
'This article may not exactly replicate the final version published in the APA journal. It is not the copy of record.'

The Reliability, Validity, and Accuracy of Self-Reported Absenteeism from Work: A Meta-Analysis

Gary Johns and Mariella Miraglia

Concordia University, Montreal, Quebec

Authors’ Note

Department of Management, John Molson School of Business, Concordia University, 1455 de Maisonneuve Blvd. West, Montreal QC H3G 1M8, Canada.

Voice: (514) 848-2424 ext. 2914; Fax: (514) 848-4292; E-mail: gjohns@jmsb.concordia.ca; m_miragl@jmsb.concordia.ca
Abstract

Due to a variety of access limitations, self-reported absenteeism from work is often employed in research concerning health, organizational behavior, and economics, and it is ubiquitous in large scale population surveys in these domains. Several well established cognitive and social-motivational biases suggest that self-reports of absence will exhibit convergent validity with records-based measures but that people will tend to underreport the behavior. We used meta-analysis to summarize the reliability, validity, and accuracy of absence self-reports. The results suggested that self-reports of absenteeism offer adequate test-retest reliability and that they exhibit reasonably good rank order convergence with organizational records. However, people have a decided tendency to underreport their absenteeism, although such underreporting has decreased over time. Also, self-reports were more accurate when sickness absence rather than absence for any reason was probed. It is concluded that self-reported absenteeism might serve as a valid measure in some correlational research designs. However, when accurate knowledge of absolute absenteeism levels is essential, the tendency to underreport could result in flawed policy decisions.

Keywords: absence, absenteeism, self-reported absenteeism, meta-analysis, self-reports
The Reliability, Validity, and Accuracy of Self-Reported Absenteeism from Work: A Meta-Analysis

The use of self-reports has generated considerable controversy among researchers in health psychology, organizational psychology, and organizational behavior (e.g., Chan, 2009; Kammeyer-Mueller, Steel, & Rubenstein, 2010; Spector, 1994; Stone et al., 2000). On one hand, such reports are relatively easy to obtain and often provide information that is hard to access by any other means. On the other hand, self-reports have been criticized as being prone to a variety of biases encompassing inaccurate memory, inflated self-presentation, and self-delusion. In addition, the exclusive use of self-reports has been implicated in inflating relationships among variables due to common method variance (Podsakoff, MacKenzie, & Podsakoff, 2012). The controversy surrounding self-reports signals the need for research concerning their psychometric properties. However, the content of self-reports varies substantially, ranging from inner states and traits to descriptions of work experiences (e.g., justice, bullying) to accounts of overt work behaviors. In turn, this content itself varies in accessibility to memory, degree of threat and social desirability, extent of verifiability, and other features pertinent to the quality of response (Tourangeau & Yan, 2007). Hence, dedicated research into the properties of specific kinds of self-reports is necessary (Donaldson & Grant-Vallone, 2002; Stone et al., 2000). In this study we use meta-analytic techniques to probe the reliability, validity, and accuracy of self-reports of absenteeism from work.

The quality of self-reported absenteeism data is an important issue. Such data are common in health research (e.g., Borritz, Rugulies, Christensen, Villadsen, & Kristensen, 2006;
Gosselin, Lemyre, & Corneil, 2013) and not uncommon in organizational psychology and organizational behavior research (e.g., Avery, Volpone, McKay, King, & Wilson, 2012; Shapira-Lishchinsky & Tsemach, 2014). Below, we note trends in sampling which portend an increase in such usage. In fact, most absenteeism measured in labor economics, epidemiology, occupational medicine, and medical clinical trials is self-reported. To the extent that self-reported absenteeism is compromised, incorrect inferences may be made about employee productivity, employee health, the efficacy of pharmaceutical interventions, and other important policy-related matters.

Johns (1994a; 2003) has discussed reasons why self-reports of absenteeism are utilized. For one thing, some organizations do not do a thorough job of recording absenteeism, especially for certain occupational groups. For instance, managerial absence data may be hard to obtain or recognized as contaminated due to granting days off in lieu of extensive overtime work. Even when well recorded, extant data might not meet a researcher’s particular needs. For example, absenteeism specifically due to health problems, stress, bullying, child care, elder care, or substance abuse may not be denoted as such in company records. Furthermore, well recorded data is no guarantee that access to it is available. Organizations may view such data as proprietary, unions may balk at granting access to members’ attendance records, and employees themselves may be reluctant to self-identify so as to allow access. All of these reactions stem from the well established mildly deviant connotations of absenteeism (Johns, 1994a; 2003).

When survey respondents are approached outside of the workplace and work in varied employment settings, no economical alternative to self-report is available, and comparability across employers would be difficult to achieve in any event. This applies to the aforementioned epidemiological research, most clinical trials, and a wide variety of public surveys that measure
absence, including the US Bureau of Labor Statistics Current Population Survey, the US Panel Study of Income Dynamics (e.g., Gifford, 2013), the Statistics Canada Labour Force Survey (e.g., Dabboussy & Uppal, 2012), the European Survey on Working Conditions (e.g., Muckenhuber, Burkert, Dorner, Großschädl, & Freidl, 2013), the Australian Workplace Barometer (e.g., McTernan, Dollard, & LaMontagne, 2013), and Columbia’s Continuous Household Survey (e.g., Restrepo & Salgado, 2013). Also, we note an increasing use of “respondents for hire” accessed via online panels and crowdsourcing platforms, using SurveyResponse, Amazon’s Mechanical Turk, and the like (Behrend, Sharek, Meade, & Wiebe, 2011; Mason & Suri, 2012; Stanton & Weiss, 2002). Whatever its various merits, this sampling practice precludes the use of records-based employer data.

Two decades ago, Johns (1994a) published a mainly qualitative review of the use of self-reported absenteeism in research studies. Much of that review remains instructive, including an extensive discussion of theoretical perspectives, detailed descriptions of various measures, and various suggestions for obtaining quality self-reports of absence. In the current research we provide a quantitative update of the reliability, validity, and accuracy of self-reported absenteeism, with reliability referring to test-retest stability, validity referring to correlation with records-based data, and accuracy referring to mean degree of deviation from records-based data. In brief, Johns (1994a) a) uncovered two markedly different estimates of test-retest reliability, b) reported convergent validity coefficients ranging from .30 to .92, with a sample size weighted but otherwise uncorrected average of .64 for 6 time lost coefficients, and c) found some evidence of underreporting in a number of individual studies. With the luxury of 20 more years of research history and advanced electronic search techniques, we considerably supplemented Johns’s (1994a) original data and employed meta-analysis to a) estimate the time-standardized
reliability of self-reported absenteeism employing reliability generalization (Vacha-Haase, 1998; Viswesvaran, & Ones, 2000), b) estimate the population convergent validity of absence self-reports using state-of-the-art reliability corrections, and c) quantify the degree of accuracy of self-reports. Possible moderators of these psychometric characteristics were also examined.

Although there are many factors that might affect the psychometric quality of self-reported absenteeism, two stand out—memory and motives. From a cognitive viewpoint, Johns (1994a, p. 585) noted that the “salience of the behavior, motives to remember, and motives to rehearse one’s experience will often be low.” These factors suggest that memory for absenteeism will be faulty, but not randomly faulty. This is because of the mildly deviant connotations of the behavior (Robinson & Bennett, 1995; Runcie, 1988), which bias people to “forget” or redefine some absence incidents so as to maintain a positive self-image. Evidence for the generalized negativity of absenteeism is exemplified by the following research findings: Probationary employees exhibit minimal absence compared to prevailing organizational norms (Dello Russo, Miraglia, Borgogni, & Johns, 2013); absence behavior is a common source of conflict between employees and managers (Johns, 1997); survey respondents sometimes report guilt or shame concerning the behavior (Hackett, Bycio, & Guion, 1989; Knapstad, Øverland, Henderson, Holmgren, & Hensing, 2014), even relabeling it to enhance social acceptability (e.g., “mental health day,” Hackett et al., 1989); people report that their own attendance is better than that of their colleagues (Harrison & Shaffer, 1994; Johns & Xie, 1998). This last finding is an example of the better-than-average effect, the tendency to see oneself as superior to one’s peers, especially in the domain of negative behavior (Sedikides & Alicke, 2012). This effect, along with the selective self-memory effect, in which people show poor recall for negative information about themselves, are prime examples of the pervasiveness with which people both assume and
project a positive rather than negative self-image (Sedikides & Alicke, 2012). The established negative connotations of absenteeism, in combination with the much-demonstrated robustness of these two psychological biases, give rise to the following hypothesis, which speaks to the accuracy (Johns, 1994a) of the self-report:

_Hypothesis 1:_ People have a tendency to underreport their own absenteeism compared to corresponding data from organizational records.

Although it was expected that people would tend to underreport their absence, such underreporting should occur with reference to their actual absence behavior. As such, there are both cognitive and motivational reasons to expect that under-reports might nonetheless be reasonably well correlated with organizational records. First, given the pervasive robustness of anchoring effects on human judgment (Furnham & Boo, 2011), one’s current level of absenteeism should provide a potent cognitive anchor for the self-report (Nisbett & Ross, 1980). In other words, even when underreporting, people will do so in light of their current level of absence, and those with higher or lower absence will adjust their self-reports relative to this higher or lower anchor. Such anchoring suggests that the complete distortion of one’s absence record is unlikely, a point reinforced by the false consensus effect. Under this bias, people tend to view their own behavior as reasonable and typically appropriate (Marks & Miller, 1987). This means that even high absentees will see their own behavior as relatively normal (if a bit better than that of others) and will be disinclined to radically underreport. Some evidence for this expectation can be seen in the demonstrated validity of overt integrity tests, in which some job applicants openly express supportive attitudes regarding dishonest and counterproductive (e.g., absenteeism) work behaviors (Ones, Viswesvaran, & Schmidt, 1993). The anchoring and false
consensus biases both suggest that actual and self-reported absence will exhibit similar rank orders even if they have different means. Hence, it was predicted that:

_Hypothesis 2_: Self-reports of absenteeism will exhibit convergent validity with regard to records-based absence data.

We only had two _a priori_ notions concerning possible moderators of the effects predicted by Hypotheses 1 and 2. These had to do with the time span of the self-report recall period and the corresponding aggregation period for the records-based data. Especially given that absenteeism is a low base rate behavior, it was expected that an increasing time span would facilitate the validity of the self-report, due to increasing reliability of both measures. In other words, as both measures sample more instances in which absence might occur, they should become more reliable. On the other hand, the distortions of memory noted in support of Hypothesis 1 should be most operative as the time span of recall increases. Basic memory decay combined with motives to forget suggest that longer reporting periods will foster inaccuracy.

_Hypothesis 3_: a) The validity of absence self-reports will increase with the reporting time span, while b) their accuracy will decrease.

In addition to the continuously measured reporting period we also explored the possible moderating effects of two other continuous variables (proportion of women in the sample, year of publication) and three categorical variables: absence index (frequency, time lost); absence cause (sickness, general absence); source of records-based measure (employing organization, insurance register).

Given what we knew to be a paucity of reliability estimates for self-reported absence, we declined to offer a specific hypothesis about just how reliable the self-reports would be.
However, we expected that they would exhibit enough reliability to underpin the hypothesized accuracy and validity effects.

**Method**

**The meta-analytic databases**

Building upon the studies reported in Johns (1994a), we searched *PsychInfo, MedLine, Web of Science, Business Source Complete, ProQuest Business*, and *Google Scholar* using the keywords *accuracy or validity or validation or reliability of self-reported absenteeism or absence* or *sick leave or sick calls; registries or records or recorded (of) absenteeism or absence or sick leave or sick calls*. We also manually searched the reference lists of all the articles found as well as an extensive number of absenteeism articles collected by the first author. As a result, 57 studies were identified.

To be included in the meta-analyses, a study had to compare a self-reported measure of absenteeism with its records-based counterpart (mostly employer records; occasionally state, provincial, or national insurance registers). After this screening, 43 studies were retained. Moreover, the articles had to report the zero-order correlation between the self-reported and recorded measures or the difference between the means of the two variables with its standard deviation or other quantitative information that allowed the calculation of these indexes (e.g., confidence interval, standard errors). Additionally, studies had to clearly indicate the time frame over which absenteeism was aggregated in order to precisely correct for measurement error. Two studies were discarded because of incommensurate comparisons, since they matched self-reported and recorded absenteeism referring to different reasons (health problems versus general causes, Short et al., 2009) or to different absence indexes (time lost versus frequency of absence, Beehr & Gupta, 1978). In considering multiple statistics from the same study, we used the
following criteria: when a study reported effect sizes or correlations separately for men and women, we averaged them between the two groups (Ferrie et al., 2005; Stapelfeldt, Jensen, Andersen, Fleten, & Nielsen, 2012; Voss, Stark, Alfredsson, Vingård, & Josephson, 2008); when a study reported statistics for multiple absence time periods we retained the 12 month data or that which best approximated that figure, since 12 months was the most common time frame among the studies (Grøvle et al., 2012; Revicki, Irwin, Reblando, & Simon, 2004; Severens, Mulder, Laheij, & Verbeek, 2000); where independent samples were investigated (Kessler et al., 2003), we coded these separately. This screening process resulted in a final set of 30 studies (27 published articles, 2 unpublished master’s theses, 1 unpublished conference paper), containing 19 zero-order correlations for the convergent validity meta-analysis of correlations and 21 effect sizes for the accuracy meta-analysis using $d$. All zero-order correlations and effect sizes were based on independent samples except for the inclusion of Gaziel (2004), which reported both voluntary and involuntary absenteeism. Eleven studies presented data used in both meta-analyses. In addition, 4 studies allowed us to calculate the diagnostic odds ratio (Glas, Lijmer, Prins, Bonsel, & Bossuyt, 2003), an overall measure of diagnostic accuracy expressing the strength of association between the questionnaire measuring absenteeism and actual absence behavior. These studies, which also speak to convergent validity, were analyzed separately, as in Darr and Johns (2008).

**Coding process and moderator variables**

Two independent raters (one of the authors and a PhD student familiar with meta-analysis and absenteeism research) coded the self-report reliabilities, the correlation between self-reports and records-based data, and the means of the two absenteeism measures. The respective inter-
rater reliabilities were 1.00, 1.00, .92 (self-report means), and .90 (records-based means). Disagreements were resolved through discussion.

As noted earlier, we did not have strong expectations for moderators except for the possibility that the absenteeism aggregation period would condition the results. Hence, aggregation period was measured continuously and examined for any connection with the psychometric properties of the self-reports. Other continuously coded variables included the year of publication, to identify any psychometric trends over time, and gender mix (proportion of women in the sample), as women tend to exhibit more absenteeism than men (Côté & Haccoun, 1991). Several categorical measurement properties were also examined as potential moderators. Absences were coded as reflecting time lost (total hours, days, or weeks missed) or frequency (total inceptions, irrespective of duration). Reasons for absence were coded as reflecting sickness absenteeism (including accidents and mental health) or general, unattributed, absenteeism. Source of the records-based data was coded as emanating from organizational records or insurance registers. The inter-rater reliabilities for the continuous moderators were 1.00. Interrater agreement (Cohen’s kappa) for the categorical moderators was 1.00 for absenteeism measure and the source of records-based data and .78 for absence reasons.

Meta-analytic procedure

Three primary meta-analyses were performed: a) an analysis of the reliability of self-reported absenteeism employing reliability generalization (Vacha-Haase, 1998; Viswesvaran & Ones, 2000), b) an analysis of correlations between self-reported and recorded absenteeism, testing the convergent validity of self-reports, c) an analysis of Cohen’s $d$, which expresses the mean difference between the absence measures, to verify the accuracy of self-reports in terms of deviation from their objective counterparts.
The random effects meta-analytic procedure (Hunter & Schmidt, 2004) was employed, correcting for sampling error in every case and measurement error in the validity and accuracy analyses, using reliability artifact distributions to estimate a true population parameter. To correct for unreliability in the records-based measures, we used the reliability coefficients from the recent meta-analysis by Johns and Al Hajj (2014) which reported a sample size weighted population reliability of .53 for total time lost, .52 for frequency of absence, and .53 for the combined indexes \((K = 95, N = 25,587)\). To correct for unreliability in the self-reports we conducted an original bare-bones meta-analysis (Hunter & Schmidt, 1990, 2004) of the four test-retest self-report absence reliability coefficients identified in our search (Blau, Tatum, & Cook, 2004; Dalton & Mesch, 1991; Laestadius, Ye, & Dimberg, 2008; Tang & Hammontree, 1992). Like Ones, Viswesvaran, and Schmidt (2003) and Berry, Lelchook, and Clark (2012), we equated the reliabilities for the length of the absence aggregation period used in each study with the Spearman-Brown prophecy formula. Since the most common period among the studies was 12 months, we transformed the reliability coefficients measured on shorter or longer absenteeism time frames to a 12 month base. Then, we meta-analyzed the square roots of these converted coefficients and obtained the population estimate of self-report reliability by squaring the resulting meta-analytic mean (see Johns & Al Hajj, 2014; Vacha-Haase, 1998; Viswesvaran & Ones, 2000). For the record, we also discovered one study that reported an alpha reliability coefficient for self-reported absenteeism (Gaudine & Gregory, 2010). Incorporating this study into our reliability analysis resulted in no material changes in any of the results to be reported. Thus, we retained only the test-retest coefficients, since these correspond in type to the reliability estimates used for records-based absence.
To test the convergent validity of self-reported absenteeism a meta-analysis was conducted that included the 19 zero-order correlations from the primary studies corrected for unreliability in the records-based measure only. We used the Johns and Al Hajj (2014) meta-analytic reliability for combined indexes tailored to the absence aggregation period used in each primary study, based on the Spearman-Brown formula. Second, a parallel analysis was conducted correcting for measurement error in both absenteeism measures. We also performed a sensitivity analysis removing the study by Ferrie et al. (2005), because it had a sample size of 7,995, almost three times the sum of the remaining studies.

Regarding the meta-analysis for accuracy, we first calculated the standardized mean difference (Cohen’s $d$) between self-report and records-based absenteeism measures for each of the 21 primary studies. Following Dunlap, Cortina, Vaslow, and Burke (1996), Morris and DeShon (2002), and Hunter and Schmidt (2004), $d$ was calculated with the formula for correlated designs, as the ratio of the difference between means (i.e., records-based minus self-report measure) and the standard deviation of this difference, which is equal to \( \frac{S_r - S_{SR}}{\sqrt{2(1-r)}} \) with \( S_r \) representing the standard deviation for records-based absenteeism, \( S_{SR} \) representing the standard deviation for self-reports, and \( r \) representing the correlation between the two measures. In this case, the sample size weighted estimated mean correlation from the previous meta-analysis was employed. For three studies the standard deviation was derived from the confidence interval and in one case $d$ was computed from a $t$ statistic, as indicated in Dunlap et al. (1996, p. 171). Absenteeism expressed in weeks or hours was converted into days. Finally, to correct for sampling error and unreliability in both measures, we transformed the $d$ values to correlations, performed the meta-analysis on the correlations, and transformed the outcomes back to $d$, as

The same steps used in the previous analysis of correlations were followed. Thus, the first meta-analysis included all of the 21 studies corrected only for unreliability in the records-based absenteeism measure. Then, this analysis was replicated correcting for measurement error in both the self-reported and recorded measures. Third, a sensitivity analysis was performed omitting the study by Voss et al. (2008), given its $N$ of 4,477, almost the half of the total sample size of the other studies.

We applied the 75% rule concerning variance accounted for by artifacts (Hunter & Schmidt, 2004) and the width of the 80% credibility intervals (Whitener, 1990) as rules of thumb to detect heterogeneity across study results. For categorical moderators (absence indexes, absence causes, source of records-based measure) meta-analyses were performed for each subset separately for $r$ and $d$. To correct for measurement error in the records-based variable we used the time lost reliability coefficient from Johns and Al Hajj (2014) for the meta-analyses assessing time lost, while their frequency reliability estimate was employed for the meta-analyses assessing frequency. Correlations were employed to verify the moderating role of the three continuous variables, absenteeism aggregation period, year of publication, and gender mix. Therefore, the absenteeism aggregation period, expressed in months, the publication year, or the percentage of women in the sample was correlated with the zero-order correlations between self-report and recorded measures as well as Cohen’s $d$. The resulting observed correlations were then corrected for sampling error to obtain population correlations. That is, the correlation between $r$ and the moderator was corrected using the ratio of the variance of population correlation $\rho$ to the variance of $r$ from the previous meta-analysis, and the correlation between $d$
and the moderator was corrected using the ratio of the variance of population effect size $\delta$ to the variance of $d$ from the previous meta-analysis, applying the formulas in Hunter and Schmidt (2004, pp. 93, 301).

**Results**

**Reliability of self-reports**

As seen in Table 2, combining the data from the four studies that reported test-retest reliabilities, the time-standardized sample size weighted population reliability of self-reported absenteeism was .79 (the square of $\tilde{r} = .89; K = 4, N = 300, 95\% CI = .84, .94$).

**Validity of self-reports**

Table 3 reports the results from the meta-analyses of the correlations between self-reports and records. The sample size weighted validity coefficient was .73, or .66 with the deletion of Ferrie et al. (2005). For reference, Cohen (1988) described $r$s of .10 as small, .30 as medium, .50 as large, and .70 as very large. Whether we corrected for measurement error in the records-based measures only or in both measures, the population correlation estimate exceeded 1 (respectively, $\rho = 1.09, \rho = 1.23, K = 19, N = 11,479$), indicating high convergence between the two absenteeism indicators and providing support for Hypothesis 2. Eliminating Ferrie et al. (2005), the population correlation increased ($\rho = 1.43, K = 18, N = 3,484$). All the validities were significant, as the nonzero positive 95% confidence interval revealed.

The variance accounted for by artifacts was lower than 75% (Hunter & Schmidt, 2004) and the 80% credibility interval was quite wide (Table 3), indicating the possible presence of moderators. As for the continuous variables, the corrected correlation of absenteeism aggregation period with the association between self-report and records-based measures of absence was positive ($r_{corr} = .17$), as expected from Hypothesis 3a, although not significant due to the low $K$. 
The corrected correlation of the year of publication with the association between self-reported and record-based measures was .12, and that for percentage of women was .67 ($p < .05$). Thus, the convergent validity of the self-report increased with the proportion of women in the study.

Among the categorical variables, the estimated population correlation was higher for time lost ($ho = 1.51$, $K = 14$, $N = 3,103$) than for the frequency index ($ho = .99$, $K = 4$, $N = 381$), and the confidence intervals did not overlap, indicating a moderating effect (Table 4). We re-ran the frequency meta-analysis without Mueller, Wakefield, Price, Curry, and McCloskey (1987), since it was the only study adopting a single-day frequency measure. The resulting validity estimate was larger ($ho = 1.15$, $K = 3$, $N = 262$) than the previous one, but the confidence interval still showed no overlap with time lost. The subgroup analyses for absenteeism reasons did not exhibit overlapping confidence intervals (Table 4) and highlighted that the estimated population correlation was higher when individuals were asked to recall absenteeism for sickness reasons ($ho = 1.72$, $K = 6$, $N = 2,103$) than for any general cause ($ho = 1.16$, $K = 9$, $N = 919$). To investigate the possible effect due to how absences were recorded, a further meta-analysis was conducted that compared those studies retrieving absenteeism from organizational records with those based on insurance register data. The population correlation was greater for organizational records, but only two insurance register studies were available for comparison.

A further meta-analysis was conducted on 4 investigations published in medical journals aimed at evaluating the performance of instruments measuring sickness absenteeism or identifying employees at risk for health problems (Agius et al., 1994; Burdorf, Post, & Bruggeling, 1996; Duijts, Kant, Landeweerd, & Swaen, 2006; Fredriksson, Toomingas, Torgén, Bildt Thorbjörnsson, & Kilbom, 1998). The studies presented indexes of sensitivity (the proportion of individuals with sickness absenteeism correctly identified by the instrument) and
specificity (the proportion of persons without sickness absenteeism correctly identified by the instrument) which allowed calculation the diagnostic odds ratio (DOR). With the formula presented in Glas et al. (2003) and Jones, Darzi, and Athanasiou (2011) we calculated the DOR for each study and then converted the log DOR to \( d \) (Borenstein et al., 2009, p. 47) and, finally, to point biserial correlations, as in Darr and Johns (2008), to compare the results with the meta-analysis of correlations assessing convergent validity. The results corroborate the previous meta-analysis, as the population correlation corrected for sampling error and measurement error in both variables was .99 \( (K = 4, N = 7,753, SD_δ = .11, 95\% \ CI = .95, 1.03) \). A sensitivity analysis excluding the study by Duijts et al. (2006) because of its large sample size \( (N = 6,780) \) provided a population estimate of 1.15 \( (K = 3, N = 973, SD_δ = .16, 95\% \ CI = 1.10, 1.20) \).

**Accuracy of self-reports**

The results from the meta-analyses of \( d \) are reported in Table 5. All the 95% confidence intervals excluded zero, suggesting significant effects. For the overall meta-analysis, the estimated true mean difference was .28 \( (K = 21, N = 9,799) \) after correction for measurement error in both variables. This is indicative of an overall degree of inaccuracy in self-reports. When we excluded the study by Voss et al. (2008) given its disproportionate \( N \), the population mean difference estimate increased to .34 \( (K = 20, N = 5,322) \). For reference, Cohen (1988) described \( ds \) of .20 as small, .50 as medium, and .80 as large.

Of the 21 samples included in the overall meta-analysis, 18 supported the idea that employees underreport their absences, with a mean tendency of 2.03 days \( (SD = 2.19) \), ranging from .05 to 8.40. However, three studies indicated a tendency for individuals to overestimate their absence behavior by .52 (Revicki et al., 1994), 1.39 (Goldberg & Waldman, 2000) and .30 days (Gaziel, 2004) \( (M = .74, SD = .58) \). The effect size corrected for sampling error and
unreliability in both measures for the 18 studies indicating underreporting was unchanged at .28
\(SD = .46, K = 18, N = 9,571, 95\% CI = .22, .35\). This supports Hypothesis 1.

In addition, of the seven studies not included in the meta-analyses of \(d\) due to missing
information, several presented data in line with the underreporting hypothesis, ranging from .4
days a year (Dalton & Mesch, 1991) to 1 day (Ferrie et al., 2005), to 5.7 days (Bouwmans et al.,
2013). In 2 independent samples, Kessler et al. (2003) also concluded that self-reports
underestimated the numbers of work hours missed (6.4 per month) and the numbers of days
missed (1.3 per month). However, two studies documented the tendency for self-reports to
overestimate insurance-reported data for duration of sick leave and disability (Dasinger,
Both focused on the accuracy of insurance register records to address whether the cessation of
benefits corresponded to the full recovery and return to work of employees. The findings
confirmed the authors’ expectations that insurer-compensated days are lower than the actual time
lost from work and, hence, that employees receive fewer benefits than they might expect given
their actual absenteeism, prompting calls for a reconsideration benefits policies (Dasinger et al.,
1999).

As with the previous meta-analysis, the variance explained by sampling and measurement
error fell below the 75% guideline, and the 80% credibility intervals were wide. Thus, we tested
for possible moderators. Considering the continuous variables, the corrected correlation of the
absenteeism aggregation period with the standardized mean difference between self-reported and
recorded measures was positive, as predicted by Hypothesis 3b, and moderate \(r_{corr} = .29\),
though again not significant given the low \(K\). However, the corrected correlation between
publication year \(d\) was decidedly negative \(r_{corr} = -.71\), indicating that more recent studies
featured greater accuracy in self-reports. The corrected correlation between the proportion of women in the sample and \( d \) was .12.

Among the categorical moderators, the population difference was smaller for time lost estimates \((\delta = .32, K = 17, N = 5,028)\) compared to frequency estimates \((\delta = 1.25, K = 3, N = 294)\) (Table 6). However, the latter analysis produced negative standard deviation in the corrected correlation, probably attributable to the small numbers of studies (Hunter & Schmidt, 2004). Studies that avoided casual attributions for absenteeism produced a large effect size \((\delta = .79, K = 10, N = 1,942)\), while studies focusing sickness absence showed an effect smaller than the overall meta-analysis \((\delta = .26, K = 8, N = 3,054)\). In terms of the source of the recorded absence measure, the mean difference was larger for those studies retrieving absence from insurance registers \((\delta = .67, K = 3, N = 344)\) than for those based on employer records \((\delta = .31, K = 16, N = 4,866)\).

**Publication Bias Check**

It is possible that research showing inaccurate or invalid self-reports of absenteeism is less likely to be published. To examine possible publication bias, we used multiple advanced methods, as suggested by Kepes, Banks, McDaniel, and Whetzel (2012) and Harrison, Banks, Pollack, O’Boyle, and Short (in press). First, we adopted the funnel plot technique (Sterne & Egger, 2005; Sterne, Gavaghan, & Egger, 2005), which plots the magnitude of the effect sizes of individual studies on the X axis and their precision (i.e., the inverse of the standard error) on the Y axis. This way, larger effect sizes will lie on the right side of the plot with smaller effects on the left. Moreover, more precise samples, having bigger sample sizes and lower standard errors, will be displayed on the top of the plot while less precise studies will be at the base. This implies that asymmetry in the studies’ distribution denotes the presence of publication bias, since small
samples with non-significant effects (i.e., those less likely to get published) have been excluded from the analyses. Second, we integrated the funnel plots with the trim and fill method (Duval & Tweedie, 2000) which tests the degree of symmetry in the distribution and, if asymmetrical, imputes the number of missing samples, through a process that initially “trims” the asymmetric studies from the right-hand side of the plot and inserts them back on the correct side (i.e., “fills”) to gain symmetry. Moreover, it re-calculates the meta-analytic mean effect size based on the primary and the newly imputed studies, providing an estimation of the effect not influenced by publication bias. Finally, we used Egger’s intercept test (Egger, Smith, Schneider, & Minder, 1997), considered the most powerful among the correlation/regression-based tests (Borenstein et al., 2009; Kepes et al., 2012), in which publication bias is inferred by a regression model for which the intercept is expected to pass through the origin, and thus approach zero, indicating a symmetrical distribution.

Regarding the meta-analysis for convergent validity (i.e., $r, K = 19$), the funnel plot was symmetrical, the trim and fill method suggested that no studies were missing, and Egger’s intercept test was not significant ($\beta_0 = .12; p = .48$). Hence, we concluded that there was no evidence for publication bias. For the meta-analysis for accuracy (i.e., $d, K = 21$), the funnel plot exhibited marginal asymmetry and the trim and fill procedure resulted in a mean effect size of .14 (95% CI .10-.18), slightly lower than the original effect size of .17 (Table 5), with 3 studies missing. However, the difference in magnitude was not of a notable size in that the original effect size was still included in the 95% CI of the imputed effect size. Thus, we can consider the original meta-analysis robust with respect to publication bias (Kepes et al., 2012), especially given that Egger’s test revealed a nonsignificant intercept ($\beta_0 = .83; p = .48$).
For sake of precision, we also performed the publication bias analyses separately in moderator subgroups (Table 6), since asymmetry can be caused by moderating effects. We re-ran the analyses only for the three subgroups based on 10 studies or more (i.e., time lost, general absenteeism, and organizational records), as recommended by Kepes et al. (2012). Asymmetry was reduced for the subgroups with time lost studies and with organizational records studies, where only one and two samples, respectively, were missing. In the subgroup including general absenteeism studies, the samples were symmetrically distributed and no missing study was imputed through the trim and fill procedure. Finally, Egger’s intercept test was not significant in any of the three subgroups.

**Discussion**

The purpose of this study was to provide a quantitative summary of the psychometric properties of self-reported absenteeism, which is employed in a wide variety of research endeavors. So, are self-reports of absenteeism psychometrically sound? In support of Hypothesis 1, we affirmed that people have a tendency to under-report their absenteeism compared to the data found in organizational records. Thus, in a mean difference sense, self-reports tend to be inaccurate. In support of Hypothesis 2, despite under-reporting, we found that self-reports exhibited reasonable rank order convergence with records-based data. Thus, they exhibit convergent validity. In support of Hypotheses 3a and 3b, there was some tendency for validity to be enhanced and for accuracy to decrease when the self-report spanned a longer period of time. Below, these and some other emergent moderator findings are discussed in more detail. First, since adequate reliability is necessary to establish validity, accuracy, and moderation, we elaborate on the reliability of absence self-reports.
Clearly, we do not have a very good understanding concerning the reliability of self-reported absenteeism. On the surface, our population estimate of .79 is promising, nearly equivalent to the unweighted average of .77 that Freund and Kasten (2012) reported for 54 self-evaluations of cognitive ability. However, it is based on only 4 test-retest coefficients and a small total N, and it is much higher than the reliability of records-based data. Furthermore, the sources of variance that underlie repeated queries about the occurrence of behavior mentally aggregated over time are sketchy. Two of our coefficients mirrored the kinds of measures that are seen for virtually all records-based reliability estimates—respondents estimated their absenteeism for the past 6 months (Tang & Hammontree, 1992) or past year (Blau et al., 2004) and then did the same 6 months or a year later. These are clearly stability indices that treat variation over time as error. On the other hand, Laestadius et al. (2008) and Dalton and Mesch (1991) asked respondents to report twice on the previous year’s absence, with respective intervals of 5 days and 6 weeks (respective raw reliabilities of .98 and .91). Such measures are more akin to parallel forms than stability measures, and they might suffer from transient error due to variations in mood or attention but not changes in the underlying behavior over time (Le, Schmidt, Harter, & Lauver, 2010). Despite this uncertainty, we hasten to note that, Hippocratically, our reliability estimate has “done no harm,” in that it led to small corrections compared to those for the records-based measures, which were separately reported. Also, it is apparent that self-reported absence is reliable enough to enable reasonable inferences about (in)accuracy and validity, points to which we now turn.

In support of Hypothesis 1, it can be stated with some assurance that people tend to underreport their own absenteeism compared to organizational records, particularly when reporting periods are longer (confirming Hypothesis 3b) and absence for any reason is queried.
This tendency is reduced, but not erased, by focusing specifically on absence due to sickness. This might be because sickness is both a salient and socially acceptable reason for missing work, except in cases where people are reticent to self-disclose chronic illness (Munir, Leka, & Griffeths, 2005). Thus, people might be more inclined to both remember and report it in questionnaires. Also, sickness absence frequently requires medical certification, an effortful event that might cue memory for the absentee and facilitate accurate reportage in organizational records. Additionally, people may be more attuned to literally enumerating sick days, since it is these that are most inclined to be contractually specified and compensated. Absence for reasons other than sickness may be estimated rather than enumerated, a sure opportunity for motivated forgetting (Johns, 1994a), as discussed in the introduction.

The tendency to underreport decreased over the time span of the data collection. It is possible that this corresponds to increased job insecurity over the years in the face of globalization, organizational restructuring, and economic uncertainty. Given job insecurity, people may be more motivated to keep accurate account of their attendance behavior, since excessive absenteeism is an objective criterion for laying people off. Also, there is some evidence that social science survey response rates have declined in recent years (e.g., Baruch & Holtom, 2008), and it is possible that those who still respond are among the more assiduous in their reporting.

Cohen (1988) would denote the overall population estimate for under-reporting ($\delta = .28$) to be between small and medium. However, some of our subgroup analyses resulted in considerably stronger effects (e.g., $\delta = .79$ for general absenteeism), and it bears emphasis that even smallish effects can translate into large aggregate estimation errors when, for instance, days of absence are translated into dollars of productivity loss.
The assurance from the meta-analytic results must be qualified by what we had to work with. Specifically, the modal aggregation period for the accuracy analyses was 12 months and the mean was 8.53 months. Suffice it to say that the raw \( ds \) for the lowest aggregation periods in our sample (1 to 1.5 months) were as large or larger than our average corrected population coefficients (Voss et al., 2008, included). Thus, within the range of aggregation studied, a short time frame does not seem to guarantee accuracy. Unfortunately, we do not know whether a very short one week reporting period (e.g., Statistics Canada Labour Force Survey, Dabboussy & Uppal, 2012; Columbia’s Continuous Household Survey, Restrepo & Salgado, 2013) forestalls underreporting, but underestimation is likely in the Australian Workplace Barometer (McTernan et al., 2013), which uses a one month reporting period, and the European Survey on Working Conditions and the Panel Study on Income Dynamics, both of which employ one year query periods.

In support of Hypothesis 2, the convergent validity of self-reported absenteeism, uncorrected for measurement error, ranges between .66 and .73, depending on whether the large dataset of Ferrie et al. (2005) is included. Correction for measurement error resulted in validity estimates greater than unity. In a rank order sense, people who miss more days report more absence and vice versa. This is reassuring in light of the general tendency to underreport, although reasonable minds will disagree about how much convergence is necessary to use self-reports in lieu of records, even ignoring common method variance concerns for self-reports. Despite potential memory issues, a longer aggregation period (as predicted in Hypothesis 3a) and focusing on sickness absence boosted validity. Also, validities were higher in samples with greater proportions of women, although samples with more women were marginally more inclined to exhibit under-reporting. Restriction of range issues may underpin both effects. Since
men are less inclined to be absent (Côté & Haccoun, 1991), restriction of range would attenuate the validity coefficient for men. At the same time, more days of absence on average afford women more opportunity to forget an absence event.

Our results are notable in light of some other convergent validities of self-assessments appearing in the meta-analytic literature. For instance, Jaramillo, Carrillat, and Locander (2005) reported corrected validities of self-reports of sales performance to be .19 against managerial ratings and .34 for objective performance measures. Similarly, Freund and Kasten (2012) reported a corrected association of .33 between self-estimated and psychometrically assessed cognitive ability, and Carpenter, Berry, and Houston (2014) reported a corrected association of .26 between self- and other-rated organizational citizenship behavior. By these standards, the validity results for absenteeism are encouraging, and surprising for a skewed, low base rate behavior.

Much health psychology, organizational psychology, and organizational behavior research employs correlational designs. When access to records is impossible, such designs might profit from the convergence between self-reports and records-based data, and the tendency to underreport might not be a serious issue. In other words, carefully measured self-reports (see below) are better than having no absence data at all in research in which absolute accuracy is not imperative. For instance, studying the relationship between stress and absence requires that people are rank ordered appropriately in terms of absence (i.e. a valid measure), and a bit of under-reporting should not affect the results. However, other research designs might suffer badly from inaccurate reporting because they place a premium on knowing exact absence levels. As a case in point, a plethora of studies in occupational medicine have attempted to tie dollar figures to the impact of various illnesses on productivity loss due to work absenteeism and
presenteeism (working while ill) (Hemp, 2004; Johns, 2010; Schultz & Edington, 2007).

Virtually all of this research is based on self-reports, and one of the most common findings is the tendency for more aggregate productivity to be lost due to going to work ill than to absenteeism; that is, sample-wise, people lose more productivity to presenteeism because the behavior is more common than absence. To the extent that absence is underreported, as per our results, this discrepancy is exaggerated, and policy decisions stemming from it may be flawed.

Although much more research is needed on the subject, several tactics should enhance the psychometric quality of absence self-reports (Johns, 1994a). In line with our introduction, these encompass facilitating memory and tempering motivation to misrepresent one’s attendance history. Given the results for accuracy, confirming Hypothesis 1, a particular challenge is getting people to report all of their absenteeism. In every case, care should be taken to carefully define absenteeism for the respondent in light of the researcher’s goals, as careful definitions preclude the tendency to discount or redefine absences (“I wasn’t really absent that day, the car wouldn’t start”). In particular, the moderation results suggest specifically querying sickness absence if the study concerns health. Also, potential threat should be countered by reminding respondents that almost everyone has been absent occasionally, and it is thus a normal part of working life. For retrospective estimates, our results for Hypothesis 3 suggest care in timing the probe, as validity increases with time while accuracy decreases. If individual-level stability and convergent validity are most important, longer retrospective periods are desirable. If absolute accuracy is important and data are mainly employed in descriptive aggregated form, shorter periods are called for, although they do not fully preclude under-reporting. In all cases, a free response format should be used (fill in the blank with days missed), since there is copious evidence that probes of frequency of behavior are badly biased by fixed numerical scale formats,
which connotes how much behavior is “normal” (Schwarz, 1999). Although somewhat cumbersome to implement, aided recall has been used to have respondents systematically enumerate their absences, beginning with the most recent (Nicholson, & Payne, 1987; Shapira-Lishchinsky & Rosenblatt, 2009). Finally, one of the most promising ways to avoid errors in reporting due to memory would be to employ daily diary experience sampling techniques using mobile phone technology (Fisher & To, 2012; Uy, Foo, & Aguinis, 2010).

As for limitations, the incidental use of absence data might be missed in computerized searches, and we thus might have overlooked a couple of relevant data points despite an additional intensive manual search. Also, only a limited number of studies have examined the psychometric properties of self-reported absence, and the research designs are almost by necessity survey-based, lacking the control that might be achieved in experiments and clinical trials. Thus, our results can be viewed as promising but modifiable with further research. We are acutely aware that some organizational absence records may have their own flaws and thus represent less than the gold standard for judging validity and accuracy. This said, users of self-reports (and journal reviewers) often indicate that they would have preferred records-based data. Nevertheless, research on the accuracy of organizational absence records would be welcome. We suspect that such records might be more accurate in blue collar, hourly paid, and unionized settings and less accurate in managerial and professional ranks, where unrecorded days off might be given in compensation for unpaid overtime work or stressful assignments. Also, our results tentatively suggest that insurance registers are a poor surrogate for employer records, given that self-reports over-estimate the former and underestimate the latter.

A research agenda for absence self-reports would identify the situational and personal variables that are associated with better and poorer reports. The former might include variations
in absence control systems and contractual provisions for sick days. The latter might include personality variables and personal conceptions of health. Again, motives and memory should underlie any anticipated effects. Our finding that queries specifically having to do with sickness absence were more accurate is of great interest. Although our speculations on the reasons for this given above are feasible, the documented subjectivity of much day-to-day illness (e.g., Wikman, Marklund, & Alexanderson 2005) might suggest that the requirement for illness attributions would reduce accuracy. Hence, more work is needed to understand the psychometric impact of making attributions about various causes of absenteeism. Variations in reporting periods need to be systematically examined. Although psychometric logic provided an underpinning for Hypothesis 3, we were limited in our empirical inferences to the rather restricted variability in periods shown in Table 1. On a closing note, it would be very helpful if research using records-based data would also measure self-reports as a matter of course. This very simple step would greatly increase the opportunity to better understand the dynamics of self-reports.

Johns (1994a) stressed the need for more research devoted to understanding the psychometric properties of self-reported absence, and it is gratifying to see some progress on this front in recent years. Indeed, a considerable number of the articles in our meta-analyses were dedicated attempts to study such properties. This said, more research is warranted, especially in the case of large scale public surveys, for which we have essentially no psychometric evidence. Important policy decisions are made using such surveys, and it is essential to understand the quality of the obtained responses.
References

References marked with a double dagger (‡) indicate studies included in the meta-analysis of self-report reliability; references marked with an asterisk (*) indicate studies included in the meta-analysis of correlations; references marked with a dagger (†) indicate studies included in the meta-analysis of d; references marked with a circumflex (^) indicate studies included in the meta-analysis of odds ratios.


VALIDITY AND ACCURACY OF SELF-REPORTED ABSENTEEISM


VALIDITY AND ACCURACY OF SELF-REPORTED ABSENTEEISM


†Severens, J. L., Mulder, J., Laheij, R. J. F., & Verbeek, A. L. M. (2000). Precision and accuracy in measuring absence from work as a basis for calculating productivity costs in The


<table>
<thead>
<tr>
<th>Meta-analysis</th>
<th>Study</th>
<th>N</th>
<th>Effect size</th>
<th>S-R measure</th>
<th>S-R reason</th>
<th>Records-based source</th>
<th>Time in months</th>
</tr>
</thead>
<tbody>
<tr>
<td>Validity</td>
<td>Adler et al., 2005</td>
<td>433</td>
<td>$r = 0.16$</td>
<td>Time Lost</td>
<td>Sickness</td>
<td>Org. records</td>
<td>3</td>
</tr>
<tr>
<td>Validity (OR)</td>
<td>Agius et al., 1994</td>
<td>265</td>
<td>$\bar{r} = 0.66$</td>
<td>Dichotomy</td>
<td>Sickness</td>
<td>Org. records</td>
<td>12</td>
</tr>
<tr>
<td>Validity and reliability</td>
<td>Blau et al., 2004</td>
<td>87</td>
<td>$r = 0.70$</td>
<td>Frequency</td>
<td>General</td>
<td>Org. records</td>
<td>12</td>
</tr>
<tr>
<td>Validity (OR)</td>
<td>Burdorf et al., 1996</td>
<td>404</td>
<td>$\bar{r} = 0.71$</td>
<td>Time Lost</td>
<td>Sickness</td>
<td>Org. records</td>
<td>6</td>
</tr>
<tr>
<td>Reliability</td>
<td>Dalton &amp; Mesch, 1991</td>
<td>62</td>
<td>rel. $= 0.91$</td>
<td>Time Lost</td>
<td>Sickness</td>
<td>Org. records</td>
<td>12</td>
</tr>
<tr>
<td>Validity (OR)</td>
<td>Duijts et al., 2006</td>
<td>6780</td>
<td>$\bar{r} = 0.38$</td>
<td>Frequency</td>
<td>Sickness</td>
<td>Org. records</td>
<td>4</td>
</tr>
<tr>
<td>Validity</td>
<td>Ferrie et al., 2005</td>
<td>7995</td>
<td>$r = 0.77$</td>
<td>Time Lost</td>
<td>Sickness</td>
<td>Org. records</td>
<td>12</td>
</tr>
<tr>
<td>Validity (OR)</td>
<td>Fredriksson et al., 1998</td>
<td>304</td>
<td>$\bar{r} = 0.72$</td>
<td>Time Lost</td>
<td>Sickness</td>
<td>Insurance register</td>
<td>12</td>
</tr>
<tr>
<td>Validity, accuracy, and reliability</td>
<td>Gaudine &amp; Gregory, 2010</td>
<td>192</td>
<td>$r = 0.90$; $a = 0.89$</td>
<td>Time Lost</td>
<td>General</td>
<td>Org. records</td>
<td>12</td>
</tr>
<tr>
<td>Validity and accuracy</td>
<td>Gaziel, 2004 – Measure 1</td>
<td>148</td>
<td>$r = 0.91; d = -2.45$</td>
<td>Time Lost</td>
<td>Voluntary Involuntary</td>
<td>Org. records</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td>Measure 2</td>
<td></td>
<td>$r = 0.84; d = -0.08$</td>
<td>Time Lost</td>
<td>General</td>
<td>Org. records</td>
<td>6</td>
</tr>
<tr>
<td>Validity and accuracy</td>
<td>Goldberg &amp; Waldman, 2000</td>
<td>98</td>
<td>$r = 0.60$</td>
<td>Time Lost</td>
<td>General</td>
<td>Org. records</td>
<td>6</td>
</tr>
<tr>
<td>Accuracy</td>
<td>Grøvle et al., 2012</td>
<td>44</td>
<td>$d = 0.14$</td>
<td>Time Lost</td>
<td>Sickness</td>
<td>Insurance register</td>
<td>12</td>
</tr>
<tr>
<td>Validity and accuracy</td>
<td>Harrison &amp; Shaffer, 1994</td>
<td>112</td>
<td>$r = 0.69$</td>
<td>Frequency</td>
<td>General</td>
<td>Direct observation</td>
<td>1.5</td>
</tr>
<tr>
<td></td>
<td>Innes, 1985</td>
<td>30</td>
<td>$r = 0.92$</td>
<td>Time Lost</td>
<td>General</td>
<td>Org. records</td>
<td>10</td>
</tr>
<tr>
<td>Validity and accuracy</td>
<td>Johns, 1994b</td>
<td>129</td>
<td>$r = 0.59$</td>
<td>Time Lost</td>
<td>General</td>
<td>Org. records</td>
<td>12</td>
</tr>
<tr>
<td>Validity</td>
<td>Kessler et al., 2003 – Sample 1</td>
<td>441</td>
<td>$r = 0.71$</td>
<td>Time Lost</td>
<td>Sickness</td>
<td>Org. records</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>Sample 2</td>
<td>847</td>
<td>$r = 0.66$</td>
<td>Time Lost</td>
<td>Sickness</td>
<td>Org. records</td>
<td>12</td>
</tr>
<tr>
<td>Reliability</td>
<td>Laestadius et al., 2008</td>
<td>91</td>
<td>rel. $= 0.98$</td>
<td>Time Lost</td>
<td>Sickness</td>
<td>Org. records</td>
<td>12</td>
</tr>
</tbody>
</table>
### Validity and Accuracy of Self-Reported Absenteeism

<table>
<thead>
<tr>
<th>Validity and Accuracy</th>
<th>Study Authors, Year</th>
<th>Sample Size</th>
<th>Correlation</th>
<th>Measure</th>
<th>Measure Basis</th>
<th>Study Sample Size</th>
<th>Effect Size</th>
</tr>
</thead>
<tbody>
<tr>
<td>Validity</td>
<td>Linton, 1995</td>
<td>66</td>
<td>$r = .95$</td>
<td>Time Lost</td>
<td>Sickness</td>
<td>Insurance register</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>$d = 2.94$</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Validity</td>
<td>Merkesdal et al., 2005</td>
<td>234</td>
<td>$r = .83$</td>
<td>Time Lost</td>
<td>Sickness</td>
<td>Insurance register</td>
<td>18</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>$d = .48$</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Validity</td>
<td>Morison, 1985</td>
<td>89</td>
<td>$r = .76$</td>
<td>Time Lost</td>
<td>General Org. records</td>
<td>6</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Mueller et al., 1987</td>
<td>119</td>
<td>$r = .30$</td>
<td>Frequency General</td>
<td>Org. records</td>
<td>6</td>
<td></td>
</tr>
<tr>
<td>Validity and Accuracy</td>
<td>Newsome, 1993</td>
<td>63</td>
<td>$r = .54$</td>
<td>Frequency General</td>
<td>Org. records</td>
<td>12</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>$d = .77$</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Accuracy</td>
<td>Parker &amp; Kulik, 1995</td>
<td>46</td>
<td>$d = .13$</td>
<td>Time Lost</td>
<td>Sickness</td>
<td>Org. records</td>
<td>6</td>
</tr>
<tr>
<td></td>
<td>Rees &amp; Cooper, 1993</td>
<td>182</td>
<td>$r = .96$</td>
<td>Time Lost</td>
<td>Sickness</td>
<td>Org. records</td>
<td>6</td>
</tr>
<tr>
<td>Accuracy</td>
<td>Revicki et al., 1994</td>
<td>36</td>
<td>$d = -1.13^a$</td>
<td>Time Lost</td>
<td>Sickness</td>
<td>Org. records</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>$d = -.98^b$</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Accuracy</td>
<td>Severens et al., 2000</td>
<td>144</td>
<td>$d = .26$</td>
<td>Time Lost</td>
<td>Sickness</td>
<td>Org. records</td>
<td>12</td>
</tr>
<tr>
<td>Accuracy</td>
<td>Shapira-Lishchinsky &amp; Rosenblatt, 2009</td>
<td>1106</td>
<td>$d = .22$</td>
<td>Time Lost</td>
<td>General</td>
<td>Org. records</td>
<td>5</td>
</tr>
<tr>
<td>Validity and Accuracy</td>
<td>Spector, 1987</td>
<td>66</td>
<td>$r = .58$</td>
<td>Time Lost</td>
<td>Org. records</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>$d = .03$</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Reliability</td>
<td>Stapelfeldt et al., 2012</td>
<td>2311</td>
<td>$d = .15$</td>
<td>Time Lost</td>
<td>Sickness</td>
<td>Org. records</td>
<td>12</td>
</tr>
<tr>
<td>Accuracy</td>
<td>Tang &amp; Hammontree, 2007</td>
<td>60</td>
<td>rel. = .35</td>
<td>Time Lost</td>
<td>Sickness</td>
<td>Org. records</td>
<td>6</td>
</tr>
<tr>
<td>Accuracy</td>
<td>van Poppel et al., 2002</td>
<td>173</td>
<td>$d = .01$</td>
<td>Time Lost</td>
<td>Sickness</td>
<td>Org. records</td>
<td>12</td>
</tr>
<tr>
<td>Accuracy</td>
<td>van Poppel et al., 2002</td>
<td>173</td>
<td>$d = .01$</td>
<td>Time Lost</td>
<td>Sickness</td>
<td>Org. records</td>
<td>12</td>
</tr>
<tr>
<td>Accuracy</td>
<td>Voss et al., 2008</td>
<td>4477</td>
<td>$d = .14$</td>
<td>Time Lost</td>
<td>Sickness</td>
<td>Org. records</td>
<td>12</td>
</tr>
<tr>
<td>Accuracy</td>
<td>Winkler, 1980</td>
<td>57</td>
<td>$d = 2.70$</td>
<td>Time Lost</td>
<td>General</td>
<td>Org. records</td>
<td>12</td>
</tr>
</tbody>
</table>

**Note.** $OR = $ Odds ratio; $N = $ study sample size; S-R = self-report; Org. records = Organizational records; $r = $ correlation reported in the study; $\bar{r} = $ correlation calculated from $OR$; $d = $ effect size calculated from means and standard deviation of the self-report and records-based measures; rel. = test-retest reliability for the self-report measure; $\alpha = $ Cronbach’s alpha for the self-report measure. $^a$ The effect size was used in the subgroup meta-analysis for sickness absenteeism; $^b$ The effect size was used in the subgroup meta-analysis for general absenteeism.
Table 2

Meta-analysis for reliability of self-reports

<table>
<thead>
<tr>
<th></th>
<th>K</th>
<th>N</th>
<th>(\bar{r})</th>
<th>SD.(\bar{r})</th>
<th>Reliability</th>
<th>80% CV</th>
<th>95% CI</th>
</tr>
</thead>
<tbody>
<tr>
<td>Self-report reliabilities</td>
<td>4</td>
<td>300</td>
<td>.89</td>
<td>.10</td>
<td>.79</td>
<td>.76, 1.02</td>
<td>.84, .94</td>
</tr>
</tbody>
</table>

*Note.* K = total number of samples; N = total sample size across studies; \(\bar{r}\) = sample-size-weighted mean of square root of reliabilities; SD.\(\bar{r}\) = standard deviation of \(r\); Reliability = squared \(\bar{r}\); 80% CV = 80% credibility interval; 95% CI = 95% confidence interval.
Table 3

Meta-analyses for validity of self-reports

<table>
<thead>
<tr>
<th></th>
<th>K</th>
<th>N</th>
<th>$\bar{r}$</th>
<th>SD.$\bar{r}$</th>
<th>$\rho$</th>
<th>SD.$\rho$</th>
<th>$%$ var.</th>
<th>80% CV</th>
<th>95% CI</th>
</tr>
</thead>
<tbody>
<tr>
<td>Corrected for records-based</td>
<td>19</td>
<td>11,479</td>
<td>.73</td>
<td>.14</td>
<td>1.09</td>
<td>.26</td>
<td>1.17</td>
<td>.76, 1.42</td>
<td>1.08, 1.10</td>
</tr>
<tr>
<td>measurement error</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Corrected for measurement</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1.23</td>
<td>.37</td>
<td>.76</td>
<td>.76, 1.70</td>
<td>1.21, 1.24</td>
</tr>
<tr>
<td>error in records and self-</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>reports</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Without Ferrie et al. (2005)</td>
<td>18</td>
<td>3,484</td>
<td>.66</td>
<td>.23</td>
<td>1.43</td>
<td>.93</td>
<td>1.17</td>
<td>.23, 2.63</td>
<td>1.38, 1.48</td>
</tr>
</tbody>
</table>

Note. K = total number of samples; N = total sample size across studies; $\bar{r}$ = sample-size-weighted mean of correlations; SD.$\bar{r}$ = standard deviation of $\bar{r}$; $\rho$ = estimated population correlation (weighted average correlation corrected for unreliability in self-report and records-based variables); SD.$\rho$ = standard deviation of $\rho$; $\%$ var. = percent of variance accounted for by sampling and measurement error; 80% CV = 80% credibility interval; 95% CI = 95% confidence interval.
Table 4

Categorical moderators for validity

<table>
<thead>
<tr>
<th>Variable</th>
<th>K</th>
<th>N</th>
<th>(\bar{r})</th>
<th>SD.(\bar{r})</th>
<th>(\rho)</th>
<th>SD.(\rho)</th>
<th>% var.</th>
<th>80% CV</th>
<th>95% CI</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Self-report measure</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Time lost</td>
<td>14</td>
<td>3,103</td>
<td>0.67</td>
<td>0.23</td>
<td>1.51</td>
<td>0.96</td>
<td>0.28</td>
<td>2.74</td>
<td>1.46, 1.56</td>
</tr>
<tr>
<td>Frequency</td>
<td>4</td>
<td>381</td>
<td>0.55</td>
<td>0.17</td>
<td>0.99</td>
<td>0.57</td>
<td>6.20</td>
<td>0.26, 1.72</td>
<td>.85, 1.14</td>
</tr>
<tr>
<td>Frequency without Mueller et al. (1987)</td>
<td>3</td>
<td>262</td>
<td>0.66</td>
<td>0.07</td>
<td>1.15</td>
<td>0.62</td>
<td>3.65</td>
<td>0.35, 1.95</td>
<td>1.01, 1.29</td>
</tr>
<tr>
<td><strong>Self-report reason</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sickness absenteeism</td>
<td>6</td>
<td>2,103</td>
<td>0.61</td>
<td>0.25</td>
<td>1.72</td>
<td>1.40</td>
<td>0.64</td>
<td>-0.7, 3.52</td>
<td>1.63, 1.81</td>
</tr>
<tr>
<td>General absenteeism</td>
<td>9</td>
<td>919</td>
<td>0.66</td>
<td>0.18</td>
<td>1.16</td>
<td>0.37</td>
<td>6.40</td>
<td>0.67, 1.64</td>
<td>1.09, 1.22</td>
</tr>
<tr>
<td><strong>Records-based source</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Organizational records</td>
<td>15</td>
<td>3,072</td>
<td>0.63</td>
<td>0.23</td>
<td>1.48</td>
<td>0.99</td>
<td>1.23</td>
<td>0.21, 2.74</td>
<td>1.42, 1.53</td>
</tr>
<tr>
<td>Insurance register</td>
<td>2</td>
<td>300</td>
<td>0.86</td>
<td>0.05</td>
<td>1.19</td>
<td>0.57</td>
<td>0.35</td>
<td>0.47, 1.92</td>
<td>1.15, 1.24</td>
</tr>
</tbody>
</table>

Note. K = total number of samples; N = total sample size across studies; \(\bar{r}\) = sample-size-weighted mean of correlations; SD.\(\bar{r}\) = standard deviation of \(\bar{r}\); \(\rho\) = estimated population correlation (weighted average correlation corrected for unreliability in self-report and records-based variables); SD.\(\rho\) = standard deviation of \(\rho\); % var. = percent of variance accounted for by sampling and measurement error; 80% CV = 80% credibility interval; 95% CI = 95% confidence interval.
### Table 5

**Meta-analyses for accuracy of self-reports**

<table>
<thead>
<tr>
<th>Corrected for records-based measurement error</th>
<th>K</th>
<th>N</th>
<th>d</th>
<th>SDₐ</th>
<th>δ</th>
<th>SDδ</th>
<th>% var.</th>
<th>80% CV</th>
<th>95% CI</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>21</td>
<td>9,799</td>
<td>.17</td>
<td>.30</td>
<td>.25</td>
<td>.41</td>
<td>9.58</td>
<td>-.14,.38</td>
<td>.19,.30</td>
</tr>
<tr>
<td>Corrected for measurement error in records and self-reports</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>.28</td>
<td>.47</td>
<td>9.62</td>
<td>-.16,.44</td>
<td>.21,.35</td>
</tr>
<tr>
<td>Without Voss et al. (2008)</td>
<td>20</td>
<td>5,322</td>
<td>.19</td>
<td>.41</td>
<td>.34</td>
<td>.67</td>
<td>9.27</td>
<td>-.26,.59</td>
<td>.25,.44</td>
</tr>
</tbody>
</table>

**Note.** K = total number of samples; N = total sample size across studies; d = sample-size-weighted mean effect size; SDₐ = standard deviation of d; δ = estimated population effect size (weighted average effect size corrected for unreliability in self-report and records-based variables); SDδ = standard deviation of δ; % var. = percent of variance accounted for by sampling and measurement error; 80% CV = 80% credibility interval; 95% CI = 95% confidence interval.
### Table 6

**Categorical moderators for accuracy**

<table>
<thead>
<tr>
<th>Variable</th>
<th>K</th>
<th>N</th>
<th>d</th>
<th>SD(d)</th>
<th>(\delta)</th>
<th>SD(\delta)</th>
<th>% var.</th>
<th>80% CV</th>
<th>95% CI</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Self-report measure</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Time lost</td>
<td>17</td>
<td>5,028</td>
<td>.17</td>
<td>.41</td>
<td>.32</td>
<td>.68</td>
<td>8.15</td>
<td>-.27, .59</td>
<td>.23, .42</td>
</tr>
<tr>
<td>Frequency</td>
<td>3</td>
<td>294</td>
<td>.42</td>
<td>.16</td>
<td>1.25</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>.49, 2.91</td>
</tr>
<tr>
<td><strong>Self-report reason</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sickness absenteeism</td>
<td>8</td>
<td>3,054</td>
<td>.18</td>
<td>.27</td>
<td>.26</td>
<td>.41</td>
<td>13.00</td>
<td>-.27, .78</td>
<td>.14, .37</td>
</tr>
<tr>
<td>General absenteeism</td>
<td>10</td>
<td>1,942</td>
<td>.33</td>
<td>.32</td>
<td>.79</td>
<td>.54</td>
<td>24.03</td>
<td>.09, 1.49</td>
<td>.58, 1.02</td>
</tr>
<tr>
<td><strong>Records-based source</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Organizational records</td>
<td>16</td>
<td>4,866</td>
<td>.15</td>
<td>.38</td>
<td>.31</td>
<td>.61</td>
<td>9.66</td>
<td>-.23, .54</td>
<td>.21, .42</td>
</tr>
<tr>
<td>Insurance register</td>
<td>3</td>
<td>344</td>
<td>.65</td>
<td>.57</td>
<td>.67</td>
<td>1.12</td>
<td>5.74</td>
<td>-.77, 2.10</td>
<td>.34, 1.05</td>
</tr>
</tbody>
</table>

**Note.** K = total number of samples; N = total sample size across studies; d = sample-size-weighted mean effect size; SD\(d\) = standard deviation of d; \(\delta\) = estimated population effect size (weighted average effect size corrected for unreliability in self-report and records-based variables); SD\(\delta\) = standard deviation of \(\delta\); % var. = percent of variance accounted for by sampling and measurement error; 80% CV = 80% credibility interval; 95% CI = 95% confidence interval.
Footnote

1Various conditions influence the extent to which estimated validity coefficients can exceed unity, a subject that has received scholarly attention since the early 1900s (Charles, 2005; Muchinsky, 1996). In our particular case, two factors are contributory. One is the relatively high starting point for the reliability correction, as the basic bare bones meta-analysis yielded a sample weighted $r$ of .73. Sampling error accompanying the correction process dictates that a high population coefficient will occasion even higher values for half of the meta-analytic estimates of this figure, and some may exceed 1.0 (Hunter & Schmidt, 2004). The other is the low reliability for records based absenteeism data (.53). In a simulation, Zimmerman and Williams (1997) demonstrated that low reliabilities were a salient cause of overcorrection. Given this, we conducted a sensitivity analysis in which we simulated the reliability of the records based data to be .60, .70, and .80 using both individual corrections (tailored to time frame of each study) and omnibus artifact distributions. The lowest $\rho$ value obtained was .92, indicating that self-reports and records based data are in any event highly correlated. It bears emphasis that the exact population correlation is not of paramount interest in this particular case given the more applied focus on the viability of the self-report as a substitute for organizational data. Also, we agree with Charles (2005) that some imprecision in correction is better than assuming perfect measurement in the face of evidence to the contrary. For some caveats concerning corrections applied in the face of low criterion reliability see Le Breton, Scherer, and James (in press).