VALUE IN HEALTH 21 (2018) 650–657

Available online at www.sciencedirect.com

ScienceDirect

journal homepage: www.elsevier.com/locate/jval

Agreement among the Productivity Components of Eight Presenteeism Tests in a Sample of Health Care Workers

Angus H. Thompson, PhD[1,2,*], *Arianna Waye, PhD*[1]

[1]*Institute of Health Economics, Edmonton, Alberta, Canada;* [2]*School of Public Health, University of Alberta, Edmonton, Alberta, Canada*

ABSTRACT

Background: Presenteeism (reduced productivity at work) is thought to be responsible for large economic costs. Nevertheless, much of the research supporting this is based on self-report questionnaires that have not been adequately evaluated. **Objectives:** To examine the level of agreement among leading tests of presenteeism and to determine the inter-relationship of the two productivity subcategories, amount and quality, within the context of construct validity and method variance. **Methods:** Just under 500 health care workers from an urban health area were asked to complete a questionnaire containing the productivity items from eight presenteeism instruments. The analysis included an examination of test intercorrelations, separately for amount and quality, supplemented by principal-component analyses to determine whether either construct could be described by a single factor. A multitest, multiconstruct analysis was performed on the four tests that assessed both amount and quality to test for the relative contributions of construct and method variance. **Results:** A total of 137 questionnaires were completed. Agreement among tests was positive, but modest. Pearson *r* ranges were 0 to 0.64 (mean = 0.32) for Amount and 0.03 to 0.38 (mean = 0.25) for Quality. Further analysis suggested that agreement was influenced more by method variance than by the productivity constructs the tests were designed to measure. **Conclusions:** The results suggest that presenteeism tests do not accurately assess work performance. Given their importance in the determination of policy-relevant conclusions, attention needs to be given to test improvement in the context of criterion validity assessment.

Keywords: health, presenteeism, productivity, testing, work.

Introduction

Presenteeism refers to decreased productivity and/or below-normal work quality when physically present at work [1]. Although there are still discussions about definitions [2], presenteeism is usually measured by one of a number of self-report instruments. It has received more attention in recent decades because of reports that presenteeism causes significant losses to businesses and to the economy as a whole [3,4]. In fact, presenteeism has been described as a "silent" but significant source of productivity loss that can cost organizations much more than absence from work [4,5].

Unfortunately, the use of self-report data, rather than more expensive direct measures, carries with it a number of problems. Presenteeism estimates require the measurement of work outputs that are often not clearly specified [2,6], are often difficult to quantify, and are not easily compared across disparate work roles and conditions. Moreover, self-reports of presenteeism are likely influenced by bias due to method variance [2]. Method variance occurs when something inherent to the structure or presentation of the questionnaire produces the apparent statistical association between test items, as opposed to such being due to the relationship of the respective terms.

Within some of the presenteeism instruments, work productivity has been broken down into two components, quantity (i.e., amount of output) and quality (excellence of output) [7–10]. Notably, neither component is very clearly defined, thus making them vulnerable to imprecision-related errors and biases of several sorts, including method bias. Analyses such as Campbell and Fiske's multimethod, multitrait approach [11], however, allow the "teasing out" of such method variance when faced with concepts such as amount and quality of work that are likely related, but not identical.

Although rarely addressed in the context of presenteeism, self-reports generally show poor validity [12]. In particular, self-reports of performance are notoriously inaccurate [13] and are vulnerable to a form of social desirability that reflects a general tendency of humans to place themselves in a good light in comparison with others [14,15]. It seems highly likely that our ego involvement in work performance would make us particularly susceptible to this form of bias.

Conflicts of interest: Neither author has any conflict of interest with regard to the content of this article.

* *Address correspondence to:* Angus H. Thompson, Institute of Health Economics, 1200–10405 Jasper Avenue, Edmonton, Alberta T5J 3N4, Canada.

E-mail: g60thomp@gmail.com

1098-3015$36.00 – see front matter

https://doi.org/10.1016/j.jval.2017.10.014

In contrast to the measurement of absenteeism (the number of work-days missed), the estimation of presenteeism is somewhat more complicated and generally requires a subjective interpretation. Mattke et al. [6] reviewed 17 self-report tests of lost productivity and found a variation in formats, ranging from generic instruments to tests that were disorder-/condition-specific. The authors concluded that most of the instruments under study had been validated to some extent. It should, however, be noted that criterion validity, the association between self-report and a benchmark (criterion standard), was generally not assessed or was unimpressive in magnitude. Similar conclusions have been supported in two more recent systematic reviews [16,17]. Nonetheless, self-report tests of presenteeism continue to serve as the mainstay of publicly reported productivity studies, presumably because of lower costs and ease of data collection.

A final issue is that the reviews of presenteeism instruments in the last decade or so (i.e., from 2004 onward) [6,16–21] are based on studies that used all or most of the questions from each test and have thus included items that did not measure lost productivity. Rather, most of the items were often about the work environment, personal limitations, health, mental well-being, and other matters that are thought to provide relevant context. For example, one of the most frequently used self-report instruments, the general health version of the Work Productivity and Activity Impairment Questionnaire (WPAI) comprises six items [22,23], but only one of these can be said to measure presenteeism (also see [24]). That is, the respondents are presented with an 11-point scale (0–10) to answer a single question about the extent that health problems had affected work productivity. The remaining five questions covered employment status, time absent from work for health or for other reasons, hours actually worked, and the effect of health issues on nonwork activities. This issue similarly pertains to several presenteeism tests, but none of the aforementioned review studies excluded the nonpresenteeism questions of any of the tests in their analyses. Thus, the findings of these reviews [i.e., 6,16–21] may be biased because they were focused to a meaningful degree on nonpresenteeism test questions rather than on those items that specifically addressed productivity loss.

The goal of this study was to assess agreement among the more prominent presenteeism instruments using only those items from each test that deal directly with productivity. Specifically, the objectives were 1) to determine the inter-relationships of the productivity components of eight presenteeism instruments and 2) to assess the interplay of the two productivity subcategories, amount and quality, to allow the examination of the contribution of method variance to estimates of construct validity.

Methods

Participants

Eligibility for the study was restricted to persons actively employed/engaged as health care workers by Alberta Health Services within the Edmonton Zone of the Province of Alberta, Canada, who willingly chose to participate in the study. Participants were recruited via the managers/supervisors of enough clinical units to produce a subject pool comprising 494 health care workers. That is, the pool comprised about one-third of the full complement of 1477 persons listed as health care employees in the zone. Respondents were recruited in two waves. The first group completed a self-report paper-and-pencil questionnaire that was distributed to workers via their supervisors. Each questionnaire was placed in a sealable, addressed envelope along with an information sheet that explained the procedure, confidentiality provisions, testing protocol, and the voluntary nature of participation. It was clearly stated that individuals could choose to not participate without consequence and that the employer would not see individual responses (the exception being the data specialist who extracted activity log data, who had access to such information in any case under strict confidentiality rules). Furthermore, respondents were asked to provide their names, activity log number (associated with on-the-job recording of patient-related work activity), and the name of the clinic where they were employed so as to allow database matching for a later study that is not reported here. It was made clear that such identifying information would be stripped from the file once the database records were merged.

The second recruitment wave involved Internet administration of the questionnaire. This happened because of a request from one of the health service areas (nutrition/dietary) where it was suggested that an online delivery would likely produce greater participation. Minimal changes in format were incorporated to accommodate the differences in mode of delivery. Notably though, the confidentiality component for the online version included the additional warning that responses were likely to be stored on a server in the United States and thus subject to the laws of that country, particularly those that allowed government access to any such recorded information. Notification of the survey involved group distribution of an invitation via email to health care workers within the study region, including an advance notice to those within the clinical area in question.

As a consequence of the wish of the hosts of this study to have their managers involved in the distribution of questionnaires and of our wish to use a questionnaire collection system that was anonymous, we did not seek personalized data on eligible nonparticipants.

Completion of the first wave of data collection (i.e., paper-and-pencil questionnaire completions) took place between April 11 and May 12, 2014. Data from the second wave (online completions) were collected between June 16 and July 26, 2014.

The Questionnaire

The findings from a systematic review of the psychometric properties of existing presenteeism tests [17,25] provided the information on test quality and utilization that served as the base for the selection of the instruments for the present study. Sixteen discrete instruments were examined in the review. The two that showed the highest frequency of appearance in the literature (the WPAI and the Health and Work Performance Questionnaire [HPQ] at 56% and 14% of studies, respectively) were retained for the present investigation (see Table 1). Note that the general health version of the WPAI was used here. Also retained were the Valuation of Lost Productivity Questionnaire (VOLP), Work Productivity Short Inventory (WPSI), Health and Work Questionnaire (HWQ), Lam Employment Absence and Productivity Scale (LEAPS), Endicott Work Productivity Scale (EWPS), and Quantity and Quality (Q-Q) on the basis of the relative strength of their psychometric properties (reliability, validity, and sensitivity to change). Note that the six-item version of the Stanford Presenteeism Scale, although it showed psychometric strength [25], was excluded because none of its items were deemed to be measures of productivity.

Presenteeism test items

The appropriate test items from the eight selected instruments were embedded in one questionnaire to be used in this study. The study questionnaire also included sections on the work environment, general health, mental health, and several other

Table 1 – Presenteeism instruments included in the study, with Amount (A) and Quality (Q) items that were selected for analysis.

Test	Name/Reference	No. of items analyzed/administered
WPAI[X]	Work Productivity and Activity Impairment Questionnaire, general health [22,23]	1/1 (11-point scale)
	A: During the past 7 days, how much did any of your health problems affect your productivity while you were working?[*]	
HPQ[X]	WHO Health and Work Performance Questionnaire [26,27]	1/4 (11-point scale)
	A: Using the 0 to 10 scale, how would you rate your *overall* job performance on the days you worked during the *past week*?	
VOLP[X]	Valuation of Lost Productivity Questionnaire [28]	1/5 (11-point scale)
	A: In the past 7 days, to what extent was your performance at work affected by your health while you were working?	
WPSI[X]	Work Productivity Short Inventory [29]	1/4 (% hours sick)[†]
	A: During a typical 8-hour workday, estimate the total hours you were unproductive because of any health-related symptoms.	
HWQ[X]	Health and Work Questionnaire [7]	2/6 (11-point scales)
	A: How would you describe the overall amount of your work this week? ["My worst ever" to "My best possible"]	
	Q: How would you describe the overall quality of your work this week? ["My worst ever" to "My best possible"]	
LEAPS[X]	Lam Employment Absence and Productivity Scale [8]	3/3 (5-point scales)
	A: Over the past week, how often at work were you bothered by getting less work done?	
	QP: Over the past week, how often at work were you bothered by doing poor quality work? ["QP": Poor work]	
	QE: Over the past week, how often at work were you bothered by making more mistakes? ["QE": work Errors]	
EWPS[X]	Endicott Work Productivity Scale [9]	2/2 (5-point scales)
	A: During the past week, how frequently did you notice that your productivity for the time spent is lower than expected?	
	Q: During the past week, how frequently did you have to do a job over because you made a mistake or your supervisor told you to do a job over?	
Q-Q[X]	Quantity and Quality [10]	2/2 (10-point scales)
	A: Could you indicate how much work you actually performed during regular hours on an average day over the past week as compared with normal on the scale below?	
	Q: Could you indicate the quality of the work you performed on an average day over the past week as compared with normal?	

Note. "X" was added to the name of each scale here to indicate that an extract was used, rather than the whole instrument.
[*] WPAI instructions were revised by adding the two-word phrase "any of," that is, "… how much did *any of* your health problems …."
[†] The denominator for the calculation of the weekly percentage was taken from the work diary section of the questionnaire.

factors thought to be related to presenteeism. These factors will be examined in a separate article and will not be discussed further here.

As noted earlier, most of these tests contain a number of questions that do not specifically denote presenteeism. Rather, for example, several gather information on potential work limitations such as handicapping and/or health conditions, mental well-being, and/or the work environment. Consequently, although many of these context items were included in the study questionnaire, only those pertaining to productivity loss were subjected to statistical analysis here. This reduced the number of items retained for analysis to a relatively small range of one to three items per test (see Table 1). Four of the test instruments addressed only the amount of productivity (the WPAI, HPQ, VOLP, and WPSI). For them, presenteeism was described by a single, generally stated question covering, respectively, productivity, performance, performance while working, and unproductivity. These will be treated as measures of quantity (Amount) for the remainder of this article. Three of the four remaining instruments (the HWQ, EWPS, and Q-Q) included an item that clearly and appropriately referred to quantity and another that measured quality. The latter was defined by either the term "quality" itself (the HWQ and the Q-Q) or a reference to "work errors" (the EWPS). The LEAPS, in addition to its quantity item, had two quality items, one literally identifying quality and the other naming work errors (see Table 1). Both the LEAPS quality items were retained for an initial analysis, with a view to a determination of whether the two would be best collapsed into a single measure.

Recall periods

Presenteeism test respondents are typically asked to make a retrospective rating about some aspect of their work over a particular recent time period—called the "recall period." An example using a 1-week recall period would be "Please estimate the percentage of time that you were fully productive at work during the past week." Any recall period can be chosen, but accuracy of recall tends to be higher at shorter time periods, whereas representations of day-to-day work are better over longer intervals. Specifically, a 7-day recall was found to produce greater accuracy than a 4-week period [30]. The eight tests of interest here use various recall periods, but they generally fall within the range of 1 week to 1 month. For consistency, a recall period of 7 days was assigned to each test. This involved an adjustment of the recall period used with four of the eight instruments. These were the HPQ (previously 4 weeks), LEAPS (2 weeks), Q-Q (1 day), and WPSI (which provided a choice of three recall periods: 2 weeks, 3 months, and 12 months).

Permissions

All but two of the eight tests were in the public domain. Of the two proprietary tests, one, the WPAI, does not require usage permission for research purposes, and the test question from the other, the WPSI (which is no longer being marketed), was taken from a published version of the test protocol [24]. Nonetheless, to ensure that this would not be an issue, only those items relevant to presenteeism have been extracted for use from each test and the wording of one proprietary test has been altered to make its items indistinguishable from a number of similar tests without loss of its original meaning (see the WPAI, Table 1). Our alteration of the recall period and our presenteeism-focused abridgment of many of the instruments have resulted in changes in the structure of nearly all the tests. Thus, to make it clear that the protocols being used here are adaptations of the original instruments, the superscript of the letter "X" will be appended to the acronym for each test as appropriate from this point forward

(i.e., when discussing analyses and findings). To be clear, however, we will be interpreting our findings in a manner that will implicate the original tests in the usual way, but the reader will be aware of the caveat related to the adaptations and may thus judge the findings within that context.

Test valence

Some items were worded in a positive direction (i.e., higher scores meaning higher productivity) and some were presented with a negative valence (i.e., higher scores meaning higher "unproductivity"). Valence is indicated when appropriate through the remainder of the article. An exception is in Table 2, in which all correlations are presented as positive relationships to aid in comprehensibility (in terms of meaning, all but one inconsequential score proved to be unidirectional; i.e., higher productivity in one test was always associated with either higher productivity or lower unproductivity in another).

Latin Square Design

Within the study questionnaire, the productivity components of all eight tests were presented in eight different sequences to balance for order and position effects. Thus, there were eight forms of the questionnaire that did not differ at all in terms of content, but only in the presentation order of the eight test components. The presentation orders were determined using Williams' application of a Latin square design [31] to the matter. The purpose of this was to ensure that across the set of eight orderings, each test 1) appeared only once in each of the eight positions, 2) followed immediately after each of the other seven tests only once (noting, of course, that the test in the first position will not have a precedent), and 3) preceded any other test 50% of the time and followed 50% of the time. We were not able to apply this design to the Web-based segment of survey administration, and thus the balanced ordering pertains only to the 106 paper-and-pencil respondent records. One of the Williams orderings (i.e., T2, T3, T1, T4, T8, T5, T7, T6) [31] was arbitrarily chosen for the 31 online respondents.

Analyses of Potential Bias

First, an analysis of potential distribution bias was undertaken. The eight forms of the questionnaire were physically arranged so that they would be given to potential participants in a repeatable form 1 to form 8 rotation (to increase the likelihood of an equal distribution). Nevertheless, this part of the process was not operated by the research team, but rather by health care system officials. Thus, it was decided to assess the ultimate assortment of completions across four important worker demographic variables (profession, work status, age, and sex).

The second bias-related analysis comprised a comparison of responses to the paper-and-pencil (hard copy) and online modes of the administration of the survey questionnaire. This was carried out with the awareness of an overarching confound present in the data. That is, in addition to the obvious difference in the mode of questionnaire administration, the online respondents were involved primarily in nutrition/dietary work, whereas those completing the paper-and-pencil questionnaire were engaged in several other areas of health care. Nonetheless, although there was no reason to necessarily expect a detrimental level of bias due to these differences, we determined that a comparison of productivity ratings could be illuminating and might help to point to more definitive future investigations.

Third, we faced an issue regarding the two versions of quality contained in the LEAPS. That is, if responses to them were similar, "quality" would be counted twice and thus over-represented. Nevertheless, if they were responded to in a very different way, retaining them both would bring into question the decision to view quality among the other three instruments as members of a single category when one of them is described in terms of work errors. To aid in our selection of the most appropriate option, we examined the correlation between these LEAPS subscores and conducted a subsidiary principal-component analysis to compare the loadings of these two variables.

Analysis of Study Objectives

Three sequentially related analyses were undertaken to assess the study objectives. First, a correlational analysis (Pearson *r*) was conducted on the Amount and Quality measures to provide an examination of the inter-relatedness of self-rated productivity across all eight presenteeism tests (thus forming a base for our subsequent investigations). Second, the Amount and Quality values were independently subjected to a principal-component analysis to determine whether either construct could meaningfully be described by a single factor across their corresponding eight- and four-test arrays. Finally, an adaptation of Campbell and Fiske's technique for assessing the interplay of multiple methods and multiple traits [11] (referred to from here on as "multitest" and "multiconstruct," respectively) was applied to the data from the four instruments that provided ratings for the constructs of both amount and quality of work (i.e., the HWQ^{X}, $LEAPS^{X}$, $EWPS^{X}$, and $Q\text{-}Q^{X}$). Thus, to allow an examination of method variance among these instruments, overall variance values were summarized using mean *Z* scores of the correlations in question. These means were then converted back to correlation coefficients and tested by an adaptation of Steiger and Brown's procedure for the comparison of composite interdependent values [32].

Statistical analyses were conducted using SPSS 14.0 (Armonk, New York (IBM Corporation)) for Windows. This study was approved by the Health Research Ethics Board of the Faculty of Medicine and Dentistry, University of Alberta.

Results

Sample Description

There were 137 questionnaires completed (27.7% of those deemed to be eligible), comprising 106 paper-and-pencil forms (32.7% of those eligible) and 31 online submissions (18.2% of those eligible). Note that these completion figures are response rates, which may include nonparticipation for almost any reason in their calculation and are thus relatively conservative in comparison with the preferred (but here infeasible) cooperation rates, which are based on noncompletions due to refusals or unfinished questionnaires. The sample included therapists (occupational therapists (OTs) and physiotherapists (PTs): 42%), nurses (registered psychiatric nurses (RPNs) and licensed practical nurses (LPNs): 21%), dieticians (19%), cardiology technologists (10%), and counselors (7%). Sixty-six percent worked full-time, 25% were part-time, and 9% were casual or temporary employees. Sample members were primarily involved in patient care: 70% reported that they spent more than 60% of their work time involved in duties involving individual patients, with 23% at 40% to 60%, and only 7% endorsing the less than 40% range.

Missing Values

Overall, 4.5% of the potential test responses were missing, primarily attributable to 21 cases missing VOLP data (15.3%) and 12 cases missing WPSI scores (8.8%). This was caused in large part by an administrative error that, although rectified when discovered, resulted in the items in the question being skipped early in the data collection process. The remaining six Amount and all

Table 2 – Pearson intercorrelations between Amount and Quality measures of presenteeism*.

Measure	Amount								Quality		
	WPAI^X (–)	HPQ^X	VOLP^X (–)	WPSI^X (–)	HWQ^X	LEAPS^X (–)	EWPS^X (–)	Q-Q^X	HWQ^X	LEAPS^{X†} (–)	EWPS^X (–)
Amount											
HPQ^X	0.18‡										
VOLP^X (–)	0.64§	0.18									
WPSI^X (–)	0.33§	0.13	0.33§								
HWQ^X	0.09	0.58§	0.17	0.16							
LEAPS^X (–)	0.26‖	0.32§	0.41§	0.16	0.28§						
EWPS^X (–)	0.42§	0.30§	0.49§	0.18	0.23‖	0.50§					
Q-Q^X	0.23‖	0.39§	0.37§	0.11	0.41§	0.39§	0.39§				
Quality											
HWQ^X	0.16	0.56§	0.24‖	0.23‖	0.51§	0.23‖	0.32‖	0.18‡			
LEAPS^{X†} (–)	0.11	0.23‖	0.12	0.12	0.11	0.28§	0.26‖	0.11	0.23‖		
EWPS^X (–)	0.03	0.10	0.13	0.00	0.04	0.04	0.28§	0.04	0.03	0.30	
Q-Q^X	0.28§	0.51§	0.36§	0.19‡	0.43§	0.22‡	0.39§	0.71§	0.38§	0.29§	0.14

Note. Negative valences are shown in parentheses, with a "(–)" indicating a negative orientation, that is, amount or quality of unproductivity. EWPS, Endicott Work Productivity Scale; HPQ, Health and Work Performance Questionnaire; HWQ, Health and Work Questionnaire; LEAPS, Lam Employment Absence and Productivity Scale; Q-Q, Quantity and Quality; VOLP, Valuation of Lost Productivity Questionnaire; WPAI, Work Productivity and Activity Impairment Questionnaire; WPSI, Work Productivity Short Inventory.

* Values are shown as positive correlations, with one trivial exception—HWQ Amount × EWPS Quality at 0.04 is nonconsonant. All other cases are semantically consistent; unproductivity on one test is associated with unproductivity or low productivity on another.

† The LEAPS Quality values used in these calculations are composites of the mean of the two LEAPS quality questions (LEAPS "Poor quality" and LEAPS "Errors").

‡ $P < 0.05$.

§ $P < 0.001$.

‖ $P < 0.01$.

four Quality measures were unaffected by this error, showing missing values ranging from 2.2% to 4.4% (mean = 3.0%). SPSS was set to apply pairwise deletion of cases when handling missing values.

Bias Analysis

An analysis to assess potential distribution bias across the eight forms of the paper-and-pencil questionnaire found no statistically significant results. Although there was some nominal deviation from the expected average distribution (12.5% for each of the eight orderings), the profiles did not show statistically significant differences across professions (three categories: $X^2 = 11.16$; $df = 14$; $P = 0.67$), work status (three categories: $X^2 = 13.80$; $df = 14$; $P = 0.47$), age (four categories: $X^2 = 24.03$; $df = 21$; $P = 0.29$), or sex ($X^2 = 7.24$; $df = 7$; $P = 0.41$). This indicates that we found no meaningful bias in the dissemination of the questionnaire.

The exploration of the influence of the mode of survey delivery (hard copy vs. online) showed no differences in mean productivity ratings. None of the eight Amount scores nor any of the four Quality scores showed statistical significance (at $P < 0.05$) when an independent samples t test was applied to the data. Further elucidation of the delivery mode issue is presented here with the findings on method variance.

With regard to the two LEAPS versions of productivity quality, it was determined that the use of a single score, the mean of the two, would best serve the interests of this study. First, the measures of poor quality work and work errors had more in common ($r = 0.57$) than either had with LEAPS Amount (0.34 and 0.16, respectively). Second, the subsidiary principal- component analysis of the two LEAPS^X Quality scores along with the quality measures from the remaining three tests produced near to identical factor loadings for the two LEAPS^X measures of quality (0.78 and 0.76). Overall, the principal component accounted for 41.5% of the variance. When we reran the analysis with the composite LEAPS^X score, the resultant principal component produced a very similar factor loading profile while accounting for 42.7% of the variance (discussed in the next section and presented in Table 3). Thus, the use of the composite form of LEAPS^X Quality allowed us to work with a model that involved a single Amount score and a single Quality score for each of the four tests that produced additional simplicity without loss of information.

Analysis of Study Objectives

Table 2 presents Pearson intercorrelations of the Amount and Quality measures of productivity. First, with one inconsequential exception (underlined in Table 2), all the relationships are semantically in agreement in direction (after correcting for test differences in valence). Second, a high proportion of the coefficients are statistically significant at P less than 0.05 (42 of 66; 64%), well above the expectation if chance was the determinant (the interdependence of the measures notwithstanding). Broken down, the three sectors in Table 2 all show high proportions of significant findings; 71% for Amount × Amount correlations, 50% for Quality × Quality, and 59% for the Amount × Quality associations. Importantly, in contrast to the widespread commonality among the test results, the overall magnitude of the relationships was relatively modest. Only a single correlation among the Amount × Amount associations ($r = 0.64$ for WPAI^X × VOLP^X) placed in the "strong" range (0.60–0.79) [33], with the overall mean (based on Z-score transformations) showing only a weak association ($r = 0.32$). Similarly weak was the mean of the Quality × Quality intercorrelations at 0.25. Notably and surprisingly, the correlations between the different constructs (Amount and Quality) are in the same range, averaging 0.23.

Table 3 – Unrotated principal-component analyses for Amount and Quality.

Measure	Amount	Quality
WPAI^{X} (–)	0.64	–
HPQ^{X}	–0.59	–
VOLP^{X} (–)	0.74	–
WPSI^{X} (–)	0.42	–
HWQ^{X}	–0.55	0.65
LEAPS^{X} (–)	0.68	–0.72
EWPS^{X} (–)	0.73	–0.48
Q-Q^{X}	–0.67	0.74
Eigenvalue	3.21	1.71
% of variance	40.1%	42.7%

EWPS, Endicott Work Productivity Scale; HPQ, Health and Work Performance Questionnaire; HWQ, Health and Work Questionnaire; LEAPS, Lam Employment Absence and Productivity Scale; Q-Q, Quantity and Quality; VOLP, Valuation of Lost Productivity Questionnaire; WPAI, Work Productivity and Activity Impairment Questionnaire; WPSI, Work Productivity Short Inventory.

To refine our understanding of these relationships, we applied a principal-component analysis to the data separately for Amount and Quality. Table 3 shows that the principal component for Amount included meaningful loadings on all eight tests (ranging from 0.42 to 0.73), accounting for 40.1% of the variance. Intercorrelations among the Quality items of the four tests showed a similar pattern, with loadings ranging from 0.47 to 0.74, accounting for 42.7% of the variance.

Method Variance

The analyses to this point indicate that the tests do share some common variance that appears to be related to the general notion of productivity. The commonality, however, was nearly as strong between the two different productivity subconstructs (i.e., Amount and Quality). Further refinement of our understanding of this latter issue was pursued through our multitest, multiconstruct analysis of the four instruments measuring both amount and quality. This analysis is based on the expectation that strong psychometric test properties should produce three outcomes [11]: 1) the largest correlations should be found between measures of the same construct even though they will have been assessed by different tests (e.g., HWQ^{X} Amount with LEAPS^{X} Amount); 2) correlations between the constructs (Amount and Quality) should be much lower, even though some positive level of association between them might reasonably be expected (they are both within the higher order construct of "productivity"); and 3) the association between the two constructs should, in the absence of method variance, be about the same for same-test and different-test values (higher correlations for the same-test comparisons would then be indicative of method variance).

This is not at all what we found. The highest values overall went to the four measures that are most indicative of method variance (shown in boldface in Table 4). These represent the relationship between Amount and Quality within the same test (mean $r = 0.47$; also see Table 5). The construct validity values, designated in Table 4 by the superscripts "A" and "Q" for, respectively, Amount (mean $r = 0.37$) and Quality (mean $r = 0.23$), not only fell short of being higher overall than the just-mentioned same-test different-construct values, but were also lower. Applying the adaptation of Steiger and Brown's procedure for the comparison of composite interdependent values [32], the difference, overall, between the method-laden correlations and the construct validity mean values proved to be significant for quality ($Z = 2.40$; $P < 0.01$) but not for Amount ($Z = 1.05$; $P = 0.11$). Furthermore, the association between Amount and Quality across different tests (mean $r = 0.20$) is not the same as it is within the same test (0.47, as noted earlier), but is meaningfully lower ($Z = 2.86$; $P < 0.005$). Overall, then, none of Campbell and Fiske's three criteria for adequate construct validity were satisfactorily met, suggesting that any differences must be attributed to within-test influences. This indicates that the structural attributes of these instruments and their mode of administration (i.e., method variance) contribute more to the ultimate test score than does the nature of productivity, whether amount or quality.

A comparison of these values broken down by mode of administration lessens the aforementioned concern about possible bias due to the confound between the two levels of this variable. The pattern of relationships derived from the multitest, multiconstruct analysis of online and hard copy modes shows much the same pattern as the overall analysis described in the preceding paragraph (see Table 5). That is, for both modes, the mean Amount and Quality construct validity correlations (for different-tests same-constructs) did not, statistically speaking, exceed their respective same-method different-construct

Table 4 – Multitest, multiconstruct analysis of the Pearson correlations depicting the inter-relationships between the four tests of both Amount (A) and Quality (Q).

Test	HWQ^{X}		LEAPS^{X} (–)		EWPS^{X} (–)		Q-Q^{X}	
	Amount	Quality	Amount	Quality	Amount	Quality	Amount	Quality
HWQ^{X} A								
HWQ^{X} Q	**0.51***							
LEAPS^{X} A (–)	–0.28^{A*}	–0.23†						
LEAPS^{X} Q (–)	–0.11	–0.23^{Q†}	**0.28**†					
EWPS^{X} A (–)	–0.23^{A†}	–0.32*	0.50^{A*}	0.26†				
EWPS^{X} Q (–)	0.04	–0.03^{Q}	0.04	0.30^{Q*}	**0.28**†			
Q-Q^{X} A	0.41^{A*}	0.18	–0.39^{A*}	–0.11	–0.39^{A*}	–0.04		
Q-Q^{X} Q	0.43*	0.38^{Q*}	–0.22†	–0.29^{Q*}	–0.39*	–0.14^{Q}	**0.71***	

EWPS, Endicott Work Productivity Scale; HWQ, Health and Work Questionnaire; LEAPS, Lam Employment Absence and Productivity Scale; Q-Q, Quantity and Quality.

* $P < 0.001$.

† $P < 0.01$.

Table 5 – Mean multitest, multiconstruct correlations for online and paper-and-pencil (hard copy) test completion modes.

Survey mode	Amount-Quality mismatch within tests	Amount across tests	Quality across tests	Amount-Quality mismatch across tests
Online	0.28	0.32	0.17	0.04
Hard copy	0.52	0.39	0.26	0.24

(method variance) values. In fact, three of the four (excepting the Amount value for the online mode) were nominally lower). Interestingly, although the patterns of the two modes told the same story, correlations were nominally greater across the board for the hard copy mode than for online delivery. Nevertheless, these differences did not prove to be statistically significant.

Discussion

The data presented here indicate that a sample of leading tests of presenteeism did show some measurable commonality. There are significant correlations among items measuring amount and and among items measuring quality. There are, however, two serious caveats. First, the overall level of shared variance is weak and thus too low to allow any hopeful conclusions about construct validity for either amount or quality. Second, the overbearing influence of method variance among the four instruments that measure both amount and quality suggests that much of whatever is being characterized by such commonality is not productivity, but rather the consequences of the self-report form of measurement.

The strengths of this study include its comparison of the leading presenteeism instruments in the literature. To our knowledge, no other study has pitted this many tests of presenteeism against each other in a single investigation of a common task. Furthermore, a number of methodological issues have been uniquely addressed by analyzing the output of only those test items that deal with productivity, standardizing the recall period, and counterbalancing the order of presentation of these test items within the questionnaire. Finally, the use of an adaptation of Campbell and Fiske's multitrait, multimethod technique of summarizing data allowed us to shed light on the troublesome influence of method variance on data purportedly representing construct validity.

Having said this, our study has a number of limitations, some are the consequence of the just-mentioned methodological modifications. First, the alteration of any test can be criticized because removing items, even when they are deemed to be unnecessary for the task at hand, will change the context and may thus alter respondent behavior. We have addressed this, as noted earlier, by including "context" items in our questionnaire for several of the tests although these items were excluded from the statistical analysis for this study. Nevertheless, the complement of context items may have been incomplete because inclusion was not always feasible when a large number of items were involved. It was beyond the scope of this study to pursue the matter further, but the effects of item context would be of interest for future investigation in any case. Second, our setting of the recall period to 7 days for all instruments required a change in the protocols of four of the tests that used other durations. Although it cannot be argued that changing the recall period will have no effect, it seems equally indefensible to assume that any of the instruments in question will stop performing because of a variation in the recall period. As in the case of most limitations, this may need to be determined by a comparative study involving several instruments and several periods of recall. A third issue that may limit generalizations from the present study is that it is not necessarily a given that high intercorrelations are to be expected among presenteeism measures from different instruments, particularly when they may differ on the involvement of health conditions or other factors. It is not clear how this potential issue can explain the strong influence of method variance (a within-test variable), but it is an interesting question nonetheless. This could be addressed by comparing responses to the various health-specific forms of tests such as the WPAI on a sample of workers exhibiting a representative selection of the conditions in question (a similar approach could be followed on by converting health-neutral instruments into new health-specific forms for research purposes). Finally, the aforementioned possibility of effects because of differences between the hard copy and online presentations of the survey was not strongly supported by the data, but the fact that a confound was nonetheless present (online and hard copy respondents also differed on their work duties) points to an unanswered question that could serve as a useful topic for further research.

Our study, like many others, cannot make definitive pronouncements about the ultimate utility of self-report tests of presenteeism without comparing their performance to some acceptable criterion standard. The literature cited earlier has made it clear that these instruments, individually or as a class, have not been properly examined for criterion validity. Given the importance of presenteeism and its assessment, this should be an area of high priority.

Conclusions

The initial objective of this project was to examine the level of construct validity among leading self-report presenteeism instruments. The findings clearly show some support for construct validity because there is visible commonality across most tests. This, however, proved to be a weak association overall, and moderate at best. Furthermore, it was found that method variance appears to be stronger than construct variance, suggesting that these tests may not be measuring productivity, but rather some attribute that is inherent to their structure and/or operation. This begs the question of whether any one, or more, of the existing tests can adequately stand in for direct measures of productivity on the job. At this juncture, the data clearly point to the need for investigations of the criterion validity of self-report presenteeism instruments.

Acknowledgments

Any benefit that may result from this investigation will be due, in large part, to the 137 individuals who contributed their time, professional opinions, and personal information to make this a useful endeavor. We are indebted to Alberta Health Services (AHS) for hosting this study. Several people within the AHS administration provided important support during the completion of this project, including Stephen Gould (AHS Vice President—Human

resources) and Dr Kathryn Todd (Vice President—Research) for pointing us in the right direction and Cindy Gerdes for her effectiveness in guiding the project through a complicated system. There were many others within AHS who made a difference. These were Andrew Switzer, Christine Whitford, Amy Manchak, Mishaela Houle, Lois Stephaniuk, Mei Tom, Rae Emogene, Jennifer Yelland, Noella Inions, Peggy Mann McKeown, Michelle Fry, Chris Ackerman, and Monique Duval. Kathleen LaBranche played a major role in the extraction and preparation of activity log data. Gary Pitcher set up the online response collection component. We are particularly grateful to Philip Jacobs and Egon Jonsson of the Institute of Health Economics for their support for this project.

Source of financial support: This study was funded by the Alberta Depression Initiative.

REFERENCES

[1] Koopman C, Pelletier KR, Murray JF, et al. Stanford Presenteeism Scale: health status and employee productivity. J Occup Environ Med 2002;44:14–20.
[2] Johns G. Presenteeism: a short history and a cautionary tale. In: Houdmont J, Leka S, Sinclair R,eds. Contemporary Occupational Health Psychology: Global Perspectives on Research and Practice. Chichester, UK: Wiley-Blackwell, 2012. p. 204–20.
[3] Stewart WF, Ricci JA, Chee E, Morganstein D. Lost productive work time costs from health conditions in the United States: results from the American Productivity Audit. J Occup Environ Med 2003;45:1234–46.
[4] Collins JJ, Baase CM, Sharda CE, et al. The assessment of chronic health conditions on work performance, absence, and total economic impact for employers. J Occup Environ Med 2005;47:547–57.
[5] Aronsson G, Gustaffsson K. Sickness presenteeism: prevalence, attendance-pressure factors, and an outline of a model for research. J Occup Environ Med 2005;47:958–66.
[6] Mattke S, Balakrishnan K, Bergamo G, et al. A review of methods to measure health-related productivity loss. Am J Manag Care 2007;13:211–7.
[7] Shikiar R, Halpern MT, Rentz AM, Khan ZM. Development of the Health and Work Questionnaire (HWQ): an instrument for assessing workplace productivity in relation to worker health. Work 2004;22:219–29.
[8] Lam RW, Michalak EE, Yatham LN. A new clinical rating scale for work absence and productivity: validation in patients with major depressive disorder. BMC Psychiatry 2009;9:78.
[9] Endicott J, Nee J. Endicott Work Productivity Scale (EWPS): a new measure to assess treatment effects. Psychopharmacol Bull 1997;33:13–6.
[10] Brouwer WB, Koopmanschap MA, Rutten FF. Productivity losses without absence: measurement validation and empirical evidence. Health Policy 1999;48:13–27.
[11] Campbell DT, Fiske DW. Convergent and discriminant validation by the multitrait-multimethod matrix. Psychol Bull 1959;56:81–105.
[12] Podsakoff PM, Organ DW. Self-reports in organizational research: problems and prospects. J Manag 2012;12:531–44.
[13] Podsakoff PM, MacKenzie SB, Podsakoff PN. Sources of method bias in social service research and recommendations on how to control it. Ann Rev Psychol 2012;63:539–69.
[14] Brown JD. Evaluations of self and others: self-enhancement biases in social judgments. Soc Cogn 1986;4:353–76.
[15] Alicke MD, Govorun O. The better-than-average effect. In: Alicke MD, Dunning DA, Kruger JL,eds. The Self in Social Judgment. New York, NY: Psychology Press, 2005. p. 85–106.
[16] Noben CYG, Evers SMAA, Nijhuis FJ, et al. Quality appraisal of generic self-reported instruments measuring health-related productivity changes: a systematic review. BMC Public Health 2014;14:115.
[17] Ospina MB, Dennett L, Waye A, et al. A systematic review of the measurement properties of instruments assessing presenteeism. Am J Manag Care 2015;21:e171–85.
[18] Abma FI, van der Klink JJL, Terwee CB, et al. Evaluation of the measurement properties of self-reported health-related work-functioning instruments among workers with common mental disorders. Scand J Work Environ Health 2012;38:5–18.
[19] Lofland JH, Pizzi L, Frick KD. A review of health-related workplace productivity loss instruments. Pharmacoeconomics 2004;22:165–84.
[20] Prasad M, Wahlqvist P, Shikiar R, et al. A review of self-report instruments measuring health-related work productivity: a patient-reported outcomes perspective. Pharmacoeconomics 2004;22:225–44.
[21] Roy J-S, Desmeules F, MacDermid JC. Psychometric properties of presenteeism scales for musculoskeletal disorders: a systematic review. J Rehabil Med 2011;43:23–31.
[22] Reilly MC, Zbrozek AS, Dukes EM. The validity and reproducibility of a work productivity and activity impairment instrument. Pharmacoeconomics 1993;4:353–65.
[23] Reilly Associates. The Work Productivity and Activity Impairment Questionnaire—Specific Health Problem (WPAI:SHP V2.0) 2010: Available from: http://www.reillyassociates.net/WPAI_SHP.html. [Accessed December 17, 2012].
[24] Lynch W, Riedel JE. Measuring Employee Productivity: A Guide to Self-Assessment Tools. Scottsdale, AZ: Institute for Health and Productivity Management, 2001:35.
[25] Thompson AH, Ospina MB, Dennett L, et al. A systematic review of the measurement properties of self-report instruments that assess presenteeism. Available from: http://www.ihe.ca/advanced-search/a-systematic-review-of-the-measurement-properties-of-self-report-instruments-that-assess-presenteeism. [Accessed May 15, 2016].
[26] Kessler RC, Barber C, Beck AL, et al. The World Health Organization Health and Work Performance Questionnaire (HPQ). J Occup Environ Med 2003;45:156–74.
[27] Kessler RC, Ames M, Hymel PA, et al. Using the WHO Health and Work Performance Questionnaire (HPQ) to evaluate the indirect workplace costs of illness. J Occup Environ Med 2004;46(Suppl. 6):S23–37.
[28] Zhang W, Bansback N, Kopec J, et al. Measuring time input loss among patients with rheumatoid arthritis: validity and reliability of the Valuation of Lost Productivity questionnaire. J Occup Environ Med 2011;53:530–6.
[29] Goetzel RZ, Ozminkowski RJ, Long SR. Development and reliability analysis of the Work Productivity Short Inventory (WPSI) instrument measuring employee health and productivity. J Occup Environ Med 2003;45:743–62.
[30] Stewart WF, Ricci JA, Leotta C. Health-related lost productive time (LPT): recall interval and bias in LPT estimates. J Occup Environ Med 2004;46(Suppl. 6):S12–22.
[31] Williams EJ. Experimental designs balanced for the estimation of residual effects of treatments. Aust J Sci Res 1949;Ser. A2:149–68.
[32] Steiger JH, Browne MW. The comparison of interdependent correlations between optimal linear composites. Psychometrika 1984;49:11–24.
[33] Evans JD. Straightforward Statistics for the Behavioral Sciences. Pacific Grove, CA: Brooks/Cole Publishing, 1996.