Journal of Business and Psychology (2020) 35:697–712
https://doi.org/10.1007/s10869-019-09653-8

ORIGINAL PAPER

A Psychometric Assessment of OCB: Clarifying the Distinction Between OCB and CWB and Developing a Revised OCB Measure

Alexandra A. Henderson[1] · Garett C. Foster[2] · Russell A. Matthews[3] · Michael J. Zickar[4]

Published online: 29 October 2019

Abstract

This study was performed to (1) assess the appropriateness of using negatively worded items in organizational citizenship behavior (OCB) scales, (2) psychometrically demonstrate the construct distinctness of OCB and counterproductive work behavior (CWB), and (3) report on a revised, short-form OCB scale. Leveraging classical test theory (CTT) and item response theory (IRT), we demonstrate that the negatively worded items from a popular OCB scale (Williams and Anderson 1991) do not measure OCB, but rather a unique construct (CWB). CTT analyses (factor analyses) indicate that the negatively worded items load onto a unique factor when the scale is analyzed on its own and load onto a CWB factor when the scale is analyzed with a CWB scale. Additionally, IRT analyses indicate that the negatively worded items exhibit lower discrimination parameters and higher levels of local independence than the positively worded items, and similar discrimination parameters and levels of local independence as the CWB items. In turn, IRT analyses were used to identify the best items from the OCB scale to create a revised, short-form scale. The short-form scale showed comparable or improved convergent and discriminant validity and internal consistency reliability, as well as similar patterns of psychometric information yielded from IRT analyses, compared to the original scale. In short, the revised measure better aligns with conceptual definitions of OCB, demonstrates acceptable psychometric characteristics, and, given its reduced length, is of more practical value to researchers wishing to assess this construct within different types of research designs (e.g., longitudinal, multi-source).

Keywords Organizational citizenship behavior · Measurement · Item response theory · Scale

✉ Alexandra A. Henderson
Alexandra.Henderson@zu.ac.ae

Garett C. Foster
Fostergc@umsl.edu

Russell A. Matthews
Ramatthews2@ua.edu

Michael J. Zickar
Mzickar@bgsu.edu

[1] College of Business, Zayed University, PO Box 144534, Abu Dhabi, UAE

[2] Department of Psychology, University of Missouri – St. Louis, One University Boulevard, 423 Sadler Hall, St. Louis, MO 63121-4400, USA

[3] Department of Management, University of Alabama, 361 Stadium Dr, Tuscaloosa, AL 35487, USA

[4] Department of Psychology, Bowling Green State University, Bowling Green, OH 43403, USA

Scholars define organizational citizenship behavior (OCB) as discretionary behavior that promotes the effective functioning of the organization by contributing to the maintenance and enhancement of the social and psychological context that supports task performance (Organ 1988, 1990). The topic of OCB has sustained traction within the organizational literature, as researchers have identified it as a critical dimension of performance (Motowidlo and Kell 2013) that has important consequences for organizations (Podsakoff et al. 2009). Indeed, Podsakoff et al. (2009) found that OCB contributes to increased performance ratings, decreased withdrawal behaviors, increased unit performance, and increased customer satisfaction.

In light of the ongoing interest in OCB, it is important that the construct is measured correctly. Within the OCB literature, there are numerous measures of the construct that range in the number and types of dimensions. However, one of the most frequently used measures was developed by Williams and Anderson (1991), with approximately 6000 citations (Google Scholar). Unlike Organ (1988, 1990), who

conceptualized a five, and later seven, factor models of OCB based on the specific types of behaviors, Williams and Anderson conceptualized a two-factor model based on the target of the behaviors. Specifically, the Williams and Anderson scale is a 14-item measure predicated on the two-factor conceptualization of OCB—(1) *individually directed OCB* (OCB-I) and (2) *organizationally directed* OCB (OCB-O). This conceptualization of OCB is often preferred by researchers, as there are concerns about the distinguishability of some of the organ's factors (e.g., MacKenzie et al. 1991; Podsakoff and MacKenzie 1994). Additionally, the two-factor conceptualization parallels the conceptualization of another related construct, counterproductive work behavior (CWB).

Despite the popularity of the two-factor Williams and Anderson (1991) measure, problems with the scale exist, which may limit the resulting validity of study results that utilize this scale. Jepsen and Rodwell (2006) and Yun et al. (2007)) demonstrated that the Williams and Anderson scale had poor confirmatory factor analysis model fit, particularly the OCB-O dimension. Similarly, other researchers have demonstrated lower internal reliability scores for the OCB-O subscale, compared to the OCB-I subscale (Byrne 2005; Mayer and Gavin 2005). Despite these potential limitations, the Williams and Anderson scale appears to offer good content coverage of the OCB domain and has contributed greatly to advancing our understanding of the nature and role of OCB in the workplace.

It is our belief that the three negatively worded items of the Williams and Anderson (1991) scale, which are all considered OCB-O items, may be the primary cause of the differences in the reliability and validity of the two subscales. The authors oftentimes include negatively worded items as a means of providing an attention check to prevent careless responding. Williams and Anderson followed what was, at the time, best practices in writing negatively worded items; specifically, the items were written to tap into very low levels of the construct via poor behaviors (e.g., "I took undeserved work breaks"), rather than simply writing a positively worded item and adding a negative qualifier, such as "not." However, researchers have since shown that negatively worded items can create additional (i.e., methodological) factors within a scale (e.g., Magazine et al. 1996). Typically, an additional factor resulting from negative wording does not present a theoretical issue, unless the negative wording introduces other potential biases (i.e., confounds the criterion space of the scale; Bandalos 2018).

We contend that the three negatively worded items in the Williams and Anderson (1991) scale may introduce such criterion confounding biases. Specifically, we argue that these negatively worded items form their own unique factor, not because they are negatively worded, but because they are measuring CWB. We also argue that OCB and CWB are distinct constructs, rather than behaviors that lie on opposite ends of the same behavioral continuum—a proposition that has, to date, only received correlational support (Dalal 2005). We intend to provide evidence for these arguments using psychometric analyses, including factor analysis and item response analysis. Expected results would demonstrate that the inclusion of negatively worded items in OCB scales could be problematic, as they may introduce both psychometric and theoretical contamination.

Furthermore, the Williams and Anderson (1991) scale, as well as most other OCB scales, can also be critiqued for its length. The 14-item scale might be perceived as burdensome by many researchers, as the length may lead to excessive participant attrition and survey costs, especially in longitudinal research (Fisher et al. 2016). Indeed, it is critically important to create psychometrically sound short measures, especially as research questions and methods become more complex (Fisher and To 2012).

The Current Study

The purpose of the current study is threefold. First, we seek to demonstrate the inappropriateness of using negatively worded items in OCB scales, by demonstrating that these items are more representative of another construct (CWB). Second, we seek to psychometrically demonstrate the construct distinctness of OCB and CWB, as the current literature has only supported this proposition with correlational evidence (Dalal 2005). Finally, we present a revised, short-form OCB scale that demonstrates improved psychometric properties and construct validity. Such a measure, we suggest, will be particularly advantageous for scholars interested in examining OCB within complex research designs (e.g., longitudinal, multisource) that preclude the use of lengthy measures (e.g., Ford et al. 2018). To achieve these aims, we leverage classical test theory (CTT) and item response theory (IRT).

For the current study, we analyze and revise the Williams and Anderson (1991) scale at the overall scale level and at the subscale level (OCB-I and OCB-O). Although the OCB dimensions are conceptually distinct, researchers have demonstrated that the dimensions are empirically similar, in that they have equivalent relationships with predictors and correlates (LePine et al. 2002; Podsakoff et al. 2009). As such, it could be argued that the distinction between these dimensions does not need to be made in research. However, despite this empirical evidence, these same researchers have cautioned that it may be premature to conclude that OCB-I and OCB-O have the exact same causes and effects (Podsakoff et al. 2009). Indeed, there may be unstudied variables that have different relationships with the two dimensions, which would be masked if only the overall scale is utilized. Furthermore, researchers may feel the need to distinguish between these two

dimensions in practice, as this distinction may provide more conceptual clarity. Thus, we take the middle of the road approach by assessing the scale at the overall and subscale levels to provide researchers with either option depending on their research needs.

Methodological Approach

To examine the proposed phenomena, we strategically leveraged two complimentary methods: classical test theory (CTT) and item response theory (IRT). CTT and its related methods, such as factor analysis, analyze the covariance between items or scales to extract structural relations. This factor solution reflects how many underlying constructs are influencing the items. The appropriateness of the solution yielded by factor analysis is informed by both numerical output (e.g., item loadings) and by substantive interpretability. Conversely, IRT uses raw response data within items to estimate the properties of individual items. One of the assumptions of IRT is known as local independence, or a lack of covariation among items when controlling for levels of the construct being measured. Violations of local independence cause inflation of item and test information estimates (Ip 2010). Local dependencies can also be used as a diagnostic tool by indicating that subsets of items are more related to each other than the factor structure would suppose (Chen and Thissen 1997). As such, local dependence can give insights into the nature of items that may represent a distinct construct or methodological artifact.

Specific then to the current study, within an IRT perspective, we suggest that local dependency will exist for the three negatively worded items within the OCB measure. In turn, within a CTT perspective, we expect that the three negatively worded OCB items will load distinctly on a secondary factor. While some may suggest that this secondary factor is an empirically derived method factor (DiStefano and Motl 2006), as we will discuss in more detail shortly, we expect that this secondary factor (i.e., the negatively worded OCB items) will also account for items found within an established and empirically validated measure of CWB (Bennett and Robinson 2000).

Again, CTT and IRT methods serve complimentary purposes and are most useful when used in conjunction with one another. In general, CTT is better suited for examining how items relate to each other (i.e., their structure) whereas IRT is better for examining and diagnosing the properties of individual items (e.g., item discrimination). For a full discussion and comparison of CTT and IRT, see Zickar and Broadfoot (2009).

Analysis of Negatively Worded OCB Items

Factor analysis was utilized to assess the underlying factors of the Williams and Anderson (1991) scale. As noted previously, while there is empirical reason to treat it as unitary construct (LePine et al. 2002; Podsakoff et al. 2009), the scale is based on the two-dimensional conceptualization of OCB, which includes individually directed OCB (OCB-I) and organizationally directed OCB (OCB-O). As such, we would expect a two-factor model to fit the data well in a confirmatory factor analysis (CFA). However, the Williams and Anderson scale includes three negatively worded OCB-O items, which might be better modeled by a third factor. Indeed, previous research has shown that negatively worded items can create additional factors within a scale when conducting exploratory factor analyses (e.g., Magazine et al. 1996). As such, in replication of previous factor analysis of the Wlliams and Anderson scale (e.g., Yun et al. 2007), we expect a three-factor model—an OCB-I factor, an OCB-O factor, and a negatively worded OCB-O factor—to fit the data better than a two-factor model.

- *Hypothesis 1: In addition to the two OCB factors (OCB-I & OCB-O), modeling the negatively worded OCB-O items as a third factor results in improved model fit.*

Researchers have provided correlational evidence that OCB and CWB may be distinct constructs. For example, researchers have found that OCB and CWB are only moderately correlated ($r = -.32$) and have differential relationships with antecedents, including positive and negative affect (Dalal 2005; Dalal et al. 2009). These researchers argue that if OCB and CWB are at opposite ends of the same spectrum, then they should have strong negative correlations and should have similar (yet opposite) relationships with other constructs. Additionally, Sackett et al. (2006) found that the correlations of interpersonal and organizational facets within each construct are typically larger than the correlations between interpersonal and organizational facets between constructs. Despite this body of evidence, somewhat surprisingly, there has yet to be a more in depth investigation to clarify the construct distinctiveness of OCB and CWB *at a measurement level*. Conclusions in research are only as reliable and valid as the data from which they are drawn (Borsboom 2006; Thorndike 1904), and as OCB and CWB research matures and begins to deal with more fine-grained and nuanced questions, accurate measurement becomes even more important.

As with the Williams and Anderson (1991) OCB scale, the Bennett and Robinson (2000) CWB scale is based on a two-factor conceptualization, which includes individually directed CWB (CWB-I) and organizationally directed CWB (CWB-O). If OCB and CWB are unique constructs, as suggested by

previous research (e.g., Dalal 2005), then the four-factor model specification—OCB-I, OCB-O, CWB-I, and CWB-O—should show mediocre model fit when the two scales are analyzed together, as the negatively worded OCB items should load poorly on the OCB-O factor. Furthermore, based on the similarities between the negatively worded OCB items and CWB items, as noted previously, we would expect the negatively worded OCB items to more appropriately load onto the CWB-O factor. Thus, an alternative four-factor model—OCB-I, OCB-O without negative worded items, CWB-I, and CWB-O with the negatively worded OCB-O items—should show superior model fit.

- *Hypothesis 2: When analyzed with a two-factor CWB scale (CWB-I and CWB-O), modeling the negatively worded OCB-O items with the CWB-O factor results in improved model fit.*

Although factor analysis is often used to identify different factors, it does not provide concrete support for construct discrimination (i.e., construct distinctness). Indeed, in the context of the current program of research, if four factors are identified, this may still be the result of methods factors (positive and negative wording), rather than the presence of distinct constructs. In other words, the factors may simply represent interpersonally and organizationally directed behaviors that are either positively or negatively worded.

As noted earlier, IRT is a useful tool for identifying items that do not behave properly and therefore do not accurately measure their intended construct. As such, IRT may be a useful method for identifying whether or not the additional factors in a factor analysis are the result of methods or distinct constructs. If OCB and CWB are the same construct, then negatively worded items will still differentiate between respondents who have low levels of the construct and therefore have good item discrimination parameters; these items will also not exhibit greater statistical similarity to one another than they do to the rest of the items after accounting for levels of the trait (i.e., they will not exhibit issues related to local independence). However, if OCB and CWB are distinct constructs, and the negatively worded OCB items are actually measuring CWB (instead of OCB), then we expect the three negatively worded OCB items to exhibit lower discrimination parameters, since the items are unable to differentiate along the latent trait being measured, and higher levels of local independence, than the positively worded OCB items.

- *Hypothesis 3: The negatively worded OCB items exhibit inflated levels of local dependence.*
- *Hypothesis 4: The negatively worded OCB items exhibit lower discrimination parameters.*

A Revised, Short-Form OCB Measure

In addition to the three negatively worded items, there may be other items in the Williams and Anderson (1991) scale that demonstrate weak discrimination parameters, which would mean that the items have little utility in providing information about a respondent's level of OCB behaviors. Removing these poor performing items would have the benefit of shortening the scale while still maintaining, overall, a measure that demonstrates strong psychometric characteristics. Indeed, shorter scales are becoming more appealing due to their ability to shorten survey length, thus reducing participant attrition and survey costs (Fisher et al. 2016). In order to balance scale reliability (longer scales have higher reliabilities), SEM requirements (a suggested minimum of three items per latent factor), and length (shorter scales have lower attrition and cost), we seek to retain the six items (three OCB-I and three OCB-O) with good discrimination parameters ($a > 0.80$) and locations spanning the same range of the latent continuum as the full scale, which will yield comparable test information. Again, as noted earlier, we seek to retain enough OCB-I and OCB-O items to create usable subscales, should researchers want to measure a specific dimension.

Although the reduction from 14 items to six items may not seem inherently substantial, the eight additional items puts participants under additional minutes of unnecessary cognitive demands that could substantially increase participant attrition, especially in repetitive surveys (e.g., daily diary surveys). If the OCB scale (as well as other scales in the survey) can be reduced by approximately half (while maintaining the reliability and validity of the original scale), then this would be extremely valuable for reducing participant attrition in these types of situations. Furthermore, as noted by Lapierre et al. (2018), it is commonplace when conducting research in an organizational context for organizational stakeholders to require scholars to administer as few items as possible (to minimize the amount of time employees are "off the line"). As a result, scholars often trim existing scales to fit length requirements, and unfortunately, the item trimming process is done without sound empirical reasoning (Stanton et al. 2002). The resulting short-form measure reported here serves as a proactive response to the need to apply short measures in an organizational data collection context.

In a review of scale reduction techniques, Stanton et al. (2002) highlighted the importance of IRT as a tool for judging the loss of information due to eliminating items. We will also seek to ensure that this revised, short-form scale has comparable or improved convergent and discriminant validity and

internal consistency reliability, as well as similar patterns of psychometric information yielded from IRT analyses.

Methods

Two large, heterogeneous field samples of working adults were used to test our study hypotheses and to develop a revised, short-form OCB scale. In sample 1, we tested hypotheses 1 (i.e., that the negatively worded OCB-O items constitute an separate factor). In sample 2, we replicated and extended these findings to examine if these items more accurately reflect and relate to CWB (hypothesis 2). In turn, both samples were used to assess item discrimination and local dependence (hypotheses 3 and 4) and to determine the best items to retain for a short-form scale. Finally, in sample 2, we compared the convergent and discriminant validity of the original and short-form scales, comparing their relationships with related constructs (Bandalos 2018).

Sample 1

Participants and Procedure

Participants were recruited via a peer nomination sampling method. Specifically, faculty at a variety of higher education institutions was asked to provide the survey to their students. In turn, students were requested to recruit working adults to complete the survey (students were not allowed to participate). Students received nominal extra credit for their efforts.

The survey had 1157 initial participants; however, 134 were removed for working less than 24 h, and 162 were removed for not having a supervisor (a condition from the larger study). To ensure effortful responding, we reviewed survey response times. There was no immediate indication that any respondents completed the survey in an unreasonably short period of time (i.e., respondents took at least 5 min to complete the survey, with most averaging 10–20 min). However, 32 participants were removed for having incomplete data (i.e., they did not finish the entire survey). Collectively then, the final sample consisted of 829 participants, was 35.8% male, with a mean age of 37.5 (SD = 12.9), and worked approximately 42.7 h per week (SD = 8.9).

Measures

OCB was measured using the original Williams and Anderson (1991) scale. Conceptually, the scale measures seven OCB-I behaviors and seven OCB-O behaviors. Items are reported in Table 1. Participants were asked to indicate how frequently they engaged in the behaviors over the previous month on a five-point scale ranging from *1 = never* to *5 = many times*. The internal consistency reliability of the overall OCB scale was $\alpha = 0.76$, whereas the internal consistency of OCB-I dimension was $\alpha = 0.81$ and the OCB-O dimension was $\alpha = 0.49$ (positively worded items only was $\alpha = 0.55$).

Sample 2

Participants and Procedure

Participants were recruited through Amazon's Mechanical Turk. To ensure data quality, only US participants with a 96% approval rate (i.e., 96% of their prior tasks had been approved) and who had previously completed at least 1000 tasks were allowed to participate. Additional inclusion criteria consisted of (1) at least 24 h of work per week and (2) working for the same employer for at least 1 month. Participants who met the inclusion criteria were requested to complete the larger study questionnaire across five separate time points with a 1-month lag between each assessment. Validation questions (e.g., "In order to show that you are carefully reading the interview questions, please leave this item blank") were also used to ensure effortful responding. The time 1 survey had 987 respondents; of these, only 924 respondents were retained (based on inclusion criteria and effortful responding). For the purpose of the current study, only these initial responses (i.e., time 1 respondents) were included in our analyses.

The sample was 52.3% male, with a mean age of 35.1 (SD = 10.1), and worked approximately 40.5 h per week (SD = 7.2). Approximately 45% of participants reported working in management, professional, and related occupations; 26% in sales and office occupations; 17% in service occupations; 8% in production, transportation, and material moving occupations; and 4% in natural resources, construction, and maintenance occupations. These figures align with recent BLS data (Bureau of Labor Statistics 2017), with slightly higher representation of management and lower representation of production and transportation, which supports the generalizability of this sample to the US working population.

Measures

Organizational Citizenship Behavior *OCB* was again measured with the Williams and Anderson (1991) scale. The internal consistency reliability of the overall OCB scale was $\alpha = 0.80$, whereas the internal consistency of OCB-I dimension was $\alpha = 0.85$ and the OCB-O dimension was $\alpha = 0.58$ (positively worded items only was also $\alpha = 0.58$).

Counterproductive Work Behavior *CWB* was measured using an adapted version of the Bennett and Robinson (2000) scale (Matthews and Ritter 2016; study 3). The adapted scale measures four CWB-I behaviors and four CWB-O behaviors. Sample items include "Said something hurtful to someone at

Table 1 Williams and Anderson (1991) OCB scale

Item 1	I helped others who have been absent.
Item 2	I helped others who have heavy workloads.
Item 3	I helped orient new people even though it is not required.
Item 4	I assisted my supervisor with his/her work (when not asked).
Item 5	I took time to listen to co-workers' problems and worries.
Item 6	I took a personal interest in other employees.
Item 7	I passed along information to co-workers.
Item 8	My attendance at work was above the norm.
Item 9	I gave advance notice when I was unable to come to work.
Item 10	*I took undeserved work breaks.*
Item 11	*A great deal of my time was spent on personal phone/email/other communications.*
Item 12	*I complained about insignificant things at work.*
Item 13	I conserved and protected organizational property.
Item 14	I adhered to informal rules devised to maintain order.

Items 1–7 = interpersonally directed OCB (OCB-I); items 8–14 = organizationally directed OCB (OCB-O); italicized items are negatively worded

work" and "Taken an additional or longer break than is acceptable at your workplace." Participants were asked to indicate how frequently they engaged in the behaviors over the previous month on a five-point scale ranging from *1 = never* to *5 = many times*. The internal consistency reliability of the overall CWB scale was $\alpha = 0.81$, while the internal consistency of CWB-I dimension was $\alpha = 0.81$ and the CWB-O dimension was $\alpha = 0.79$.

Construct Validity Measures To examine construct validity of the short-form OCB measure, data were collected on three theoretically and empirically established OCB antecedents (Bolino and Turnley 2005; Chen and Chiu 2009; Schappe 1998). *Affective commitment* ($\alpha = 0.94$; e.g., "I feel emotionally attached to my organization") was assessed with the three highest loading items from Griffin et al. (2007). *Autonomy* ($\alpha = 0.83$, e.g., "I have the freedom to decide what I do on my job") was assessed with a four-item measure (Thompson and Prottas 2006). *Workload* ($\alpha = 0.84$; e.g., "How often does your job require you to work very hard?") was assessed with a five-item measure (Spector and Jex 1998).

Results

Factor Structure Analyses We first conducted a CFA on the Williams and Anderson (1991) OCB scale from sample 1 using the lavaan package (Rosseel 2012) in R (R Core Team 2017). We tested two separate models to examine hypothesis 1. First, we fit the original two-factor structure, allowing the factors to correlate. Next, we fit a three-factor solution in which the negatively worded OCB-O items loaded onto their own factor; no other changes to the factor structure were made. The two-factor structure demonstrated poor fit, while the three-factor structure demonstrated improved fit ($\Delta\chi^2(2) = 159.30$, $p < .001$), as well as improved CFI, SRMR, and RMSEA (Table 2), which meet minimum conventions for moderate fit. These results support hypothesis 1: negatively worded OCB-O items are best modeled as their own separate factor.

Next, we replicated these findings with sample 2. Again, two separate models were tested: the original two-factor structure and the three-factor structure supported by data from sample 1 (Table 2). The three-factor structure showed improved fit ($\Delta\chi^2$ (2) = 274.07, $p < .001$) and improved relative fit statistics, providing further support that the three-factor model is a more appropriate representation of the data.

We then conducted a final set of CFAs on both the OCB and CWB scales with the sample 2 data. As reported in Table 3, when items were loaded onto their original conceptual scale factor for which they were developed, model fit was poor, with relative fit statistics falling below conventional levels (e.g., CFI < 0.90). However, in support of hypothesis 2, the alternative model, which placed the three negatively worded OCB-O items onto the CWB-O factor, showed improvement, with all relative fit statistics meeting minimum cutoffs. It should be noted that because these two models involved the same number of items, latent constructs, and factor loadings (only the specific loadings changed), they are not nested, so a direct test of change in model fit cannot be conducted. In lieu of this, Table 3 also reports BIC, which can be used to compare non-nested models (Kline 2016). In further support of hypothesis 2, the alternative model yielded a lower BIC (BIC = 45,335.35) compared to the published model (BIC = 45,770.07), supporting the likelihood of this model over the published model.

Collectively then, the CTT analyses provided support for the argument that the Williams and Anderson (1991) OCB

Table 2 Confirmatory factor analysis of the two- and three-factor models of the Williams and Anderson (1991) OCB scale

	Sample 1		Sample 2	
	2-factor	3-factor	2-factor	3-factor
χ^2(df)	426.33 (76)***	267.03 (74)***	569.01 (76)***	294.936 (74)***
CFI	0.843	0.914	0.847	0.931
SRMR	0.067	0.052	0.075	0.044
RMSEA 90% CI	(0.070, 0.084)	(0.051, 0.066)	(0.080, 0.093)	(0.052, 0.066)
BIC	45,532.81	45,098.08	34,072.98	33,812.45

CFI = confirmatory fit index, SRMR = standardized root mean square residual, RMSEA = root mean square error of approximation, 90% CI = 90% confidence interval, BIC = Bayesian information criterion

***$p < .001$

scale does not operate as originally conceptualized, likely due to the inclusion of the three negatively worded items. The moderate fit of the hypothesized alternative model, however, suggests that deeper investigation is warranted, and IRT may provide useful information to this end.

Item Level Analyses The sample 1 OCB items were also assessed using IRT in IRTPRO 2.1 (Cai et al. 2011) to better understand the measurement properties of individual items. Items were modeled using Samejima's (1969) graded response model (GRM), the most commonly used IRT model for polytomous response data in the organizational sciences (Foster et al. 2017). The GRM contains two types of item parameters: a discrimination parameter (a) that indicates how well the item differentiates between similar people, and a set of threshold parameters (b) that indicate how much of the trait a respondent needs to have in order to select the next highest level of endorsement (e.g., how much OCB is required to select *Strongly Agree* to an item instead of *Agree*). For an item with k response options, there are $k - 1$ threshold parameters; as such, the set of threshold parameters for the current items, which have five response options, contain four threshold parameters for each item.

The scale meets the assumption of sufficient unidimensionality because the first factor explains most of the variance among the items (i.e., the first factor is the prepotent factor according to the EFA; Drasgow and Parsons 1983). Model fit was assessed using a chi-square to degrees of freedom ratio, where ratios of 3.0 or less are indicative of acceptable fit, and ratios of less than 2.0 are indicative of excellent fit (Drasgow and Hulin 1990); the average χ^2/df ratio was 1.32 and no item had a ratio greater than 2.0 (see Table 4). Based on these fit criteria, all items were well fit by the GRM. The parameter estimates from sample 1 for all items are reported in Table 4.

Chen and Thissen (1997) proposed a standardized local dependence (LD) χ^2 statistic for assessing violations of the local independence assumption and suggest that values exceeding 10.0 should have their item content investigated for substantive similarities that could cause the mathematical indications of local dependence. In support of hypothesis 3, values for the negatively worded items greatly exceeded 10.0 ($LD_{10,11} = 27.7$, $LD_{10,12} = 28.1$, $LD_{11,12} = 16.1$), indicating excessive covariation among the items beyond what is expected by the model. Unexpectedly, items 5 and 6 also exhibited large covariation ($LD_{5,6} = 24.4$), suggesting these items also violate local independence. Examination of these items reveals that the content of both relates to personal discussions that do not necessarily relate to work.

Examining item parameter estimates further revealed that the three negatively worded items did not differentiate among people on the latent OCB construct, as evidenced by the low discrimination parameters ($a < 0.35$), which fall below recommended cutoffs for discrimination parameters (Embretson and Reise 2000; Zickar 2012), providing support for hypothesis 4 (see Table 4). Because of the low discrimination parameters, the items give little information at the item level across any level of the construct being measured, which suggests one of two things: (1) they are either poor items in general or (2) they are poor items specifically for the latent OCB construct. In either case, our IRT analysis suggests that these three items

Table 3 Confirmatory factor analysis of the published and alternative models of the Williams and Anderson (1991) and Bennett & Robinson (2004) scales

	Published	Alternative
χ^2(df)	1170.50 (203)***	735.78 (203)***
CFI	0.842	0.913
SRMR	0.092	0.050
RMSEA 90% CI	(0.071, 0.079)	(0.051, 0.060)
BIC	45,770.07	45,335.35

Published model has four factors (OCB-I, OCB-O, CWB-I, CWB-O) with items loading as described in original scale development articles; alternative model loads negatively worded OCB-O items onto CWB-O factor

CFI = confirmatory fit index, SRMR = standardized root mean square residual, RMSEA = root mean square error of approximation, BIC = Bayesian information criterion

***$p < .001$

Table 4 Item fit statistics, parameter estimates, and item-total correlations for the Williams and Anderson (1991) scale

	χ^2/df Ratio	*a*	*b1*	*b2*	*b3*	*b4*	*ITC*
Sample 1							
Item 1	1.23	1.82	− 2.08	− 1.26	− 0.33	0.88	0.54
Item 2	1.27	1.96	− 2.47	− 1.64	− 0.52	0.76	0.56
Item 3	1.56	1.79	− 1.84	− 1.29	− 0.33	0.77	0.51
Item 4	1.20	1.47	− 1.94	− 1.26	− 0.11	1.09	0.48
Item 5	1.29	1.85	− 3.09	− 1.94	− 0.78	0.54	0.54
Item 6	1.57	1.28	− 2.34	− 1.66	− 0.68	0.90	0.39
Item 7	1.36	1.75	− 2.93	− 2.21	− 1.10	0.18	0.52
Item 8	0.89	1.07	− 3.65	− 2.86	− 1.37	0.04	0.41
Item 9	1.33	0.94	− 3.00	− 1.94	− 1.24	0.19	0.31
Item 10	1.53	0.35	− 10.85	− 7.76	− 4.52	− 1.09	0.16
Item 11	1.50	0.14	− 20.17	− 13.90	− 6.07	4.09	0.05
Item 12	1.34	− 0.09	35.75	23.13	8.35	− 8.30	− 0.04
Item 13	1.21	1.19	− 3.34	− 2.59	− 1.21	0.25	0.47
Item 14	1.25	1.16	− 3.22	− 2.31	− 0.90	0.81	0.41
Sample 2							
Item 1	1.11	2.20	− 1.74	− 1.00	0.11	1.27	0.62
Item 2	1.20	2.57	− 2.04	− 1.14	0.01	1.15	0.64
Item 3	1.48	1.64	− 1.27	− 0.67	0.42	1.63	0.58
Item 4	1.13	1.59	− 1.76	− 0.79	0.35	1.64	0.54
Item 5	0.93	2.26	− 2.27	− 1.30	− 0.14	1.09	0.64
Item 6	1.31	1.74	− 2.29	− 1.19	− 0.02	1.47	0.55
Item 7	0.96	1.60	− 3.30	− 2.27	− 0.87	0.64	0.53
Item 8	1.30	0.92	− 4.02	− 2.65	− 1.16	0.46	0.33
Item 9	1.07	0.82	− 2.71	− 1.89	− 1.10	0.26	0.32
Item 10	1.25	− 0.26	0.74	− 5.47	− 10.50	− 17.75	− 0.04
Item 11	1.38	0.01	− 113.71	117.67	307.21	548.49	0.09
Item 12	1.15	− 0.04	13.60	− 24.90	− 57.93	− 105.77	0.08
Item 13	1.41	1.15	− 2.74	− 1.75	− 0.47	1.09	0.40
Item 14	1.44	0.98	− 3.91	− 2.83	− 1.32	0.71	0.37

χ^2/df ratio represents item fit to the model; a is the discrimination parameter; b1–b4 are threshold parameters; *ITC* is the item-total correlation

should be removed. All other items showed acceptable levels of discrimination ($a > 0.94$), supporting their retention.

Additionally, item-total correlations (ITCs) were calculated for the original, full-form OCB scale. The results, shown in Table 4, reflect item-total correlations computed with the item in question removed (e.g., the ITC for item 1 is the correlation between scores on item 1 and the total scale score calculated from items 2 through 14), in accordance with common recommendations (e.g., Allen and Yen 1979). The ITCs for item 10 ($ITC = 0.16$) is small, only half as large as the next smallest value, and the ITCs for items 11 and 12 (0.05 and − 0.04, respectively) are negligibly small. In sum, these items do not relate well to the full scale score. This further suggests that the negatively worded items do not behave properly alongside the rest of the OCB scale.

The psychometric behavior of items determined by the IRT parameters can by visualized via item response functions (IRFs), which relate a respondent's level of the measured trait (labeled Theta on the abscissa) to the likelihood of them given a specific response to that item (on the ordinate); Fig. 1 contrasts the IRFs for Item 1 on the left, which demonstrated good psychometric behavior, and item 10 on the right, which demonstrated poor characteristics. These properties can also be visualized via item information functions (IIFs), which show how much information an item gives about a respondent; this information naturally varies across the latent continuum as a function of the item's discrimination and threshold parameters. Figure 2 contrasts the IIFs for Items 1 (left) and 10 (right) as examples of a good and poor information, respectively. These figures are representative of all IRFs and IIFs, with the negatively worded item figures mirroring those of item 10 and the positively worded item figures mirror those of item 1.

Additionally, the lowest item threshold parameters of the negative-worded items were all extremely out of bounds ($b_1 =$

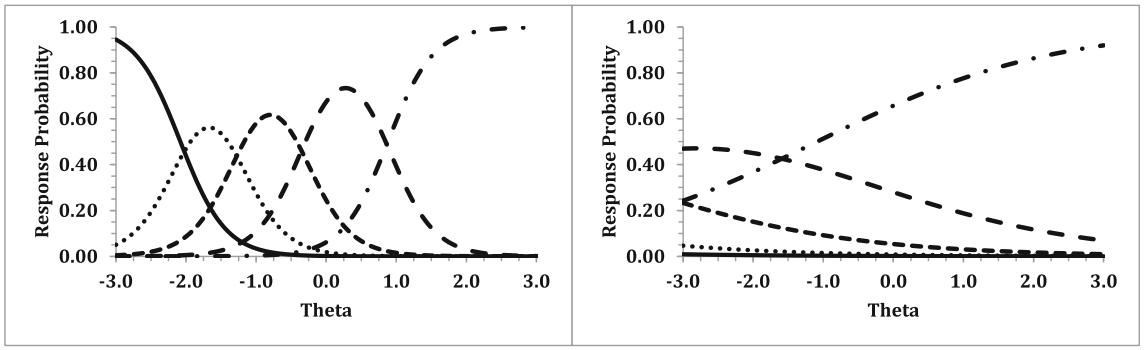

Fig. 1 Item response functions for item 1 (left) and item 10 (right) from the sample 1 data. Note: Item 1 (left) exhibits good IRF form with distinguishable, ordered peaks which all fall within the normal range on the latent continuum; item 10 (right) contains the same number of trace lines corresponding to the same 5-point response scale but has no discernible peaks and little interpretable differentiation in the normal range

− 10.85, − 20.17, and 35.75, respectively), and the majority of the remaining threshold parameters of these items would be considered out of bounds (only the fourth threshold parameter of item 10 falls within a reasonable range; $b_4 = -1.09$). Given that the continuum's scale is distributed normally around zero, good items have parameter estimates falling between − 3.0 and 3.0 (Embretson and Reise 2000; Zickar 2012); the estimates from the negatively worded items can therefore be deemed excessively negative and make the items functionally useless as they do not provide any information in the normal range. It should also be noted that, in addition to being out of bounds, the lowest threshold parameter for item 12 ($b_1 = 35.75$) is also at the opposite end of the latent continuum from where it should be, indicating that the item is not functioning as intended.

Furthermore, in addition to being excessively far from zero, the estimated threshold parameters for item 12 in sample 1 and items 10 and 12 in sample 2 (discussed in greater detail below) are in the reverse order of what would be expected. That is, higher thresholds fall at the negative end of the latent continuum, rather than the positive end, even after properly reverse coding the items. These out-of-bounds results match the negative discrimination parameters for these items. Because of the excessively low absolute value of the discrimination parameters, it is likely that any variability in item responses is more attributable to random noise than a psychometrically sound signal, thus providing further evidence that these items do not properly model the latent OCB trait.

In general, however, the OCB scale has good test information across all levels of the theta continuum, with only slightly lower levels of information at the highest levels (see Fig. 3), though the lack of discrimination among the negatively worded items demonstrates that these items do not contribute to this positive quality of the scale. The test information indicates that the scale is useful for our model testing purposes, as it can differentiate between people across all levels of OCB.

Replication Using Sample 2 Results for sample 2, reported in Table 4, mirrored those of sample 1 with even more evidence for removal of items 10, 11, and 12. Although all 14 items showed acceptable fit for the GRM (average χ^2/df ratio = 1.22), the three items in question yielded uninterpretable parameter estimates (see Table 4). All other items showed acceptable discrimination parameters ($a > 0.82$) and good coverage of the latent continuum. The local dependence statistics for items 10, 11, and 12 once again exceed acceptable values ($LD_{10,11} = 35.9$, $LD_{10,12} = 33.0$, $LD_{11,12} = 20.3$), and as with sample 1, items 5 and 6 showed unexpectedly high local dependence ($LD_{5,6} = 20.1$). Additionally, items 13 and 14 also showed excess covariation ($LD_{13,14} = 22.8$). Item content for these items, however, does not readily yield an explanation. In addition to the IRT analysis, sample 2 data also replicated the

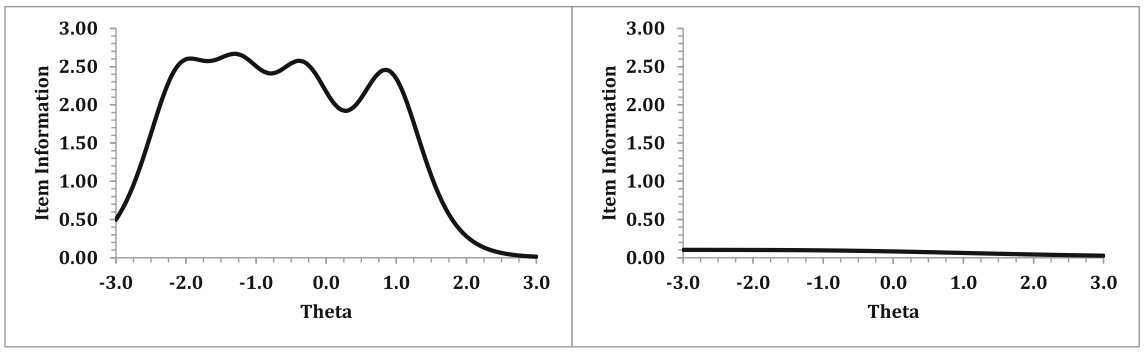

Fig. 2 Item information curves for item 1 (left) and item 10 (right)

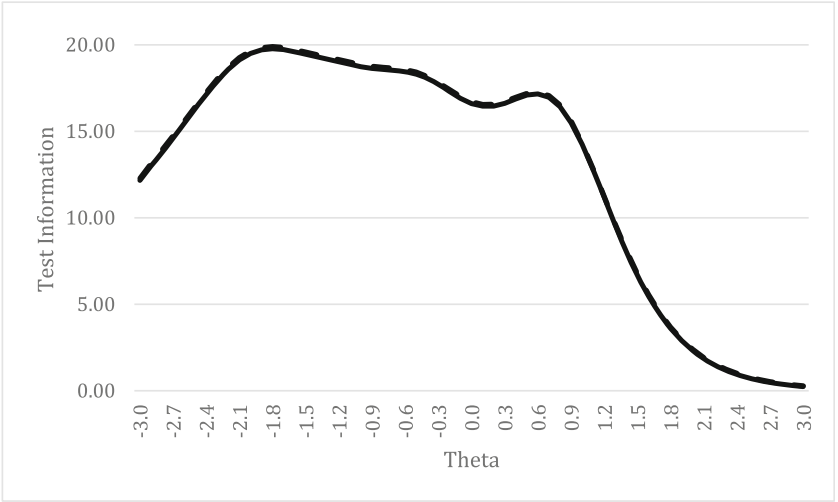

Fig. 3 Test information function for the original (dotted) and 11-item revised (solid) scales from sample 1

findings of the item-total correlations, with the negatively worded items having negligibly small values.

Item analysis was also performed on an ad hoc scale consisting of the negatively worded OCB items and the CWB scale. Results of the IRT analysis show good psychometric properties of the items, with all discrimination parameters above 1.00, threshold parameters within the acceptable range (except for the highest thresholds for each item, which all exceeded positive 3.00) and in the correct order, and all item fit statistics showing acceptable fit (below 2.00 for all items). Furthermore, the ITCs for these items closely match those of the CWB items, suggesting that they relate closely to the rest of the items measuring CWB. This provides evidence from yet another angle that these negatively worded OCB items are more appropriately interpreted as CWB items. The results are shown in Table 5.

Post Hoc Analysis Based on the results reported above, we conducted two post hoc two-factor CFAs, removing the three negatively worded OCB items. In sample 1, the resulting 11-item scale showed acceptable, but not exceptional, fit ($\chi^2(43) = 201.38$, CFI = 0.919, SRMR = 0.044, RMSEA 90% CI = (0.058, 0.076), BIC = 26,947.38). The removal of the three items also resulted in improved internal consistency reliability ($\alpha = 0.82$). Similar results were found with sample 2; fit statistics met minimum cutoffs ($\chi^2(43) = 264.18$, CFI = 0.926, SRMR = 0.044, RMSEA 90% CI = (0.068, 0.085), BIC = 27,006.33), and internal consistency improved to $\alpha = 0.84$. As such, it appears that researchers should avoid including negatively worded items in OCB scales. IRT results on the revised, full-form (11-item) version of the scale yielded item parameter estimates equal to the full scale (within the standard error of each estimate) and no new issues with local dependence, although the inflated LD statistics for items 5 and 6 in samples 1 and 2 and items 13 and 14 in sample 2 did persist. Finally, the test information functions for the original full-form and revised full-form (11-item) versions of the scale in sample 2 are functionally identical for the information functions from sample 1 (Fig. 3).

Scale Revision Per our third contribution, to develop a revised, psychometrically valid, short-form scale, we utilized the IRT output from both samples to identify the six items best suited for retention. Because one of the initial goals of this short-form scale was to retain balanced assessment of both OCB domains, three items were selected from the seven OCB-I items and three were selected from the remaining OCB-O items. Of the remaining four OCB-O items, 13 and 14 demonstrated local dependence in sample 2 and are therefore not good candidates, though one of the two items could be

Table 5 Item fit statistics, parameter estimates, and item-total correlations treating all CWB items and negatively worded OCB items as a single scale

	χ^2/df ratio	*a*	*b1*	*b2*	*b3*	*b4*	*ITC*
OCB Item 10	1.28	1.77	−0.14	1.17	2.10	3.28	0.58
OCB Item 11	1.24	1.02	−0.78	0.92	2.18	3.70	0.39
OCB Item 12	1.04	1.29	−0.60	1.08	2.41	4.12	0.47
CWB Item 1	1.33	1.57	1.57	2.62	3.53	4.62	0.45
CWB Item 2	1.02	2.28	1.88	2.35	2.81	3.71	0.43
CWB Item 3	1.44	1.61	1.02	2.29	3.42	4.56	0.51
CWB Item 4	0.89	1.58	0.97	2.04	2.98	4.34	0.49
CWB Item 5	1.23	1.86	0.32	1.25	2.23	3.04	0.54
CWB Item 6	1.14	2.07	0.30	1.54	2.43	3.38	0.58
CWB Item 7	1.15	2.06	0.02	1.05	1.94	2.83	0.62
CWB Item 8	0.96	2.22	0.11	1.15	1.94	2.75	0.63

χ^2/df ratio represents item fit to the model; *a* is the discrimination parameter; *b1–b4* are threshold parameters; *ITC* is the item-total correlation; CWB items 1–4 = interpersonally directed CWB (CWB-I); CWB items 5–8 = organizationally directed CWB (CWB-O); OCB items 10–12 = organizationally directed OCB (OCB-O)

retained without issue. Likewise, items 5 and 6 from the OCB-I items had local dependence issues in both samples. Again, to ensure stronger psychometric characteristics, only one of these two items was considered for retention. Based on discrimination and location of the threshold parameters that provide more uniformly distributed information, items 6 and 13 were retained.

The final set of six items, chosen based on their item parameters, can be found in Table 6 (note that parameters were similar across both samples and could be reasonably expected to not change when equated across samples, so only parameters for sample 2 are reported; for full results, contact the first author). These items show good discrimination ($a > 0.80$) and span the latent continuum (b_k ranges from −4.02 to 1.64). Test information for the revised scale is found in Fig. 4. Data from sample 2 also indicates that the revised, short-form scale also shows good internal consistency on its own ($\alpha = 0.70$) and when using the Spearman-Brown correction to project the reliability onto an 11-item version ($\alpha = 0.81$). The revised, short-form OCB-O subscale (three-item) demonstrates poor internal consistency ($\alpha = 0.48$) but becomes slightly more acceptable ($\alpha = 0.68$) when applying the Spearman-Brown prophecy formula to project the revised three-item OCB-O subscale onto the original seven-item length. While still falling short of traditional cutoffs, this alpha does exceed the value of the original, full-form subscale ($\alpha = 0.58$). This is an important note when considering that internal consistency estimates of reliability, such as alpha, capture measurement error due to the items themselves, specifically agreement or lack thereof among them (Allen and Yen 1979). Thus, this indicates that the items retained in the revised scale are more internally consistent than those of the original scale (which requires the use of the Spearman-Brown formula as shorter scales have inherently lower internal consistencies; Cortina 1993). This metric therefore provides more evidence of the need to remove the negatively worded items as originally postulated. It should be noted, however, that this theoretical interpretation is not used to suggest that the internal consistency of the revised scale is good; it is simply to provide a more holistic understanding of the psychometric information provided.

Revised, Short-Form Scale Construct Validity Assessment Construct validity of the revised, short-form scale was assessed by running regressions to compare the relationships between the original and revised scales with three theoretically and empirically established OCB antecedents: affective commitment, autonomy, and workload (Bolino and Turnley 2005; Chen and Chiu 2009; Schappe 1998). Means, standard deviations, internal consistency reliabilities, and correlations for all variables are reported in Table 7. In total, six regressions were performed (original full-form OCB scale, revised short-form OCB scale, original full-form OCB-I subscale, revised short-form OCB-I subscale, original short-form OCB-O subscale, revised short-form OCB-O subscale) and the corresponding scales were compared (see Table 8). The revised, short-form scale demonstrated comparable, if not stronger, relationships with each of these antecedents compared to the original scale (revised: affective commitment $\beta = 0.25$, $p < .001$; autonomy $\beta = 0.12$, $p < .01$; quantitative workload $\beta = 0.28$, $p < .001$; original: affective commitment $\beta = 0.25$, $p < .001$; autonomy $\beta = 0.07$, $p < .05$; quantitative workload $\beta = 0.27$, $p < .001$). The revised, short-form scale also demonstrated a weaker relationship with CWB compared to the original scale (revised: CWB $\beta = -0.21$, $p < .001$; original: CWB $\beta = -0.34$, $p < .001$). Thus, the revised, short-form scale demonstrates similar (if not improved) construct validity compared to the original scale, by demonstrating comparable relationships with known antecedents and improved discriminant validity with a theoretically unrelated construct (CWB).

Discussion

In this study, we applied a variety of psychometric analyses to demonstrate that negatively worded OCB items measure a unique construct (CWB), not OCB. Indeed, when factor analyzed, the negatively worded items loaded onto a unique, third factor. Furthermore, these negatively worded OCB items loaded onto a CWB factor when factor analyzed with a CWB scale. Additionally, the IRT analyses revealed that the negatively worded items exhibit lower discrimination parameters and higher levels of local independence than the positively worded items. Further, when the three negatively worded items were analyzed with a CWB scale, the items showed acceptable discrimination. Finally, we present a revised, short-form scale that can be utilized in the future measurement of OCB.

Research Implications

Our results have several critical research implications. First, our results indicate that there may be measurement error, and in turn biased results, in the current OCB literature. Indeed, in studies using OCB scales with negatively worded items, the reported estimates of relationships may be under or overestimated, as evidenced by the original Williams and Anderson (1991) scale's smaller relationships with theoretical antecedents and larger relationships with theoretically unrelated constructs (e.g., CWB), compared to a revised scale (i.e., a scale without negatively worded items). As such, studies that have utilized such scales should be interpreted with caution

Table 6 Final retained OCB items and parameter estimates for retained items from sample 2

		a	*b1*	*b2*	*b3*	*b4*
OCB-I items						
Item 2	I helped others with heavy workloads.	2.57	− 2.04	− 1.14	0.01	1.15
Item 4	I assisted my supervisor with his/her work when not asked.	1.59	− 1.76	− 0.79	0.35	1.64
Item 6	I took a personal interest in other employees.	1.74	− 2.29	− 1.19	− 0.02	1.47
OCB-O items						
Item 8	My attendance at work was above the norm.	0.92	− 4.02	− 2.65	− 1.16	0.46
Item 9	I gave advance notice when I was unable to come to work.	0.82	− 2.71	− 1.89	− 1.10	0.26
Item 13	I conserved and protected organizational property.	1.15	− 2.74	− 1.75	− 0.47	1.09

OCB = organizational citizenship behavior (I = interpersonal; O = organizational); *a* is the discrimination parameter; *b1–b4* are threshold parameters

and should perhaps be reanalyzed (or replicated) without the negatively worded items in order to better understand the underlying OCB construct. Furthermore, our results suggest that issues related to the replication crisis (Schooler 2014) in the OCB literature may be partially explained by the operationalization of OCB based on measures with negatively worded items in one study, and a unidirectional measure of OCB in another study. The resulting implication is that researchers performing meta-analytic OCB research should perhaps consider this measurement issue as a moderator.

Second, the results of this study provide a concrete example of the potential pitfalls of including negatively worded items in scales, in general (Bandalos 2018). Although negatively worded items have the potential to reduce biases (e.g., careless responding; Bandalos 2018), these items can also, as seen with the OCB measure, inadvertently introduce a unique construct into the measure. We would suggest that the risk of construct contamination far outweighs the limited protection negatively worded items may provide in terms of preventing careless responding. As such, we encourage researchers think about the theoretical implications of adding negatively worded items. If these items could introduce a unique construct, then it is advised that negatively worded items not be used. In these situations, researchers may want to consider alternative approaches to preventing or screening for careless responding (e.g., instructed items such as "Select 'Strongly Agree' for this answer").

However, if researchers do wish to include negative items, there are several options for mitigating the negative impact the items exert on the psychometric properties of the scale. When using CTT techniques, such as confirmatory factor analysis (CFA), researchers can correlate error terms among the items and use the resultant scale scores instead of simple sums or averages of items. Alternatively, researchers can use testlet response theory (TRT; Wainer et al. 2007) to estimate latent trait scores. A specialized form of IRT, TRT adds an additional parameter, gamma (γ), which models interdependence among subsets of items—referred to as testlets—and treats any such covariation as a substantively unimportant random variable. Use of TRT helps prevent overestimation of discrimination parameters and item and test information. Regardless of which of these techniques is used, it is important for researchers to consider, and empirically assess if at all possible, that the negatively worded items may be substantively inappropriate, as was found in the present study. If that is the case, then none of the above techniques will fully alleviate the issues of negatively worded items, and it is recommended to remove them completely.

Third, the results of this study provide further psychometric evidence that OCB and CWB items are distinct. Specifically, we utilize IRT to examine the properties of each item and their relationship to the latent construct under evaluation, OCB. IRT results overwhelmingly indicate that negatively worded items are not measuring the OCB construct. These findings support propositions based on correlational results (Dalal 2005) that OCB and CWB are not opposite ends of the same continuum, but rather distinctly unique constructs.

Finally, and perhaps most importantly, we present a revised, short-form OCB measure that can be used in future studies on OCB (see Table 6 for a list of the retained items). This short-form scale is an improvement from the original Williams and Anderson (1991) measure in two primary ways. First, the short-form scale demonstrates improved psychometric properties. It only includes items with high factor loadings and good internal consistency. Additionally, the item and test information functions of the short-form mirror those of the full scale almost identically, albeit at a lower level (which is always the case with shortened scales). Second, the short-form scale demonstrates improved construct validity, as it has

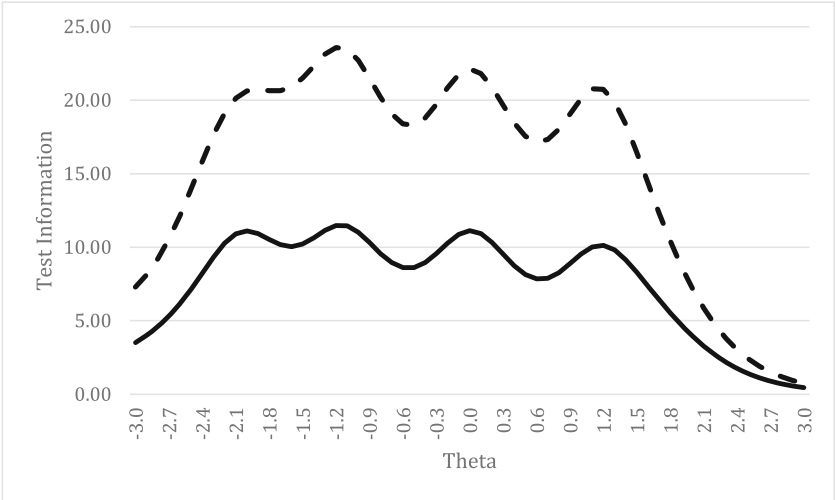

Fig. 4 Test information function for the original (dotted) and six-item short-form (solid) scales from sample 2

stronger associations with related constructs and weaker correlations with unrelated constructs. As such, this revised measure should produce more valid study results, as well as reduce survey length and participant attrition, within the field of organizational research.

Limitations, Future Directions, and Conclusion

As with all studies, our research has certain limitations, as well as areas of opportunity for future research. For example, the original Williams and Anderson (1991) OCB scale was intended to be used as a supervisor report measure of employee behavior. However, in our study, consistent with how OCB measures have been applied in more recent research (Bal et al. 2010; Li and Thatcher 2015), all responses were employee self-report. Indeed, not all OCB needs to be, or should be, supervisor report, as LePine et al. (2002) argue that the conceptual framework should determine the source of OCB ratings. Although we expect similar findings for supervisor report, further analyses would need to be performed to confirm measurement equivalence across reporter. In any case, we are confident that our study provides a useful, psychometrically sound, short-form measure that can be used for self-report purposes, and perhaps other-report purposes.

One additional limitation is that the reduction of items in the short-form measure has reduced the number of behaviors in the construct space that are being assessed. Thus, someone could perform an act of OCB not measured by the scale, thus making it appear that the person did not engage in OCB. Although this is an important limitation, all scales arguably do not cover the entire construct space. Based on our psychometric evaluation of the scale, it is our belief that this revised scale includes the best items for assessing general OCB. Furthermore, the short-form scale should enable researchers to perform more effective longitudinal research on OCB by reducing survey length and participant attrition.

Finally, we acknowledge that the internal consistency reliability of the revised OCB-O subscale does not meet the typical standards of 0.70 (Schmitt 1996). However, we do not believe this is a major issue, as numerous psychometric analyses indicate that the subscale is reliable and valid. Specifically, IRT and CFA analyses demonstrate that the items behave properly and adhere to the more formal requirements and modeling capabilities of these more informative psychometric analyses. Indeed, many researchers have denounced the overreliance and overemphasis on internal consistency reliability (Cortina 1993; Schmitt 1996; Sijtsma 2009). For example, Cortina (1993) advises that some measures with low levels of alpha may still be useful and that presenting only alpha is not sufficient. Additionally, as noted above, when applying the appropriate corrections, the revised scale is more internally consistent than the original version. However, if a researcher is particularly concerned about the OCB-O scale internal consistency, then we suggest utilizing the overall scale, which demonstrated sufficient internal consistency.

Beyond that, based on the results of this study, researchers should be aware that negatively worded items might not be measuring their intended constructs and should take steps to conceptually and empirically assess whether these negative items measure their intended constructs. Furthermore, researchers should assess scales using appropriate psychometric techniques (CTT and IRT) to potentially modify and improve these scales for future research.

Conclusion Based on two large-scale field samples that relied on demonstratively different recruitment methods, it is our hope that the revised, short-form OCB scale will enable researchers to better measure and study OCB in organizational research and further expand our understanding of its nomological network. The revised scale is shorter, which is more appealing for inclusion in surveys, and demonstrates improved psychometric properties and construct validity, which

Table 7 Means, standard deviations, correlations, and internal consistency reliability for sample 2

			Correlations											
Variable	*M*	*SD*	1	2	3	4	5	6	7	8	9	10	11	12	13
1. Original full-form OCB	3.66	0.58	(0.80)												
2. Original full-form OCB-I	3.39	0.80	0.88**	(0.85)											
3. Original full-form OCB-O	3.93	0.59	0.77**	0.38**	(0.58)										
4. Revised full-form OCB	3.56	0.70	0.96**	0.94**	0.62**	(0.84)									
5. Revised short-form OCB	3.56	0.73	0.93**	0.84**	0.68**	0.95**	(0.70)								
6. Revised short-form OCB-I	3.31	0.87	0.81**	0.93**	0.35**	0.86**	0.85**	(0.70)							
7. Revised short-form OCB-O	3.82	0.85	0.76**	0.50**	0.81**	0.75**	0.85**	0.44**	(0.48)						
8. Affective Commitment	3.29	1.13	0.34**	0.33**	0.21**	0.33**	0.33**	0.35**	0.21**	(0.94)					
9. Autonomy	3.46	0.91	0.18**	0.21**	0.06	0.21**	0.22**	0.23**	0.14**	0.48**	(0.83)				
10. Quantitative Workload	3.10	0.88	0.19**	0.23**	0.06	0.22**	0.21**	0.23**	0.12**	− 0.09+	− 0.12**	(0.84)			
11. CWB	1.47	0.51	− 0.35**	− 0.15**	− 0.49**	− 0.20**	− 0.22**	− 0.17**	− 0.21**	− 0.21**	− 0.06	0.16**	(0.81)		
12. CWB-I	1.25	0.49	− 0.17**	− 0.04	− 0.27**	− 0.08+	− 0.08+	− 0.05	− 0.09**	− 0.08+	− 0.04	0.18**	0.77**	(0.81)	
13. CWB-O	1.69	.73	− 0.38**	− 0.18**	− 0.51**	− 0.23**	− 0.26**	− 0.20**	− 0.24**	− 0.24**	− 0.05	0.10*	0.90**	0.42**	(0.79)

N = 924. Scale reliabilities listed along the diagonal; original full-form OCB includes 14 items, revised full-form OCB includes 11 items (does not include the three negatively worded OCB-O items), and revised short-form OCB includes six items. Original full-form OCB-I and OCB-O include seven items each, revised short-form OCB-I and OCB-O include three items each

OCB = organizational citizenship behavior (I = interpersonal; O = organizational), CWB = counterproductive work behavior (I = interpersonal; O = organizational)

**$p < .001$; *$p < .01$; +$p < .05$

Table 8 Comparison of the relationship between OCB and antecedents/CWB using the original and revised OCB scales for Sample 2

	Original full-form scale				Revised short-form scale			
Antecedents	*B*	*SE*	*β*	*t*	*B*	*SE*	*β*	*t*
	OCB							
Aff commitment	0.13	0.02	0.25	7.69***	0.16	0.02	0.25	7.26***
Autonomy	0.05	0.02	0.07	2.22*	0.09	0.03	0.12	3.44**
Quant workload	0.18	0.02	0.27	9.43***	0.23	0.03	0.28	9.21***
CWB	− 0.38	0.03	− 0.34	− 11.49***	− 0.29	0.04	− 0.21	− 6.79***
R^2	0.27				0.21			
F(df)	83.53***(4, 917)				60.53***(4, 917)			
	OCB-I							
Aff commitment	0.21	0.02	0.30	8.91***	0.24	0.03	0.31	9.27***
Autonomy	0.09	0.03	.010	2.89**	0.10	0.03	0.11	3.10**
Quant workload	0.26	0.03	0.28	9.32***	0.28	0.03	0.28	9.32***
CWB-I	− 0.11	0.05	− 0.07	− 2.24*	− 0.13	0.05	− 0.07	− 2.33*
R^2	0.19				0.20			
F(df)	54.18***(4, 916)				57.47***(4, 916)			
	OCB-O							
Aff commitment	0.05	0.02	0.09	2.85**	0.09	0.03	0.12	3.33**
Autonomy	0.01	0.02	0.01	.25	0.08	0.03	0.09	2.49*
Quant workload	0.08	0.02	0.11	4.03***	0.16	0.03	0.17	5.26***
CWB-O	− 0.40	0.02	− 0.50	− 17.19***	− 0.26	0.04	− 0.22	− 6.85***
R^2	0.28				0.11			
F(df)	89.97***(4, 917)				28.47***(4, 916)			

OCB = organizational citizenship behavior (I = interpersonal, O = organizational), Aff commitment = affective commitment, Quant workload = quantitative workload, CWB = counterproductive work behavior (I = interpersonal, O = organizational)

*$p < .05$; **$p < .01$; ***$p < .001$

should improve the measurement and study of this important organizational construct.

References

Allen, M. J., & Yen, W. M. (1979). *Introduction to measurement theory*. Waveland Press.

Bal, P. M., Chiaburu, D. S., & Jansen, P. G. W. (2010). Psychological contract breach and work performance: Is social exchange a buffer or an intensifier? *Journal of Managerial Psychology, 25*, 252–273.

Bandalos, D. L. (2018). *Measurement theory and applications for the social sciences*. New York, NY: Guilford Press.

Bennett, R. J., & Robinson, S. L. (2000). Development of a measure of workplace deviance. *Journal of Applied Psychology, 85*, 349–360.

Bolino, M. C., & Turnley, W. H. (2005). The personal costs of citizenship behavior: The relationship between individual initiative and role overload, job stress, and work-family conflict. *Journal of Applied Psychology, 90*, 740–748.

Borsboom, D. (2006). The attack of the psychometricians. *Psychometrika, 71*, 425–440.

Bureau of Labor Statistics (2017). *Current population survey: Employed persons by occupation, sex, and age* (cpsaat09). Retrieved from https://www.bls.gov/cps/tables.htm .

Byrne, Z. S. (2005). Fairness reduces the negative effects of organizational politics on turnover intentions, citizenship behavior and job performance. *Journal of Business and Psychology, 20*, 175–200.

Cai, L., Du Toit, S. H. C., & Thissen, D. (2011). *IRTPRO: Flexible, multidimensional, multiple categorical IRT modeling (version 4.1) [computer software]*. Chicago, IL: Scientific Software International.

Chen, C. C., & Chiu, S. F. (2009). The mediating role of job involvement in the relationship between job characteristics and organizational citizenship behavior. *The Journal of Social Psychology, 149*, 474–494.

Chen, W. H., & Thissen, D. (1997). Local dependence indices for item pairs using item response theory. *Journal of Educational and Behavioral Statistics, 22*, 265–289.

Cortina, J. M. (1993). What is coefficient alpha? An examination of theory and applications. *Journal of Applied Psychology, 78*, 98–104.

Dalal, R. S. (2005). A meta-analysis of the relationship between organizational citizenship behavior and counterproductive work behavior. *Journal of Applied Psychology, 90*, 1241–1255.

Dalal, R. S., Lam, H., Weiss, H. M., Welch, E. R., & Hulin, C. L. (2009). A within-person approach to work behavior and performance: Concurrent and lagged citizenship-counterproductivity associations, dynamic relationships with affect and overall job performance. *The Academy of Management Journal, 52*, 1051–1066.

DiStefano, C., & Motl, R. W. (2006). Further investigating method effects associated with negatively worded items on self-report surveys. *Structural Equation Modeling, 13*, 440–464.

Drasgow, F., & Hulin, C. L. (1990). Item response theory. In M. D. Dunnette & L. M. Hugh (Eds.), *Handbook of industrial and*

organizational psychology (*2nd ed*) (Vol. 1, pp. 577–636). Palo Alto, CA: Consulting Psychologists Press.

Drasgow, F., & Parsons, C. K. (1983). Application of unidimensional item response theory models to multidimensional data. *Applied Psychological Measurement, 7*, 189–199.

Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists*. Mahwah, NJ: Lawrence Erlbaum Associates, Inc., Publishers.

Fisher, C. D., & To, M. L. (2012). Using experience sampling methodology in organizational behavior. *Journal of Organizational Behavior, 33*, 865–877.

Fisher, G. G., Matthews, R. A., & Gibbons, A. M. (2016). Developing and investigating the use of single-item measures in organizational research. *Journal of Occupational Health Psychology, 21*, 3–23.

Ford, M. T., Wang, Y., Jin, J., & Eisenberger, R. (2018). Chronic and episodic anger and gratitude toward the organization: Relationships with organizational and supervisor supportiveness and extrarole behavior. *Journal of Occupational Health Psychology, 23*, 175–187.

Foster, G. C., Min, H., & Zickar, M. J. (2017). Review of item response theory practices in organizational research: Lessons learned and paths forward. *Organizational Research Methods, 20*, 465–486.

Griffin, M. A., Neal, A., & Parker, S. K. (2007). A new model of work role performance: Positive behavior in uncertain and interdependent contexts. *Academy of Management Journal, 50*, 327–347.

Ip, E. H. (2010). Empirically indistinguishable multidimensional IRT and locally dependent unidimensional item response models. *British Journal of Mathematical and Statistical Psychology, 63*, 395–416.

Jepsen, D. M., & Rodwell, J. J. (2006). A side by side comparison of two organizational citizenship behavior models and their measures: Expanding the construct domain's scope. *Proceedings of the 11th Annual conference of Asia Pacific Decision Sciences Institute*, Hong Kong.

Kline, R. B. (2016). *Principles and practice of structural equation modeling*. New York, NY: Guildford Press.

Lapierre, L. M., Matthews, R. A., Eby, L. T., Truxillo, D. M., Johnson, R. E., & Major, D. A. (2018). Recommended practices for academics to initiate and manage research partnerships with organizations. *Industrial and Organizational Psychology, 11*, 543–581.

LePine, J. A., Erez, A., & Johnson, D. E. (2002). The nature and dimensionality of organizational citizenship behavior: A critical review and meta-analysis. *Journal of Applied Psychology, 87*, 52–65.

Li, A., & Thatcher, S. M. B. (2015). Understanding the effects of self and teammate OCB congruence and incongruence. *Journal of Business and Psychology, 30*, 641–655.

MacKenzie, S. B., Podsakoff, P. M., & Fetter, R. (1991). Organizational citizenship behavior and objective productivity as determinants of managerial evaluations of salesperson's performance. *Organizational Behavior and Human Decision Processes, 50*, 123–150.

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*, 241–250.

Matthews, R. A., & Ritter, K. J. (2016). A concise, content valid, gender invariant measure of workplace incivility. *Journal of Occupational Health Psychology, 21*, 352–365.

Mayer, R. C., & Gavin, M. B. (2005). Trust in management and performance: Who minds the shop while the employees watch the boss? *The Academy of Management Journal, 48*, 874–888.

Motowidlo, S. J., & Kell, H. J. (2013). Job performance. In N. W. Schmitt & S. Highhouse (Eds.), *Handbook of psychology: Industrial and organizational psychology* (Vol. 12, pp. 82–103). Hoboken, NJ: John Wiley & Sons, Inc..

Organ, D. W. (1988). *Organizational citizenship behavior: The good soldier syndrome*. Lexington, MA: Lexington Books.

Organ, D. W. (1990). The motivational basis of organizational citizenship behavior. In B. M. Staw & L. L. Cummings (Eds.), *Research in organizational behavior* (Vol. 12, pp. 43–72). Greenwich, CT: JAI Press.

Podsakoff, P. M., & MacKenzie, S. B. (1994). Organizational citizenship behaviors and sales unit effectiveness. *Journal of Marketing Research, 31*, 351–363.

Podsakoff, N. P., Whiting, S. W., Podsakoff, P. M., & Blume, B. D. (2009). Individual- and organizational-level consequences of organizational citizenship behaviors: A meta-analysis. *Journal of Applied Psychology, 94*, 122–141.

R Core Team (2017). R: A language and environment for statistical computing (Version 3.4.3) [Computer software].Vienna, Austria: R Foundation for Statistical Computing.

Rosseel, Y. (2012). lavaan: An R package for structural equation modeling. *Journal of Statistical Software, 48*, 1–36.

Sackett, P. R., Berry, C. M., Wiemann, S. A., & Laczo, R. M. (2006). Citizenship and counterproductive behavior: Clarifying relations between the two domains. *Human Performance, 19*, 441–464.

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. Psychometrika Monograph Supplement, No. 17

Schappe, S. P. (1998). The influence of job satisfaction, organizational commitment, and fairness perceptions on organizational citizenship behavior. *The Journal of Psychology, 132*, 277–290.

Schmitt, N. (1996). Uses and abuses of coefficient alpha. *Psychological Assessment, 8*, 350–353.

Schooler, J. W. (2014). Metascience could rescue the 'replication crisis'. *Nature, 515*, 9. https://doi.org/10.1038/515009a.

Sijtsma, K. (2009). One the use, the misuse, and the very limited usefulness of Cronbach's alpha. *Psychometrika, 74*, 107–120.

Spector, P. E., & Jex, S. M. (1998). Development of four self-report measures of job stressors and strain: Interpersonal conflict at work scale, organizational constraints scale, quantitative workload inventory, and physical symptoms inventory. *Journal of Occupational Health Psychology, 3*, 356–367.

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

Thompson, C. A., & Prottas, D. J. (2006). Relationships among organizational family support, job autonomy, perceived control, and employee well-being. *Journal of Occupational Health Psychology, 11*, 100–118.

Thorndike, E. L. (1904). *An introduction to the theory of mental and social measurements*. New York, NY: Science Press.

Wainer, H., Bradlow, E. T., & Wang, X. (2007). *Testlet response theory and its applications*. Cambridge: New York, NY.

Williams, L. J., & Anderson, S. E. (1991). Job satisfaction and organizational commitment as predictors of organizational citizenship and in-role behaviors. *Journal of Management, 17*, 601–617.

Yun, S., Takeuchi, R., & Liu, W. (2007). Employee self-enhancement motives and job performance behaviors: investigating the moderating effects of employee role ambiguity and managerial perceptions of employee commitment. *Journal of Applied Psychology, 92*, 745–756.

Zickar, M. J. (2012). A review of recent advances in item response theory. In J. J. Martocchio, A. Joshi, & H. Liao (Eds.), *Research in Personnel and Human Resources Management Research in Personnel and Human Resources Management, Volume 31* (pp. 145–176). Emerald Group Publishing Limited.

Zickar, M. J., & Broadfoot, A. A. (2009). The partial revival of a dead horse? Comparing classical test theory and item response theory. In C. E. Lance & R. J. Vandenberg (Eds.), *Statistical and methodological myths and urban legends: Doctrine, verity and fable in the organizational and social sciences* (pp. 37–59). New York, NY: Taylor & Francis Group.