the observed and expected item means for the scalar models along with the item's residuals (Table 18). In study one, the observed and expected means for each item were similar across groups.

For instance, the observed ($M = 4.25$) and expected ($M = 4.27$) means for item one in the control group were similar to the observed ($M = 4.31$) and expected ($M = 4.27$) means for the vignette group prior to score adjustment. This was not the case for the item means after using the vignettes to adjust scores for the second study. In the second study, the item means *across groups* were quite different. Take for example, the first item. The observed ($M = 4.25$) and expected ($M = 4.27$) means for this item in the control group

were substantially different from the observed (M = 3.48) and expected (M = 3.52) means for the vignette group after score adjustment.

Importantly, the observed and expected means were very similar within both the control group and the vignette group after adjustment. Because of the similarity between the observed and expected means, scalar invariance was supported even though the latent mean between the two groups was statistically significant. Scalar invariance would not have been supported if some, but not all of the items were adjusted downward. If this were the case, then the observed and expected means within groups would differ substantially.

As previously mentioned, this systematic effect of rescaling was demonstrated in Figure 13 through Figure 17. Notice how students in the control group and the vignette group before adjustment tended to place themselves in response categories that represented higher levels of effort. Adjusting scores for the vignette group tended to place student in response categories that represented more moderate levels of effort. As such, it is clear from the difference in item means between the control group and the vignette group that there was non-uniform DIF for all the items across groups due to adjusting scores in the vignette group. Since all items contain uniform DIF across groups, this group difference in item intercepts manifested in a statistically significant latent mean difference.

The effect that adjusting scores had on the item means can be explained by how the participants in the vignette group rated the anchoring vignettes. Table 13 shows that most participants (79.6%) rated the high vignette with a 5 (*strongly agree*). Notice that in Figure 13 through Figure 17 most participants in the vignette group rated themselves as

either a 4 (agree) or 5 (*strongly agree*). If a participant rated the high vignette as a 5 (*strongly agree*), which is the highest possible response option, then there was no way for the participant's score to be adjusted to a 5 (*more than the high vignette*) no matter how the participant rated herself/himself on the self-assessment. In this way, most scores were adjusted to either a 3 (*more than the low vignette, but less than the high vignette*) or 4 (*equal to the high vignette*).

Since most participants rated the high vignette as a 5 and most participants' adjusted scores were relative to how they rated the high vignette, the adjustment of scores approximated a simple linear transformation. In this case, metric non-invariance was not found because adjusting scores did not account for any DIRC. Recall that how participants rate the anchoring vignettes sheds light on their interpretation of the response categories. Since the high vignette was predominately used to adjust scores and since the participants largely agreed on how to rate this vignette, no DIRC was accounted for. If there was more of a variety of how participants endorsed the high vignette then adjusting scores would have impacted the rank order among cases. However, since there was no DIRC among participants with regards to the high vignette, the rank order among cases was not affected by adjusting scores.

It is clear that using anchoring vignettes had an impact on the scores, which could influence interpretations and inferences made based on those scores. However, it is unclear whether the vignette adjusted or the unadjusted scores are a more adequate reflection of student effort. The implication of this finding will be examined in the discussion section.

CHAPTER FIVE

Discussion

There were two main goals of the current research. The first goal was to propose using invariance testing to directly evaluate how the use of anchoring vignette impacts DIRC. Specifically, invariance testing was used to determine whether just being exposed to the anchoring vignettes influenced scores. Then invariance testing was proposed for determining whether DIRC was removed from a sample as a result of using anchoring vignettes. If DIRC was removed from the sample then non-uniform DIF would be detected between an experimental and control group. The second goal was to demonstrate this method by examining the impact of using anchoring vignettes on the factor structure of the SOS effort subscale. Below are summaries of the substantive findings for both studies, followed by implications and limitations for using invariance testing to evaluate anchoring vignettes and for interpreting the SOS effort subscale. Directions for future research are also discussed.

Summary of Substantive Findings

Study 1

The first study was designed to determine whether just viewing and responding to anchoring vignettes (without adjustment to scores) impacted the factor structure of the SOS effort subscale. Measurement invariance was examined between the control group and the vignette group before score adjustment. Strict measurement invariance (configural, metric, and scalar invariance, and residuals) was supported. Additionally, no difference were found between the group's latent means, which would be expected if evaluating random samples of students who experienced the same testing procedures.

This finding supported the conclusion that there was no effect of just viewing and responding to the anchoring vignettes. Moreover, the results of this study aided the interpretation of the second study because it appeared as though just viewing the anchoring vignettes did not substantially impact the factor structure or latent means for the vignette group.

Study 2

The second study investigated whether using anchoring vignettes to adjust scores on the SOS effort subscale resulted in items that were more salient to the factor possibly due to the removal of DIRC. Measurement invariance was examined between the control group and the vignette group after score adjustment. There were two expected outcomes of these analyses: adjusting scores using anchoring vignettes would either 1) result in more salient items or 2) anchoring vignettes would have no impact on the factor structure. If the first expected outcome were the case, then metric invariance would not have been supported, which would indicate that non-uniform DIF, possibly due to DIRC, was present in the sample. Alternatively, if adjusting scores had no impact on the factor structure, which may possibly be due to a lack of DIRC in the sample, then complete invariance, coupled with minimal latent mean differences, would have been supported

Neither of these expected outcomes were fully supported. Using anchoring vignettes to adjust SOS effort scores did not result in items that were more salient to the factor, but it did have an effect on the latent means. Complete measurement invariance (configural, metric, and scalar invariance), as well as error invariance, was supported. However, a statistically significant difference between the group's latent means was also found. This latent mean difference was explained by the fact that most participants rated

the high vignette as a 5 and rated the self-assessment as either a 4 or a 5. Since most participants rated the high vignettes with the highest response category, few participants could rate themselves as higher than the high vignettes. So most scores were transformed downward in much the same way.

Using Anchoring Vignettes with the SOS Effort Subscale

Implications of Results

If using anchoring vignettes had corrected for DIRC, then one would expect that metric non-invariance would have been found between the control group and the vignette group after adjustment. However, there was no evidence that DIRC was controlled for, as metric invariance had been supported. Instead, it was apparent that when scores were adjusted using anchoring vignettes, they were transformed downward. This result was due to the fact that most participants rated the high vignette as a 5 (*strongly agree*) and also rated themselves as either a 4 (*agree*) or 5 (*strongly agree*), which resulted in most participants adjusted scores becoming either 3 (more than low vignette and less than high vignette) or 4 (equal to high vignette). It is clear that DIRC was not corrected for because most participants endorsed the high vignette in the same way. This vignette was most pertinent to adjusting high scoring participants, which most were. Although the use of anchoring vignettes to adjust the SOS effort subscale did not result in the removal of DIRC, it did result in substantially different observed scores, which resulted in very different latent means. This difference has implications for how the SOS effort subscale scores would be interpreted. To be more precise, since the participants' endorsement of the high vignette was consistent, the difference in observed scores did not affect the rank order of cases, but it did impact the interpretation of scores in an absolute way.

It is clear that using anchoring vignettes had an impact on the scores, which could influence interpretations and inferences made based on those scores. However, it is unclear whether the vignette adjusted or the unadjusted scores are a more adequate reflection of student effort. Although it is difficult to know which is true, I present two possibilities for what the results may mean for how scores could be interpreted.

One possibility is that instead of correcting for DIRC, the anchoring vignettes corrected for a tendency to over report level of effort. A tendency to over report may be evaluated by using the researcher's conceptualization of the level of effort illustrated by each vignette as a criterion. As shown in Table 13, most (58.0%) participants rated the low anchoring vignette with a 2 (*disagree*), which was the response category that the vignette was designed to represent. A total of 32.5 percent of participants rated it either a 3 (*neutral*, 23.7%) or a 4 (*agree*, 8.8%), which indicates that only a moderate proportion of participants over rated this vignette. However, the tendency to over rate the high vignette is much more clear as only 20.1% rated it as a 4 (*agree*), which was the response category that the vignette was thought to represent. The large majority (79.2%) over rated the high vignette as a 5 (*strongly agree*).

This tendency to over report the level of effort could be specific to the sample used in these studies. Recall that the participants were incoming first-year undergraduate students. First-year students have been shown to be more motivated on assessments than their upperclassmen peers (Sundre & Finney, 2002). As such, it may be the case that first-year students use the response categories differently from their sophomore, junior, or senior peers. If this were the case, then DIRC would be present, as evidenced by non-uniform DIF, in a sample containing both first-year and upperclassmen students.

Another possibility is that the anchoring vignettes were problematic. The anchoring vignettes in these studies were designed based on the current research about student effort. However, it is possible that the participants simply did not interpret the anchoring vignettes in the intended way because of how they were designed. For instance, potential issue with the anchoring vignettes was that they both depicted female students. It is possible that participants had different standards for the anchoring vignettes than they did for the self-assessment, either because they had different standards of effort for women or because some participants did not relate to the vignettes. To avoid this issue, King, Murray, Salomon, and Tandon (2009) recommended using either gender neutral names or acronyms when designing the anchoring vignettes.

An additional concern about the design of the anchoring vignettes is their content. Even though the anchoring vignettes were informed by current research on student effort, it possible that they were written to be too extreme. In other words, it is possible that the vignette did not represent the levels that they were thought to represent. Take the high effort vignette for example. This vignette involves a student who reads each item carefully and responds to each item in a thoughtful manner. This item was intended to represent a 4 (*agree*) response category, where a participant may rate oneself as higher than this vignette by selecting a 5 (*strongly agree*) for the self-assessment items. However, it is difficult to operationalize a level of effort that is higher than this vignette. In this way, more development of the anchoring vignettes is necessary before adjusted scores should be used to make inferences about the population.

Limitations

Two main limitations for using the results of these studies to draw conclusions about using anchoring vignettes with the SOS effort subscale are that the anchoring vignettes may be underdeveloped and the results are based on a single sample. The fact that the anchoring vignettes may be underdeveloped is a limitation of the interpretation of scores because it is possible that the scores were inadvertently adjusted downward. As a result, there is insufficient evidence to suggest that the SOS effort subscale was enhanced by adjusting scores using the current anchoring vignettes. Additional modification to the anchoring vignettes must occur before their use with the SOS effort subscale is justified.

With regard to the sample, there is a risk that the findings may not generalize to other samples. Replication of results with multiple independent samples is needed in order to judge whether the findings are stable across samples (MacCallum, Roznowski, & Necowitz, 1992). For example, the present study consisted of incoming first-year college students who tend to be more motivated to perform well on low-stakes assessments than upper classmen (Sundre & Finney, 2002). It is important to replicate this study on independent samples who are known to have lower levels of motivation.

Another limitation of using anchoring vignettes in the context of the SOS effort subscale is that the assumptions of anchoring vignettes were not yet assessed. Specific means of testing the assumptions of anchoring vignettes in the context of self-reported effort will be discussed in the subsequent section.

Future Research

A priority for future research will be to continue development on the anchoring vignettes. Before the anchoring vignettes can be used in practice, a stronger argument for

their content is necessary. After the content of the effort anchoring vignettes is better established, invariance testing should be conducted across multiple independent samples to establish whether findings are stable across samples. It is also important to test the effect of anchoring vignettes on the factor structure of the SOS across a variety of contexts. This is because the SOS is administered to a variety of populations. For example, it is used internationally (e.g., Saudi Arabia, China, Canada: Abdelfattah, 2010; Ip, Liu, Chien, Lee, Lam, & Lee, 2012; Smith, Given, Julien, Ouellette, & DeLong, 2013) and across different types of institutions (Sessoms & Finney, 2015). So, evaluation of the effort anchoring vignettes should be completed using samples that represent these populations to determine whether results can be generalized across those various contexts.

The assumptions of anchoring vignettes should also be evaluated using various methods. As previously mentioned, even if results support the position that DIRC was removed it does not necessarily mean that the anchoring vignettes functioned correctly. In order to build a strong body of evidence to justify the use of the effort anchoring vignettes, the assumptions of anchoring vignettes should be thoroughly assessed using other methods. Specifically, future research could assess vignette equivalence and response consistency by conducting cognitive interviews, which could help determine whether participants actually interpreted the effort anchoring vignettes in a consistent manner.

Future research may also investigate external validity evidence by examining the relationship between the SOS effort scale and external criteria. For example, one could compare the correlations between the SOS effort scores (before and after adjustment) and

RTE. As previously mentioned, RTE refers to the amount of time a participant spends reading and responding to assessment items. It is a good relatively objective proxy measure for effort because it is not self-reported and is usually collected without the student's awareness. So if using anchoring vignettes with the SOS effort subscale 1) removes DIRC and 2) results in more accurate scores then one would expect that the relationship between the SOS and RTE would be stronger when the SOS scores were adjusted than when they were not adjusted.

Finally, future research could examine how the effort anchoring vignettes affect how the SOS scale correlates with variables that theoretically should or should not relate to effort. For example, personality factors such as conscientiousness and agreeableness have been shown to relate to self-reported effort (Barry & Finney, 2016; Barry, Horst, Finney, Brown & Kopp, 2010; Horst, 2010). However, it is unclear why these variables relate to effort as there is no current theoretical explanation. One possibility is that these personality factors are more related to interpretation and use of response scale than related to effort. If this were the case then the relationship between the SOS and personality factors would decrease after score adjustment.

Using Measurement Invariance Testing to Evaluate Anchoring Vignettes Implications for Other Research Designs

So far in the literature, the evaluation of anchoring vignettes has provided indirect methods for determining whether the use of anchoring vignettes to adjust scores function to remove DIRC from a sample. Currently, vignettes are evaluated by assessing the assumptions of anchoring vignettes using what may be considered as indirect methods. In

response, I have proposed that anchoring vignettes should also be directly evaluated using measurement invariance procedures.

The main purpose of the current set of studies was to demonstrate that anchoring vignettes can be evaluated by conducting invariance testing procedures within a CFA framework. Through these procedures the impact of both just viewing the vignettes and using the vignettes to adjust scores was directly investigated. It was expected that if DIRC was present in the sample then invariance testing between adjusted and non-adjusted groups would result in non-uniform DIF. The presence of non-uniform DIF due to DIRC would indicate that the raw scores may have been limited in comparability. As such, adjusting scores using anchoring vignettes would allow for more comparable scores by placing the scores on a common metric.

In this set of studies invariance testing was conducted twice: once between the control group and the vignette group before score adjustment and once between the control group and the vignette group after score adjustment. Although this specific design was used in the current set of studies, invariance testing to evaluate anchoring vignettes is not limited to just this design. For instance, if a control group is not available then one may simply compare a single group before and after score adjustment. This design would be similar to the one presented in this study, however one would need to acknowledge the possibility that just being exposed to the anchoring vignettes may have affected the scale in addition to the score adjustment. In other words, with this design you may find support for measurement invariance simply because a participant's interpretation of the response categories has already been influenced by reading the vignettes. One may therefore use a single sample to evaluate measurement invariance before and after scores are adjusted

with the vignettes, while recognizing that a failure to find non-invariance may be attributed to the absence of a control condition. Additional research is needed to investigate the extent to which reading vignettes influences interpretations of response categories prior to adjusting the scores.

A quasi-experimental design investigating pre-existing groups may also be employed to investigate measurement invariance. This design may be particularly appealing if these groups reflect distinct populations thought to exhibit differences in DIRC. The factor structure of both groups could be compared before and after scores are adjusted to evaluate anchoring vignettes. If one assumes that DIRC exists between two populations then one would expect non-uniform DIF between each group prior to adjusting the scores with anchoring vignettes. This should disappear however after using the vignettes to adjust scores from both groups. When assuming that DIRC exists between two populations, a failure to eliminate DIF after adjusting the scores suggests the vignettes are in need of modification since they failed to be effective.

Challenges with using Measurement Invariance to Evaluate Anchoring Vignettes

A delimitation of using the invariance testing procedures described in this paper is that it assumes that a reflective measurement model is appropriate. This strategy should only be used when a reflective model is theoretically appropriate. To elaborate, before conducting any procedures one must decide whether the theoretical model of interest is a reflective or formative measurement model two (Edwards & Bagozzi, 2000). Reflective models are those for which a latent variable is theorized to be a common cause of patterns in item responses. Formative models, on the other hand, consist of a latent composite that in some sense consists of independent, but correlated, variables. If the construct is

formative then invariance testing, should be used in a formative framework (e.g. constraining paths from observed variables to the latent variable; disturbance term of the latent variable, etc.), rather than in a reflective framework (e.g. CFA).

Additionally, using invariance testing in a CFA framework is theory driven approach of evaluating anchoring vignettes in that it requires that the anchoring vignettes were designed using theory to map onto response categories. Some other approaches in the literature for selecting vignettes are more data driven (i.e. King et al., 2004) as opposed to construct driven. For instance, King et al. (2004) recommends writing several anchoring vignettes and using entropy analysis to determine which vignettes should be used for score adjustment. Entropy analysis is akin to item discrimination because it represents the capacity of the vignettes to differentiate among various levels of a construct (King & Wand, 2007). In other words, this procedure requires a researcher to assess which combination of vignettes maximizes entropy while minimizing the number of interval cases (i.e., cases who rated the vignettes in unexpected ways) without examining the dimensionality of adjusted scores. This data driven approach is different than the theoretically driven approach outlined in previous sections because the selection of anchoring vignettes involves making decisions based on data rather than ensuring that the anchoring vignettes theoretically correspond onto the response scale. While both approaches are viable, it is recommended for a researcher to be consistent in their choice of approach. For instance, if the literature on a construct is well developed then a theory driven approach may be advantageous as it would reduce the risk of capitalizing on chance. However, if a construct is not well developed then a data driven approach with cross validation may be most useful toward contributing to the literature.

Finally, using invariance testing to evaluate anchoring vignettes involves the same delimitations as other structural equation modeling techniques. A main consideration is that invariance testing is not a confirmatory technique in that hypotheses are supported, but never proven. As such, even if non-uniform DIF is found, it may be reasonable to expect that it was caused by DIRC, but the result does not prove that anchoring vignettes controlled for DIRC. Another general issue with using CFA models is that the models must be over identified in order to estimate them. The implication of this requirement is that a scale must have at least three items in order to model the construct using a CFA framework. As such, it would be inappropriate to use this method of invariance testing to examine anchoring vignettes that were designed for a scale with less than three items.

Future directions

The intention behind proposing the use of invariance testing to evaluate anchoring vignettes was to encourage researchers and practitioners to use this method within their contexts of interest. By no means should invariance testing with anchoring vignettes be limited to studying only effort in higher education. This method could be used in a variety of contexts with instruments that measure a latent trait. For instance, current research on anchoring vignettes could benefit from not only establishing that assumptions had been met, but by also providing some direct evidence that DIRC may have been removed from the sample as a result of adjusting scores.

Future research is also not limited to only testing the invariance of CFA models. Invariance across groups could be assessed across full structural models. The procedures for testing invariance for a full structural model involves first examining the measurement model (e.g. configural, metric, scalar) followed by constraining structural

parameters to determine if they are equivalent between groups (i.e. Beta, Gamma, and Exogenous Factor correlations; Byrne, 1998). This method may be useful for a more in-depth evaluation of how anchoring vignettes affect how the scale relates to other variables of interest.

Tables

Table 1

Demographic Information for Participants

	Anchoring Vignettes Group *n* (%)	*No Anchoring Vignettes Group* *n* (%)	*Total* *n* (%)
Gender			
Female	187 (62.1%)	222 (55.2%)	409 (58.2%)
Male	114 (37.9%)	180 (44.8%)	294 (41.8%)
Ethnicity			
American Indian	5 (1.7%)	4 (1.0%)	9 (1.3%)
Asian	27 (9.0%)	23 (5.7%)	50 (7.1%)
Black	24 (8.0%)	12 (3.0%)	36 (5.1%)
Hispanic	13 (4.3%)	31 (7.7%)	44 (6.3%)
Pacific Islander	4 (1.3%)	1 (0.2%)	5 (0.7%)
White	240 (79.7%)	348 (86.6)	588 (83.6%)
Not Specified	13 (4.3%)	17 (4.2%)	30 (4.3%)
Total *n*	301	402	703
Age: Mean (*SD*)	18.46 (0.50)	18.44 (0.34)	18.45 (0.44)

Note. Control group $n = 402$ and vignette group $n = 274$.

Table 2

Descriptive Statistics for Student Opinion Scale Across all Groups Data

	Control Group				Vignette Group before Adjustment				Vignette Group after Adjustment			
Item	*M*	*SD*	Skew	Kurt	*M*	*SD*	Skew	Kurt	*M*	*SD*	Skew	Kurt
1	4.25	0.72	-1.37	4.14	4.31	0.69	-1.24	3.45	3.48	0.76	-0.99	1.21
2	4.28	0.81	-1.32	2.28	4.33	0.79	-1.45	2.93	3.53	0.78	-0.73	0.49
3	3.50	1.12	-0.49	-0.50	3.43	1.11	-0.47	-0.46	2.80	0.98	-0.26	-0.62
4	4.03	0.92	-0.97	0.80	3.96	0.96	-1.17	1.42	3.22	0.89	-0.70	0.41
5	4.30	0.84	-1.28	1.78	4.28	0.83	-1.43	2.73	3.50	0.78	-0.77	0.81

Note. Control group $n = 402$ and vignette group $n = 274$. *M* = item mean, *SD* = item standard deviation, Skew = item skew, and Kurt = item kurtosis.

Table 3

Correlation Matrices and Descriptive Statistics for Student Opinion Scale before Adjustment of Vignette Group

						Vignette Group			
Items	1	2	3	4	5	*M*	*SD*	Skew	Kurt
1	-	0.509	0.364	0.329	0.400	4.31	0.69	-1.24	3.45
2	0.580	-	0.534	0.575	0.607	4.33	0.79	-1.45	2.93
3	0.406	0.513	-	0.598	0.391	3.43	1.11	-0.47	-0.46
4	0.318	0.416	0.503	-	0.587	3.96	0.96	-1.17	1.42
5	0.348	0.353	0.329	0.417	-	4.28	0.83	-1.43	2.73
Control Group									
M	4.25	4.28	3.50	4.03	4.30				
SD	0.72	0.81	1.12	0.92	0.84				
Skew	-1.37	-1.32	-0.49	-0.97	-1.28				
Kurt	4.14	2.28	-0.50	0.80	1.78				

Note. Control group $n = 402$ and vignette group $n = 274$. *M* = item mean, *SD* = item standard deviation, Skew = item skew, and Kurt = item kurtosis.

Table 4

Fit Indices for the Unidimensional Model of Effort before Adjustment of Vignette Group

Group	χ^2	*p*	df	SRMR	$RMSEA_{SB}$	CI	TLI_{SB}	CFI_{SB}
Control Group (*n* = 402)	24.67	<0.001	5	0.052	0.10	0.064 - 0.14	0.90	0.95
Vignette Group (*n* = 274)	26.41	<0.001	5	0.048	0.12	0.076 - 0.17	0.94	0.97

Note. χ^2 = chi-square; df = degrees of freedom; *p* = probability value for $\chi 2$ test; SRMR= standardized root mean square residual; $RMSEA_{SB}$= root mean square error of approximation; CI= 90% confidence interval for $RMSEA_{SB}$; TLI_{SB}= Tucker-Lewis Index; CFI_{SB}=comparative fit index.

Satorra-Bentler correction to χ^2, fit indexes, and standard errors was used for all analyses.

The degrees of freedom for both groups was obtained by subtracting 15 observations from 10 estimated parameters (5 errors, 4 path coefficients, 1 factor variance).

Table 5

Test of Invariance across SOS Groups before Adjustment of Vignette Group

	χ^2	df	*p*	$\Delta\chi^2$	Δdf	*p*	SRMR	$RMSEA_{SB}$	CI	TLI_{SB}	CFI_{SB}	ΔCFI_{SB}
Configural	50.93	10	<.001				0.052	0.11	0.08 - 0.14	0.94	0.97	
Metric	55.69	14	<.001	7.660	4	.105	0.069	0.09	0.07 - 0.12	0.96	0.97	0.00
Scalar	66.53	18	<.001	7.047	4	.133	0.069	0.09	0.07 - 0.11	0.96	0.96	0.01
Error Variance	72.11	23	<.001	7.422	5	.191	0.065	0.08	0.06 - 0.10	0.97	0.97	0.01

Note. $\Delta\chi^2$ = chi-square difference; Δdf = degrees of freedom difference; Δp-value = probability value for the $\Delta\chi2$ test; SRMR = standardized root mean square residual; $RMSEA_{SB}$ = root mean square error of approximation; CI = 90% confidence interval for $RMSEA_{SB}$; TLI_{SB} = Tucker-Lewis Index; CFI_{SB} = comparative fit index, ΔCFA = change in comparative fit index. Satorra-Bentler correction to χ^2, fit indexes, and standard errors was used for all analyses.

The degrees of freedom for the configural model was obtained by subtracting 30 observations from 20 estimated parameters (10 errors, 8 path coefficients, 2 factor variances). The degrees of freedom for the metric model was obtained by subtracting 30 observations from 16 estimated parameters (10 errors, 4 path coefficients, 2 factor variances). The degrees of freedom for the scalar model was obtained by subtracting 40 observations (30 variances and covariances, 10 means) from 22 estimated parameters (10 errors, 5 item intercepts, 4 path coefficients, 2 factor variances, 1 latent mean difference). The degrees of freedom for the errors constrained model was obtained by subtracting 40 observations (30 variances and covariances, 10 means) from 17 estimated parameters (5 errors, 5 item intercepts, 4 path coefficients, 2 factor variances, 1 latent mean difference

Table 6

Correlation Residuals for Models Run Separately for Each Group before Adjustment of Vignette Group

Items	1	2	3	4	5
1	-	0.068	-0.049	-0.091	0.000
2	-0.073	-	-0.011	-0.040	-0.044
3	0.000	0.000	-	0.097	-0.021
4	0.090	0.0390	-0.093	-	0.104
5	0.000	-0.015	0.098	-0.037	-

Note. The upper half of the matrix is the correlation residuals for the control group (n = 402) and the lower half of the matrix is correlation residuals for the vignette group (n = 274).

Table 7

Correlation Residuals for Metric Model before Adjustment of Vignette Group

Items	1	2	3	4	5
1	-	0.082	-0.033	-0.119	-0.024
2	0.019	-	0.034	-0.061	-0.059
3	0.078	-0.063	-	0.097	-0.026
4	0.090	-0.060	0.006	-	0.068
5	0.030	-0.095	-0.045	-0.029	-

Note. The upper half of the matrix is the correlation residuals for the control group (n = 402) and the lower half of the matrix is correlation residuals for the vignette group (n = 274).

Table 8

Observed and Expected Means with Residuals across both Groups before Adjustment of Vignette Group

	Control Group					Vignette Group				
	1	2	3	4	5	1	2	3	4	5
Observed mean	4.25	4.28	3.50	4.03	4.30	4.31	4.33	3.43	3.96	4.28
Expected mean	4.27	4.30	3.47	4.00	4.29	4.27	4.31	3.48	4.00	4.29
Residual	-0.02	-0.02	0.03	0.03	0.01	0.04	0.02	-0.05	-0.04	-0.01

Note. Control group n = 402 and vignette group n = 274. M = item mean, SD = item standard deviation, Skew = item skew, and Kurt = item kurtosis.

Table 9

Correlation Residuals for Errors Constrained Model before Adjustment of Vignette Group

Items	1	2	3	4	5
1	-	0.117	-0.002	-0.079	-0.012
2	0.018	-	0.002	-0.100	-0.105
3	-0.063	-0.015	-	0.056	-0.076
4	-0.092	0.037	0.128	-	0.025
5	0.021	0.128	-0.038	0.164	-

Note. The upper half of the matrix is the correlation residuals for the control group (n = 402) and the lower half of the matrix is correlation residuals for the vignette group (n = 274).

Table 10

Latent Mean Difference, Significance Test, and Effect Size before Adjustment of Vignette Group

	LMDE	OMDE
Estimated latent mean difference	0.00	
Effect size associated with latent mean difference	0.00	
Observed mean difference		0.04
Mean observed source: Control group		4.07
Mean observed source: Vignette group		4.06
Effect size associated with observed mean difference		0.01

Note. Control group, n = 402. Vignette group, n = 274. LMDE= Latent mean difference estimation. OMDE= Observed mean difference estimation.

Table 11

Correlation Matrices and Descriptive Statistics for Student Opinion Scale Groups Data after Adjustment of Vignette Group

						Vignette Group			
	1	2	3	4	5	*M*	*SD*	Skew	Kurt
1	-	0.633	0.514	0.507	0.458	3.48	0.76	-0.99	1.21
2	0.580	-	0.567	0.566	0.562	3.53	0.78	-0.73	0.49
3	0.406	0.513	-	0.632	0.431	2.80	0.98	-0.26	-0.62
4	0.318	0.416	0.503	-	0.591	3.22	0.89	-0.70	0.41
5	0.348	0.353	0.329	0.417	-	3.50	0.78	-0.77	0.81
Control Group									
M	4.25	4.28	3.50	4.03	4.30				
SD	0.72	0.81	1.12	0.92	0.84				
Skew	-1.37	-1.32	-0.49	-0.97	-1.28				
Kurt	4.14	2.28	-0.50	0.80	1.78				

Note. Control group $n = 402$ and vignette group $n = 274$. *M* = item mean, *SD* = item standard deviation, Skew = item skew, and Kurt = item kurtosis.

Table 12

Correlation Matrices and Descriptive Statistics for Student Opinion Scale before and after Adjustment

						Vignette Group before Adjustment			
Items	1	2	3	4	5	*M*	*SD*	Skew	Kurt
1	-	0.509	0.364	0.329	0.400	4.31	0.69	-1.24	3.45
2	0.633	-	0.534	0.575	0.607	4.33	0.79	-1.45	2.93
3	0.514	0.567	-	0.598	0.391	3.43	1.11	-0.47	-0.46
4	0.507	0.566	0.632	-	0.587	3.96	0.96	-1.17	1.42
5	0.458	0.562	0.431	0.591	-	4.28	0.83	-1.43	2.73
Vignette Group after Adjustment									
M	3.48	3.53	2.80	3.22	3.50				
SD	0.76	0.78	0.98	0.89	0.78				
Skew	-0.99	-0.73	-0.26	-0.70	-0.77				
Kurt	1.21	0.49	-0.62	0.41	0.81				

Note. Vignette group $n = 274$. *M* = item mean, *SD* = item standard deviation, Skew = item skew, and Kurt = item kurtosis.

Table 13

Frequency of Anchoring Vignette Endorsement by Response Category

Response Option	Low Effort Anchoring Vignette *n* (%)	High Effort Anchoring Vignette *n* (%)
1 (Disagree Strongly)	26 (9.5%)	0 (0.0%)
2 (Disagree)	159 (58.0%)	0 (0.0%)
3 (Neutral)	65 (23.7%)	1 (0.4%)
4 (Agree)	24 (8.8%)	55 (20.1%)
5 (Strongly agree)	0 (0.0%)	218 (79.6%)

Note. Vignette group $n = 274$.

Table 14

Fit Indices for the Unidimensional Model of Effort after Adjustment of Vignette Group

Group	χ^2	df	*p*	SRMR	$RMSEA_{SB}$	CI	TLI_{SB}	CFI_{SB}
Control Group (*n* = 402)	24.67	5	<.001	0.052	0.10	0.06 - 0.14	0.90	0.95
Vignette Group (*n* = 274)	21.58	5	<.001	0.037	0.11	0.06 - 0.16	0.96	0.98

Note. χ^2 = chi-square; df = degrees of freedom; *p* = probability value for $\chi 2$ test; SRMR= standardized root mean square residual; $RMSEA_{SB}$= root mean square error of approximation; CI= 90% confidence interval for $RMSEA_{SB}$; TLI_{SB}= Tucker-Lewis Index; CFI_{SB}=comparative fit index.

Satorra-Bentler correction to χ^2, fit indexes, and standard errors was used for all analyses.

The degrees of freedom for both groups was obtained by subtracting 15 observations from 10 estimated parameters (5 errors, 4 path coefficients, 1 factor variance).

Table 15

Correlation Residuals for Student Opinion Scale Separate Groups Data after Adjustment of Vignette Group

	1	2	3	4	5
1	-	0.068	-0.049	-0.091	0.000
2	0.068	-	-0.011	-0.040	-0.044
3	0.000	-0.013	-	0.097	-0.021
4	-0.045	-0.058	0.069	-	0.104
5	-0.034	0.017	-0.065	0.058	-

Note. The upper half of the matrix is the correlation residuals for the control group (n = 402) and the lower half of the matrix is correlation residuals for the vignette group (n = 274).

Table 16

Test of Invariance across SOS Groups after Adjustment of Vignette Group

	χ^2	df	*p*	$\Delta\chi^2$	Δdf	*p*	SRMR	$RMSEA_{SB}$	CI	TLI_{SB}	CFI_{SB}	ΔCFI_{SB}
Configural	46.74	10	<.001				0.052	0.10	0.08 - 0.14	0.95	0.98	
Metric	47.02	14	<.001	4.547	4	0.337	0.062	0.09	0.06 - 0.11	0.97	0.98	0.00
Scalar	68.66	18	<.001	20.37	4	<.001	0.083	0.09	0.07 - 0.11	0.96	0.97	0.01
Error Variance	84.59	23	<.001	16.00	5	.007	0.072	0.09	0.07 - 0.11	0.97	0.96	0.01

Note. $\Delta\chi^2$ = chi-square difference; Δdf = degrees of freedom difference; Δp-value = probability value for the $\Delta\chi$2 test; SRMR = standardized root mean square residual; $RMSEA_{SB}$ = root mean square error of approximation; CI = 90% confidence interval for $RMSEA_{SB}$; TLI_{SB} = Tucker-Lewis Index; CFI_{SB} = comparative fit index, ΔCFA = change in comparative fit index. Satorra-Bentler correction to χ^2, fit indexes, and standard errors was used for all analyses.

The degrees of freedom for the configural model was obtained by subtracting 30 observations from 20 estimated parameters (10 errors, 8 path coefficients, 2 factor variances). The degrees of freedom for the metric model was obtained by subtracting 30 observations from 16 estimated parameters (10 errors, 4 path coefficients, 2 factor variances). The degrees of freedom for the scalar model was obtained by subtracting 40 observations (30 variances and covariances, 10 means) from 22 estimated parameters (10 errors, 5 item intercepts, 4 path coefficients, 2 factor variances, 1 latent mean difference). The degrees of freedom for the errors constrained model was obtained by subtracting 40 observations (30 variances and covariances, 10 means) from 17 estimated parameters (5 errors, 5 item intercepts, 4 path coefficients, 2 factor variances, 1 latent mean difference).

Table 17

Correlation Residuals for Metric Model after Adjustment of Vignette Group

	1	2	3	4	5
1	-	0.082	-0.033	-0.119	-0.024
2	0.056	-	0.034	-0.061	-0.059
3	-0.016	-0.037	-	0.097	-0.026
4	-0.018	-0.035	0.071	-	0.068
5	-0.012	0.033	-0.071	0.095	-

Note. The upper half of the matrix is the correlation residuals for the control group (n = 402) and the lower half of the matrix is correlation residuals for the vignette group (n = 274).

Table 18table

Observed and Expected Means with Residuals across both Groups after Adjustment of Vignette Group

	Control Group					Vignette Group before Adjustment					Vignette Group after Adjustment				
	1	2	3	4	5	1	2	3	4	5	1	2	3	4	5
Observed mean	4.25	4.28	3.50	4.03	4.30	4.31	4.33	3.43	3.96	4.28	3.48	3.53	2.80	3.22	3.50
Expected mean	4.22	4.30	3.59	4.03	4.25	4.27	4.31	3.48	4.00	4.29	3.52	3.51	2.73	3.22	3.55
Residual	0.03	-0.02	-0.09	0.00	0.05	0.04	0.02	-0.05	-0.04	-0.01	-0.04	0.02	0.07	0.00	-0.05

Note. Control group $n = 402$ and vignette group $n = 274$.

Table 19

Correlation Residuals for Errors Constrained Model after Adjustment of Vignette Group

	1	2	3	4	5
1	-	0.057	-0.059	-0.140	-0.041
2	0.081	-	-0.012	-0.092	-0.085
3	0.028	0.016	-	0.044	-0.065
4	0.036	0.039	0.157	-	0.034
5	0.046	0.092	0.013	0.191	-

Note. The upper half of the matrix is the correlation residuals for the control group (n = 402) and the lower half of the matrix is correlation residuals for the vignette group (n = 274).

Table 20

Latent Mean Difference, Significance Test, and Effect Size after Adjustment of Vignette Group

	LMDE	OMDE
Estimated latent mean difference	0.70*	
Effect size associated with latent mean difference	1.30	
Observed mean difference		0.81*
Mean observed source: Control group		4.07
Mean observed source: Vignette group		3.26
Effect size associated with observed mean difference		1.17

Note. Control group, n = 402. Vignette group, n = 274. LMDE= Latent mean difference estimation. OMDE= Observed mean difference estimation.

* p <.001

Figures

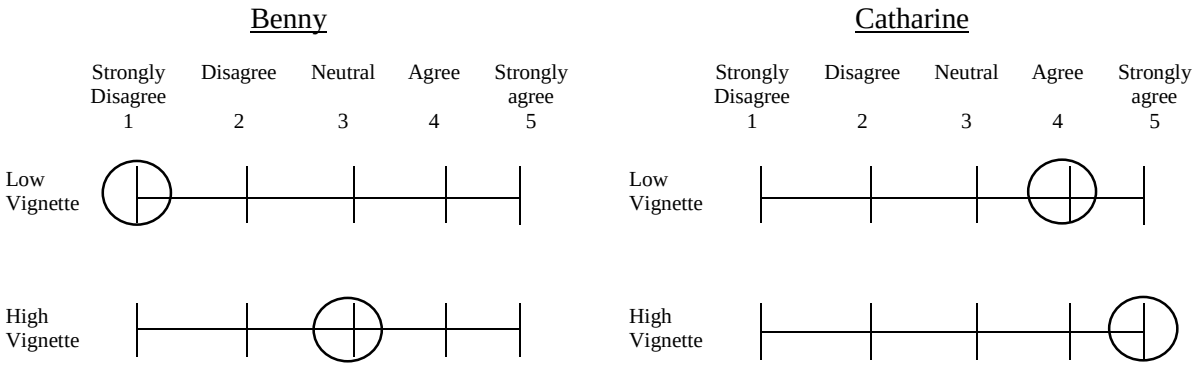

Figure 1. Illustration of how ratings can indicate DIRC. DIRC indicated by differential ratings of the anchoring vignettes by participants.

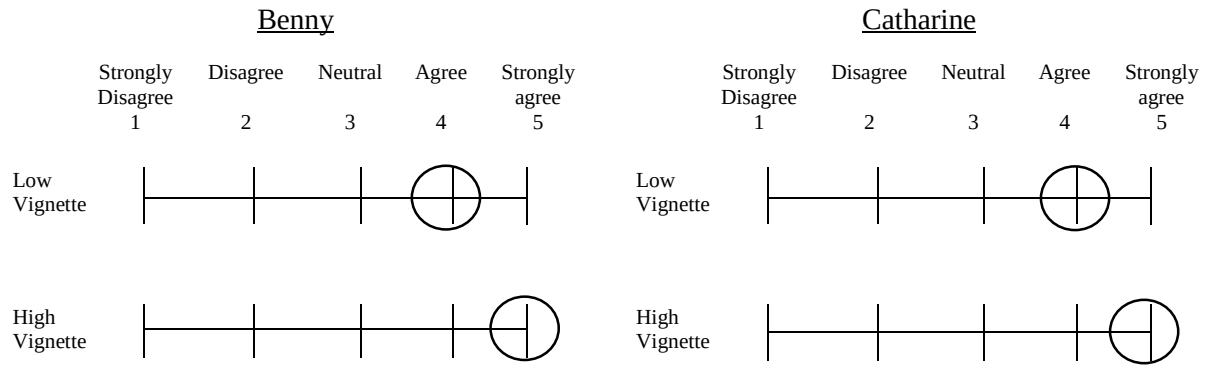

Figure 2. Illustration of how ratings can indicate absence of DIRC. Absence of DIRC is indicated by consistent ratings of the anchoring vignettes by participants.

Benny's Scalar Solution

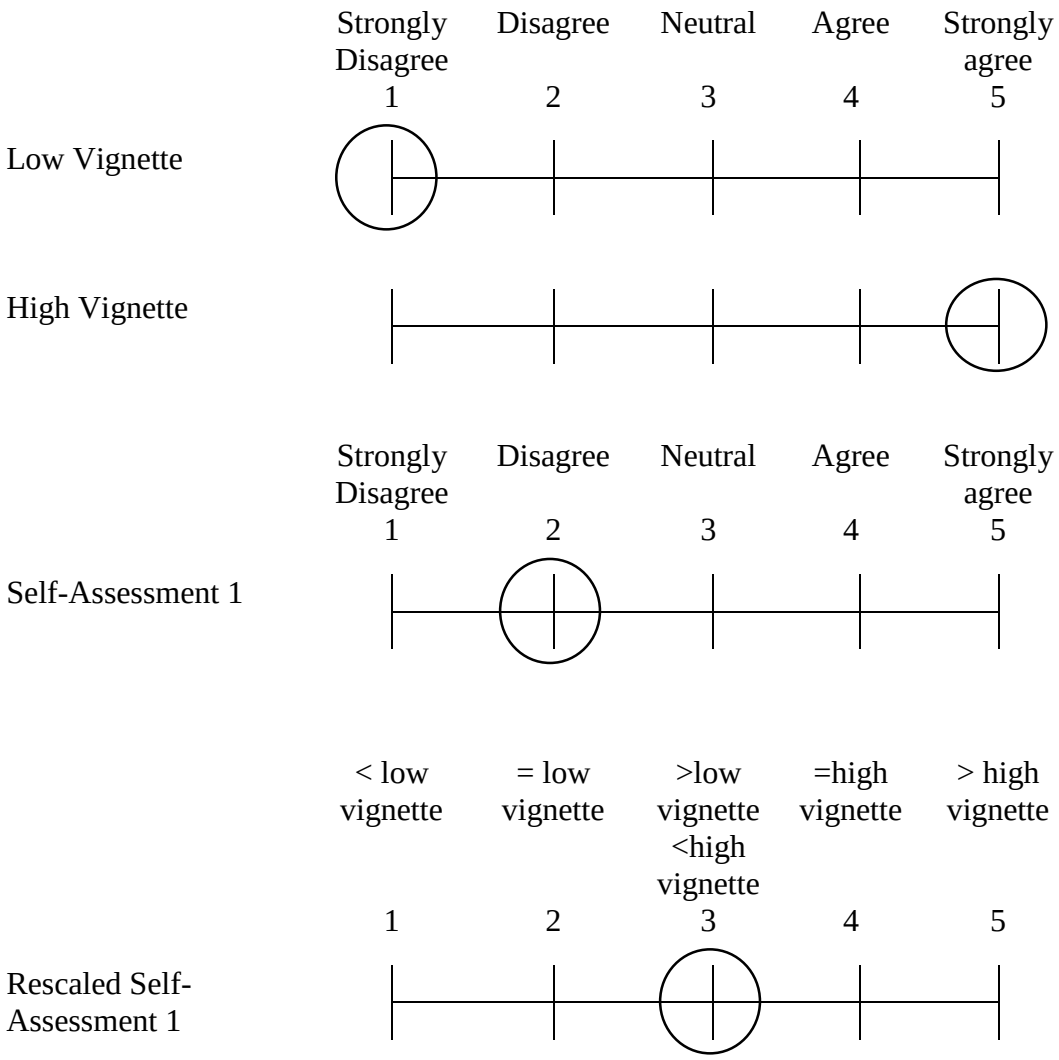

Figure 3. Scalar example of non-parametric scoring for anchoring vignettes. Adapted from "PISA 2012 Technical Report" by Programme for International Student Assessment, 2014, retrieved from Organization for Economic Co-operation and Development website: http://www.oecd.org/pisa/pisaproducts/PISA-2012-technical-report-final.pdf.

Catharine's Interval Solution

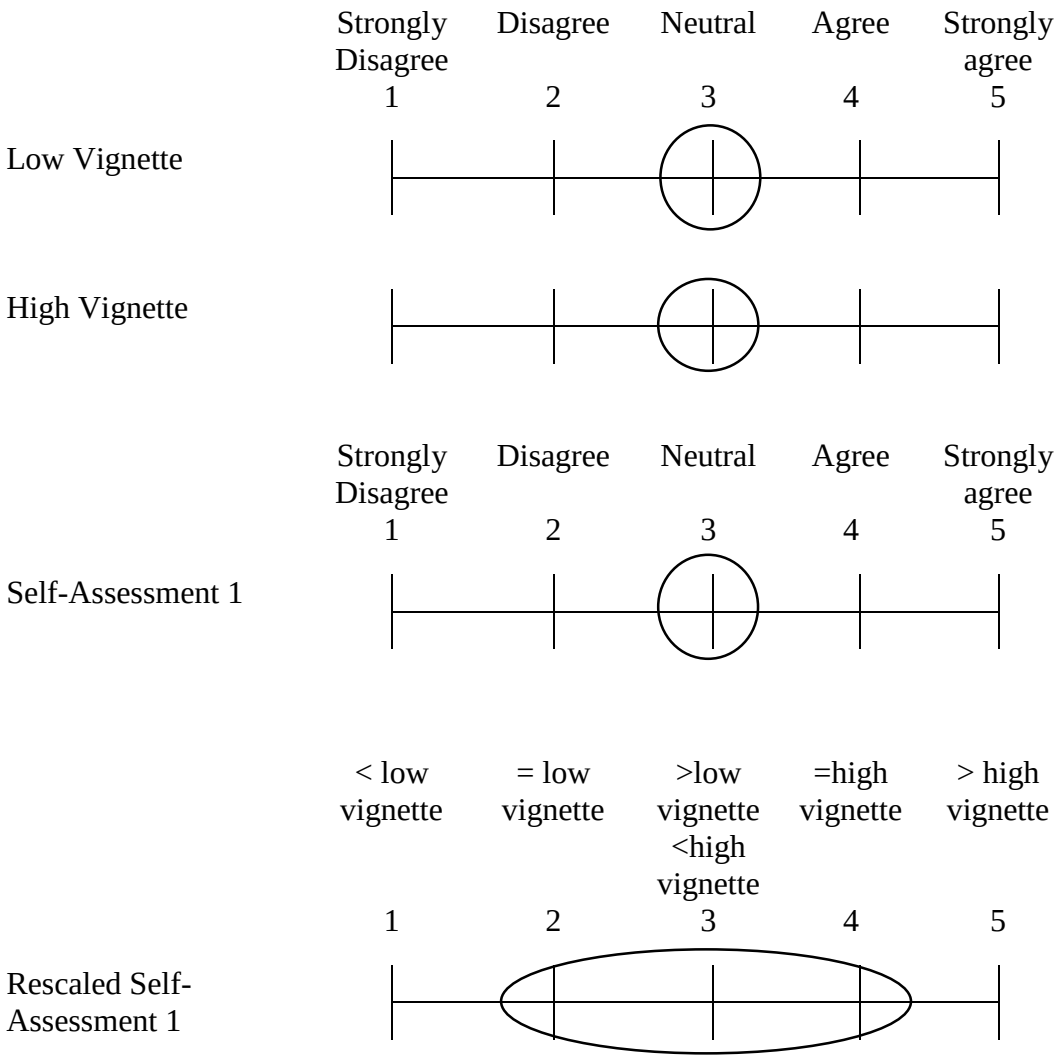

Figure 4. Interval example of non-parametric scoring for anchoring vignettes. Adapted from "PISA 2012 Technical Report" by Programme for International Student Assessment, 2014, retrieved from Organization for Economic Co-operation and Development website: http://www.oecd.org/pisa/pisaproducts/PISA-2012-technical-report-final.pdf.

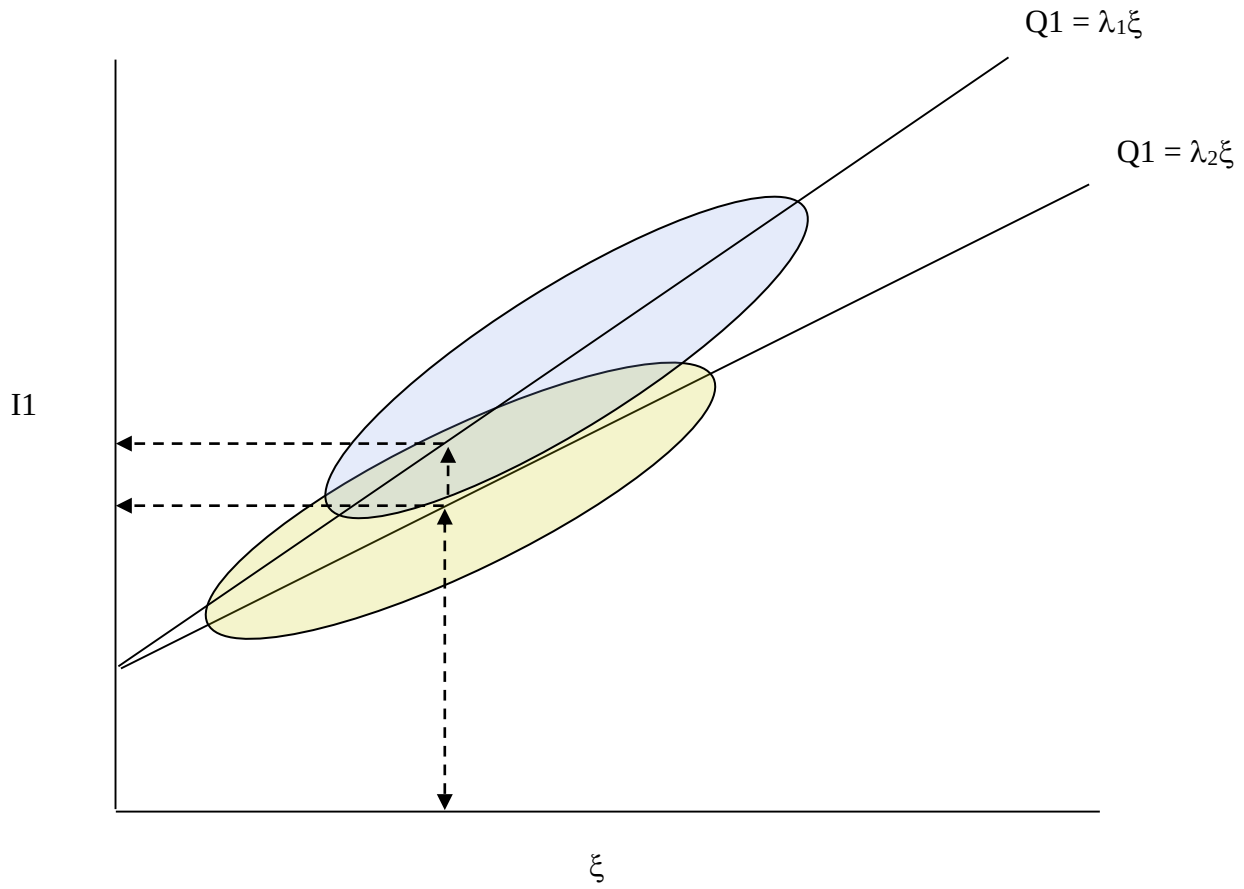

Figure 5. Non-uniform DIF or metric non-invariance at the item level. Two cases who possess an equal amount of the trait will have different observed scores. When comparing a control group to a vignette group, this is an indication that DIRC was present in the control group, but not present in the vignette group. This can occur if adjusting scores changed the rank order of cases within the vignette group.

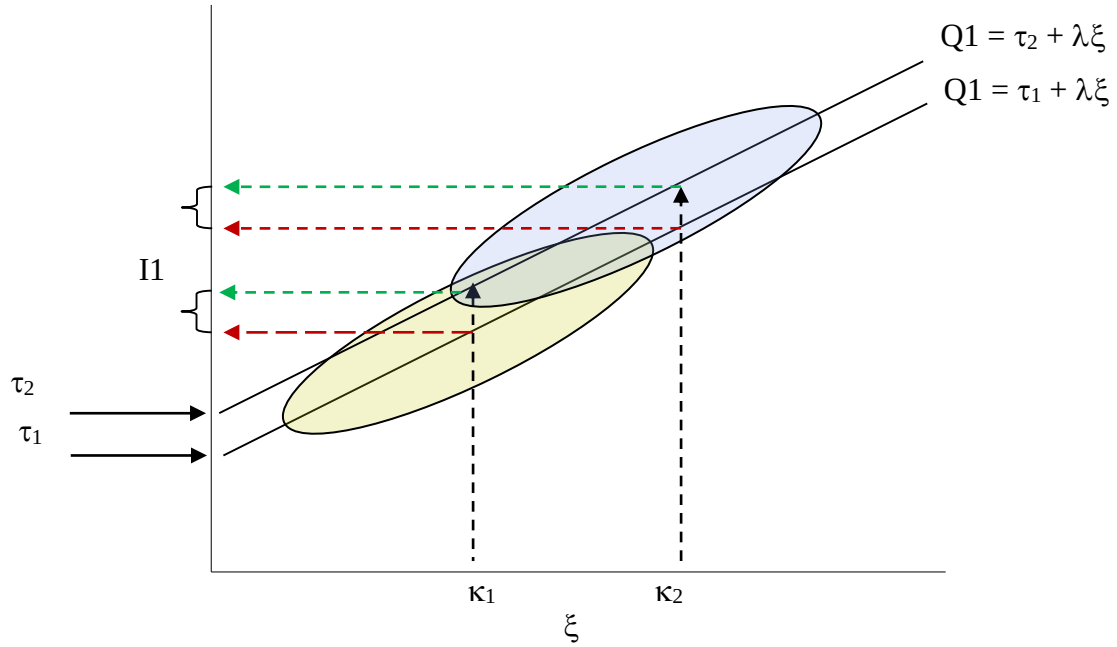

Figure 6. Uniform DIF or scalar non-invariance at the item level. Observed differences on the item reflect different amounts of the trait and systematic bias. When comparing a control group to a vignette group, this is an indication that DIRC was not present in the control group or the vignette group, but that the vignette group's adjusted scores were systematically different from the control group's scores. This can occur if adjusting scores did not change the rank order of cases within the vignette group.

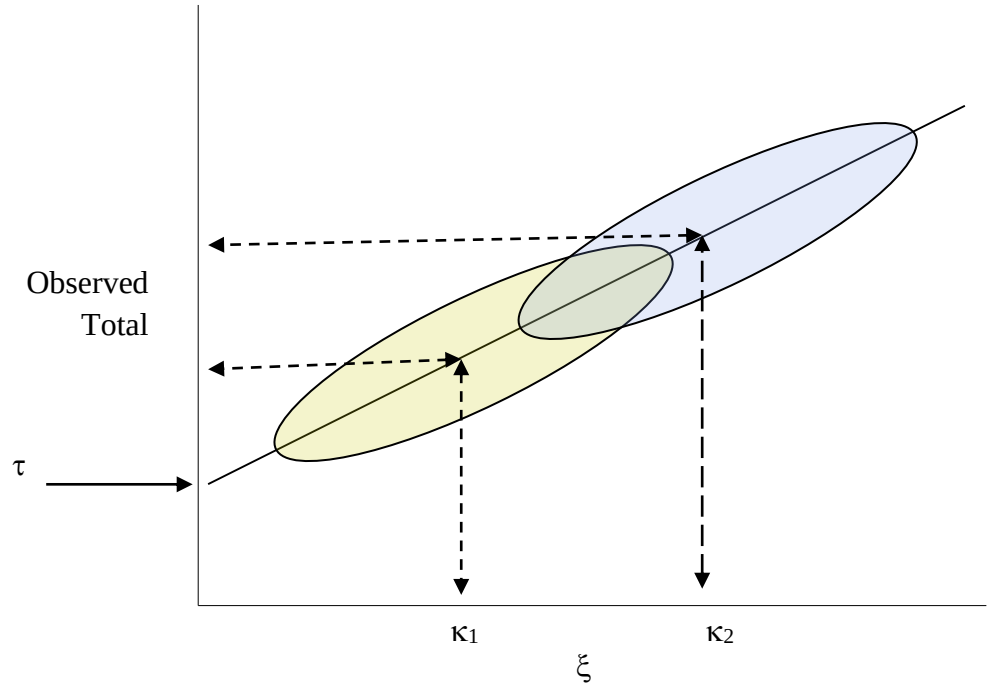

Figure 7. No DIF or complete invariance (configural, metric, and at least partial scalar invariance) at the scale level, with latent mean difference. Assuming that the control and vignette groups are actually equivalent, observed differences on the scale level reflect differences in item intercepts for all items across groups.

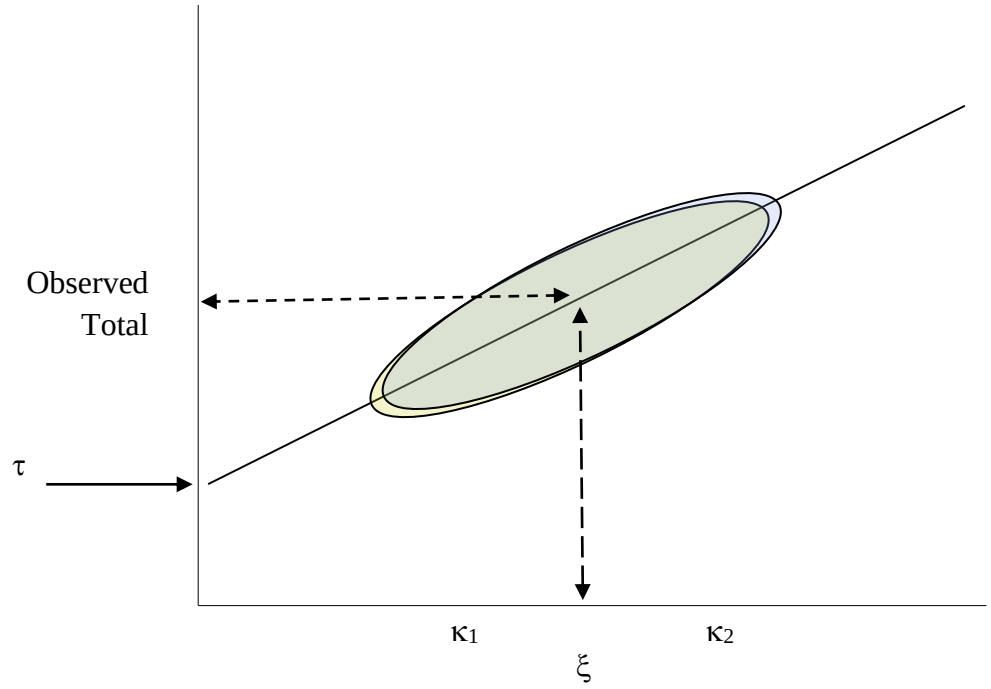

Figure 8. No DIF or complete invariance (configural, metric, and at least partial scalar invariance) at the scale level, with equivalent latent mean. Assuming that the control and vignette groups are actually equivalent, using anchoring vignettes did not make the factor structure for the vignette group any different from the control group. In this case, the anchoring vignettes had no impact on the factor structure or score values. This could occur if participants rated anchoring vignettes in the same way and in the way that the researcher operationalized them.

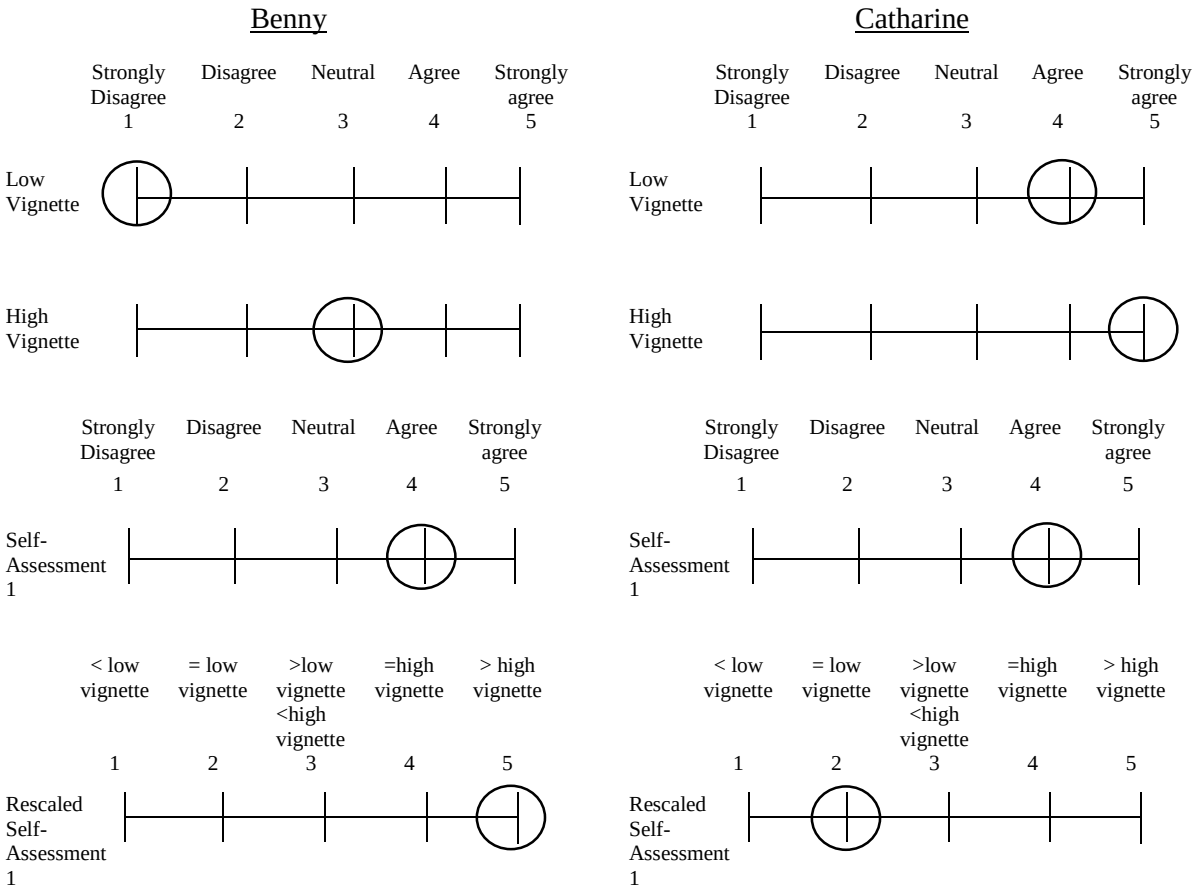

Figure 9. Illustration of how ratings could result in non-uniform DIF. Case rank order changes because the response scale for rating anchoring vignettes is used differently.

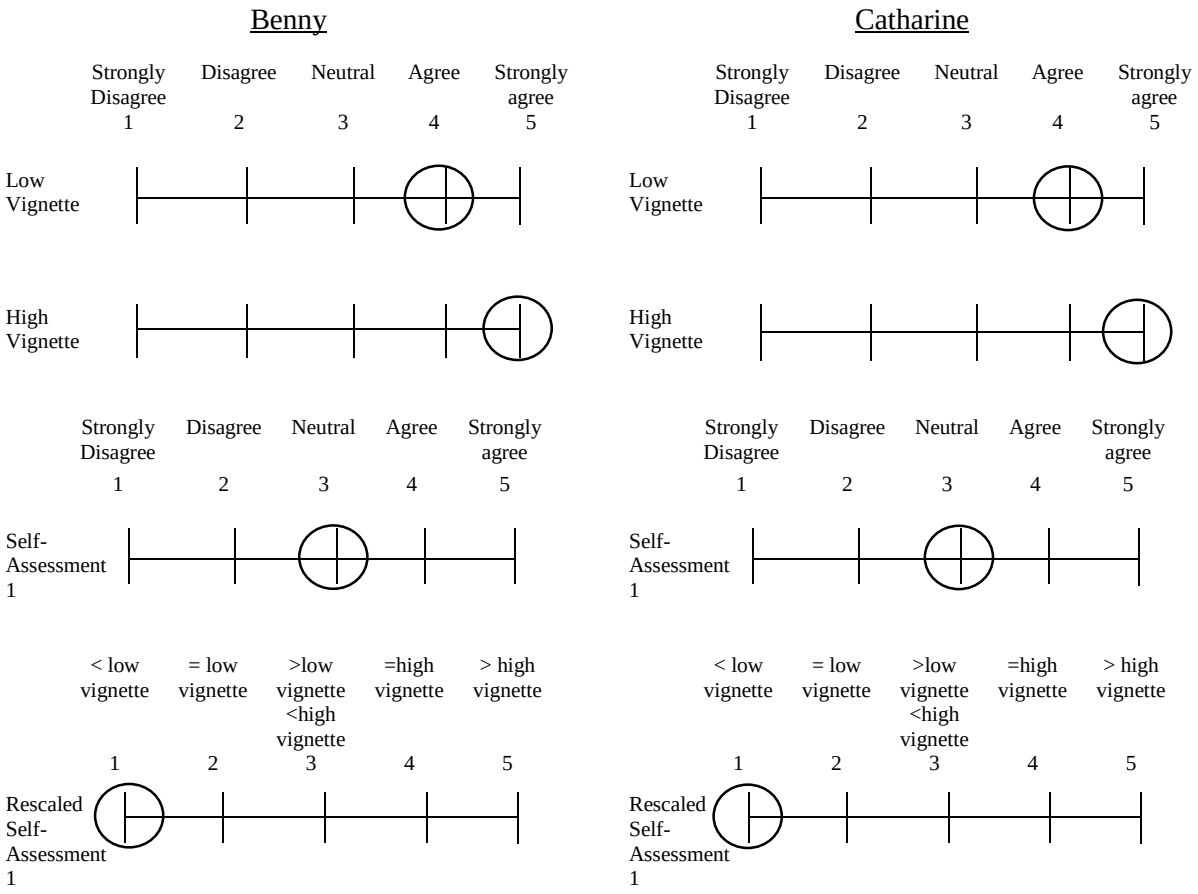

Figure 10. Illustration of how ratings could result in uniform DIF. Scores are adjusted downward while the rank order of each case remains equal.

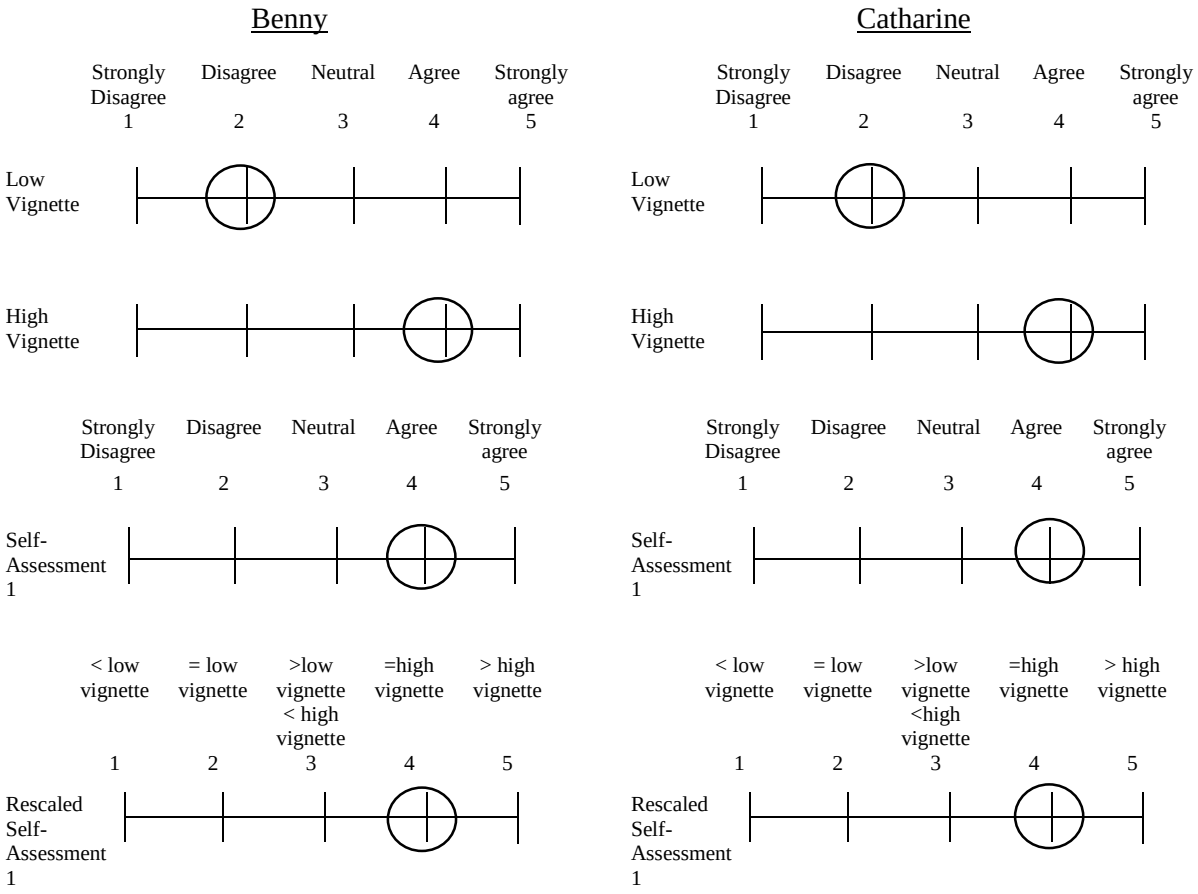

Figure 11. Demonstration of no DIF (complete invariance). In this scenario, the anchoring vignettes had no impact on factor structure or inferences made from scores. Anchoring vignettes are not needed since each participant is already using the scale in the same way (i.e., vignettes are rated the same between each participant) and these ratings are consistent with how each vignette was operationalized by the researcher.

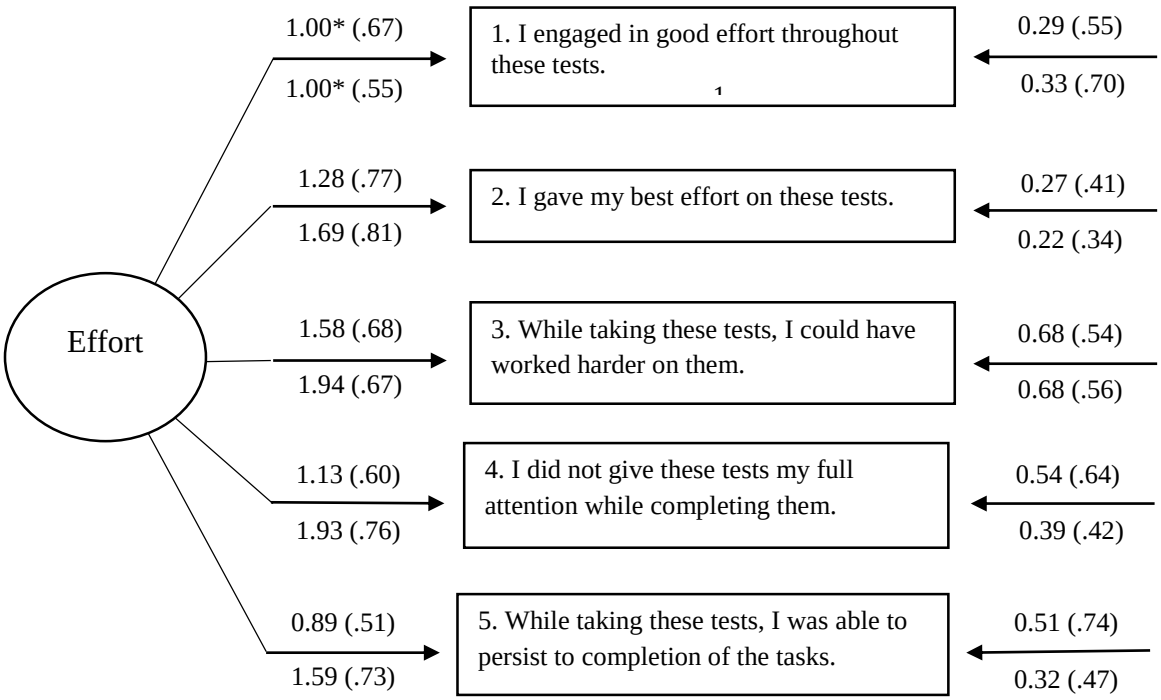

Figure 12. Factor pattern coefficients and error variances from the unidimensional model of effort estimated separately for two groups (before adjustment of vignette group).

Values above arrows pertain to the control group (n = 402) and values below the arrows pertain to the vignette group (n = 274). Values in the parentheses are standardized. The parameters marked with an asterisk were fixed to 1.00. All unstandardized path coefficients were statistically significant at the .05 level.

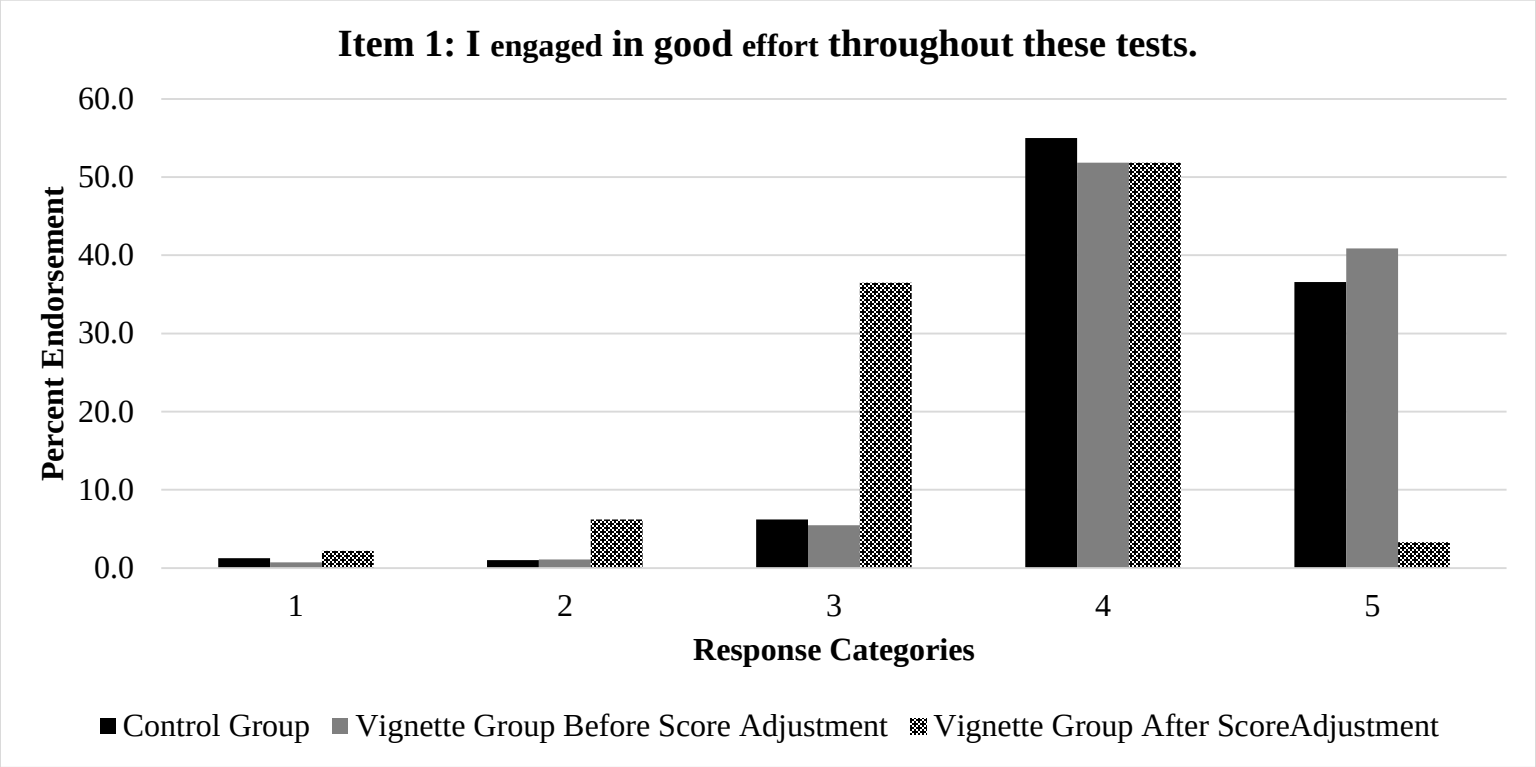

Figure 13. Percent of response category endorsement across groups for Item 1.

Note. Control group, *n* = 402. Vignette group, *n* = 274.

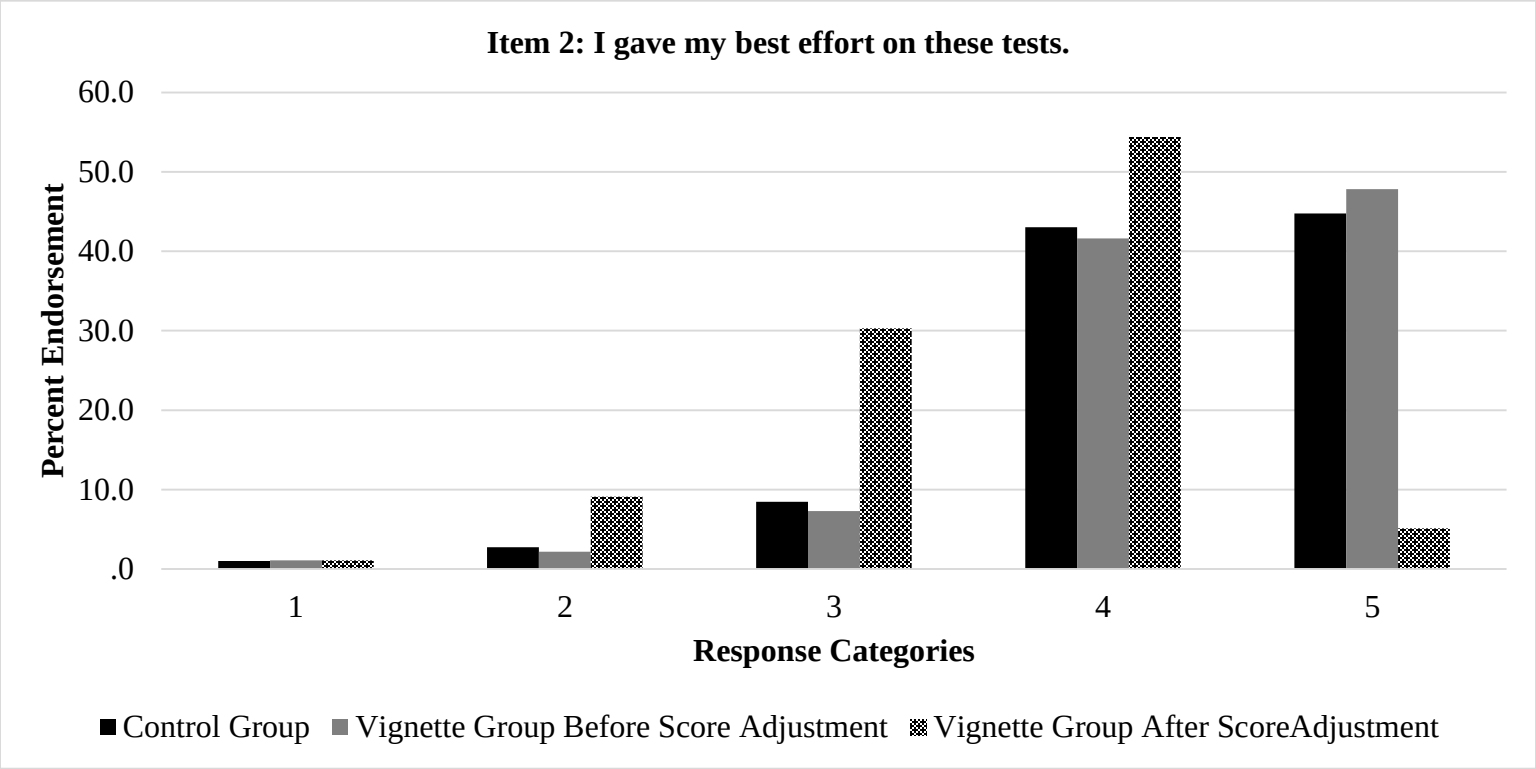

Figure 14. Percent of response category endorsement across groups for item 2.

Note. Control group, n = 402. Vignette group, n = 274.

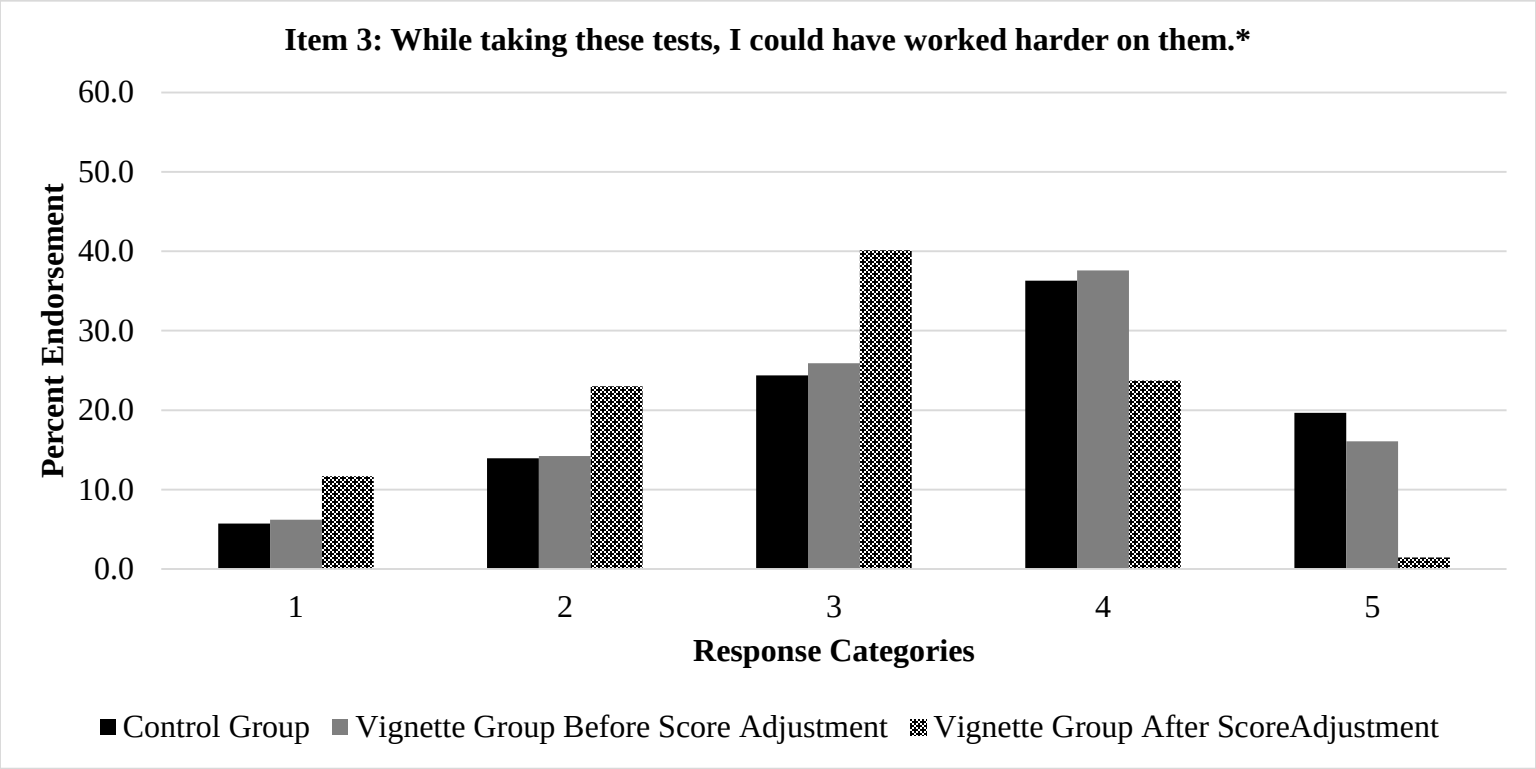

Figure 15. Percent of response category endorsement across groups for item 3.

Note. Control group, *n* = 402. Vignette group, *n* = 274. *Reverse coded.

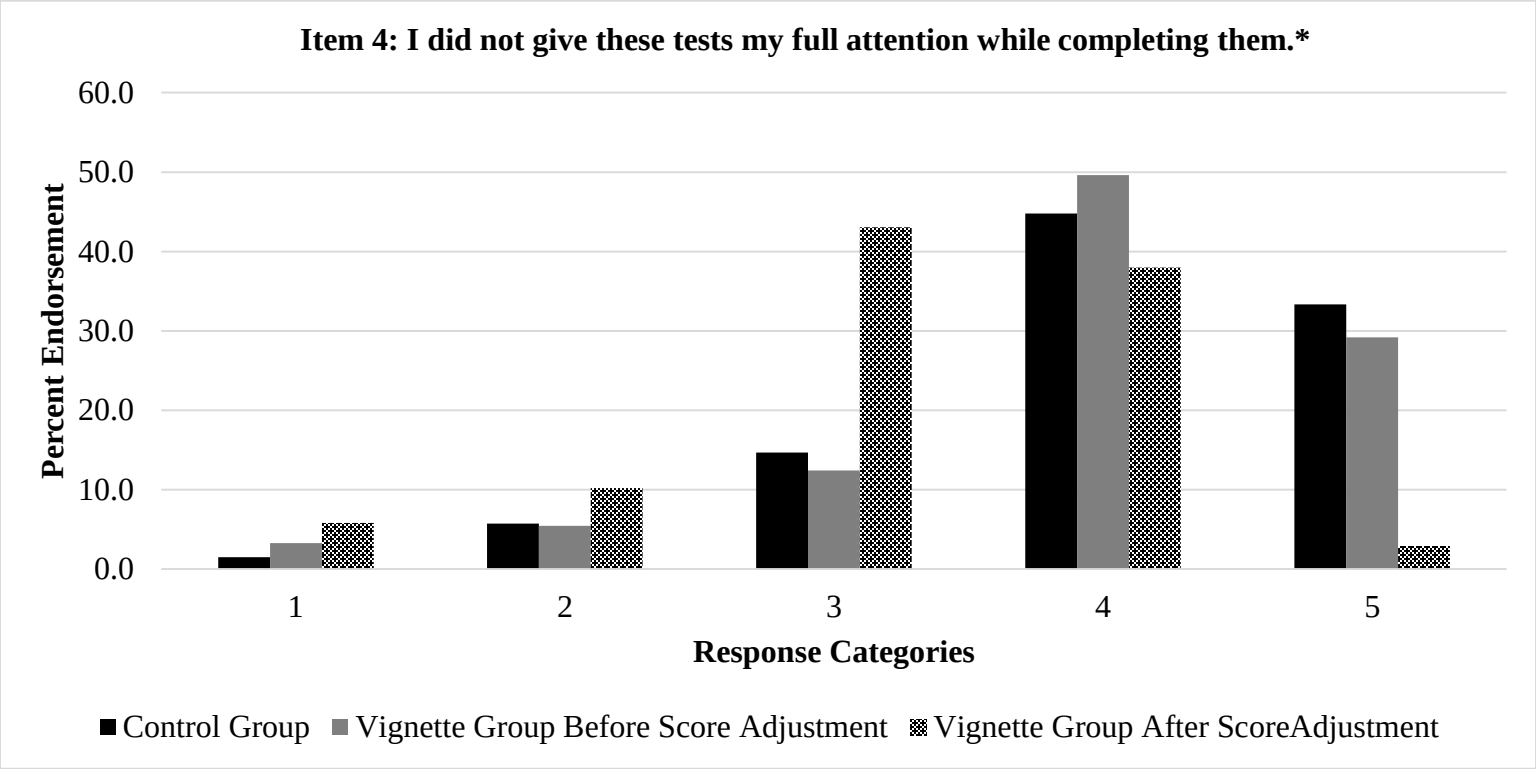

Figure 16. Percent of response category endorsement across groups for item 4.

Note. Control group, *n* = 402. Vignette group, *n* = 274. *Reverse coded.

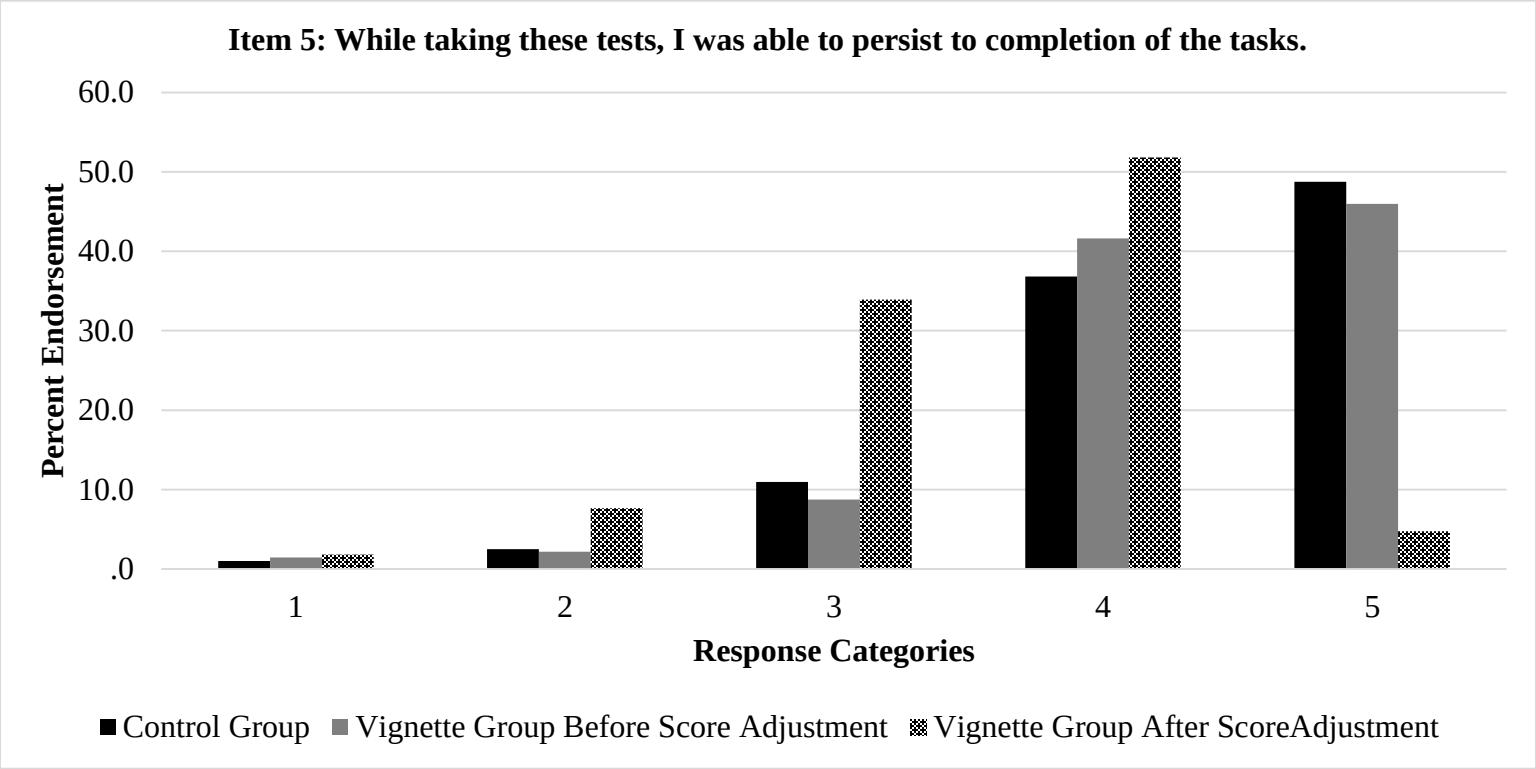

Figure 17. Percent of response category endorsement across groups for item 5.

Note. Control group, *n* = 402. Vignette group, *n* = 274

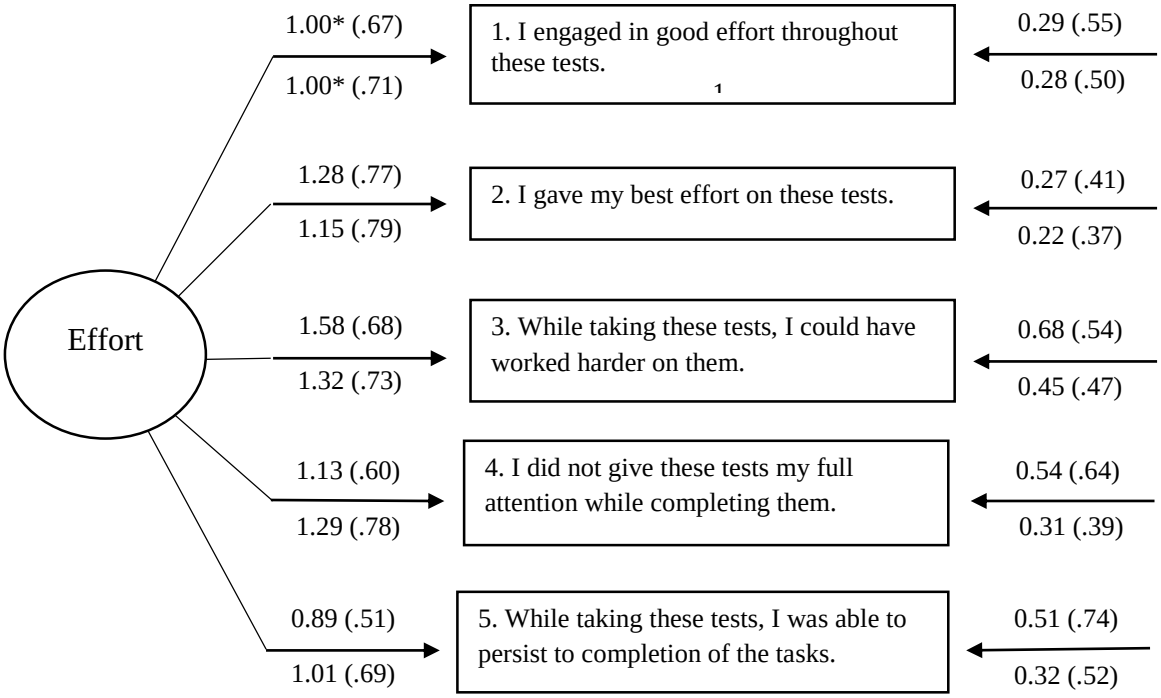

Figure 18. Factor pattern coefficients and error variances from the unidimensional model of effort estimated separately for two groups (after adjustment of vignette group).

Values above arrows pertain to the control group (n = 402) and values below the arrows pertain to the vignette group (n = 274). Values in the parentheses are standardized. The parameters marked with an asterisk were fixed to 1.00. All unstandardized path coefficients were statistically significant at the .05 level.

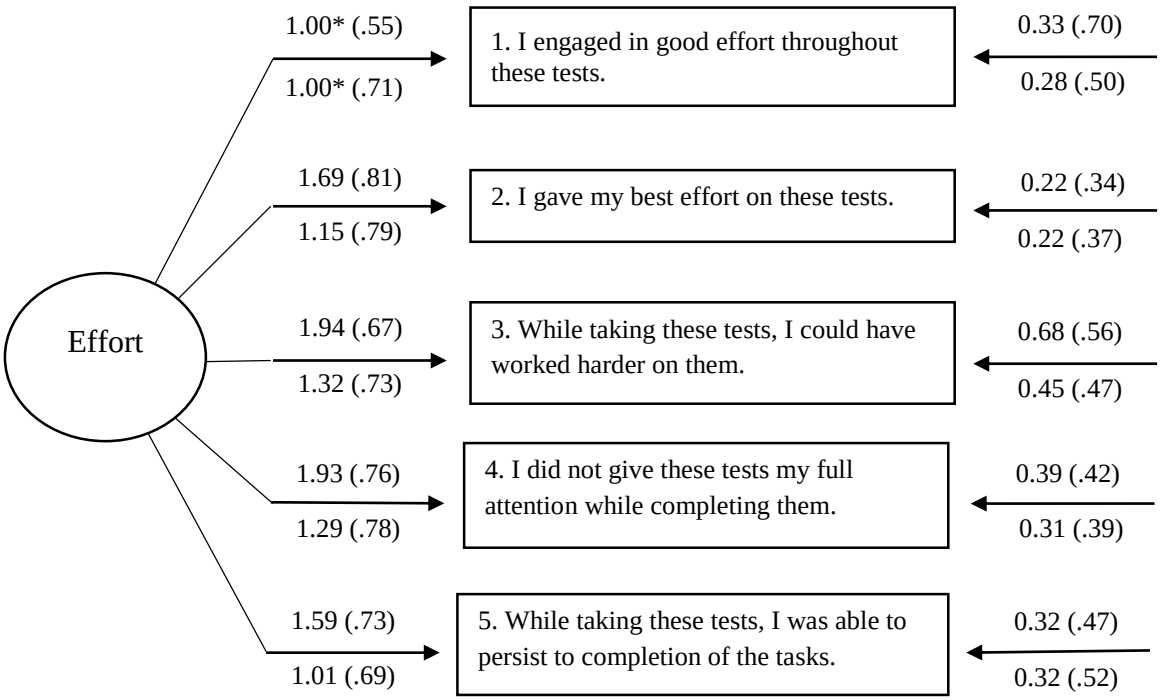

Figure 19. Factor pattern coefficients and error variances from the unidimensional model of effort estimated separately for vignette group (before and after adjustment).

Values above arrows pertain to the vignette group before adjustment (n = 274) and values below the arrows pertain to the vignette group (n = 274). Values in the parentheses are standardized. The parameters marked with an asterisk were fixed to 1.00. All unstandardized path coefficients were statistically significant at the .05 level.

Appendices

Appendix A
Effort Anchoring Vignettes

Below you will read a description of three students. Read each of the descriptions of these students. Then let us know to what extent to you agree with the final statement by using the scale below:

1	2	3	4	5
Disagree Strongly	**Disagree**	**Neutral**	**Agree**	**Strongly agree**

Dena read each item carefully before providing an answer. She thought about each question, and the most appropriate response, before providing an answer. **Dena tried her best on these assessments.**

Jessica read each item carefully before providing an answer. If a question was mentally taxing or looked like it may take too much time, she gave a response without giving it much thought. **Jessica tried her best on these assessments.**

Appendix B
Student Opinion Scale

Please think about the test that you **just completed**. Mark the answer that best represents how you feel about each of the statements below.

1	2	3	4	5
Disagree Strongly	**Disagree**	**Neutral**	**Agree**	**Strongly agree**

1. Doing well on these tests was important to me.[1]
2. I engaged in good effort throughout these tests.
3. I am not curious about how I did on these tests relative to others.
4. I am not concerned about the scores I receive on these tests.
5. These were important tests to me.
6. I gave my best effort on these tests.
7. While taking these tests, I could have worked harder on them.
8. I would like to know how well I did on these tests.
9. I did not give these tests my full attention while completing them.
10. While taking these tests, I was able to persist to completion of the tasks.

Note. Importance items: 1, 3, 4, 5, 8. Effort items: 2, 6, 7, 9, 10.

References

Abdelfattah, F. (2010). The relationship between motivation and achievement in low-stakes examinations. *Social Behavior and Personality*, *38*, 159-168.

Barry, C. L., & Finney, S. J. (2016). Modeling change in effort across a low-stakes testing session: A latent growth curve modeling approach. *Applied Measurement in Education*, *29*, 46-64.

Barry, C. L., Horst, S. J., Finney, S. J., Brown, A. R., & Kopp, J. P. (2010). Do examinees have similar test-taking effort? A high-stakes question for low-stakes testing. *International Journal of Testing*, *10*, 342–363.

Benson, J. (1998). Developing a strong program of construct validation: A test anxiety example. *Educational Measurement: Issues and Practice*, *17*, 10-17, 22.

Bentler, P. M. (1998, March 10). Kurtosis, residuals, fit indices. Message posted to SEMNET electronic mailing list, archived at http://bama.au.edu.archives/semnet.html.

Brown, T. A. (2015). *Confirmatory factor analysis for applied research*. Guilford Publications.

Bryant, F. B., & Satorra, A. (2012). Principles and practice of scaled difference chi-square testing. *Structural Equation Modeling: A Multidisciplinary Journal*, *19*, 372-398.

Buckley, J. (2008). *Survey context effects in anchoring vignettes*. Retrieved from The Society for Political Methodology website: http://polmeth. wustl. edu/media/Paper/surveyartifacts. pdf.

Byrne, B. M. (1998). Structural equation modeling with LISREL, PRELIS and SIMPLIS: Basic concepts, applications, and programming. Mahwah, NJ: Lawrence Erlbaum.

Cheung, G. W., & Rensvold, R. B. (1999). Testing factorial invariance across groups: A reconceptualization and proposed new method. *Journal of Management, 25*, 1-27.

Cheung, G. W., & Rensvold, R. B. (2002). Evaluating goodness-of-fit indexes for testing measurement invariance. *Structural equation modeling, 9*, 233-255.

Cole, J. S., & Osterlind, S. J. (2008). Investigating differences between low-and high-stakes test performance on a general education exam. *The Journal of General Education, 57*, 119-130.

d'Uva, T. B., Lindeboom, M., O'Donnell, O., & Van Doorslaer, E. (2011). Slipping anchor? Testing the vignettes approach to identification and correction of reporting heterogeneity. *Journal of Human Resources, 46*, 875-906.

d'Uva, T. B., Van Doorslaer, E., Lindeboom, M., & O'Donnell, O. (2008). Does reporting heterogeneity bias the measurement of health disparities? *Health Economics, 17*, 351-375.

Demars, C. E. (2007). Changes in rapid-guessing behavior over a series of assessments. *Educational Assessment, 12*, 23-45.

Dorans, N. J., & Holland, P. W. (1993). *DIF detection and description: Mantel-Haenzel and standardization*. In P. W. Holland & H. Wainer (Eds.), Differential item functioning (pp. 35-66). Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Finney, S. J., & DiStefano, C. (2006). Nonnormal and categorical data in structural equation models. In G.R. Hancock & R.O. Mueller (Eds.). *A second course in structural equation modeling*. Information Age: Greenwich, CT.

Grol-Prokopczyk, H., Freese, J., & Hauser, R. M. (2011). Using anchoring vignettes to assess group differences in general self-rated health. *Journal of Health and Social Behavior, 52,* 246-261.

Groot, W., 2000. Adaptation and scale of reference bias in self-assessments of quality of life. *Journal of Health Economics, 19*, 403-420.

Hancock, G. R. (1997). Structural equation modeling methods of hypothesis testing of latent variable means. *Measurement and Evaluation in Counseling and Development, 30,* 91 - 105.

Hancock, G. R. (2003). Fortune cookies, measurement error, and experimental design. *Journal of Modern Applied Statistical Methods, 2,* 293–305.

Horst, S. J. (2010). *A mixture-modeling approach to exploring test-taking motivation in large-scale low-stakes contexts*. (Unpublished doctoral dissertation). James Madison University, Harrisonburg, VA.

Hu, L. T., & Bentler, P. M. (1998). Fit indices in covariance structure modeling: Sensitivity to underparameterized model misspecification. *Psychological methods*, *3*, 424.

Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling, 6,* 1-55.

Hussar, W. J., & Bailey, T. M. (2013). Projections of Education Statistics to 2021. NCES 2013-008. *National Center for Education Statistics.*

Ip, W. Y., Lui, M. H., Chien, W. T., Lee, I. F., Lam, L. W., & Lee, D. (2012). Promoting self-reflection in clinical practice among Chinese nursing undergraduates in Hong Kong. *Contemporary nurse*, *41*, 253-262.

Jöreskog, K. G., & Sörbom, D. (1996). LISREL 8: User's reference guide. Scientific Software International.

Kapteyn, A., Smith, J. P., & van Soest, A. (2007). Vignettes and self-reports of work disability in the United States and the Netherlands. *The American Economic Review, 97*, 461-473.

Kennedy, K., & Ishler, J. C. (2008). The changing college student. *Academic advising: A comprehensive handbook*, 123-141.

Kerkhofs, M., & Lindeboom, M. (1995). Subjective health measures and state dependent reporting errors. Health economics, 4(3), 221-235.

King, G., Murray, C. J., Salomon, J. A., & Tandon, A. (2004). Enhancing the validity and cross-cultural comparability of measurement in survey research. *American Political Science Review, 98*, 191-207.

King, G., Murray, C. J., Salomon, J. A., & Tandon, A. (2009) "Anchoring vignettes: Frequently asked questions." *Retrieved March 14* (2016) from http://gking.harvard.edu/files/gking/files/vfaq.pdf?m=1360791098.

Kline, R. B. (2011).*Principles and Practice of Structural Equation Modeling* (3rd Edition). New York: Guilford.

Kong, X. J., Wise, S. L., & Bhola, D. S. (2007). Setting the response time threshold parameter to differentiate solution behavior from rapid-guessing behavior. *Educational and Psychological Measurement*.

Kong, X. J., Wise, S. L., & Bhola, D. S. (2007). Setting the response time threshold parameter to differentiate solution behavior from rapid-guessing behavior. *Educational and Psychological Measurement*.

Kristensen, N., & Johansson, E. (2008). New evidence on cross-country differences in job satisfaction using anchoring vignettes. *Labour Economics,15,* 96-117.

Lau, A. R., Swerdzewski, P. J., Jones, A. T., Anderson, R. D., & Markle, R. E. (2009). Proctors matter: Strategies for increasing examinee effort on general education program assessments. *The Journal of General Education,58,* 196-217.

Liu, O. L. (2011). Outcomes assessment in higher education: Challenges and future research in the context of voluntary system of accountability. *Educational Measurement: Issues and Practice, 30,* 2-9.

MacCallum, R. Roznowski, M., & Necowitz, L. B. (1992). Model modifications in covariance structure analysis: The problem of capitalization on chance. *Psychological Bullentin, 111,* 490-504

Marsh, H. W. (1994). Confirmatory factor analysis models of factorial invariance: A multifaceted approach. *Structural Equation Modeling, 1,* 5 – 34.

Messick, S. (1989). Meaning and values in test validation: The science and ethics of assessment. *Educational researcher, 18,* 5-11.

Millsap, R. E. & Meredith, JW. (2007). Factorial invariance: Historical perspectives and new problems. In R. Cudeck & R. C. MacCallum (Eds.), *Factor Analysis at 100: Historical developments and future directions* (pp. 131- 152). Mahwah, NJ: Lawrence Erlbaum.

Murray, C. J. L., Tandon, A., Salomon, J., & Mathers, C. D. (2000). Enhancing cross-population comparability of survey results. Global Programme on Evidence for Health Policy Series, 35.

Olsson, U. H., Foss, T., Troye, S. V., & Howell, R. D. (2000). The performance of ML, GLS, & WLS estimation in structural equation modeling under conditions of misspecification and nonnormality. *Structural Equation Modeling, 7*, 557-595.

Pascarella, E. T. (2006). How college affects students: Ten directions for future research. *Journal of College Student Development, 47*, 508-520.

Peracchi, F., & Rossetti, C. (2012). Heterogeneity in health responses and anchoring vignettes. *Empirical Economics, 42*, 513-538.

Pintrich, P. R. (1989). The dynamic interplay of student motivation and cognition in the college classroom. *Advances in Motivation and Achievement: Motivation Enhancing Environments, 6*, 117–60.

Pintrich, P. R. (2002). The role of metacognitive knowledge in learning, teaching, and assessing. *Theory into practice*, *41*, 219-225.

Pintrich, P. R., & DeGroot, E. V. (1990). Motivational and self-regulated learning components of classroom academic performance. *Journal of Educational Psychology, 82*, 33–40.

Programme for International Student Assessment, (2014). *PISA 2012 Technical Report*. Retrieved from Organization for Economic Co-operation and Development website: http://www.oecd.org/pisa/pisaproducts/PISA-2012-technical-report-final.pdf, p.

Raykov, T. (1997). Estimation of composite reliability for congeneric measures. *Applied Psychological Measurement, 21*, 173–184.

Raykov, T. (2004). Behavioral scale reliability and measurement invariance evaluation using latent variable modeling. *Behavior Therapy, 35*, 299–331.

Rensvold, R. B., & Cheung, G. W. (2001). Testing for metric invariance using structural equation models: Solving the standardization problem. *Research in management*, *1*, 21-50.

Rice, N., Robone, S., & Smith, P. (2011). Analysis of the validity of the vignette approach to correct for heterogeneity in reporting health system responsiveness. *The European Journal of Health Economics, 12*, 141-162.

Salomon, J. A., Tandon, A., & Murray, C. J. (2004). Comparability of self-rated health: cross sectional multi-country survey using anchoring vignettes. *BMJ, 328*, 258.

Sessoms, J., & Finney, S. J. (2015). Measuring and modeling change in examinee effort on low-stakes tests across testing occasions. *International Journal of Testing*, *15*, 356-388.

Smith, J. K., Given, L. M., Julien, H., Ouellette, D., & DeLong, K. (2013). Information literacy proficiency: Assessing the gap in high school students' readiness for undergraduate academic work. *Library & Information Science Research*, *35*, 88-96.

Steedle, J. T. (2014). Motivation filtering on a multi-institution assessment of general college outcomes. *Applied Measurement in Education*, *27*, 58-76.

Steenkamp, J., & Baumgartner, H. (1998). Assessing measurement invariance in cross-national consumer research. *Journal of Consumer Research, 25*, 78 – 90.

Sundre, D. L., & Finney, S. J. (2002). *Enhancing the validity and value of learning assessment: Furthering the development of a motivation scale*. Paper presented at the annual meeting of the American Educational Research Association, New Orleans.

Sundre, D. L., & Kitsantas, A. (2003). An exploration of the psychology of the examinee: Can examinee self-regulation and test-taking motivation predict consequential and nonconsequential test performance? *Contemporary Educational Psychology, 29*, 6–26.

Sundre, D. L., & Thelk, A. D. (2007). The Student Opinion Scale (SOS): A Measure of Examinee Motivation: Test Manual.

Sundre, D. L., & Wise, S. L. (2003, April). '*Motivation filtering': An exploration of the impact of low examinee motivation on the psychometric quality of tests*. Paper presented at the annual meeting of the National Council on Measurement in Education, Chicago, IL.

Sundre, D.L. (1999, April). *Does examinee motivation moderate the relationship between test consequences and test performance?* Paper presented at the annual meeting of the American Educational Research Association, Montreal (ERIC Document Reproduction Service No. ED432588).

Swerdzewski, P. J., Finney, S. J., & Harmes, J. C. (2007, October). *Examinee motivation in low-stakes testing: Two approaches to identifying data from low-motivated students in an applied assessment context.* Paper presented at the annual meeting of the Northeastern Educational Research Association, Rocky Hill, Conn.

Thelk, A. D. (2006). *Examinee awareness of performance expectation and its eff ects on motivation and test scores.* Unpublished doctoral dissertation, James Madison University.

Thelk, A. D., Sundre, D. L., Horst, S. J., & Finney, S. J. (2009). Motivation matters: Using the Student Opinion Scale to make valid inferences about student performance. *The Journal of General Education, 58*, 129-151.

Thompson, M. S., & Green, S. B. (2006). Evaluating between-group differences in latent variable means. In G. R. Hancock & R. O. Mueller (Eds.), *A second course in structural equation modeling.* Greenwich, CT: Information Age.

Van Doorslaer, E., & Jones, A. M. (2003). Inequalities in self-reported health: validation of a new approach to measurement. *Journal of health economics, 22*, 61-87.

Van Soest, A., Delaney, L., Harmon, C., Kapteyn, A., & Smith, J. P. (2011). Validating the use of anchoring vignettes for the correction of response scale differences in subjective questions. *Journal of the Royal Statistical Society: Series A (Statistics in Society), 174*, 575-595.

Wand, J., King, G., & Lau, O. (2011). Anchors: Software for anchoring vignette data. *Journal of Statistical Software, 42*, 1-25.

Wise, S. L. (2006). An investigation of the differential effort received by items on a low-stakes computer-based test. *Applied Measurement in Education*, *19*, 95-114.

Wise, S. L., & DeMars, C. E. (2005). Low examinee effort in low-stakes assessment: Problems and potential solutions. *Educational Assessment, 10*, 1-17.

Wise, S. L., & Kong, X. (2005). Response time eff ort: A new measure of examinee motivation in computer-based tests. *Applied Measurement in Education, 18,* 162–83.

Wolf, L. F., & Smith, J. K. (1993, April). *The effects of motivation and anxiety on test performance*. Paper presented at the annual meeting of the American Educational Research Association, Atlanta.

Yu, C., & Muthén, B (2002, April). Evaluation of model fit indices for latent variable models with categorical and continuous outcomes. Paper presented at the annual meeting of the American Educational Research Association, New Orleans.