Cox et al., *Cogent Psychology* (2017), 4: 1323988
https://doi.org/10.1080/23311908.2017.1323988

cogent psychology

CLINICAL PSYCHOLOGY & NEUROPSYCHOLOGY | RESEARCH ARTICLE

Effects of augmenting response options of the MMPI-2-RF: An extension of previous findings

Andrew Cox[1], Seth C. Courrégé[2]*, Abigail H. Feder[2] and Nathan C. Weed[2]

Received: 28 December 2016
Accepted: 24 April 2017
Published: 10 May 2017

*Corresponding author: Seth C. Courrégé, Department of Psychology, Central Michigan University, Mount Pleasant, MI, USA
E-mail: seth.courrege@gmail.com

Reviewing editor:
Kulbir Singh Birak, University Campus Suffolk, UK

Additional information is available at the end of the article

Abstract: The purpose of this study was to investigate the effects of augmenting the response options of the Minnesota Multiphasic Personality Inventory-Second Edition-Restructured Form (MMPI-2-RF). Numerous investigations indicate that scores on scales with more response options tend to possess better psychometric properties than those with fewer response options. A previous investigation by Cox and colleagues compared the psychometric performance of the MMPI-2 Restructured Clinical scales using the standard response format to a version using an augmented, four-point response format. Scores from the augmented version demonstrated superior internal consistency compared to the standard form. Scores from the augmented version failed to demonstrate superior convergent validity compared to the standard form. The current study replicates and expands these findings to all the MMPI-2-RF scales. The augmented version took approximately 3 minutes longer to complete, but participants felt the augmented response format allowed them to describe themselves more accurately. As in the previous study, internal consistency was superior for scores on the augmented version, but these gains did not lead to increased convergent validity. No order effects were observed. Potential explanations for this counterintuitive finding are discussed, and recommendations are made for future investigations in response option augmentation.

ABOUT THE AUTHORS

The co-authors of this project are members of the Psychological Assessment Laboratory at Central Michigan University. Led by Kyunghee Han, PhD, and Nathan Weed, PhD, the Psychological Assessment Laboratory conducts research on psychometric measures used in applications of clinical psychology. Most commonly, we conduct research on aspects of assessment with the MMPI, the most widely used psychological test in the world. Recent projects have focused on the substance abuse scales of the MMPI-2-RF, the Hindi translation of the MMPI-2, and q-sort applications of MMPI-2-RF research. The present manuscript is adapted from the doctoral dissertation of the first author, Andrew Cox, PhD, currently affiliated with Dominion Behavioral Healthcare located in Richmond, VA.

PUBLIC INTEREST STATEMENT

The various forms of the Minnesota Multiphasic Personality Inventory (MMPI) are some of the most widely used psychological tests in the world, employed in a variety of clinical and occupational settings. The most recent update, the Minnesota Personality Inventory-2nd Edition-Restructured Form (MMPI-2-RF) is used to evaluate a wide range of personality and psychopathological phenomena, and to plan treatment. Because of its widespread use and real world impact, research evaluating its applications and improving its utility is important. This study reports on attempts to improve the accuracy of the test without making the test longer, by comparing the test's traditional true-false format with a multiple choice format. Previous attempts to use a multiple-choice format with this test have had mixed results. Our hope is that we can identify which format is more effective with this test and why, which may lead to improvement in the reliability and accuracy of results of the test.

cogent oa

© 2017 The Author(s). This open access article is distributed under a Creative Commons Attribution (CC-BY) 4.0 license.

Subjects: Psychological Science; Testing, Measurement and Assessment; Psychometrics/ Testing & Measurement Theory; Test Development, Validity & Scaling Methods; Mental Health; Psychiatry & Clinical Psychology - Adult; Clinical Testing & Assessment

Keywords: MMPI; MMPI-2-RF; response format; assessment

1. Introduction

When designing a rating scale, test authors must choose from a wide variety of response formats. Examinees may be asked to make a mark on a line to express the extent of their agreement, check a box to indicate the presence or absence of a characteristic, rank statements based on accuracy, select a verbal description from a series of responses, or to use any number of other methods. The Minnesota Multiphasic Personality Inventory (MMPI; Hathaway & McKinley, 1940) presents examinees with a series of statements, asking them to mark each as either true or false as applied to them. This response format was retained for the revised version of the test (MMPI-2; Butcher, Dahlstrom, Graham, Tellegen, & Kaemmer, 1989) and is used in the most recent edition, the Minnesota Multiphasic Personality Inventory-Second Edition-Restructured Form (MMPI-2-RF; Tellegen & Ben-Porath, 2008). A great deal of evidence, however, suggests that this response format may not be optimal.

The purpose of this study was to test an alternative response format for the MMPI-2-RF and to compare its psychometric performance to the original. The methodology replicates that of Cox et al. (2012). The alternative format being tested increases the number of response options in a process known as "augmentation." Data simulation studies show when response augmentation is tested under the most tightly controlled conditions, clear psychometric benefits accrue as response options are added to a scale. Data from actual participants supports the results of these studies, providing further evidence of the benefits of response augmentation.

An open question remains regarding whether an optimal number of response options for psychological scales exists. Although Likert's (1932) research advanced the measurement of self-reported psychological variables, he failed to determine how many response categories should be used when constructing scales of this kind. Symonds (1924) was the first major writer to consider this issue. Before then, a variety of authors had used any number of response categories on their scales, including Galton (1883) with nine, Pearson (1907) and Webb (1915) with seven, Downey (1921) with eleven, and Plant (1922) with ten. Symonds concluded that "Apparently the construction of rating scales has proceeded quite without consideration as to the reason for constructing scales with one rather than another number of classes" (Symonds, 1924, p. 456).

Most authors have settled on fewer response options than early writers, with Preston and Colman reporting in 2000 that most Likert scales use five to seven response options. Although disagreement lingers about the ideal number of response categories for Likert-type scales, it seems to be taken for granted that five to seven response options is an appropriate range. Theorists generally agree that scales possessing more response options, all else being equal, should possess better psychometric properties. In his popular text on summated rating scale construction, Spector (1992) devoted six sentences to the topic. Citing Nunnally, he stated that a point of diminishing returns may be reached with the addition of response options, and that "generally between five and nine choices are optimal for most uses" (Nunnally, 1978, p. 21).

1.1. Studies supporting augmentation

Conflicting findings have emerged regarding the optimal number of response options for Likert scales. The problem is largely one of comparability. There is no guarantee that the ideal number of response categories for a marketing survey would be the same for a teacher rating scale or a scale of psychopathology. Each scale has a unique configuration of psychometric properties that vary across settings, making it difficult to state conclusively the "best" number of response choices. Some investigators attempted to address this concern by taking the specific scale out of the equation.

cogent psychology

Instead of collecting data from participants, these investigators used computer-generated data to fill in response patterns. This technique is called the "Monte Carlo" method due to its use of randomly generated data within defined parameters.

Lissitz and Green (1975) were the first investigators to use this method to determine the effect of the number of response categories on a scale. The authors concluded a five-point scale is optimal for most instruments and settings, and there seems to be little utility in adding additional response options. A similar study was conducted by Jenkins and Taber (1977) that lent support to the conclusions of Lissitz and Green (1975), in that the psychometric properties of a two-point scale can be enhanced by increasing its response options up to five points. Cicchetti, Shoinralter, and Tyrer (1985) used similar methods to examine enhancements in inter-rater reliability that result from increasing a scale's response options. Paralleling earlier studies, increases in reliability leveled off around five response options, with little improvement noted beyond seven. The most recent Monte Carlo investigation into augmentation was conducted by Lozano, García-Cueto, and Muñiz (2008). The findings of this and other Monte Carlo investigations are consistent. As the number of response options per item increases, so do many of its psychometric properties. This effect was demonstrated for internal consistency, retest reliability, inter-rater reliability, criterion validity, and clarity of factor structure. These investigations provide concrete recommendations regarding the appropriate number of response categories for Likert scales. Generally, four or five categories appear to be acceptable, with relatively little benefit of having more than seven categories. Although no ideal number of response options was found, these studies show two-point response scales consistently performed poorest psychometrically when compared to scales with more response options. In fact, Lozano et al. (2008) concluded, "from a psychometric perspective, it is advisable for questionnaires to avoid using such formats" (Lozano et al., 2008, p. 78).

Although computer studies lend strong support to the case for augmentation, these simulations may not accurately represent the way real examinees respond to Likert scales. Fortunately, many empirical investigations using real-participant data (known as "*in vivo*") have been conducted to address this concern. Cox et al. (2012) provides comprehensive lists of *in vivo* studies that either support or fail to support the findings of Monte Carlo studies. Cox et al. (2012) note that most of the studies that failed to find psychometric benefit from augmentation suffered from methodological or interpretive flaws. An *in vivo* study by Komorita and Graham (1965) found that when a set of items is maximally internally consistent (i.e. the items measure almost exactly the same thing), there is little benefit from augmentation. Otherwise, response option augmentation functions as theorized and produces similar effects to Monte Carlo studies. Subsequent studies (see Cox et al., 2012 for a comprehensive list) applied the original findings of Komorita and Graham to specific domains of interest, with similar conclusions. While none of these studies have addressed the MMPI directly, together they form a strong argument in favor of augmentation based on numerous observations of the same essential principle.

1.2. Previous attempts to augment the MMPI-2 and MMPI-2-RF

Cox et al. (2012) applied this principle to the MMPI-2 and conducted a project to test the effects of augmenting the response format of the MMPI-2 Restructured Clinical (RC) scales on their reliability and validity. The experimental format contained four response options identical to those of the Personality Assessment Inventory (PAI; Morey, 1991): "very true/mainly true/slightly true/false, not at all true." The research literature reviewed above supported the use of four response choices. Additionally, the mid-point was excluded based on Nunnally's (1978) warning that the presence of a neutral step may increase central tendency bias, which is the propensity for examinees to give responses closer to the middle of the scale than to the extremes. This choice was particularly important as the standard dichotomous MMPI-2 response format does not include a mid-point response option.

In Cox et al. (2012) the RC scales (Tellegen et al., 2003) were used to test the effects of augmentation, and the Lie (L) and Infrequency Psychopathology (Fp) scales were used to screen for invalid

response sets. There were several advantages to selecting the RC scales for augmentation. The intercorrelations between many of the RC scales are much lower than they are for the Clinical Scales (CS). As a result, each scale may be seen as a relatively independent (though still intercorrelated) testing ground for augmentation. Also, although the RC scales' internal consistency is substantially higher than that of the CS, there is still room for improvement. In the normative sample, the median Cronbach's alpha for the RC scales is .76, which is not in the ideal range of α = .80–.90 suggested by Streiner (2003). This suggests that the RC scales could benefit from augmentation.

Fp and L were selected to screen out invalid responding. Fp is sensitive to random responding, true response bias, and false response bias (Butcher et al., 2001). According to the meta-analysis by Rogers, Sewell, Martin, and Vitacco (2003), it is the single best predictor of overreporting on the MMPI-2. The L scale was found to be the single best indicator of underreporting (Graham, Watts, & Timbrook, 1991). Together these evaluated several forms of invalid responding.

Cox et al. (2012) also examined the potential effects of augmentation on the scales' convergent validity. Two scales from the Multidimensional Personality Questionnaire (MPQ; Tellegen & Waller, 2008) were selected to test these effects. Sellbom and Ben-Porath (2005) found the Alienation (Al) scale correlated r = .62 with RC6 and the Wellbeing (Wb) scale correlated r = −.72 with RC2. The authors hypothesized that augmentation would strengthen these relationships, providing evidence of enhanced convergent validity.

The results of this investigation showed the potential of augmentation to enhance the psychometric functioning of MMPI-2 scales. Cronbach's alphas and mean inter-item correlations (MIIC) increased for all RC scales as a result of increasing the number of response options. Augmentation increased convergent validity for participants who completed the augmented version first, but effects were equivocal for the combined sample. Augmentation did not appear to alter the meanings of the RC scales, as evidenced by strong correlations between both the standard and augmented scales.

A recent study by Finn, Ben-Porath, and Tellegen (2015) extended the investigation of Cox and colleagues to statistically examine the source of increased reliability in augmented response formats of the MMPI-2-RF. The authors compared a dichotomous and a balanced four-response-option ("Definitely True, Mostly True, Mostly False, and Definitely False") version of the MMPI-2-RF. The balanced response option format was used to reduce the likelihood of true biased responding due to more true response options than false response options in previous formats. They used several self-report psychopathology measures as external validity indicators for the MMPI-2-RF scales.

The results of this investigation followed a similar pattern as those in the Cox et al. (2012) study. The expanded response option showed increased scale internal consistency that correlated with increases in variability of scale scores. The increases in reliability were not accompanied by consistent or meaningful increases in scale validity, when compared to the external validity indicators of psychopathology. But, scales with skewed distributions and a small number of low-frequency items showed the most consistent increases in reliability and validity, suggesting these scales may benefit from the augmented response format or item revision. Based on their findings, Finn et al. (2015) suggested that increased internal reliability was due to increased systematic variability in responding attributable to more opportunities for spurious patterns of responding, but that increases in scale reliability did not translate to more valid scale scores.

1.3. The current study

The purpose of the current investigation was to replicate and expand upon the results of Cox et al. (2012) using tighter controls. To do so, the entire MMPI-2-RF was administered twice, using the standard format and an augmented version. In addition to replicating the effects of augmenting the RC scales, this study attempted to show the effects of augmentation with all 42 substantive (non-validity) scales of the MMPI-2-RF. We hypothesized that replicating the findings of Cox et al. (2012)

for a majority of scales would provide strong evidence of the benefits of augmentation. Because some examinees were expected to become fatigued and begin to respond carelessly as a result of filling out 338 items twice, administering the entire MMPI-2-RF allowed screening out respondents based on the nine validity indicators.

The first set of analyses focused on reliability, using the same methods of Cox et al. (2012). The indices of reliability include indicators of internal consistency and the correlation between standard and augmented scales. Augmentation was expected to produce increases in both types of reliability, consistent with the findings Cox et al. (2012).

The effects of augmentation on convergent validity were explored with additional scales from the MPQ. These scales were selected based on the strength of their relationships with MMPI-2-RF scales in a college sample, as reported in the MMPI-2-RF technical manual (Tellegen & Ben-Porath, 2008). They were also selected so all the different types of MMPI-2-RF substantive scales, the Higher-Order (H-O), Psychopathology Five (PSY-5), RC, and Specific Problem (SP) scales, could be tested for enhancements in convergent validity. We expected the selected MMPI-2-RF scales would show enhanced convergent validity, as evidenced by higher correlations with appropriate MPQ scales for augmented scales than for standard scales.

The standard MMPI-2-RF and the experimental augmented version were administered sequentially, counterbalanced. Significant scale mean differences were examined between the standard form and augmented version of the MMPI-2-RF between groups by order of administration. A consistent pattern of significant differences would indicate order effects like those found by Cox et al. (2012).

Test proctors also recorded the administration time for each form to determine if one form took longer to complete than the other. Participants were expected to take slightly longer to finish the augmented version than to finish the standard form.

2. Method

2.1. Participants

A total of 527 undergraduate students attending Central Michigan University participated in this investigation. They were recruited from the psychology subject pool and received extra credit in one of their courses to compensate them for their participation. The data from 80 of these individuals were excluded from analysis due to failure to attend both experimental sessions. The remaining 447 participants were screened for invalid responding based on criteria found in the MMPI-2-RF manual. Table 1 displays these criteria, showing the number of participants who obtained scores beyond the acceptable range for each scale. The table also shows sample *z*-scores corresponding to the standard form exclusion criteria. These *z*-scores were used to calculate exclusion criteria for the augmented version (sample *z** [mean of augmented validity scale]). Scale scores for the augmented version of the Variable Response Inconsistency (VRIN) scale were calculated by reverse-coding one item for each item pair composing the scale, taking the difference, and ignoring negative values (to account for the unidirectional nature of scored item pairs).

The final sample contained 383 participants, of whom 196 received the standard form first and 187 received the augmented version first. Participants were mostly female (78%), Caucasian (89%), and about 20 years of age (Mean = 19.8; SD = 3.5; range = 18–48). Groups did not differ significantly ($p < .05$) on any of these characteristics.

Table 1. Participants excluded by screening criteria from the MMPI-2-RF on both standard and augmented versions

Standard form		Sample *z*-scores	Augmented version	
Exclusion criteria	**Excluded cases**		**Exclusion criteria**	**Excluded cases**
CNS ≥ 15	0	n/a	aCNS ≥ 15	0
VRIN ≥ 80	19	2.10	aVRIN ≥ 22	22
TRIN ≥ 80	13	2.48	n/a	n/a
F ≥ 120	10	3.30	aF ≥ 80	6
Fp ≥ 100	20	2.14	aFp ≥ 42	13
Fs ≥ 100	14	2.32	aFs ≥ 36	19
FBS ≥ 100	1	3.30	aFBS ≥ 93	0
L ≥ 80	2	3.12	aL ≥ 47	0
K ≥ 70	2	2.46	aK ≥ 54	1
All criteria	47	n/a	All criteria	38

Notes: Exclusion criteria for validity scales of the standard form are displayed in *T*-scores. All other criteria are displayed in raw scores.

2.2. Materials

2.2.1. Minnesota multiphasic personality inventory-second edition-restructured form

The 338-item MMPI-2-RF contains 51 scales including 11 validity scales, 3 Higher Order scales, 5 Psychopathology-5 scales, 9 Restructured Clinical scales, and 23 Specific Problem scales measuring somatic, cognitive, internalizing, externalizing, and interpersonal complaints (Ben-Porath & Tellegen, 2008). The standard response format of the MMPI-2-RF was designed around a dichotomous response format in which examinees were asked to indicate whether items are true or false as applied to them.

2.2.2. MMPI-2-RF augmented version

This version of the answer sheet and test booklet was identical to the ones used with the standard MMPI-2-RF, except for the response format. This format was same one used by Cox et al. (2012), in which examinees were asked to indicate whether each item is very true, mainly true, slightly true, or false, not at all true, as applied to them.

2.2.3. Selected multidimensional personality questionnaire scales

The full version of the MPQ contains 276 dichotomous items (mostly true/false, though some present a forced choice between two alternatives) used to assess normal range personality traits. The Wellbeing (Wb), Social Potency (Sp), Social Closeness (Sc), Stress Reaction (Sr), Alienation (Al), Aggression (Ag), and Absorption (Ab) scales of this instrument were selected to measure the effects of augmentation on convergent validity.

2.3. Procedure

Participants were tested in a classroom on the campus of Central Michigan University, no more than 30 participants at a time. During the first testing session, the general nature of the experiment was explained to them (i.e. exploring differences in response formats for personality tests), informed consent was obtained, and participants completed the selected scales of the MPQ. After one week, they returned to fill out both forms of the MMPI-2-RF in counterbalanced order, as well as a short demographics questionnaire.

cogent psychology

3. Results

3.1. Order effects

Independent samples *t*-tests indicated whether there were significant differences on the scales of the MMPI-2-RF between participants that completed the standard MMPI-2-RF first and those that completed the augmented version first. Because of the large number of significance tests being conducted on this set of scales, a more stringent criterion of $p < .01$ was used to determine statistical significance. Overall, four statistically significant group differences were found out of 100 group mean comparisons, and what differences were found were small ($d = .27 - .35$). Overall, the data do not support the hypothesized order effects. Therefore, subsequent data analysis was conducted on a combined sample rather than on subsets.

3.2. Completion time and acceptability

Completion time for each form was measured to the nearest minute. The mean time to complete the standard MMPI-2-RF was 30.4 min ($SD = 7.5$), and the mean time to complete the augmented version was 33.4 min ($SD = 7.2$). This difference is significant (t (382) = 7.13, $p < .05$), and of moderate magnitude ($d = .41$).

Participants were asked several questions related to their perceptions of the MMPI-2-RF forms. Participants were asked to rate each form on a scale of one (extremely difficult) to ten (extremely easy) regarding ease of use and ability to describe oneself accurately. Participants thought both forms were easy to fill out, though they believed the standard form was easier (mean ratings: Standard = 8.47, Augmented = 6.89; $t(382) = 12.40$, $p < .05$, $d = .78$). Participants also reported that both forms allowed them to describe themselves adequately, though they believed the augmented version was better suited to this task (mean ratings: Standard = 5.75, Augmented = 8.08; $t(382) = 16.54$, $p < .05$, $d = 1.08$). While the augmented version may be somewhat more difficult for participants to use, they were more satisfied with their self-descriptions using this form.

3.3. Internal consistency

Increases in reliability were examined using the same methods of Cox et al. (2012), comparing Cronbach's alphas and MIIC between the standard true/false MMPI-2-RF and the experimental augmented version. Higher internal consistency values for the augmented version would show that augmentation increased the scales' reliability. High correlations between standard and augmented scales would indicate that augmentation did not change the basic meaning of these scales.

Table 2. Coefficient alpha, *k*, and MIIC for MMPI-2-RF validity scales for standard and augmented versions within the combined sample

Scale	Coefficient alpha		*k*	Inter-item correlations	
	Standard (*N* = 383)	Augmented (*N* = 383)		Standard (*N* = 383)	Augmented (*N* = 383)
VRIN	.107	.235	2.56	.004	.009
TRIN	.228	–	–	.008	–
F	.728	.815	1.64	.073	.124
Fp	.322	.486	1.99	.016	.059
Fs	.371	.571	2.26	.039	.084
FBS	.680	.698	1.09	.069	.076
RBS	.451	.487	1.16	.028	.042
L	.479	.663	2.14	.058	.123
K	.643	.722	1.44	.112	.157
Mean	.472	.585	1.79	.045	.084
Median	.465	.617	1.82	.039	.080

Note: An augmented version of TRIN was not calculated.

Tables 2–4 display results for Cronbach's alpha and MIIC. The tables also display the *k* statistic, which is derived from the Spearman-Brown Formula. This statistic shows the value by which the item count of the standard scale theoretically would have to be multiplied to obtain a Cronbach's alpha equal to that of the augmented scale. For example, $k = 2$ would mean the item count of the standard scale would have to be doubled (assuming all the additional items are of similar psychometric quality to the original ones) for that scale to obtain the same Cronbach's alpha as the augmented version. A value of $k < 1$ means the reliability for the standard scale was higher than that of the augmented scale; the item count of the standard scale would have to be reduced to make it equal to that of the augmented scale.

Table 2 shows the augmented version displayed higher reliability for all the Validity scales. The median Cronbach's alpha increased by about .152, and MIIC increased by about .040. To match results of augmentation, the *k* statistic shows the item count of these scales would need to be increased by over 50%.

Table 3 shows the effects of augmentation on the reliability of the H-O scales, PSY-5 scales, and RC scales. As with the Validity scales, augmentation enhanced reliability in terms of Cronbach's alpha and MIIC. The effects were somewhat less dramatic, however, with H-O scales median Cronbach's

Table 3. Coefficient alpha, *k*, and MIIC for the MMPI-2-RF higher-order scales, psychopathology-five, and restructured clinical scales for standard and augmented versions within the combined sample

Scale	Coefficient alpha		*k*	Inter-item correlations	
	Standard (*N* = 383)	Augmented (*N* = 383)		Standard (*N* = 383)	Augmented (*N* = 383)
EID	.900	.919	1.26	.180	.224
THD	.681	.740	1.33	.069	.112
BXD	.732	.783	1.32	.109	.143
H-O mean	.771	.814	1.30	.119	.160
H-O median	.732	.783	1.32	.109	.143
AGGR	.762	.797	1.23	.150	.177
PSYC	.682	.752	1.41	.069	.112
DISC	.681	.719	1.20	.098	.114
NEGE	.769	.832	1.49	.141	.198
INTR	.824	.872	1.46	.193	.255
PSY-5 mean	.743	.794	1.36	.130	.171
PSY-5 median	.762	.797	1.41	.141	.177
RCd	.886	.917	1.42	.246	.321
RC1	.787	.825	1.28	.120	.156
RC2	.728	.797	1.47	.137	.190
RC3	.748	.799	1.34	.162	.209
RC4	.722	.774	1.32	.103	.139
RC6	.605	.692	1.47	.065	.121
RC7	.850	.882	1.32	.192	.238
RC8	.737	.775	1.23	.131	.167
RC9	.754	.786	1.20	.097	.116
RC mean	.757	.805	1.23	.139	.184
RC median	.748	.797	1.32	.131	.167

alpha increasing by only about .050 and median MIIC increasing by about .035. The item count of the standard scales would need to be increased by about one third to achieve these effects. The results for the PSY-5 scales mirrors the results observed for the H-O scales. Median Cronbach's alpha increased by about .035 and median MIIC increased by about .035. Replicating these effects with the standard scales would require increasing their item count by about two-fifths. The reliability analyses conducted on the RC scales replicates the reliability analyses conducted by Cox et al. (2012). Median Cronbach's alpha increased by approximately .050 and median MIIC increased by about .035. Based on *k*, the number of items in the standard scales would have to be increased by about one-fifth to one half (with a median of one-third) to achieve equivalent results. These increases are somewhat larger than those in the Cox et al. (2012) sample; in the Cox et al. (2012) combined sample, both median Cronbach's alpha and median MIIC increased by approximately .030.

Table 4 displays Cronbach's alpha, *k*, and MIIC for the Specific Problem (SP) and Interest scales. These scales seemed to benefit from augmentation slightly more than did the other substantive scales of the MMPI-2-RF. Median Cronbach's alpha increased by about .075, and median MIIC also increased by about .075. To achieve similar results, the item count of the standard scales would have

Table 4. Coefficient alpha, *k*, and MIIC for MMPI-2-RF specific problem scales for standard and augmented versions within the combined sample

Scale	Coefficient alpha		*k*	Inter-item correlations	
	Standard (*N* = 383)	Augmented (*N* = 383)		Standard (*N* = 383)	Augmented (*N* = 383)
MLS	.626	.671	1.22	.173	.205
GIC	.652	.759	1.68	.250	.379
HPC	.631	.695	1.37	.217	.290
NUC	.575	.620	1.21	.118	.144
COG	.695	.757	1.37	.181	.236
SUI	.722	.779	1.36	.375	.462
HLP	.533	.634	1.52	.208	.275
SFD	.768	.818	1.36	.453	.536
NFC	.755	.795	1.26	.254	.303
STW	.578	.629	1.24	.165	.198
AXY	.585	.628	1.20	.231	.266
ANP	.713	.765	1.31	.265	.315
BRF	.552	.650	1.51	.109	.170
MSF	.642	.669	1.13	.172	.184
JCP	.493	.583	1.44	.154	.205
SUB	.647	.733	1.50	.190	.276
AGG	.684	.760	1.46	.190	.257
ACT	.656	.718	1.34	.193	.240
FML	.638	.733	1.56	.151	.222
IPP	.764	.803	1.26	.249	.290
SAV	.836	.863	1.23	.348	.393
SHY	.736	.798	1.42	.283	.367
DSF	.423	.484	1.28	.126	.172
AES	.594	.647	1.25	.176	.208
MEC	.630	.659	1.13	.173	.200
Mean	.645	.706	1.34	.216	.272
Median	.642	.718	1.34	.190	.266

Table 5. Correlations between standard and augmented higher order, psychopathology five, restructured clinical scales, and specific problems in the combined sample

H-O scales		PSY-5 scales		RC scales		SP scales			
Scale	***r***	**Scale**	***r***	**Scale**	***r***	**Scale**	***r***	**Scale**	***r***
EID	.918	AGGR	.802	RCd	.895	MLS	.763	JCP	.813
THD	.774	PSYC	.781	RC1	.806	GIC	.832	SUB	.886
BXD	.857	DISC	.869	RC2	.743	HPC	.793	AGG	.801
		NEGE	.840	RC3	.786	NUC	.659	ACT	.749
		INTR	.762	RC4	.881	COG	.820	FML	.804
				RC6	.750	SUI	.870	IPP	.786
				RC7	.859	HLP	.748	SAV	.814
				RC8	.825	SFD	.784	SHY	.854
				RC9	.777	NFC	.814	DSF	.692
						STW	.739	AES	.866
						AXY	.772	MEC	.879
						ANP	.819		
						BRF	.797		
						MSF	.827		
Mean	.850	Mean	.811	Mean	.814			Mean	.799
Median	.857	Median	.802	Median	.806			Median	.804

Notes: H-O = Higher Order. PSY-5 = Psychopathology-Five. RC = Restructured Clinical. SP = Specific Problem.

to be increased by about two-fifths. The hypothesis that augmentation would improve scale score reliability as measured by coefficient Cronbach's alpha and MIIC appears to be supported by the data.

3.4. Cross-version reliability

Table 5 shows the relationships between the standard and augmented scales in the combined sample. On average, the original scales and their counterparts were very strongly correlated (mean and median *r*s > .799) and comparable to the results of Cox et al. (2012). The previous investigation found a mean correlation of *r* = .833 for the RC scales, whereas the mean correlation was *r* = .814 for the RC scales in the present sample.

Some idiosyncrasies in the data were observed. Longer scales such as the H-O scales tended to correlate more highly than did shorter ones, such as the SP scales. The weakest of these

Table 6. Correlations between MMPI-2-RF and MPQ scales for the standard and augmented versions by order of administration and for the combined sample

Scales	Combined sample	
	Standard (*N* = 383)	**Augmented (*N* = 383)**
THD & Al	.365	.400
NEGE & Sr	.728	.724
RC2 & Wb*	−.651	−.635
RC8 & Ab	.519	.505
AGG & Ag	.667	.623
IPP & Sp*	−.583	−.567
SAV & Sc*	−.583	−.584
Mean	.585	.577
Median	.583	.584

Notes: All relationships are significant ($p < .05$). Means and medians are displayed in absolute magnitudes.

*The hypothesized relationship is negative.

relationships, however, was $r = .659$ between standard and augmented versions of NUC, which is still a fairly strong relationship in this context. Overall, the hypothesis that augmentation would not substantially change the meaning of scales was supported by these data.

3.5. Convergent validity

Due to the large number of comparisons being made, $p < .01$ was used as the criterion for statistical significance. Cohen's *d* was used to quantify the effect sizes of these mean differences, which were expected to be small to medium. Table 6 shows the relationships between these scale pairs. The results of these analyses appear to be the exact opposite of what was hypothesized. On average, the standard MMPI-2-RF scales correlated more strongly with the MPQ than did the augmented scales. The standard MMPI-2-RF scales correlated more strongly with the MPQ in five out of seven pairs.

4. Discussion

The results of this study were mixed. It seemed clear that the data did not support the hypothesized order effects, contrary to findings of Cox et al. (2012). It was also clear the augmented scales took slightly longer to complete, but that participants felt the augmented response format allowed them to describe themselves more accurately. The primary psychometric effects of augmentation, however, remain a concern. In this sample, the data showed increases in reliability with augmentation, with increases in internal consistency found for most scales, similar to the findings of Cox et al. (2012) and Finn et al. (2015). Furthermore, high correlations between standard and augmented versions of the same scales suggest construct equivalence, consistent with Cox et al. (2012). However, these same scale scores for which reliability increased failed to show increases in convergent validity, mirroring combined group results from Cox et al. (2012) and results from Finn et al. (2015). This finding is contrary to theoretical expectations, as well as earlier simulated and *in vivo* research. At least two different competing explanations may account for these effects, and given the limitations of this study, we cannot conclude which one, if either, is correct. The following discussion is intended to point out potential limitations in this study to guide future research in this area and to suggest methods of clarifying ambiguous findings.

With regard to reliability, it is possible that augmentation increased method variance more than it did variance attributable to the actual traits the scales were designed to measure. If true, then these scales may have appeared more reliable due to the increase in item covariance, but they actually became less valid due to the increased proportion of error variance relative to trait variance. More directly, we argue the dichotomous MPQ scales alone may not have been appropriate for testing the effects of augmentation on convergent validity. Because these scales used the same dichotomous true/false response format as did the standard scales, responses to them would be expected to generate method variance common to both of them, inflating their correlations spuriously based on a shared source of error. From a different perspective, the polytomous format would be expected to generate method variance not shared by the scales of the MPQ (e.g. response extremeness, as discussed by Peabody, 1962), putting these scales at a disadvantage in demonstrating convergent validity relative to those of the standard form. Future studies may be able to explore the possibility of confounding common method variance by manipulating the response format of external validity measures as well (e.g. through an augmented MPQ).

Given the difficulty associated with trying to compensate for the effects of method variance on reliability, it may be more fruitful to focus on estimating convergent validity more appropriately. If it could be demonstrated that augmentation improves validity, the concerns about reliability may be a moot point. Pursuing this strategy would not quantify how much of the increased variance in augmented scales is due to the method and how much is due to the trait. Nevertheless, increases in convergent validity as a result of augmentation would provide strong evidence that a greater proportion of this variance is due to accurate trait measurement, rather than error.

Other procedures may be employed to deal with estimating convergent validity. Structured interviews assessing relevant traits may be individually administered and scored to obtain data usable for

testing convergent validity. Participants could also be trained to make daily ratings of subjective distress over the course of a week, and total scores could be correlated with the scales of distress on the MMPI-2-RF (EID, RCd, etc.). Any number of behaviors could be tracked over time, such having participants keep count of the number of drinks they have and correlating the tracked behavior with RC4 and SUB. With respect to clinical populations, scale means could be compared based on diagnosis or relevant historical variables (e.g. number of suicide attempts correlated with SUI). All of these alternative procedures present unique challenges of their own. Structured interviews tend to be time consuming, requiring participants to keep track of something may be too demanding or generate data of dubious validity (e.g. some would probably fill out all the data retrospectively at the end of the week), and issues related to confidentiality and availability often make clinical populations difficult to access. Of course, every method of data collection used to test convergent validity will have some drawbacks. It is beyond the scope of this study to recommend a single method that will be ideal in all situations. Rather, it is expected that researchers will weigh the costs and benefits of each compared with the resources available to them. Additionally, future studies may consider data sources outside of the domain of self-report, through the use of collateral informants or record review.

Beyond the methodological limitations discussed above related to reliability and validity, other issues concerning augmentation remain unclear and may be fruitful to explore in future research. Only one type of experimental response format was tested in this study, and it is possible that other formats may be more appropriate for measuring psychopathology. Due to the wording of the items, the decision was made to retain the true/false scaling for which they were made. Other formats, such as agree/disagree or frequency estimates (e.g. always, often, sometimes, never), may be just as proper for the purposes of the MMPI-2-RF and might yield even better results.

It is also unclear exactly how many response options are ideal for tests of this kind. The studies reviewed in preparation for this investigation varied in their recommendations, though there was clear consensus in recommending more than two (e.g. Lozano et al., 2008). It may be useful to test several response formats in a single study to compare their psychometric quality. Due to the wide variation in the literature, however, it will probably be necessary to conduct this type of study several times with different samples to ensure the results replicate, as the optimal number of response options has not been consistent across instruments or settings. A middle response option was deliberately excluded from this study, because it was believed to introduce an additional variable not represented in the standard scales beyond the simple number of response options, and we were concerned that it might amplify central tendency bias (Nunnally, 1978). Although there are no obvious reasons that a middle option would improve the psychometric quality of a scale, this question remains open to investigation.

Finally, there are other issues not tested in this study related to the psychometric functioning of the augmented scales. Although it would have been possible to investigate factor structure using these data, in light of the failure of the augmented scales to improve convergent validity, this task seemed premature. If augmentation cannot be shown to produce more valid scores, then its effect on factor structure is only trivial. There are other psychometric properties relevant to augmentation that went unexamined in this study, such as retest reliability and discriminant and predictive validity. Again, while these qualities are important, augmented scales with stronger convergent relationships should be demonstrated, at a minimum, before these other properties are examined in any depth. Without this basic indication of improvement, it is unlikely the augmented scales would be any better in these other areas, which present more difficult methodological (e.g. multiple testing sessions) or interpretive challenges (examining a correlation matrix for relationships that should be absent). If augmentation improves convergent validity, then it will be necessary to examine their psychometric properties along a variety of dimensions. In addition, although the augmented scales correlated strongly with their standard counterparts, they did not correlate perfectly. If the MMPI-2-RF benefits from a polytomous response format, it will be necessary to re-examine some of its key correlates to better understand the construct validity of the instrument as a whole.

Acknowledgements
We would like to thank the University of Minnesota Press for permission to reproduce the items of the Minnesota Multiphasic Personality Inventory-Second Edition-Restructured Form (MMPI-2-RF) for this study. We would also like to thank Brian McCabe and Joseph McLaughlin for their assistance with data collection and entry. The lead author would like to thank his other dissertation committee members, Kyunghee Han and Stuart Quirk, for their instruction and helpful suggestions in crafting and refining this project.

Funding
The authors received no direct funding for this research.

Competing Interests
The authors declare no competing interest.

Author details
Andrew Cox[1]
E-mail: andycox81@gmail.com
Seth C. Courrégé[2]
E-mail: seth.courrege@gmail.com
ORCID ID: http://orcid.org/0000-0001-5115-3152
Abigail H. Feder[2]
E-mail: helbl1qj@cmich.edu
Nathan C. Weed[2]
E-mail: weed1nc@cmich.edu
ORCID ID: http://orcid.org/0000-0003-4195-8074
[1] Dominion Behavioral Healthcare, Richmond, VA, USA.
[2] Department of Psychology, Central Michigan University, Mount Pleasant, MI, USA.

Citation information
Cite this article as: Effects of augmenting response options of the MMPI-2-RF: An extension of previous findings, Andrew Cox, Seth C. Courrégé, Abigail H. Feder & Nathan C. Weed, *Cogent Psychology* (2017), 4: 1323988.

References

Ben-Porath, Y. S., & Tellegen, A. (2008). *Minnesota multiphasic personality inventory-2 restructured form: Manual for administration, scoring, and interpretation*. Minneapolis, MN: University of Minnesota Press.

Butcher, J. N., Dahlstrom, W. G., Graham, J. R., Tellegen, A., & Kaemmer, B. (1989). *Minnesota multiphasic personality inventory-2 (mmpi-2): Manual for administration and scoring*. Minneapolis, MN: University of Minnesota Press.

Butcher, J. N., Graham, J. R., Ben-Porath, Y. S., Tellegen, A., Dahlstrom, W. G., & Kaemmer, B. (2001). *MMPI-2 (minnesota multiphasic personality inventory-2): Manual for administration, scoring, and interpretation* (revised ed.). Minneapolis, MN: University of Minnesota Press. https://doi.org/10.1016/B0-08-043076-7/01294-8

Cicchetti, D. V., Shoinralter, D., & Tyrer, P. J. (1985). The effect of number of rating scale categories on levels of interrater reliability : A Monte Carlo investigation. *Applied Psychological Measurement, 9*, 31–36. doi:10.1177/014662168500900103

Cox, A., Pant, H., Gilson, A. N., Rodriguez, J. L., Young, K. R., Kwon, S., & Weed, N. C. (2012). Effects of augmenting response options on MMPI–2 RC scale psychometrics. *Journal of Personality Assessment, 94*, 613–619. doi:10.1080/00223891.2012.700464

Downey, J. E. (1921). The will profile. A tentative scale for measurement of the volitional pattern. *University of Wyoming Bulletin, 16*, 1–40.

Finn, J. A., Ben-Porath, Y. S., & Tellegen, A. (2015). Dichotomous versus polytomous response options in psychopathology assessment: Method or meaningful variance? *Psychological Assessment, 27*, 184–193. doi:10.1037/pas0000044

Galton, F. (1883). Inquiries into human faculty and its development, 1883. In W. Dennis (Ed.), *Readings in the history of psychology* (pp. 277–289). East Norwalk, CT: Appleton-Century-Crofts.

Graham, J. R., Watts, D., & Timbrook, R. E. (1991). Detecting fake-good and fake-bad MMPI-2 profiles. *Journal of Personality Assessment, 57*, 264–277. doi:10.1207/s15327752jpa5702_6

Hathaway, S. R., & McKinley, J. C. (1940). A multiphasic personality schedule (Minnesota): I. construction of the schedule. *The Journal of Psychology, 10*, 249–254. doi:10.1080/00223980.1940.9917000

Jenkins, Jr., G. D., & Taber, T. S. (1977). A Monte Carlo study of factors affecting three indices of composite scale reliability. *Journal of Applied Psychology, 62*, 392–398. doi:10.1037//0021-9010.62.4.392

Komorita, S. S., & Graham, W. K. (1965). Number of scale points and the reliability of scales. *Educational and Psychological Measurement, 25*, 987–995. doi:10.1177/001316446502500404

Likert, R. (1932). A technique for the measurement of attitudes. *Archives of Psychology, 22*, 5–55.

Lissitz, R. W., & Green, S. B. (1975). Effect of the number of scale points on reliability: A Monte Carlo approach. *Journal of Applied Psychology, 60*, 10–13. doi:10.1037/h0076268

Lozano, L. M., García-Cueto, E., & Muñiz, J. (2008). Effect of the number of response categories on the reliability and validity of rating scales. *Methodology, 4*, 73–79. doi:10.1027/1614-2241.4.2.73

Morey, L. C. (1991). *Personality assessment inventory – Professional manual*. Lutz, FL: Psychological Assessment Resources.

Nunnally, J. C. (1978). *Psychometric theory* (2nd ed.). New York, NY: McGraw-Hill Book Company.

Peabody, D. (1962). Two components in bipolar scales: Direction and extremeness. *Psychological Review, 69*, 65–73. doi:10.1037/h0039737

Pearson, K. (1907). On the relationship of intelligence to size and shape of head, and to other physical and mental characters. *Biometrika, 5*, 105–146. doi:10.2307/2331650.

Plant, J. S. (1922). Rating scheme for conduct. *American Journal of Psychiatry, 78*, 547–572. doi:10.1176/ajp.78.4.547

Preston, C. C., & Colman, A. M. (2000). Optimal number of response categories in rating scales: reliability, validity, discriminating power, and respondent preferences. *Acta Psychologica, 104*(1), 1–15. doi:10.1016/s0001-6918(99)00050-5

Rogers, R., Sewell, K. W., Martin, M. A., & Vitacco, M. J. (2003). Detection of feigned mental disorder: A meta-analysis of the MMPI-2 and malingering. *Assessment, 10*, 160–177. doi:10.1177/1073191103010002007

Sellbom, M., & Ben-Porath, Y. S. (2005). Mapping the MMPI-2 restructured clinical scales onto normal range personality traits: Evidence of construct validity. *Journal of Personality Assessment, 85*, 179–187. doi:10.1207/s15327752jpa8502_10

Spector, P. E. (1992). *Summated rating scale construction: An introduction*. Newbury Park, CA: Sage Publications. https://doi.org/10.4135/9781412986038

Streiner, D. L. (2003). Starting at the beginning: An introduction to coefficient alpha and internal consistency. *Journal of Personality Assessment, 80*, 99–103. doi:10.1207/s15327752jpa8001_18

Symonds, P. M. (1924). On the loss of reliability in the ratings due to coarseness of the scale. *Journal of Experimental Psychology, 7*, 456–461. doi:10.1037/h0074469

Tellegen, A., & Ben-Porath, Y. S. (2008). *Minnesota multiphasic personality inventory-2 restructured form: Technical manual*. Minneapolis, MN: University of Minnesota Press.

Tellegen, A., Ben-Porath, Y. S., McNulty, J. L., Arbisi, P. A., Graham, J. R., & Kaemmer, B. (2003). *The MMPI-2 restructured clinical (RC) scales: Development, validation and interpretation*. Minneapolis, MN: University of Minnesota Press.

Tellegen, A., & Waller, N. G. (2008). Exploring personality through test construction: Development of the multidimensional personality questionnaire. In G. J. Boyle, G. Matthews, & D. H. Saklofske (Eds.), *Handbook of personality theory and testing: Personality measurement and assessment* (vol. 2, pp. 261–292). London, United Kingdom: Sage. https://doi.org/10.4135/9781849200479

Webb, E. (1915). Character and intelligence. *British Journal of Psychology Monographs, 1*, 99.

cogent oa

© 2017 The Author(s). This open access article is distributed under a Creative Commons Attribution (CC-BY) 4.0 license.

You are free to:
Share — copy and redistribute the material in any medium or format
Adapt — remix, transform, and build upon the material for any purpose, even commercially.
The licensor cannot revoke these freedoms as long as you follow the license terms.

Under the following terms:
Attribution — You must give appropriate credit, provide a link to the license, and indicate if changes were made.
You may do so in any reasonable manner, but not in any way that suggests the licensor endorses you or your use.
No additional restrictions

You may not apply legal terms or technological measures that legally restrict others from doing anything the license permits.

***Cogent Psychology* (ISSN: 2331-1908) is published by Cogent OA, part of Taylor & Francis Group.**

Publishing with Cogent OA ensures:

- Immediate, universal access to your article on publication
- High visibility and discoverability via the Cogent OA website as well as Taylor & Francis Online
- Download and citation statistics for your article
- Rapid online publication
- Input from, and dialog with, expert editors and editorial boards
- Retention of full copyright of your article
- Guaranteed legacy preservation of your article
- Discounts and waivers for authors in developing regions

Submit your manuscript to a Cogent OA journal at www.CogentOA.com