Deskovitz et al., *Cogent Psychology* (2016), 3: 1161287
http://dx.doi.org/10.1080/23311908.2016.1161287

cogent psychology

CLINICAL PSYCHOLOGY & NEUROPSYCHOLOGY | RESEARCH ARTICLE

Interpretive reliability of two common MMPI-2 profiles

Mark A. Deskovitz[1], Nathan C. Weed[1], Cheryl Chakranarayan[1]* and John E. Williams[2]

Received: 25 October 2015
Accepted: 01 March 2016
First Published: 09 March 2016

*Corresponding author: Cheryl Chakranarayan, Department of Psychology, Central Michigan University, Mt. Pleasant, MI, USA
E-mail: chakr1c@cmich.edu

Reviewing editor:
Peter Walla, University of Newcastle, Australia

Additional information is available at the end of the article

Abstract: Users of multi-scale tests like the MMPI-2 tend not to interpret scales one at a time in a way that would correspond to standard scale-level reliability information. Instead, clinicians integrate inferences from a multitude of scales simultaneously, producing a descriptive narrative that is thought to characterize the examinee. This study was an attempt to measure the reliability of such integrated interpretations using a q-sort research methodology. Participants were 20 MMPI-2 users who responded to E-mail solicitations on professional listservs and in personal emails. Each participant interpreted one of two common MMPI-2 profiles using a q-set of 100 statements designed for MMPI-2 interpretation. To measure the "interpretive reliability" of the MMPI-2 profile interpretations, q-sort descriptions were intercorrelated. Mean pairwise interpretive reliability was .39, lower than expected, and there was no significant difference in reliability between profiles. There was also not a significant difference between within-profile and cross-profile correlations. Q-set item analysis was conducted to determine which individual statements had the most impact on interpretive reliability. Although sampling in this study was limited, implications for the field reliability of MMPI-2 interpretation are sobering.

Subjects: Diagnostic Practice & Assessment; Personality Tests & Assessments; Psychometrics/Testing & Measurement Theory

Keywords: MMPI-2; reliability; code types; q-sort

ABOUT THE AUTHORS

The co-authors of this project are members of the Psychological Assessment Laboratory at Central Michigan University. Led by Drs Kyunghee Han and Nathan Weed, 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, one of the most widely used psychological tests 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 master's thesis of the first author, Mark A. Deskovitz, PhD, affiliated with Partners in Change, LLC, and Central Michigan University.

PUBLIC INTEREST STATEMENT

The MMPI-2 is one of the most widely used psychological tests in the world, employed in a variety of clinical and occupational settings. It 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 the "field reliability" of MMPI-2 interpretations, that is, the degree to which practicing clinicians agree on the meaning of results from the MMPI-2. Results were somewhat disappointing, in that clinicians did not achieve an impressive consensus in interpreting the MMPI-2, even though the profiles were chosen to represent "classic" personality patterns with which test users should be familiar. The sample size of this study was modest, and it is unclear how well the results generalize to all practitioners, but more study of this issue is plainly indicated.

1. Introduction

Reliability is a fundamental property of measurement. Consequently, a great deal of research has focused on the reliability of clinical measurements, usually taken as the property of a set of scores on a single measuring scale. However, users of multi-scale tests like the MMPI-2 tend not to interpret scales one at a time in a way that would correspond to standard scale-level reliability research. Instead, clinicians often integrate inferences from a multitude of scales simultaneously and produce a descriptive narrative that is thought to characterize the examinee. Therefore, when evaluating the reliability of such instruments, one should consider not only the reliability of scores on independent scales, but also the reliability of this integrated descriptive account. The reliability of this integrative product, however, has not often been examined, probably primarily for lack of convenient methodology. The present study demonstrates evaluation of what we call the "interpretive reliability" of results from a major clinical assessment instrument, the MMPI-2, using a q-sort methodology.

Like most multi-scale personality inventories, the MMPI-2 is typically interpreted by integrating information inferred from scores on multiple scales. Presumably, inferences that are pointed to by multiple scales are assigned greater weight in the interpretation; contradictory inferences are downplayed. Very little, however, is known about the process by which integrated descriptive accounts are assembled by test users. Further, what little we understand about clinical judgment is underscored by a keen awareness of its limitations, including errors of adjustment, anchoring, and hypothesis confirmation, and cognitive biases (Garb, 1998). Because clinicians are generally thought to have little awareness of their cognitive processes or biases, they cannot confront or counteract the biases. Complicating the picture is that clinicians seem resistant to modify judgment processes in the face of new empirical information. Moreover, clinicians have trouble learning from experience because feedback often occurs several weeks or months following their decisions.

Given these heuristic errors and biases, one might question how accurately clinicians can perform a cognitively complex task such as test interpretation. In fact, we know little about both the processes and the accuracy of clinical judgments like those involved in MMPI-2 interpretation. Part of the reason, perhaps, is that it is difficult to operationalize clinical interpretation. Many clinical interpretations of test results come in the form of a written narrative, a form whose accuracy is difficult to appraise. The section below describes early efforts to operationalize and evaluate the accuracy of clinical interpretation.

1.1. Research on the validity of clinical test interpretation

In 1956 Meehl, cognizant of the potential for clinical error, theorized that formal methods for test interpretation would be superior to clinical methods of interpretation. Meehl described clinical inferences from personality data as being based on a "rule of thumb" method. When a clinician interprets test data, experience and skill play a role in determining personality descriptions. Thus, the clinician uses judgment to make inferences and thereby produce a descriptive narrative of the individual. In contrast, Meehl (1956) proposed use of a "cookbook" method. According to this method, any given configuration of psychometric data is associated statistically with facets of a personality description, and the closeness of this association is quantified. He argued that personality description would be improved as an automatic, mechanical, and clerical task that proceeds according to explicit rules set forth in the cookbook.

In collaboration with Meehl, Halbower (1955) developed a strategy for testing Meehl's hypothesis using q-sort descriptions. First, a clinician completed a q-sort description of a patient treated at the Minneapolis VA based on careful study of the case folder, including therapist notes and all available psychometrics except the MMPI. Halbower then constructed a miniature "cookbook" of MMPI interpretation corresponding to that patient by identifying a sample of nine patients with the same MMPI code type. A therapist then completed a Q sort based on each of these nine files. Pairwise correlations between these nine patients were then examined, and five "modal" patients were selected from this matrix by an inspectional cluster method. The mean q-sort on these five "core" patients was taken as the "cookbook" MMPI interpretation for the original patient. Halbower then produced a clinical “rule of thumb” interpretation, using an experienced clinician to interpret the patient’s MMPI directly. The two kinds of q-sorts were then each correlated with the q-sort produced from the patient’s file. In his sample, Halbower found that not one of the “rule of thumb” Q sorts were as valid as the “cookbook” Q sorts he created. The mean q-correlation was .78 for the cookbook method as compared to .48 for the rule of thumb method.

In addition to demonstrating some of the advantages that formal “cookbook” test interpretation has over clinical “rule of thumb” interpretation, Halbower showed that the q-sort technique is a promising methodology for the study of personality test interpretation in general. As it permits clinicians to describe patients using the data from a personality test that is most relevant for any given inference, it seems an ideal method for operationalizing the descriptive narrative of examinees that multi-scale test interpretation produces. This q-sort method will be described more thoroughly in the following section.

1.2. The q-sort technique

The q-sort procedure requires a judge to sort a set of items or statements into ordered categories, ranging from extremely characteristic or salient to extremely uncharacteristic (Ozer, 1993). This judgment is made with reference to some specific target. The categories into which the items are sorted are assigned a numerical value that becomes the score for the descriptive item. The number of items permitted in each category is set in advance, so the shape of the distribution is fixed and constant for all judges using the instrument. The reliability of the ratings on a q-sort instrument can be estimated using inter-judge agreement between q-sorts: the vector of scores produced on each item by one judge can be correlated with the vector of item scores by another judge.

The q-sort method has been used in a number of studies on the reliability of personality assessments, most quite long ago. For example, a study by Friedman (1957) employed the q-sort technique to estimate interrater reliability of ratings of TAT heroes. Using an 80-item q-set developed especially for TAT hero description, and 3 experienced TAT users for each of 10 TAT protocols, he found interrater q-correlations to range from .37 to .88.

Moos (1962) evaluated the effect of assessment training on the accuracy of test interpretation. As part of the study, three “experienced clinicians” each described two MMPI profiles using the q-set created by Halbower (1955) and also a q-set developed especially for interpreting Rorschach test results. Across both q-sets and MMPI profiles, MMPI interpreter q-intercorrelations ranged from .37 to .71. Similar results based on the interpretation of the Rorschach ranged from .13 to .48.

Within a study by Little and Shneidman (1954) on the validity of MMPI interpretations, 11 psychologists used a 150-item q-sort of “the usual aspects of psychological functioning presented in psychological reports” to interpret the MMPI of a single patient. Though not a focus of the study, the intercorrelations between the 11 q-descriptions were found to range from .42 to .73.

1.3. Present study

Oddly, few studies of this variety have been conducted in more recent years. Examination of the reliability of clinical interpretations has largely given way to estimates of the reliability of scores on a single scale or measure associated with clinical instruments. The present study was inspired by these classic studies of interpretive reliability and represents the first application of q-methodology to the MMPI-2, the revised version of the MMPI examined in the studies described above. As in Little and Shneidman (1954), we use q-sort interpretation to operationalize the integrative interpretation that characterizes use of the MMPI-2. Test users were recruited to perform q-sort interpretations of one of two common but distinct MMPI-2 profiles, and the resulting q-sorts were compared across test interpreters by means of q-correlation. Of particular note was whether interpretive reliability of these two profiles was sufficient to distinguish interpretation of one profile from the other.

2. Method

2.1. Participants

Participants in the study were 20 practitioners who indicated they were regular users of the MMPI-2. Although there are no formal surveys of MMPI-2 use by professional specialty, at least one expert (J. Butcher, personal communication, May 2002) is of the opinion that some 95% of MMPI-2 use is by psychologists. Consequently, participants were recruited from several professional E-mail listservs frequented by psychologists. The following listservs were targeted: Clinical Psychology and Counseling Psychology Diplomates of the American Board of Professional Psychology (ABPP); Divisions 5, 12, 17, 29, and 42 of the American Psychological Association (APA); and (APAGS), (NEWPSYCH), (MENTORS), and (PSYCHGRAD) of the APA of Graduate Students. Further, members of the Society of Personality Assessment (SPA) were individually solicited via direct E-mail. Also, an effort was made to recruit attendees of an MMPI-2 Research Symposium. Finally, personal contacts were made to peers and colleagues of the investigators.

On each listserv, an e-mail message was posted offering regular users of the MMPI-2 the opportunity to participate in a study of the reliability of the MMPI-2. To compensate participation, those taking part in the study were offered feedback about how consensually they interpreted the MMPI-2 (i.e. how closely the interpretation matches “expert” interpretation). Those volunteering to participate were directed to a web site where the study was conducted.

2.2. Instruments

The instrument used in this study to operationalize MMPI-2 interpretation was the Midwestern q-sort. A collection of 100 statements designed to reference behavior patterns assessed by the MMPI-2, the initial version of the Midwestern q-sort was developed as an aid to teach MMPI-2 interpretation (Weed, 2006). The items were written based on an inspection of descriptive statements listed as typical scale correlates in Graham (2000). The initial q-set consisted of 73 items. Later, McNeal (2000) combined Weed’s initial q-set with a q-set developed by Marks and Seeman (1963). The new list of items consisted of 169 statements. Review of the combined pool of items resulted in the removal of 30 items that were judged to be unclear or irrelevant and an additional 39 items that were considered redundant. In a third revision, McNeal (2000) subjected the 100 items to an “item by item examination of the correlations between q-sort descriptions of the same individual created by their therapist based on personal knowledge of the individuals.” Low correlations of less than .10 with external criteria were obtained for 33 items that were altered or replaced. In addition, items were rewritten to reflect observable behaviors rather than presumed internal processes or states, compound items were simplified, and wording of the q-set was simplified to allow for use by sorters who are not health care professionals.

A Windows-based computer program developed by Noland and Weed (1994) was used to facilitate the q-sort process. The computer program presented a split screen with a list of interpretive statements on the right side and seven numbered bins on the left side. The size of each bin was fixed and indicated the exact number of items to be placed in each. The bins were consecutively labeled one through seven, with respective anchors of “least descriptive” and “most descriptive.” The bins were arranged to approximate a normal distribution, holding 5, 10, 20, 30, 20, 10, and 5 items, in order. Participants were able to select a statement and drop it into one of the seven numbered bins. As each statement was selected and placed in a bin, the statements on the right side disappeared until all items were sorted. The program alerted the participant if a bin was not completely filled or if too many items were placed in a bin.

To address concerns about the program’s utility with a variety of operating systems, and in an effort to facilitate broader use of the program, Williams and Weed (2003) developed a web adaptation they called the Midwestern q-sort, which was used in the present study.

2.3. Procedures

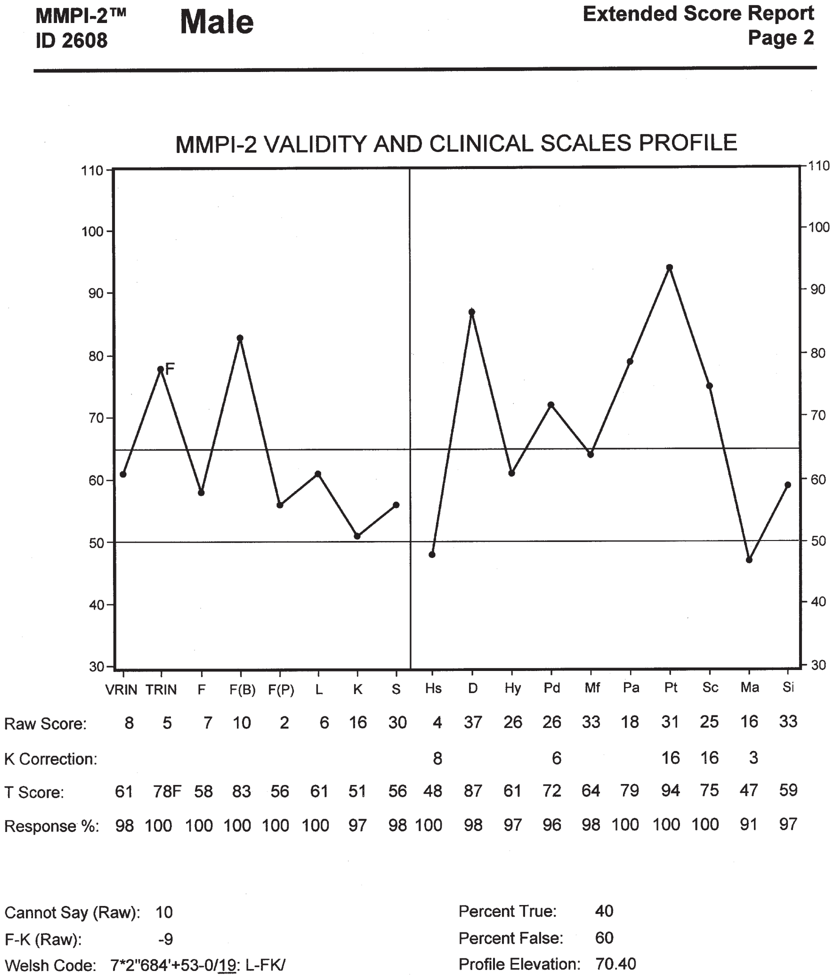
Interested participants were directed to a web site on which the study was conducted. A flowchart of the study's procedures is presented in Table 1. First, the participants were presented with a description of the study, informed consent was obtained, and the clinicians decided whether to participate. At this time, the clinicians were instructed to download a free version of Authorware, the web software on which the Midwestern q-sort ran. After the download was complete, demographic

Table 1. Matrix of Q-correlations between interpretations of common profiles

Participant	1	2	3	4	5	6	7	8	9	10	*M* 6,8
1	–	.48	.49	.12	.40	.08	.38	.40	.31	.34	.45
2	.60	–	.61	.22	.64	.42	.56	.33	.47	.45	.73
3	.26	.43	–	.27	.45	.38	.40	.20	.52	.45	.55
4	.21	.25	.38	–	.27	.28	.13	−.16	.37	.38	.33
5	.45	.55	.30	.19	–	.20	.60	.36	.36	.45	.69
6	.40	.52	.26	.04	.51	–	.27	−.02	.36	.50	.50
7	.55	.56	.37	.26	.70	.43	–	.37	.32	.38	.57
8	.37	.49	.32	.19	.37	.45	.38	–	.11	.27	.28
9	.45	.54	.22	.12	.58	.61	.53	.49	–	.52	.55
10	.41	.56	.10	.03	.49	.54	.50	.45	.63	–	.67
M 2,7	.29	.35	.51	.43	.48	.25	.56	.14	.33	.26	–

Note: Correlations below the diagonal are for the 2,7 profile; correlations above the diagonal are for the 6,8 profile.

Figure 1. MMPI-2 data interpreted by participants: 2,7 profile.



MMPI-2™
ID 2608

Male

Extended Score Report
Page 2

MMPI-2 VALIDITY AND CLINICAL SCALES PROFILE

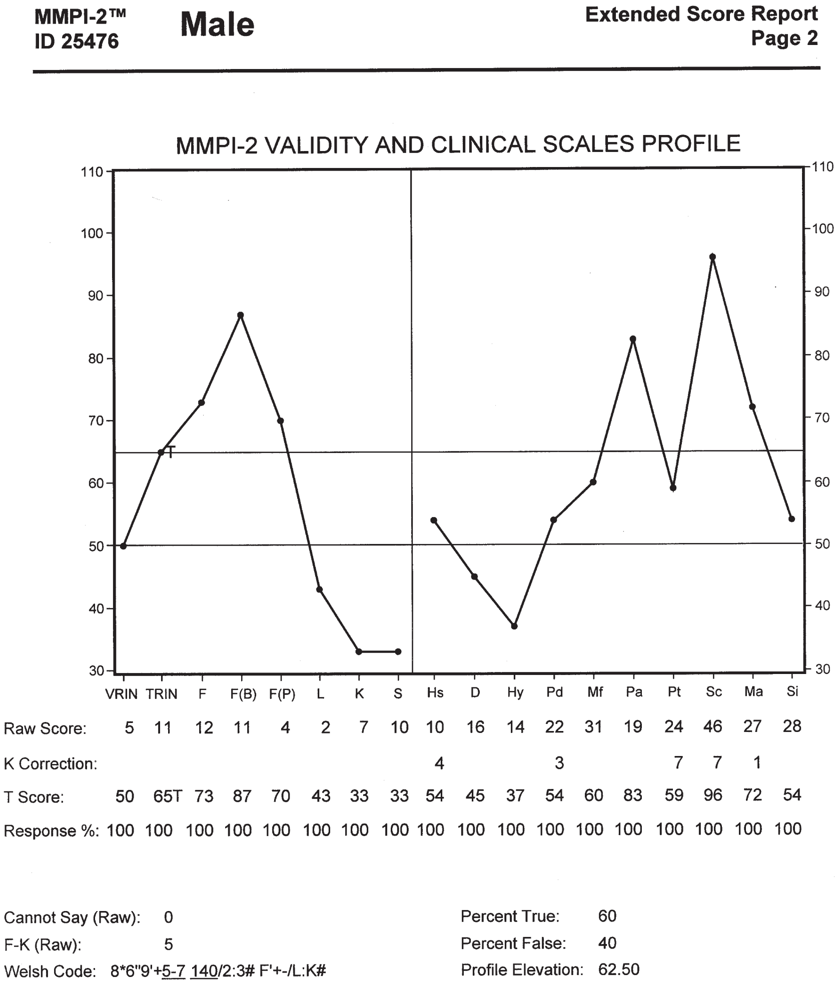
	VRIN	TRIN	F	F(B)	F(P)	L	K	S	Hs	D	Hy	Pd	Mf	Pa	Pt	Sc	Ma	Si
Raw Score:	8	5	7	10	2	6	16	30	4	37	26	26	33	18	31	25	16	33
K Correction:									8			6			16	16	3	
T Score:	61	78F	58	83	56	61	51	56	48	87	61	72	64	79	94	75	47	59
Response %:	98	100	100	100	100	100	97	98	100	98	97	96	98	100	100	100	91	97

Cannot Say (Raw):	10	Percent True:	40
F-K (Raw):	-9	Percent False:	60
Welsh Code:	7*2"684'+53-0/19: L-FK/	Profile Elevation:	70.40

information was gathered, and directions on the use of the Midwestern q-sort were given. The clinicians were then randomly presented with one of two MMPI-2 profiles containing validity scales and K-corrected clinical scales: (1) a profile characterized principally by extreme elevations on clinical scales 7 and 2, but also with clinically elevated scores on scales 4, 6, and 8 (see Figure 1); or (2) a profile characterized principally by elevations on clinical scales 8 and 6, but also with elevations on scale 9 (see Figure 2). Because the two highest elevations on these profiles represent classic patterns of MMPI-2 profile elevation, we use these two pairs of scales (2,7 and 6,8) to identify the profiles throughout the rest of this manuscript. In all, 10 participants were assigned to interpret each profile. Participants used information on the MMPI-2 profile provided to complete the q-sort.

Finally, feedback was offered regarding the participant's correlation with expert interpretation. The expert sort for each profile was generated by four research team members, who each provided q-sorts of the information produced by the computer-based test interpretation Minnesota Report authored by James N. Butcher and published by Pearson Assessments. The four q-sorts of this computerized descriptive narrative were averaged, and the mean sort served as expert feedback.

Figure 2. MMPI-2 data interpreted by participants: 6,8 profile.



MMPI-2™
ID 25476

Male

Extended Score Report
Page 2

MMPI-2 VALIDITY AND CLINICAL SCALES PROFILE

	VRIN	TRIN	F	F(B)	F(P)	L	K	S	Hs	D	Hy	Pd	Mf	Pa	Pt	Sc	Ma	Si
Raw Score:	5	11	12	11	4	2	7	10	10	16	14	22	31	19	24	46	27	28
K Correction:									4			3			7	7	1	
T Score:	50	65T	73	87	70	43	33	33	54	45	37	54	60	83	59	96	72	54
Response %:	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100

Cannot Say (Raw):	0	Percent True:	60
F-K (Raw):	5	Percent False:	40
Welsh Code:	8*6"9'+5-7 140/2:3# F'+-/L:K#	Profile Elevation:	62.50

2.4. Analysis

2.4.1. Interpretive reliability

To measure the "interpretive reliability" of the two different MMPI-2 profiles, the 10 q-sorts produced per profile were intercorrelated with one another, producing a total of 45 pairwise correlations per profile. The 45 intercorrelations per profile were averaged via the use of Fisher *Zr* transformations.

2.4.2. Item analysis

Mean q-ratings were calculated for each item to reveal characteristic interpretations for each profile.

3. Results

3.1. Interpretive reliability

The 45 intercorrelations between the 10 participants who interpreted the 2,7 profile are presented below the diagonal of Table 1; the 45 intercorrelations of the 10 participants who interpreted the 6,8 profile are presented above the diagonal of Table 1. For each profile, each of the 45 q-correlations was standardized (using *Zr*), averaged, and transformed back into *r* to yield a mean interpretive reliability estimate. As shown in Table 2, the mean *r* for the 2,7 profile was .41, with a range from .03 to .70. The mean *r* for the 6,8 profile was .36, with a range from −.16 to .64. The mean *r* across the two profiles was .38.

Also presented in Table 2 are descriptive statistics for the 100 q-correlations produced by correlating the 10 interpretations of the 2,7 profile with the 10 interpretations of the 6,8 profile. The mean *r* for these cross-profile comparisons was .26. Table 3 comprises the matrix of these cross-profile intercorrelations.

Table 2. Descriptive statistics for Q-correlations within and across profile types

Profile	*N*	Max	Min	*M*
2,7	45	.70	.03	.41
6,8	45	.64	−.16	.36
Within profile agreement	90	.70	−.16	.38
Cross profile agreement	100	.59	−.27	.26

Note: *N* represents the number of pairwise Q-correlations from which the mean pairwise *r* was calculated.

Table 3. Matrix of Q-correlations between interpretations of different profiles

Participants	1	2	3	4	5	6	7	8	9	10	*M* 6,8
1	.28	.41	.25	.19	.51	.11	.34	.30	.18	.26	.47
2	.12	.31	.30	.25	.56	.22	.35	.07	.24	.35	.50
3	.09	.02	−.05	.06	.32	−.20	.00	−.01	.05	−.05	.19
4	.10	−.10	−.04	−.11	.02	−.27	−.23	−.13	−.09	−.12	−.07
5	.41	.38	.33	.13	.58	.02	.41	.42	.34	.43	.47
6	.22	.40	.27	.22	.45	.23	.32	.30	.31	.51	.57
7	.30	.24	.21	.12	.59	−.05	.25	.20	.30	.30	.40
8	.25	.37	.32	.20	.42	.24	.32	.14	.31	.24	.44
9	.38	.44	.46	.26	.58	.20	.40	.32	.35	.49	.56
10	.35	.44	.48	.41	.53	.32	.40	.20	.46	.68	.63
M 2,7	.22	.09	.12	.15	.37	−.30	.07	.09	.24	.15	.15

Notes: Participants on the vertical axis interpreted the 2,7 profile; participants on the horizontal axis interpreted the 6,8 profile.

Statistical testing was conducted on the difference between sets of correlations using *z* tests. The difference between the pairwise reliabilities for the 2,7 profiles was not statistically significant ($p = .56$) from the pairwise reliabilities for the 6,8 profile. The magnitude of this difference, expressed as the raw difference between *Zr* values, was found to be below conventions for a "small" effect size (i.e. $q = .05$). The difference between same-profile correlations and cross-profile comparisons was also not statistically significant ($q = .13$, $p = .065$), though the effect size estimate was larger.

3.2. Item analysis

Item analysis was conducted to characterize profile interpretations and determine whether the sample exhibited idiosyncratic interpretive patterns. The items listed in Table 4 are those that received the highest and lowest mean ratings for the 2,7 profile, on a scale of +3 to −3. "Acts anxious" and "Acts depressed" obtained the highest mean ratings for the profile by far. Other highly rated items involved anxiety and depression. Conversely, lowest rated items were "Acts relaxed" and "Feels extremely happy without cause." All extreme aggregate ratings seem consistent with consensual expert interpretation of this profile.

The items listed in Table 5 are those with the highest and lowest mean ratings for the 6,8 profile. The highest rated items ("Is suspicious" and "Behaves oddly") suggest poor reality testing or thought disorder. The lowest rated items (including "Shows good judgment" and "Expresses emotions in healthy ways") reflect sound mental functioning. Again, these extreme ratings were consistent with expectations for a profile of this type.

Table 4. Highest and lowest mean item ratings for the 2,7 profile

Statement	Mean rating
Acts anxious	3
Acts depressed	2.8
Trembles, sweats, or shows other physical signs of anxiety	2.2
Cries frequently	1.7
Is optimistic	−1.5
Is hopeful	−1.6
Feels extremely happy without cause	−2.1
Acts relaxed	−2.5

Table 5. Highest and lowest mean item ratings for the 6,8 profile

Statement	Mean rating
Is suspicious	2.3
Behaves oddly	2.1
Sees or hears things that do not exist	2.1
Acts confused	1.9
Acts relaxed	−1.5
Is hopeful	−1.5
Describes self as having few problems	−1.7
Expresses emotion in healthy ways	−1.7
Shows good judgment	−1.8

4. Discussion

Although considerable variability was present, the average pairwise interpretive reliability of the q-sorts of these two classic MMPI-2 profiles was found to be .39. Mean reliabilities for the 2,7 and the 6,8 profiles were found to be .41 and .36, respectively, a difference that was not significant. For both profiles, then, the agreement observed between participants was considerably lower than expected based on previous q-methodology research on MMPI interpretation.

One might argue that considerable idiosyncrasy necessarily inheres in single-rater interpretations. Indeed, the most extremely rated items aggregated across all raters seemed quite characteristic of the typical interpretations of these profiles, and appear to differentiate the profiles quite well in terms of content. Applying the Spearman–Brown formula to the average pairwise interrater reliability of .39, our aggregate of 10 raters is associated with a reliability estimate of .86, a much more respectable result. On the other hand, we typically think of interpretive reliability as applying to an individual clinician rather than a group of clinicians. Furthermore, it is sobering that cross-profile correlations were not significantly different from within-profile intercorrelations. While it is true that the two profile types share a number of features, most notably general distress, which is surely an important dimension within the q-set, a key test of interpretive reliability is demonstrating that agreement about the meaning of a set of test results is specific to those results, and not to all possible results. Because we hand-picked these two MMPI-2 profiles to represent classic patterns in scale scores, the inability of our clinician sample to discriminate clearly between them is sobering.

Still, it is possible that results obtained underestimate true field reliability for a variety of reasons. First, most obviously, the sample is small and comprises those who cared to respond to an email invitation. It is unclear what participant characteristics our advertisements pulled for, but considering that the only compensation for participation was feedback, it is possible that the study had the greatest attraction for those who were uncertain of their interpretive skills and sought some information along those lines. On the other hand, one might speculate that participation would be limited to just those who were most interested in MMPI interpretation and therefore somewhat skilled.

A second concern about the generalizability of the results lies in the nature of the q-sort task, which is somewhat complicated and was completed individually, without supervision. It is possible that some of the participants did not understand the task sufficiently. Anecdotally, several participants indicated that they had difficulty viewing the profile simultaneous with completing the interpretation. Some participants also complained of a lack of adequate information with which to interpret the profile. Although this was also the case in previous research that produced higher reliability estimates, the frustration of completing the task alone without the opportunity for questions to be answered in real time may have affected interpretations somewhat. Motivation, too, may have varied in the absence of compensation or penalties for inaccuracy.

Yet another limitation to interpretation of our results is our reliance on a particular q-set, the Midwestern q-sort. Although it has been carefully fashioned and iteratively revised, it is possible that the q-set either omits important interpretive content or, more likely, contains too many items that are difficult to associate with particular MMPI-2 patterns, thereby introducing error variance into interpretation. Additionally, our methodology forces a highly structured interpretive process that may not match typical interpretative procedures. Consequently, the interpretive method we imposed on our participants may have resulted in estimates of field reliability that are either unfairly stingy or overly generous.

Of course, we believe it is important to consider the possibility that the present results are in fact accurately representative of generally poor field reliability of MMPI-2 interpretation. Previous research using similar methods appeared to take pains to identify expert interpreters as participants. We consider it quite plausible that expert interpreters are very consensual in their interpretations, while average test users are less able to achieve consensus and to discriminate between different profiles. There are very few empirical data available about how well the average test user is trained in MMPI interpretation, nor do we know much about the normative process by which test users develop their skills. Because profile interpretation is such a complex skill, factors as mundane as choice of textbook in graduate coursework or quality of CE instructor may produce variability in interpretive style that manifests as unreliability in test interpretation. Of course, this is not a concern that should be considered unique to interpretation of the MMPI-2. Field reliability at the level of the integrated interpretation is not well understood for any major psychological assessment instrument.

We are encouraged somewhat by recent efforts to suggest specific interpretive approaches to multi-scale personality assessment instruments. Whereas the MMPI-2 Manual (Butcher et al., 2001) provides only implicit guidance for how a full MMPI-2 protocol should be interpreted (i.e. within the context of case examples), the MMPI-2-RF Manual (Ben-Porath & Tellegen, 2008/2011) explicitly outlines a multi-step structure for interpretation. For example, substantive scales of the MMPI-2-RF are grouped thematically (e.g. scales that bear upon Somatic/Cognitive Dysfunction, scales that suggest patterns of Interpersonal Functioning), providing guidance as to which scales are likeliest to work together to shape inferences. Ben-Porath's (2012) interpretive guide to the MMPI-2-RF goes further, offering detailed examples applying this structure with the use of a worksheet. To the extent that such guidance serves to encourage uniformity in interpretive approach, the reliability of the interpretive product of assessment can only benefit. On the other hand, any given interpretive inference might necessarily entail finer decisions about which scales are most relevant to the inference, about how to weight the relevant scales, about how to resolve apparent contradictions, etc. It would be unreasonable to expect any interpretive manual or text to be able to anticipate and guide all such integrative decisions. Still, the proposal of any explicit interpretive guidelines must be considered a step forward.

It should also be noted that not all MMPI-2 applications necessarily involve inter-scale interpretive integration, or entail the production of broad descriptive accounts of the examinee. For example, some professional uses may employ strict prediction schemes involving single scales with a limited inferential scope. Other applications may involve and integrate multiple scales, but in a mechanical way that does not require any clinical judgment. Our study does not generalize to such test uses, for which interpretive reliability could be expected to be much better, and in the case of pure actuarial prediction, near perfect. Nor does our study generalize to MMPI-2 interpretation practices involving different sets of MMPI-2 scales. It is conceivable either (1) that the availability of additional scale scores would augment reliability by providing interpreters information with which to refine hypotheses, or (2) that the availability of additional scale scores would diminish reliability by offering interpreters more opportunities for idiosyncrasy.

This is difficult research to conduct. Large-scale field reliability research necessitates careful sampling, substantial sample sizes, and incentives that would attract the participation of test users who would otherwise be generating billable hours will their time. But we argue that such research is necessary for the advancement of our understanding of how instruments like the MMPI-2 can be used optimally in clinical settings. It has implications for how test interpretation should be taught, what restrictions ought to be placed on test use, and perhaps even which textbooks are most useful. Field reliability of MMPI-2 interpretations should be examined not just for two hand-picked profile types, but for profiles that are representative of clinical populations in which the MMPI-2 is used. Q-sets can be varied to suit differing clinical applications: diagnosis, generating treatment recommendations, use within a specific population, etc. Importantly, the ability to operationalize an MMPI-2 interpretation using q-methodology should be extended to studies of interpretive *validity*. How well do field interpretations of MMPI-2 profiles match clinical presentation? For what clinical phenomena, populations, and purposes are MMPI-2 interpretations valid, and to what varying degrees? How can we best train clinicians to interpret the MMPI-2 with accuracy? Which scales or sets of scales are most useful and parsimonious? How can blind MMPI-2 interpretations be combined with clinical information to optimize description and prediction? Clearly, much remains to be studied with regard to clinical interpretation of the MMPI-2 and similar clinical instruments.

Funding
The authors received no direct funding for this research.

Competing interests
The authors declare no competing interest.

Author details
Mark A. Deskovitz[1]
E-mail: desko1ma@cmich.edu
Nathan C. Weed[1]
E-mail: weed1nc@cmich.edu
Cheryl Chakranarayan[1]
E-mail: chakr1c@cmich.edu
ORCID ID: http://orcid.org/0000-0001-8055-0296
John E. Williams[2]
E-mail: john.eustis@gmail.com
[1] Department of Psychology, Central Michigan University, Mt. Pleasant, MI, USA.
[2] Department of Psychology, Central Michigan University, Canmore, Alberta, Canada.

Citation information
Cite this article as: Interpretive reliability of two common MMPI-2 profiles, Mark A. Deskovitz, Nathan C. Weed, Cheryl Chakranarayan & John E. Williams, *Cogent Psychology* (2016), 3: 1161287.

References

Ben-Porath, Y. S. (2012). *Interpreting the MMPI-2-RF*. Minneapolis: University of Minnesota Press.

Ben-Porath Y. S., & Tellegen, A. (2008/2011). *MMPI-2-RF (Minnesota multiphasic personality inventory-2 restructured form) manual for administration, scoring, and interpretation*. Minneapolis: 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*. Minneapolis: University of Minnesota Press.

Friedman, I. (1957). Objectifying the subjective—A methodological approach to the TAT. *Journal of Projective Techniques, 21*, 243–247. doi:10.1080/08853126.1957.10380778

Garb, H. N. (1998). *Studying the clinician: Judgment research and psychological assessment*. Washington, DC: American Psychological Association. http://dx.doi.org/10.1037/10299-000

Graham, J. R. (2000). *MMPI-2: Assessing personality and psychopathology* (3rd ed.). New York, NY: Oxford University Press.

Halbower, C. C. (1955). *A comparison of actuarial versus clinical prediction to classes discriminated by the MMPI* (Unpublished doctoral dissertation). University of Minnesota, Minneapolis.

Little, K., & Shneidman, E. (1954). The validity of MMPI interpretations. *Journal of Consulting Psychology, 18*, 425–428. doi:10.1037/h0061462

Marks, P. A., & Seeman, W. (1963). *Actuarial description of abnormal personality*. Baltimore, MD: Williams & Wilkins.

McNeal, T. P. (2000). *The longitudinal impact of personality assessment on therapists' understanding of clients during the course of psychotherapy* (Doctoral dissertation). Retrieved from PsycINFO: http://search.proquest.com.cmich.idm.oclc.org/psycinfo/docview/619555884/24AD79D30D684A17PQ/1?accountid=10181 (Order No. AAI9956078).

Meehl, P. (1956). Wanted—A good cook-book. *American Psychologist, 11*, 263–272. doi:10.1037/h0044164

Moos, R. (1962). Effects of training on students' test interpretations. *Journal of Projective Techniques, 26*, 310–317. doi:10.1080/08853126.1962.10381118

Noland, K. A., & Weed, N. C. (1994). *The Mississippi Q sort* [Computer software]. Oxford, MS: Author.

Ozer, D. J. (1993). The Q-sort method and the study of personality development. In D. C. Funder, R. D. Parke, C. Tomlinson-Keasey, & K. Widaman (Eds.), *Studying lives through time: Personality and development* (pp. 147–168). Washington, DC: American Psychological Association. http://dx.doi.org/10.1037/10127-000

Weed, N. C. (2006). Syndromal complexity, paradigm shifts, and the future of validation research: Comments on Nichols and Rogers, Sewell, Harrison, and Jordan. *Journal of Personality Assessment, 87*, 217–222. doi:10.1207/s15327752jpa8702_12

Williams, J. E., & Weed, N. C. (2003, June). *Using a web-based q-sort program to teach MMPI interpretation*. Paper presented at the 38th Annual Symposium on Recent Developments in the Use of the MMPI, MMPI2, and MMPIA, Minneapolis, MN.