BILINGUAL PUBLISHING CO.
Pioneer of Global Academics Since 1984

Journal of International Education and Practice
http://ojs.bilpublishing.com/index.php/jiep

ARTICLE

Investigating the Psychometric Impact of Negative Worded Items in Reading Comprehension Passages with a 3PL Cross-Classified Testlet Model

Yong Luo[1*] Junhui Liu[2]

1. National Center for Assessment, Riyadh, Saudi Arabia.
2. University of Maryland, College Park, Maryland, USA.

ARTICLE INFO

Article history
Received: 19 December 2018
Accepted: 11 March 2019
Published: 30 March 2019

Keywords:
Negative wording
Bifactor model
Cross-classified testlet model
Validation

ABSTRACT

Negative worded (NW) items used in psychological instruments have been studied with the bifactor model to investigate whether the NW items form a secondary factor due to negative wording orthogonal to the measured latent construct, a validation procedure which checks whether NW items form a source of construct irrelevant variance (CIV) and hence constitute a validity threat. In the context of educational testing, however, no such validation attempts have been made. In this study, we studied the psychometric impact of NW items in an English proficiency reading comprehension test using a modeling approach similar to the bifactor model, namely the three-parameter logistic cross-classified testlet response theory (3PL CCTRT) model, to account for both guessing and possible local item dependence due to passage effect in the data set. The findings indicate that modeling the NW items with a separate factor leads to noticeable improvement in model fit, and the factor variance is marginal but nonzero. However, item and ability parameter estimates are highly similar between the 3PL CCTRT model and other models that do not model the NW items. It is concluded that the NW items introduce CIV into the data, but its magnitude is too small to change item and person ability parameter estimates to an extent of practical significance.

1. Introduction

Negatively worded (NW) items are often recommended to be included along with positively worded (PW) ones in psychological inventories to address acquiescence (e.g., Kieruj & Moors, 2013).[34] A common practice is to reversely code NW items and treat them the same as PW ones. Such a practice, however, depends heavily upon the assumption that there is no wording effect associated with the NW format used in NW items, which is usually not the case. Studies (e.g., Chessa & Holleman, 2007)[12] have shown that the cognitive process involved in answering NW items is different than that in dealing with PW items, and consequently NW items display different psychometric properties such as lower item-total score correlation (e.g., Roszkowski & Soven, 2010).[49] Consequently, some

**Corresponding Author:*
Yong Luo,
Senior Psychometrician, National Center for Assessment, Riyadh, Saudi Arabia;
Email: jackyluoyong@gmail.com.

 DOI: https://doi.org/10.30564/ jiep.v2i1.418

researchers (e.g., van Sonderen, Sanderman, & Coyne, 2013; [56] Zhang, Noor, & Savalei, 2016) [62] caution against the use of NW items, arguing that such items introduce extraneous variance and hence pose a threat to construct validity.

Factor analysis techniques have been routinely applied to psychological inventories containing NW items to investigate the wording effect. A two-correlated-factor model is often used (Deemer & Minke, 1999;[14] Gitchel, Roessler, &Tuner, 2011;[21] Magazine, Williams, & Williams, 1996;[42] Roszkowski & Soven, 2010)[49] to model the PW and the NW items, assuming that one factor representing the negative wording effect and the other the positive wording effect. The bifactor model has also been used (e.g., Lindwall, Barkoukis, Grano, Lucidi, & Raudsepp, 2012),[37] with the general factor hypothesized to measure the latent construct of interest, and the two secondary factors represented the positive and negative wording effects. Wang, Chen, and Jin (2015)[58] decisively pointed out the logical flaw of treating the positive wording effect as a separate secondary factor in the previous bifactor modeling approach and used a different bifactor model in which only the negative wording effect is modeled as a separate secondary factor.

In contrast to the psychological measurement literature where a consensus regarding the use of NW items is lacking, in educational testing community use of the NW format is usually cautioned against (Haladyna, 2004;[22] Haladyna & Downing, 1989a;[23] 1989b;[24] Haladyna, Downing, & Rodriguez, 2002).[26] Note that the NW items in educational testing often involve adding at the stem or option level a negative word such as not or except, while in psychological measurement they can either add negative words or use vocabulary that is opposite to the measured construct (e.g., use of the word sad in a scale measuring happiness). It is further recommended that in cases where it is necessary to use in an educational test the NW format (adding a negative word), the negative word "…should be stressed or emphasized by placing it in bold type, capitalizing it, or underlining it, or all of these" (Haladyna, 2004, p. 111).[22] One of the reasons for such recommendations against the NW items is that educational tests are usually high-stakes and students are too motivated to allow acquiescence bias to materialize. Another reason, as will be discussed later, is that the addition of a negative word in educational test items tends to change item psychometric properties.

Despite the suggestion that NW items should be avoided in high-stakes testing, they are still occasionally used in some educational tests, although per Haladyna's advice (e.g., 2002;[26] 2004),[22] negative words in those items are often emphasized. Research on NW items in educational measurement often utilizes experimental studies to investigate how NW items perform in contrast to their PW counterparts, and to date, there have been no studies applying a factor analysis approach to model the effect of the NW format in educational tests.

In this study we investigate the wording effect of NW items in a high-stakes English proficiency reading comprehension test using a model similar to the bifactor model used by Wang, Chen, and Jin (2015).[58] Our model is similar to that of Wang et al. (2015)[58] in that we also model the negative wording effect as a secondary factor independent of the primary factor, which is English proficiency in this case. The similarity notwithstanding, there are several major differences between the two models. First, guessing is expected to exist in our data due to the use of MC format and as a result, we include a pseudo-guessing parameter in our model to account for guessing. Since the mathematical equivalence between the item response theory (IRT) and factor analysis (e.g., Takane & de Leeuw, 1987;[52] Kamata & Bauer, 2008) [33] does not extend to models with pseudo-guessing parameters, our model is not a factor analysis model per se. Second, instead of using a general testlet model (Li, Bolt, & Fu, 2006)[35] that is the IRT analog of a bifactor model, we use a testlet model (Wainer, Bradlow, & Wang, 2007)[57] that is a constrained version of the general testlet model (e.g., DeMars, 2006;[15] Rijmen, 2010)[47] to model the negative wording effect. We believe the testlet model is more appropriate in the current study due to its estimation of testlet variance, which allows for a straightforward interpretation of the magnitude of variance caused by the wording effect. As will be discussed later, this variance is irrelevant of the primary latent construct of interest and its magnitude indicates how much of a threat it is to the test validity. Last, since our data were drawn from a reading comprehension test, the passage effect may cause items within the same passage to be locally dependent. For NW items within passages, we hypothesize they exhibit dual local dependence due to the passage and wording effects and in order to simultaneously model both effects, a cross-classified testlet model is warranted. Consequently, we use a three-parameter logistic (3PL) cross-classified testlet response theory (CCTRT) model to answer the following research questions:

Does the NW format introduce construct irrelevant variance (CIV) into the test?

If yes, what is its magnitude?

How does failure to model such CIV affect item and ability parameter estimates?

The remainder of this paper is organized into four

DOI: https://doi.org/10.30564/ jiep.v2i1.418

sections. We start with a review of relevant studies in the educational measurement literature that investigate the wording effect of NW items. In the second section, we introduce the 3PL CCTRT model, followed by analysis of the current data in the third section. In the last section we end our paper with discussions and conclusions.

2. Studies Regarding NW Items in Educational Testing

Most studies focusing on the psychometric effect of NW items in the context of educational measurement adopt experimental designs in which the performances of two versions of a small number of items, one with NW format and the other PW format, were compared. For example, Terranova (1969)[55] found that the NW items were more difficult than their PW counterparts, and test reliability in his experiment was not affected by the NW format. Similarly, Dudycha and Carpenter (1973) [18] found that NW stems increased the item difficulty, whereas the item discrimination was not affected. In another study, Cassels and Johnstone (1984)[10] found that simply changing item stems from the NW format to the PW format with the options remaining constant lowered the item difficulty considerably, and they attributed such changes of item difficulty to the additional thinking stage required by NW format. Similar findings were also presented by Caldwell and Pate (2013)[8] that stem negation increased item difficulty. Johnstone (1983, p. 115)[31] indicated that the reason for increased item difficulty due to the NW format is that "…ideas in a negative form occupy twice as much space in the working memory as positive forms". Similarly, Abedi (2006)[1] listed the NW format as one of the linguistic features that might affect comprehension.

While the aforementioned studies consistently find that the NW format is associated with increased item difficulty, there are other studies that say otherwise. For example, Tamir (1991; 1993)[53][54] found that for items requiring low cognitive reasoning, the NW format did not affect the item difficulty; it was only when combined with requirements for high cognitive reasoning that the NW format increased item difficulty. Downing, Dawson-Saunders, Case, and Powell (1991),[17] and Rachor and Gray (1996)[44] found that the NW format had no effects upon item difficulty. In another study (Harasym, Price, Brant, Violato, & Lorscheider, 1992),[27] it was found that the NW format lowered both item difficulty and test reliability, and the researchers attributed the decrease of item difficulty to that the NW format inadvertently provided cues to the correct answer.

Other than focusing on how NW items behave in contrast to their PW counterparts, Casler (1983)[9] took a different approach by investigating whether emphasizing the negative words in the NW items alleviated the psychometric effect caused by the NW format. He found that with the negative words underlined, the NW items became easier to the high ability students and harder to the low ability ones, and the item discrimination power increased; if the negative words were capitalized, the item difficulty decreased while the item discrimination remained unchanged.

Apparently, the majority of abovementioned studies find that the NW format changes the psychometric properties of an item. Consequently, it has become a widely-accepted notion that the NW format is undesirable and should be avoided in the context of educational measurement. As stated by Haladyna (2004, p. 117):[22]

Avoid Negative Words Such as Not or Except. We should phrase stems positively, and the same advice applies to options. The use of negatives such as not and except should also be avoided in options as well as the stem. Occasionally, the use of these words in an item stem is unavoidable. In these circumstances, we should boldface, capitalize, italicize, or underline these words so that the test taker will not mistake the intent of the item.

To date, most studies regarding the psychometric effect of the NW format in educational testing literature utilized experimental designs to investigate how item properties such as item difficulty and discrimination change, and only one (Casler, 1983)[9] compared the psychometric effects caused by NW items with and without emphasizing the negative words within. There are no studies that use a bifactor model approach similar to that adopted by Wang, Chen, and Jin (2015)[58] to explore whether the NW items form a separate factor due to their NW format. We argue that such a NW factor, if existent, constitutes a source of CIV that can be a major threat to test validity (e.g., Hahadyna & Downing, 2004).[25] Messick described the role that CIV plays in educational testing as the following:

The major point here is that educational achievement tests, at best, reflect not only the psychological constructs of knowledge and skills that are intended to be measured, but invariably a number of contaminants. These adulterating influences include a variety of other psychological and situational factors that technically constitute either construct-irrelevant test difficulty or construct-irrelevant contamination in score interpretation. (Messick, 1989, p. 216)[43]

Despite the lack of consensus on how the NW format changes item difficulty, the majority of studies, as discussed previously, indicate that the NW format changes item difficulty due to the introduction of CIV. Similarly, Downing (2005, pp. 141-142)[16] stated that

 DOI: https://doi.org/10.30564/ jiep.v2i1.418

"the additional test difficulty introduced into the measure by poorly crafted and flawed item formats is an example of construct-irrelevant variance." Following his line of reasoning, we hypothesize that the NW format in items is a source of construct-irrelevant variance (CIV) and therefore, it is important to find out what its magnitude is and whether it affects model parameter estimates to an extent of practical significance.

It should be noted that since the data used in this study were drawn from a reading comprehension test, passage effect may constitute another source of CIV. In the next section we will discuss that although our interest is not in the passage effect per se, failure to model the passage effect might result in inaccurate estimation of the wording effect and consequently, we model both sources of CIV simultaneously with a CCTRT model.

3. The Testlet Model and Its Cross-Classified Extension

One pivotal assumption of item response theory (IRT) is local item independence, which can be expressed using the following equation (Reckase, 2009)[45]

$$P(\mathbf{U}=\mathbf{u}|\theta)=P(u_1|\theta)P(u_2|\theta)...P(u_I|\theta), \quad (1)$$

where $\mathbf{u}$ is a response vector to a test with I items, $P(\mathbf{U}=\mathbf{u}|\theta)$ is the probability of obtaining the response vector $\mathbf{u}$ for an examinee whose latent ability is θ, and $P(u_i|\theta)$ is the probability of obtaining a score u_i. Equation 1 states that after conditioning on the latent ability, the response to any item in the test is statistically independent of that to another item. In other words, an examinee's latent ability should be the only force that drives his or her item responses and, if there is another factor that affects the item response, this independence assumption is violated and local item dependence (LID) occurs. As listed by Yen (1993),[60] in real testing situations LID can occur due to various factors such as speededness, item or response format, and passage dependence. Numerous studies (e.g., Ackerman, 1987;[2] Chen & Thissen, 1997;[11] Zhang, 2010)[61] have shown that LID can result in biased estimation of item parameters, overestimation of test reliability, premature termination of computer adaptive testing, errors in equating, and erroneous classifications of examinees.

Due to the serious psychometric consequences that can be caused by LID, various methods (e.g., Bradlow, Wang, & Wainer, 1999[6]; Braeken, Tuerlinckx, & De Boeck, 2007[7]; Hoskens & De Boeck, 1997;[28] Rosenbaum, 1988[48]) have been proposed to address the issue of LID. Among them, a popular approach that has been applied extensively to address LID is the testlet response theory (TRT; Bradow, et al., 1999;[6] Wainer, Bradlow, & Wang, 2007),[57] which models LID among items within the same cluster by introducing a random effect parameter denoting the person specific testlet effect. The probability of answering item j correctly by person i in a 3PL TRT model is given as

$$p_i(\theta_j) = c_j + \frac{1 - c_j}{1 + e^{-a_j(\theta_i - b_j + \gamma_{id(j)})}}, \quad (2)$$

where θ_i is person i's latent ability, $\gamma_{id(j)}$ is person i's latent ability on testlet d, and a_j, b_j, and c_j are the discrimination, difficulty, and pseudo-guessing parameters of item j. For $\gamma_{id(j)}$, its variance $\sigma^2_{\gamma_{jd(i)}}$ indicates the magnitude of LID among items within the same testlet. As can be seen from equation 2, TRT only allows the modeling of one source of LID. When it is suspected that dual LID may exist, TRTM seems inadequate due to their incapability of handling more than one source of LID simultaneously.

It is not uncommon for test data to display dual LID due to the existence of two item clustering factors. For example, in PISA assessment items based on the same scenario may fall into different content categories: here scenario and content are two item clustering factors that may cause dual LID. Another example is that in language testing, items within the same listening comprehension passage may have different item formats, thus making format and passage two possible sources of LID. In the current study, dual LID is also suspected to exist due to two item clustering effects, namely the passage effect and the wording effect due to the NW format. To address the issue of dual LID, Jiao, Wang, Wan, and Lu (2009)[29] proposed a 3PL CCTRT model, which is an extension of the 3PL TRT model, to address dual LID in scenario-based science assessment items. Its equation is given as

$$p_i(\theta_j) = c_j + \frac{1 - c_j}{1 + e^{-a_j(\theta_i - b_j + \gamma_{id_1(j)} + \gamma_{id_2(j)})}}, \quad (3)$$

where $\gamma_{id_1(j)}$ is person i's latent ability on testlet d_1 caused by the first source of LID, $\gamma_{id_2(j)}$ on testlet d_2 caused by the second source of LID, and the other terms remain the same as in equation 2. As can be seen, if either source of LID has a variance of zero, equation 3 reduces to equation 2 and the 3PL CCTRT model becomes the familiar 3PL TRT model.

The difference between a TRT model and a CCTRT model can be visualized with diagrams. Assuming that we have a reading comprehension test with two passages (each has three items) and two of the six items (Items 2 and 4) have the NW format, Figure 1 provides a visual presentation of the TRT and CCTRT models that can be used to model difference sources of LID. As can be seen, the TRT model in the left panel only models LID due to the passage effect, and its CCTRT counterpart models dual LID due to both the passage and NW effects. One commonality between these two models is that there are

DOI: https://doi.org/10.30564/ jiep.v2i1.418

no arrows or curves linking the factors, which indicates that the factors are orthogonal to each other.

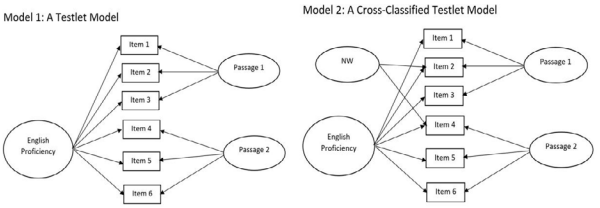


Figure 1. The traditional testlet model and the cross-classified testlet model

In addition to the 3PL CCTRT model, the Rasch version of CCTRT models has also received some attention in literature. Xie (2014)[59] proposed a cross-classified Rasch testlet model in the hierarchical generalized linear model (HGLM; Kamata 2001)[32] and investigated the consequences of failing to model the dual LID properly. She found that ignoring either source of LID leads to inaccurate estimation for item difficulty, ability parameter, and testlet effect parameters. Similar to the relation between the Rasch model and the 3PL model, Xie's model is a special case of the 3PL CCTRT model in equation 3 with the guessing parameter c_j constrained to zero and the discrimination parameter a_j to one across items. Jiao, Kamata, and Xie (2015)[30] extended Xie's model to its multilevel case and showed that the multilevel cross-classified Rasch testlet model can be accurately estimated using Markov Chain Monte Carlo (MCMC) methods implemented in OpenBUGS.

4. Methods

4.1 Data

The data were drawn from item responses of 1,839 students who took a high-stakes English proficiency test used for admission and placement purposes in the Middle East. For the purpose of this study, we focused on the reading comprehension section which consists of 40 items.

Table 1 lists descriptive statistics (p value and item total correlation) and other relevant information of these 40 items. The column named p provides the classical test theory (CTT) based item difficulty index, based on which we observe that the test is slightly difficult to the examinees since most p values are below 0.50. The column named r_{pb} provides the item total correlation values, most of which are between 0.3 and 0.5 with some lower than 0.2 (e.g., RC17).

The column named Item Type tells whether an item is a discrete one or belongs to a particular passage, and the column named NW tells whether an item is a NW one. As can be seen, 37 out of the 40 items are nested within reading comprehension passages and among them, seven items use NW format. Among these seven items, two use a negative word at the stem level and the other five at the option level, and all the negative words are bolded and capitalized for highlight purposes. Per our previous discussion, we hypothesize that these seven items display dual LID due to the passage effect and the negative wording effect.

Table 1 Item Statistics and Relevant Information

Item	p	r_{pb}	Item Type	NW	Item	p	r_{pb}	Item Type	NW
RC1	0.39	0.25	Passage 1		RC21	0.35	0.24	Passage 5	
RC2	0.50	0.44	Passage 1	Yes	RC22	0.46	0.42	Passage 5	
RC3	0.30	0.20	Passage 1		RC23	0.41	0.37	Passage 5	Yes
RC4	0.43	0.43	Passage 1		RC24	0.46	0.46	Passage 5	
RC5	0.43	0.49	Passage 1		RC25	0.43	0.45	Passage 5	
RC6	0.36	0.27	Passage 2		RC26	0.32	0.44	Passage 5	
RC7	0.33	0.33	Passage 2		RC27	0.36	0.29	Passage 5	
RC8	0.33	0.17	Passage 2	Yes	RC28	0.42	0.54	Passage 5	Yes
RC9	0.35	0.30	Passage 2		RC29	0.26	0.18	Passage 5	Yes
RC10	0.32	0.35	Passage 2		RC30	0.45	0.50	Discrete	
RC11	0.36	0.44	Passage 3		RC31	0.34	0.31	Passage 6	
RC12	0.46	0.46	Passage 3		RC32	0.35	0.37	Passage 6	
RC13	0.48	0.46	Passage 3		RC33	0.37	0.45	Passage 6	
RC14	0.45	0.43	Passage 3		RC34	0.23	0.30	Passage 6	Yes
RC15	0.27	0.32	Passage 4		RC35	0.31	0.36	Passage 6	
RC16	0.24	0.06	Passage 4		RC36	0.29	0.17	Passage 7	
RC17	0.17	0.09	Passage 4		RC37	0.32	0.19	Passage 7	
RC18	0.35	0.21	Passage 4		RC38	0.27	0.20	Passage 7	
RC19	0.35	0.39	Passage 4	Yes	RC39	0.47	0.48	Discrete	
RC20	0.36	0.39	Discrete		RC40	0.36	0.25	Discrete	

4.2 Analytic Procedure

Before we model such dual LID with a CCTRT model, it is necessary to determine which dichotomous IRT model should serve as the base model. Correspondingly, we estimate the 1PL, 2PL, and 3PL models with OpenBUGS and use Akaike's information criterion (AIC; Akaike, 1973),[3] Bayesian information criterion (BIC; Schwarz, 1978),[50] and the deviance information criterion (DIC; Spiegelhalter, Best, Carlin, & van der Linde, 2002)[51] to determine the best fitting model, which is used as the base model in the subsequent analyses. It should be noted that AIC and BIC were originally developed in the maximum likelihood estimation (MLE) framework, and here we use their Bayesian analogues that are computed with the

 DOI: https://doi.org/10.30564/ jiep.v2i1.418

posterior mean of deviance as described by Congdon (2003).[13]

After the base model is determined, we build a corresponding TRT model in which the item clustering effect due to passage dependence is modeled, and a CCTRT model in which the two item clustering effects due to both the passage dependence and the NW format are modeled simultaneously. This particular CCTRT model is treated as the true model in the current study. Similarly, we estimate the TRT and CCTRT models with OpenBUGS and compare model fit using AIC, BIC, and DIC. We also compare the item and ability parameters between different models to investigate whether failure to model any item clustering effect translates into practical significance in terms of parameter estimate differences.

4.3 Estimation

All three models were estimated via MCMC algorithm implemented in OpenBUGS. It should be noted that Stan, an emerging Bayesian software program, may be a better choice for estimating complex IRT models due to its sampling efficiency (e.g., Luo & Jiao, 2018;[40] Luo & Liang, 2019).[41] OpenBUGS was chosen in this study due to its convenient feature of computing DIC by default, while in Stan no such features exist and the users have to write their own functions to compute DIC (e.g, Luo, 2019).[38] Estimation of IRT models with MCMC methods requires the specification of prior distributions for all model parameters, and we choose priors that are commonly seen in the Bayesian IRT literature. For the 3PLM model, we assign a standard normal distribution $N(0, 1)$ as the prior for the ability parameters for model identification, and a normal distribution with unknown mean and variance as the prior for the item difficulty parameter; the unknown mean is assigned a standard normal distribution $N(0, 1)$ as the hyperprior, and we assign the distribution $\gamma(1, 1)$ as the hyperprior for the precision parameter, which is the reciprocal of the variance. We assign a truncated normal distribution $N_{+}(0, 4)$ as the prior for item discrimination parameter, and a beta distribution $\beta(5, 23)$ for the pseudo-guessing parameter. For the testlet variance parameters in the 3PL TRT and 3PL CCTRT models, we assign as the prior normal distributions with a mean of zero and unknown precisions, for which $\gamma(1, 1)$ is assigned as the hyperprior.

4.4 Model Convergence Check

Since MCMC methods are used for model estimation, it is necessary to check whether model convergence has been reached before we draw inferences from the posterior distribution. In this study we apply the Gelman and Rubin's convergence diagnostic (Gelman & Rubin, 1992),[19] which computes the potential scale reduction factor (PSRF). PSRF values close to 1 indicate model convergence and as suggested by Gelman, Carlin, Stern, and Rubin (2014),[20] PSRF value of 1.1 can be used as the cutoff value to gauge model convergence in practice. For the 1PL and 2PL models, all PSRF values converge to 1 within 2,000 iterations, and we run three parallel chains with 4,000 iterations each to be conservative. For the 3PL, 3PL TRT and 3PL CCTRT models, we run three parallel chains with 5,000 iterations each and request every 10^{th} iteration to be used for inference (thinning = 10) to reduce autocorrelation in the posterior distribution for testlet variance parameters.

5. Results

5.1 Model Comparison

As can be seen from Table 2, the 3PL model has the smallest AIC, BIC, and DIC values among the three common dichotomous IRT models, indicating that it is the best fitting model. With the 3PL model as the base model, we estimated the corresponding 3PL TRT and the 3PL CCTRT models. Regarding the comparison among the three models, AIC, BIC, and DIC values consistently indicate that the 3PL CCTRT model has the best model fit, followed by the 3PL TRT model, and the 3PL model has the worst model fit. Using Anderson's suggestion (2008)[4] that a difference of nine or greater in those information criteria constitutes strong evidence for model choice, we find that the 3PL TRT model provides a model fit considerably better than the 3PL model with the differences in AIC, BIC, and DIC being 1226, 1187, and 670 respectively; the 3PL CCTRT model fits the data noticeably better than the 3PL TRT model with the differences in AIC, BIC, and DIC being 154, 149, and 50 respectively. In other words, when we model the item clustering effect due to passage effect in the 3PL TRT model, we find strong evidence that it should be chosen over the 3PL model, which assumes that no item clustering effect exists; when we model the dual item clustering effects due to passage effect and wording effect in the 3PL CCTRT model, model fit improves considerably over that of the 3PL TRT model, which assumes that no item clustering effect exists due to the NW format.

Table 2 Model Comparison Results

Model	AIC	BIC	DIC
1PL	86200	86431	87790
2PL	84877	85329	86430
3PL	84462	85135	85700
3PL TRT	83236	83948	85030
3PL CCTRT	83082	83799	84980

DOI: https://doi.org/10.30564/ jiep.v2i1.418

5.2 Magnitude of CIV

Table 3 lists the testlet variance estimates for the 3PL TRT and 3PL CCTRT models, which are indicative of the magnitude of CIV caused by the passage effects and negative wording effects. Xie (2014)[59] found in her simulation study that failure to model one source of item clustering effect in a cross-classified model results in biased testlet variance estimates, and with the increase of the magnitude of that item clustering effect, the bias increases. One natural question here is whether failure to model the wording effect leads to an incorrect interpretation of the magnitude of the item clustering effect due to passage effect. The comparison of the testlet variance estimates between these two models indicates that despite the better model fit that the 3PL CCTRT model has over the 3PL TRT model, such a model fit advantage does not translate into practical significances regarding the testlet variance estimates. The correlation between two sets of testlet variance estimates are greater than 0.99, and the values are nearly identical. As can be seen, none of the seven passages exhibits strong item clustering effect due to passage dependence, with the largest value being approximately 0.19. For the negative wording effect, the testlet variance estimate is 0.09 with a 95% credible interval not covering zero, indicating that only a small amount of CIV has been introduced into the test due to the NW format.

Table 3 Testlet Variance Estimates

Model	T_1 1	T_1 2	T_1 3	T_1 4	T_1 5	T_1 6	T_1 7	T_2 1
3PL TRT	0.13*	0.13*	0.15*	0.19*	0.11*	0.14*	0.18*	
3PL CCTRT	0.14*	0.13*	0.15*	0.18*	0.11*	0.14*	0.18*	0.09*

Note. * indicates that the 95% credible interval of the variance estimate does not cover zero.

5.3 Item and Ability Parameter Estimate Comparison

Table 4 Item Parameter Estimates Comparison

Item	Discrimination			Difficulty			Pseudo-Guessing		
	M1	M2	M3	M1	M2	M3	M1	M2	M3
RC1	1.27	1.13	1.10	1.55	1.59	1.59	0.25	0.23	0.23
RC2	1.87	1.80	1.77	0.51	0.47	0.46	0.22	0.20	0.20
RC3	2.78	2.84	2.84	1.87	1.97	1.99	0.25	0.25	0.25
RC4	2.11	2.00	1.99	0.85	0.85	0.85	0.23	0.22	0.22
RC5	2.44	2.58	2.55	0.64	0.68	0.67	0.19	0.19	0.19
RC6	2.61	2.31	2.24	1.62	1.72	1.73	0.28	0.28	0.28
RC7	2.69	2.59	2.51	1.49	1.56	1.58	0.24	0.24	0.23
RC8	2.28	2.18	2.27	2.15	2.30	2.37	0.28	0.28	0.28
RC9	2.83	2.34	2.37	1.52	1.61	1.62	0.27	0.26	0.26
RC10	2.73	2.53	2.51	1.46	1.55	1.56	0.23	0.23	0.23
RC11	2.19	2.17	2.15	0.99	1.04	1.05	0.17	0.17	0.17
RC12	1.66	1.64	1.64	0.46	0.47	0.48	0.14	0.14	0.14
RC13	2.98	3.32	3.26	0.68	0.69	0.69	0.28	0.28	0.28
RC14	2.19	2.20	2.17	0.78	0.79	0.79	0.24	0.23	0.23
RC15	1.67	1.51	1.51	1.62	1.70	1.72	0.15	0.14	0.14
RC16	3.42	3.13	3.13	2.42	2.76	2.77	0.23	0.23	0.23
RC17	3.83	3.55	3.51	2.36	2.64	2.67	0.15	0.15	0.15
RC18	1.70	1.29	1.30	1.87	2.00	2.01	0.27	0.23	0.24
RC19	2.24	2.22	2.41	1.26	1.29	1.34	0.21	0.20	0.21
RC20	2.21	2.07	2.05	1.22	1.24	1.25	0.22	0.22	0.22
RC21	1.38	1.15	1.13	1.77	1.90	1.92	0.24	0.22	0.22
RC22	1.49	1.37	1.38	0.52	0.53	0.53	0.14	0.13	0.13
RC23	2.16	2.07	2.04	1.10	1.15	1.19	0.26	0.26	0.26
RC24	2.18	2.22	2.19	0.61	0.64	0.63	0.21	0.21	0.21
RC25	2.08	2.10	2.08	0.77	0.79	0.80	0.20	0.20	0.20
RC26	3.89	4.08	4.11	1.21	1.28	1.29	0.21	0.21	0.21
RC27	2.32	2.55	2.53	1.59	1.69	1.70	0.28	0.28	0.28
RC28	3.03	2.94	3.08	0.62	0.61	0.62	0.17	0.16	0.16
RC29	1.77	1.76	1.83	2.30	2.44	2.51	0.20	0.21	0.21
RC30	2.40	2.31	2.30	0.57	0.54	0.55	0.19	0.18	0.18
RC31	2.03	1.78	1.76	1.52	1.63	1.64	0.24	0.23	0.23
RC32	1.72	1.59	1.59	1.28	1.30	1.31	0.20	0.19	0.19
RC33	3.83	3.93	3.96	1.08	1.11	1.11	0.24	0.23	0.23
RC34	3.30	3.29	3.39	1.64	1.76	1.81	0.17	0.16	0.16
RC35	1.87	1.75	1.71	1.40	1.48	1.50	0.18	0.17	0.17
RC36	4.24	3.71	3.64	1.91	2.14	2.16	0.25	0.25	0.25
RC37	4.70	4.52	4.47	1.82	1.99	2.01	0.28	0.28	0.28
RC38	4.51	4.77	4.76	1.81	1.96	1.98	0.23	0.22	0.22
RC39	1.77	1.75	1.73	0.34	0.34	0.34	0.11	0.11	0.11
RC40	1.11	1.02	1.00	1.76	1.82	1.82	0.22	0.21	0.21

Note. M1 refers to the 3PL model, M2 the 3PL TRT model, and M3 the 3PL CCTRT model.

Table 4 lists the item parameter estimates from the three models. We observe that the three sets of item parameters are highly similar, although not identical. To further investigate whether there are any systematic patterns, we plot pairwise comparison of item and ability parameters and their corresponding standard errors in Figures 2-5. Note that the dotted line in these figures represents the regression line $y = x$, and how much a point deviates from this line indicates the magnitude of difference between two parameter estimates. Specifically, Figure 2 compares the differences in item parameter estimates and their standard errors between the 3PL and the 3PL TRT models. Such a comparison tells us the magnitude of differences regarding item parameter estimates as a result of not modeling the item clustering

 DOI: https://doi.org/10.30564/ jiep.v2i1.418

effect due to passage dependency. As can be seen, the two sets of item parameter estimates and their standard errors are similar with correlation values all greater than 0.96. Item discrimination parameter seems to be slightly overestimated in the 3PL model where the passage effect is not modeled, while its standard error appears to be slightly underestimated. Item difficulty parameter is virtually unaffected, while its standard error is somewhat underestimated. Item pseudo-guessing parameter is slightly overestimated, and its standard error is slightly underestimated.

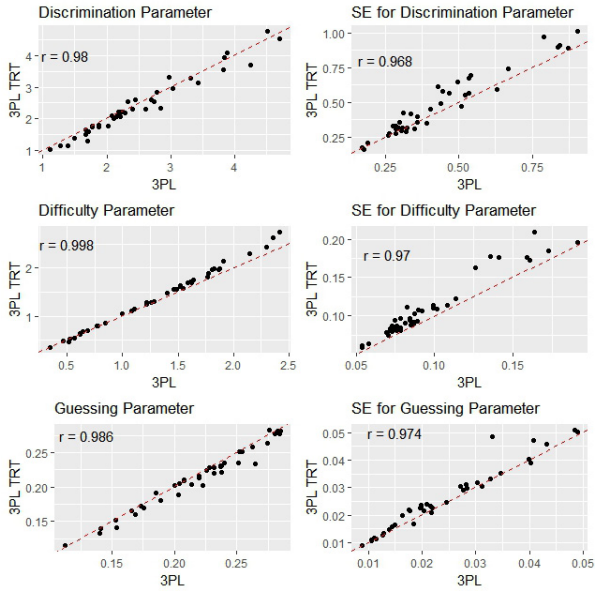
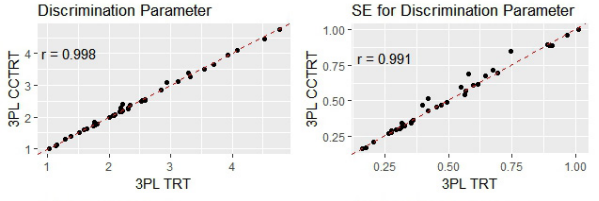


Figure 2. Item Parameter Comparison between the 3PL and 3PL TRT Models

Figure 3 compares the differences in item parameter estimates and their standard errors between the 3PL TRT and the 3PL CCTRT models. Such a comparison tells us the magnitude of differences regarding item parameter estimation as a result of modeling the item clustering effect due to passage dependency but not the item clustering effect due to the NW format. As can be seen, the two sets of item parameter estimates and their standard errors are nearly identical with the lowest correlation value being 0.991. In other words, item discrimination, difficulty, and pseudo-guessing parameters and their standard errors remain virtually the same when the item clustering effect due to the NW format is not modeled.



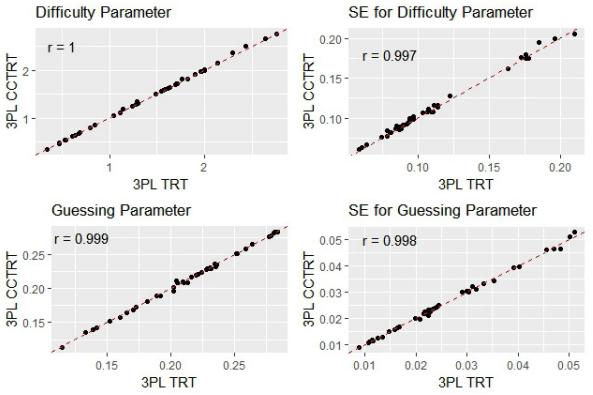


Figure 3. Item Parameter Comparison between the 3PL TRT and 3PL CCTRT Models

Figure 4 compares the differences in item parameter estimates and their standard errors between the 3PL and the 3PL CCTRT models. Such a comparison tells us the magnitude of differences regarding item parameter estimation as a result of not modeling the dual item clustering effects due to both passage dependency and the NW format. Similar to what is observed in Figure 2, item discrimination parameter is slightly overestimated and its standard error somewhat underestimated; item difficulty parameter is virtually unaffected and its standard error slightly underestimated; item pseudo-guessing parameter is slightly overestimated and its standard error virtually unaffected. However, it should be noted that the correlation values between two sets of parameter estimates and their standard errors are all extremely high with the lowest value being 0.968, which is the correlation value for standard errors of item discrimination parameter estimates.

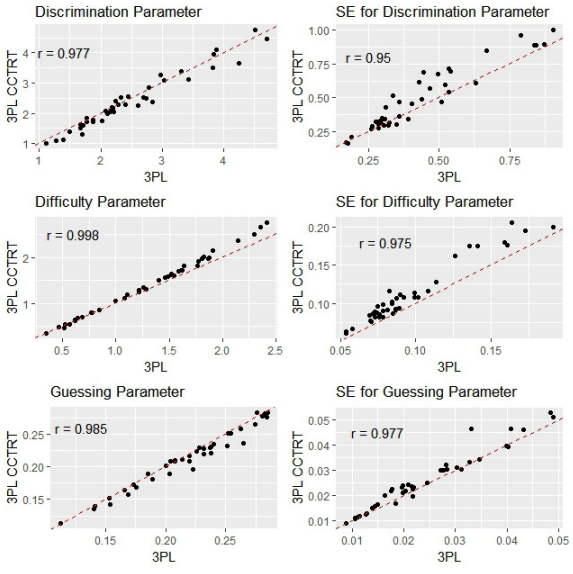


Figure 4. Item Parameter Comparison between the 3PL and 3PL CCTRT Models

DOI: https://doi.org/10.30564/ jiep.v2i1.418

As mentioned previously, the current data set was drawn from a high-stakes test that is used for admission and placement purposes. Consequently, it is critical to check whether the ability estimates and their standard errors are affected to an extent that would result in inaccurate estimation and erroneous classifications when different sources of CIV are not modeled or only partially modeled. Figure 5 provides such visual examinations. The top panel compares the differences in ability parameter estimates and their standard errors between the 3PL and the 3PL TRT models. As can be seen, the ability estimates are virtually unaffected with a correlation value of 0.998. The standard errors also seem to have a very high correlation value of 0.995, although it appears that standard errors of ability estimates from some examinees are slightly underestimated. The middle panel in Figure 5 compares the differences in estimated ability parameters and their standard errors between the 3PL TRT and the 3PL CCTRT models. Similar to what has been found in Figure 3 that item parameters are virtually the same between these two models, ability estimates and their standard errors are almost the same. The bottom panel in Figure 4 compares the differences in estimated ability parameters and their standard errors between the 3PL and the 3PL CCTRT models. Similar to the comparison between the 3PL and the 3PL TRT models, the ability estimates are virtually unaffected while some of their standard errors are slightly underestimated if item clustering effects due to the passage effect and the wording effect are not modeled.

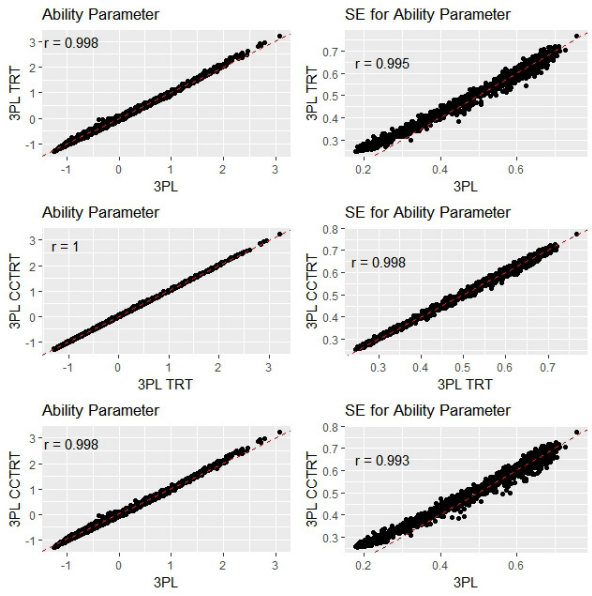


Figure 5. Ability Parameter Comparison between the Three Models

6. Discussion and Conclusions

The NW items have been extensively studied in the context of psychological instrument, especially through the lens of factor analysis to investigate whether they form a separate factor due to negative wording effect. In the educational testing literature, however, most studies focus on how the NW format affects item difficulty, and endeavors from the factor analysis perspective are scant, if not nonexistent. This study was intended to fill the gap in literature by investigating whether the NW items in a high-stakes English reading comprehension test form a separate factor and if yes, whether failure to model this particular factor leads to different item and ability parameter estimates.

Results indicate that similar to the findings in psychological instruments, the NW items in the current test does form a separate factor, of which the variance estimate is 0.09 with a 95% credible interval not covering zero. Note that the latent construct (English proficiency) that the current test is designed to measure is constrained to have a variance of one (for model identification purposes), which means that variance of the factor formed by NW items is less than one tenth of that of English proficiency. In simulation studies using the testlet model as a generating model, the variance of the testlet factor was usually generated to be 0.25 to represent small testlet effect (e.g., Li & Lissitz, 2012).[36] In this regard, it is reasonable to conclude that the NW factor has a small testlet effect.

We also investigated whether not modeling the NW factor leads to different parameter estimates. When the passage factor is modeled, not modeling the NW factor virtually makes no difference for the item and ability parameter estimates: the item and ability parameter estimates and their corresponding standard errors from the 3PL TRT and the 3PL CCTRT models have correlation values all greater than 0.99. If neither the passage factor nor the NW factor is modeled, the differences between the item and ability parameter estimates and their corresponding standard errors from the 3PL and the 3PL CCTRT models seem to be slightly bigger with all correlation values greater than 0.97, although we doubt that such differences would lead to any practical significance. The comparison of the parameter estimates between the 3PL and the 3PL TRT models also indicate that such differences between the 3PL and the 3PL CCTRT models are mainly due to the fact the passage factor is not modeled in the 3PL model.

The variance of the NW factor is small enough to be negligible. Consequently, the ability parameter and item parameter estimates are not noticeably affected. This finding is inconsistent with Wang, Chen, and Jin's finding

 DOI: https://doi.org/10.30564/ jiep.v2i1.418

(2015)[58] that for some of the NW items, the wording effect is large. Such a difference, we believe, should be attributed to the fact that the negative words used in the seven NW items in our data are both capitalized and bolded, while in their study they used a subscale of reading attitude assessment in Program for International Student Assessment (PISA) 2009 and two scales of math and science attitude assessment in Trends in International Mathematics and Science Study (TIMSS) 2011, in which the negative words within in NW items are not emphasized. Another possible cause is that among the seven NW items in our study, only two have negative words at the stem level and the other five at the option level (only one of the five items has the NW option as the correct answer). We believe that the negative words at the stem level has a more pronounced effect than at the option level, since misreading the negation at the stem level is more likely to result in an incorrect answer than at the option level. The results corroborate Haladyna's suggestion (2004)[22] that the negative words in a NW item need to be emphasized in that although the NW items with highlighted negative words introduce CIV into the current test, the magnitude of CIV is too small to cause differences of practical significance. We suspect that if the negative words in those NW items were not highlighted, the magnitude of the NW factor would be greater. However, to test such a hypothesis would require a real data set with such NW items, which are difficult, if not impossible, to find due to the popularity of Haladyna's advice.

We believe the CCTRT model can be a valuable validation tool in scenarios where more than one item clustering effect is expected to exist. Such scenarios may be common with educational testing data. For example, Baghaei and Aryadoust (2015)[5] analyzed a dataset drawn from responses to an English listening comprehension test consisting of 40 items with multiple item formats that fall under four listening passages. Suspecting that item format may constitute a source of CIV, they used a Rasch TRT model to account for such an item clustering effect and found that the testlet variance estimate for some format was large. Since their modeling approach does not account for the possibility that passage effect may form another source of CIV, the Rasch CCTRT model, we argue, is a more suitable model for their data that can simultaneous model both sources of CIV and hence produce more accurate parameter estimates. We recommend that in scenarios like this, a sensitivity analysis should be conducted to see whether the potential model improvement in the CCTRT model translates into differences in parameter estimates of practical significance. If not, we should proceed with the more parsimonious model despite its inferior model fit.

As shown previously, despite the fact that the 3PL TRT and 3PL CCTRT models have better model fit than the 3PL model, the three models lead to item and ability estimates that are highly similar and the most parsimonious one, the 3PL model, should be chosen as the ideal IRT model that combines model parsimony and practical utility. In this regard, this study can be regarded as a validation attempt to ascertain the underlying dimensionality for the data set used in the current analysis. Such an approach is similar to the bifactor approach to determining dimensionality advocated by Reise, Morizot, and Hays (2007),[46] who argue that the bifactor model always fits better than a one-factor model and hence, it is more informative to compare parameter estimates from the two models to see whether there is any difference of practical significance (if there is no noticeable difference between the two sets of parameters, then practical unidimensionality is established). Similarly, Luo and Al-Harbi (2016)[39] showed that when traditional dimensionality detection methods disagree, the bifactor approach can inform whether it makes any practical difference to proceed with the unidimensionality assumption.

To conclude, in this study we used a 3PL CCTRT model to investigate whether the NW items within in reading comprehension passages warrant modeling, and found that with the negative words highlighted, those NW items introduced CIV of negligible magnitude, a finding which supports the recommendation that if NW items have to be used, the negative words within should be accentuated. However, it should be emphasized that despite its small magnitude, the NW items do introduce into the current test CIV, which, regardless of its practical significance, still constitutes a threat to test validity. In addition, the current study used one single data set drawn from an English proficient reading comprehension test, and it is unknown whether the current findings are generalizable to the NW items used in other data sets or other tests that may or may not measure language proficiency. Consequently, we would like to recommend the judicious use of NW items in educational tests and reiterate the necessity for highlighting the negative words in cases where NW items have to be used.

Reference

[1] Abedi, J. (2006). Language issues in item development. In S. M. Downing & T. M. *Haladyna (Eds.), Handbook of test development*. Mahwah, NJ: Lawrence Erlbaum Associates.

[2] Ackerman, T. A. (1987). *The robustness of LOGIST and BILOG IRT estimation programs to violations*

  DOI: https://doi.org/10.30564/ jiep.v2i1.418

of local independence. ACT Research Report Series, 87-14. Iowa City, IA: American College Testing.

[3] Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In B. N. Petrov & F. Csaki (Eds.), *Second International Symposium on Information Theory*, (pp. 267-281). Budapest, Hungary: Akademiai Kiado.

[4] Anderson, D. R. (2008). *Model based inference in the life sciences: A primer on evidence*. New York, NY: Springer.

[5] Baghaei, P., & Aryadoust, V. (2015). Modeling local item dependence due to common test format with a multidimensional Rasch model. *International Journal of Testing*, 15(1), 71-87.

[6] Bradlow, E. T., Wainer, H., & Wang, X. (1999). A Bayesian random effects model for testlets. *Psychometrika*, 64(2), 153. doi:10.1007/bf02294533

[7] Braeken, J., Tuerlinckx, F., & De Boeck, P. (2007). Copula Functions for Residual Dependency. *Psychometrika*, 72(3), 393. doi:10.1007/S11336-007-9005-4

[8] Caldwell, D. J., & Pate, A. N. (2013). Effects of question formats on student and item performance. *American journal of pharmaceutical education*, 77(4), 71.

[9] Casler, L. (1983). *Emphasizing the negative: A note on the "not" in multiple-choice questions*. Paper presented at the meeting of the American Psychological Association.

[10] Cassels, J. R. T., & Johnstone, A. H. (1984). The effect of language on student performance on multiple-choice tests in chemistry. *Journal of Chemical Education*, 61, 613-615.

[11] Chen, W. H., & Thissen, D. (1997). Local Dependence Indexes for Item Pairs Using Item Response Theory. *Journal of Educational and Behavioral Statistics*, 22(3), 265. doi:10.3102/10769986022003265

[12] Chessa, A. G., & Holleman, B. C. (2007). Answering attitudinal questions: Modelling the response process underlying contrastive questions. *Applied Cognitive Psychology*, 21,203-225. doi:10.1002/acp.1337

[13] Congdon, P. (2003). *Applied Bayesian modelling*. New York, NY: Wiley.

[14] Deemer, S. A., & Minke, K. M. (1999). An investigation of the factor structure of the teacher efficacy scale. *The Journal of Educational Research*, 93(1), 3-10.

[15] DeMars, C. E. (2006). Application of the Bi‐Factor multidimensional item response theory model to Testlet‐Based tests. *Journal of Educational Measurement*, 43(2), 145-168.

[16] Downing, S. M. (2005). The effects of violating standard item writing principles on tests and students: the consequences of using flawed test items on achievement examinations in medical education. *Advances in health sciences education*, 10(2), 133-143.

[17] Downing, S. M., Dawson-Saunders, B., Case, S. M., & Powell, R. D. (1991). *The psychometric effects of negative stems, unfocused questions, and heterogeneous options on NBME Part I and Part II item characteristics*. Paper presented at the annual meeting of the National Council on Measurement in Education, Chicago.

[18] Dudycha, A. L., & Carpenter, J. B. (1973). Effects of item formats on item discrimination and difficulty. *Journal of Applied Psychology*, 58, 116-121.

[19] Gelman, A., & Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences. *Statistical Science*, 457-472.

[20] Gelman, A., Carlin, J. B., Stern, H. S., & Rubin, D. B. (2014). *Bayesian data analysis*. Boca Raton, FL, USA: Chapman & Hall/CRC.

[21] Gitchel, W. D., Roessler, R. T., & Turner, R. C. (2011). Gender effect according to item directionality on the perceived stress scale for adults with multiple sclerosis. *Rehabilitation Counseling Bulletin*, 55(1), 20-28.

[22] Haladyna, T. M. (2004). *Developing and validating multiple-choice test items* (3rd ed.). Mahwah, NJ: Erlbaum.

[23] Haladyna, T. M., & Downing, S. M. (1989a). A taxonomy of multiple-choice item-writing rules. *Applied Measurement in Education*, 2(1), 37-50.

[24] Haladyna, T. M., & Downing, S. M. (1989b). Validity of a taxonomy of multiple-choice item-writing rules. *Applied Measurement in Education*, 2(1), 51-78.

[25] Haladyna, T. M., & Downing, S. M. (2004). Construct‐irrelevant variance in high‐stakes testing. *Educational Measurement: Issues and Practice*, 23(1), 17-27.

[26] Haladyna, T. M., Downing, S. M., & Rodriguez, M. C. (2002). A review of multiple-choice item-writing guidelines for classroom assessment. *Applied measurement in education*, 15(3), 309-333.

[27] Harasym, P. H., Price, P. G., Brant, R., Violato, C., & Lorscheider, F. L. (1992). Evaluation of negation in stems of multiple-choice items. *Evaluation and the Health Professions*, 15, 198-220.

[28] Hoskens, M., & De Boeck, P. (1997). A parametric model for local dependence among test items. *Psychological methods*, 2(3), 261.

 DOI: https://doi.org/10.30564/ jiep.v2i1.418

[29] Jiao, H., Wang, S., Wan, L., & Lu, R. (2009, April). *Investigation of local item dependence in scenario-based science assessment*. Paper presented at the Annual Meeting of the American Educational Research Association, San Diego, CA.

[30] Jiao, H., Kamata, A., & Xie, C. (2015). A multilevel cross-classified testlet model for complex item and person clustering in item response modeling. In J. Harring, L. Stapleton, & S. Beretvas (Eds.), *Advances in multilevel modeling for educational research: Addressing practical issues found in real-world applications*. Charlotte, NC: Information Age Publishing.

[31] Johnstone, A. H. (1983). Training teachers to be aware of the student learning difficulties. In P. Tamir, A. Hofstein A, & M. Ben Peretz (Eds.), *Preservice and Inservice Education of Science Teachers*. Rehovot (Isreal) – Philadelphia (USA): Balaban International Science Services.

[32] Kamata, A. (2001). Item Analysis by the Hierarchical Generalized Linear Model. *Journal of Educational Measurement*, 38(1), 79. doi:10.1111/j.1745-3984.2001.tb01117.x

[33] Kamata, A., & Bauer, D. J. (2008). A note on the relation between factor analytic and item response theory models. *Structural Equation Modeling*, 15(1), 136-153.

[34] Kieruj, N. D., & Moors, G. (2013). Response style behavior: Question format dependent or personal style. *Quality & Quantity*, 47, 193-211. doi:10.1007/s11135-011-9511-4

[35] Li, Y., Bolt, D. M., & Fu, J. (2006). A comparison of alternative models for testlets. *Applied Psychological Measurement*, 30(1), 3-21.

[36] Li, Y., & Lissitz, R. W. (2012). Exploring the full-information bifactor model in vertical scaling with construct shift. *Applied Psychological Measurement*, 36(1), 3-20.

[37] Lindwall, M., Barkoukis, V., Grano, C., Lucidi, F., & Raudsepp, L. (2012). Method effects:The problem with negatively versus positively keyed items. Journal of Personality Assessment, 94, 196-204. doi: 10.1080/00223891.2011.645936

[38] Luo, Y. (2019). LOO and WAIC as model selection methods for polytomous items. Psychological Test and Assessment Modeling, 61(2), 161-185.

[39] Luo, Y., & Al-Harbi, K. (2016). The Utility of the Bifactor Method for Unidimensionality Assessment When Other Methods Disagree. *SAGE Open*, 6(4).

[40] Luo, Y., & Jiao, H. (2018). Using the Stan program for Bayesian item response theory. Educational and psychological measurement, 78(3), 384-408.

[41] Luo, Y., & Liang, X. (2019). Simultaneously Modeling Differential Testlet Functioning and Differential Item Functioning: Addressing Variance Heterogeneity with a Multigroup One-Parameter Testlet Model. Measurement: Interdisciplinary Research and Perspectives, 17(2), 93-105.

[42] Magazine, S. L., Williams, L. J., & Williams, M. L. (1996). A confirmatory factor analysis examination of reverse coding effects in Meyer and Allen's affective and continuance commitment scales. *Educational and Psychological Measurement*, 56(2), 241-250.

[43] Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement*. New York: American Council on Education and Macmillan

[44] Rachor, R. E., & Gray, G. T. (1996, April). *Must all stems be green? A study of two guidelines for writing multiple choice stems*. Paper presented at the annual meeting of the American Educational Research Association, New York.

[45] Reckase, M. D. (2009). *Multidimensional item response theory*. New York, NY: Springer.

[46] Reise, S. P., Morizot, J., & Hays, R. D. (2007). The role of the bifactor model in resolving dimensionality issues in health outcomes measures. *Quality of Life Research*, 16(1), 19-31.

[47] Rijmen, F. (2010). Formal Relations and an Empirical Comparison among the Bi‐Factor, the Testlet, and a Second‐Order Multidimensional IRT Model. *Journal of Educational Measurement*, 47(3), 361-372.

[48] Rosenbaum, P. R. (1988). Items bundles. *Psychometrika*, 53(3), 349. doi:10.1007/bf02294217

[49] Roszkowski, M. J., & Soven, M. (2010). Shifting gears: Consequences of including two negatively worded items in the middle of a positively worded questionnaire. *Assessment & Evaluation in Higher Education, 35*, 117-134. doi:10.1080/02602930802618344

[50] Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, 6(2), 461-464.

[51] Spiegelhalter, D. J., Best, N. G., Carlin, B. P., & Van Der Linde, A. (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society: Series B (Statistical Methodology), 64*(4), 583-639.

[52] Takane, Y., & De Leeuw, J. (1987). On the relationship between item response theory and factor analysis of discretized variables. *Psychometrika*, 52(3), 393-408.

[53] Tamir, P. (1991). Multiple choice items: How to gain the most out of them. *Biochemical Education*, 19(4), 188-192.

DOI: https://doi.org/10.30564/ jiep.v2i1.418

[54] Tamir, P. (1993). Positive and negative multiple choice items: How different are they? *Studies in Educational Evaluation*, 19, 311-325.

[55] Terranova, C. (1969). The effects of negative stems in multiple-choice test items. Dissertation Abstracts International, 30, 2390A.

[56] van Sonderen, E., Sanderman, R., & Coyne, J. C. (2013). Ineffectiveness of reverse wording of questionnaire items: Let's learn from cows in the rain. *PLoS ONE, 8*(7), 1–7. http://doi.org/10.1371/journal.pone.0068967

[57] Wainer, H., Bradlow, E. T., & Wang, X. (2007). *Testlet response theory and its applications*. Cambridge: Cambridge University Press.

[58] Wang, W. C., Chen, H. F., & Jin, K. Y. (2015). Item response theory models for wording effects in mixed-format scales. *Educational and Psychological Measurement*, 75(1), 157-178.

[59] Xie, C. (2014). *Cross-classified modeling of dual local item dependence* (Unpublished doctoral dissertation). University of Maryland, College Park, MD.

[60] Yen, W. M. (1993). Scaling performance assessments: Strategies for managing local item dependence. *Journal of educational measurement*, 30(3), 187-213.

[61] Zhang, B. (2010). Assessing the accuracy and consistency of language proficiency classification under competing measurement models. *Language Testing*, 27(1), 119-140.

[62] Zhang, X., Noor, R., & Savalei, V. (2016). Examining the effect of reverse worded items on the factor structure of the need for cognition scale. *PLoS ONE*, 11(6), 1–15. http://doi.org/10.1371/journal.pone.0157795

 DOI: https://doi.org/10.30564/ jiep.v2i1.418