In-basket Validity: A systematic review

Deborah L. Whetzel*, Paul F. Rotenberry and Michael A. McDaniel*****

*Human Resources Research Organization, 66 Canal Center Plaza, Suite 700, Alexandria, VA 22314-1591, USA. dwhetzel@humrro.org
**Management Department, West Chester University, 308 Anderson Hall, West Chester, PA 19383-2230, USA
***Department of Management, Virginia Commonwealth University, 301 West Main Street, PO Box 844000, Richmond, VA 23284-4000, USA

In-baskets are high-fidelity simulations often used to predict performance in a variety of jobs including law enforcement, clerical, and managerial occupations. They measure constructs not typically assessed by other simulations (e.g., administrative and managerial skills, and procedural and declarative job knowledge). We compiled the largest known database ($k = 31$; $N = 3,958$) to address the criterion-related validity of in-baskets and possible moderators. Moderators included features of the in-basket: content (generic vs. job specific) and scoring approach (objective vs. subjective) and features of the validity studies: design (concurrent vs. predictive) and source (published vs. unpublished). Sensitivity analyses assessed how robust the results were to the influence of various biases. Results showed that the operational criterion-related validity of in-baskets was sufficiently high to justify their use in high-stakes settings. Moderator analyses provided useful guidance for developers and users regarding content and scoring.

1. Introduction

In-baskets are relatively high-fidelity simulations that have been used to predict performance in many occupations including clerical, law enforcement, managerial, and a variety of professional jobs (Schippmann, Prien, & Katz, 1990). In typical in-baskets, examinees are given documents often found in a job incumbent's inbox, such as memos, e-mails with and without attached reports, letters, and a calendar. Examinees respond to each document by indicating the action(s) they would take. Responses can include prioritizing tasks, identifying inconsistencies in scheduling, finding mistakes in expense vouchers, delegating correspondence and other activities to a subordinate, and determining how to handle a problem employee.

Often, in-baskets have been included as part of larger assessment centers to measure constructs such as decision making, delegation, adaptability, and other dimensions needed for managerial success (Thornton & Byham, 1982; Tsacoumis, 2007). However, in-baskets are often administered in isolation. Unlike other assessment center exercises, such as role-plays and leaderless group discussions that are administered in person and in group settings, in-baskets often are completed online without human interaction. In-baskets also measure different constructs than other simulation measures. Whereas role-plays and leaderless group discussions often measure interpersonal skills (e.g., conflict management and relating to others), in-baskets often measure managerial and administrative skills (e.g., delegation and attention to detail). They also measure declarative and procedural job knowledge (e.g., knowledge of human resources [HR] processes and general investigative knowledge). In addition, in-baskets can be scored fairly easily using rating scales, checklists, or multiple-choice test items, and many lend themselves to computer scoring. Given the ease of online administration and scoring, as well as the in-basket's inherent face validity, several multinational test vendors (e.g., Aon, Psychological Services, Inc., Development Dimensions International, and Select International) administer and score in-basket exercises

International Journal of Selection and Assessment

as stand-alone simulations over the Internet. Considering that they measure different constructs than other simulation measures and their widespread use, it is important to know the in-basket's psychometric characteristics, including reliability and validity.

Similar to assessment centers, interviews, and situational judgment tests (SJTs), in-baskets are a measurement method that assess multiple constructs rather than a single construct *per se* (Arthur & Villado, 2008). Although in-baskets can and likely do measure multiple constructs, there is some consensus on constructs typically assessed. Generally, statements on constructs assessed are based on assertions by the in-basket developers. For example, it is asserted that the in-basket technique measures various administrative skills linked to managerial success (e.g., Lopez, 1966). Other studies provide empirical support for statements of the constructs assessed. For example, Meyer (1970) found that a factor analysis of in-basket ratings produced two major performance dimensions: supervision and planning/administration. Tett and Jackson (1990) found that six behaviors were reliably measured with an in-basket: delegating, seeking advice, following advice, requesting to meet with an individual, seeking nonadvisory information, and asking to be kept informed as to how a problem is developing. Upon review of studies included in this meta-analysis, there is no apparent taxonomy that describes constructs measured by in-baskets. The skills assessed generally consist of decision making, planning and organizing, and managing others, which includes delegation, directing subordinates, and consideration.

1.1. Criterion-related validity of in-baskets and their moderators

When assessing the criterion-related validity of the in-basket, it is informative to compare the in-basket's validity with that of other simulations at various levels of fidelity. The criterion-related validity of assessment centers has been well documented in meta-analytic research (Arthur, Day, McNelly, & Edens, 2003; Gaugler, Rosenthal, Thornton, & Bentson, 1987; Hermelin, Lievens, & Robertson, 2007). The first assessment center meta-analysis was conducted by Gaugler et al. (1987). Using 50 studies that contained 107 validity coefficients, they obtained an operational validity estimate of .37 for predicting job performance. Hermelin et al. (2007) provided an update of Gaugler et al.'s meta-analysis by examining assessment center validity studies conducted since 1985. Using 26 studies that contained 27 coefficients, they obtained an operational validity estimate of .28 between the overall assessment rating (OAR) and supervisor ratings of job performance. Arthur et al. (2003) investigated the validity of six assessment center dimensions: communication, consideration, drive, influencing others, organizing and planning, and problem solving. Operational validity estimates ranged from .25 for consideration to .39 for problem solving, with an estimate across dimensions of .36. However, none of the assessment center meta-analyses examined the validity of in-baskets specifically. Thus, the current article is the seminal in-basket validity meta-analysis.

In-baskets can be considered work samples in which one performs tasks that are physically and/or psychologically similar to those performed on the job (Ployhart, Schneider, & Schmitt, 2006). Roth, Bobko, and McFarland (2005) conducted a meta-analysis of the validity of work sample tests. Specifically, they estimated the observed mean validity of work samples to be .26, which increased to .33 when corrected for criterion unreliability. We note that the work samples studied in Roth et al.'s meta-analysis included a variety of physical performance tasks, such as machinist tests for general vehicle mechanics (Engel, 1970) and psychomotor tests, such as typing (West & Bolanovich, 1963) and operating a sewing machine (Blum, 1943). Because Roth et al. included only one in-basket validity study in their meta-analysis (Meyer, 1970), and because the in-basket is quite different from the other physically oriented work samples, their validity estimate says little about the validity of the in-basket. Because the current research is specific to in-baskets, it provides an important contribution to the work simulation literature, in general, and the in-basket literature, in particular.

Because the in-basket is a simulation, it can be expected to have levels of validity comparable to other simulations. The assessment center and work sample literature reviewed earlier offers operational validities in the .20s and .30s. SJTs, often described as low-fidelity simulations (Motowidlo, Dunnette, & Carter, 1990), yield validities in the mid-.20s (McDaniel, Hartman, Whetzel, & Grubb, 2007). Given the meta-analytic results for assessment centers, work samples, and SJTs, we provide the following hypothesis:

Hypothesis 1: The in-basket will have criterion-related validity comparable with other simulations, such as assessment centers, work samples, and SJTs.

The criterion-related validity of in-baskets may be influenced by a variety of factors. We investigated two categories of moderator variables with our data: (a) characteristics of in-baskets, specifically whether the content is job specific (content is developed for a particular job) or generic (content is designed for use across multiple jobs), and how in-baskets are scored (objectively vs. subjectively); and (b) features of the study, specifically whether the study was published or unpublished, and whether the study used a predictive or concurrent design. The investigation of these moderators is important as they provide guidance to developers

and users regarding content and scoring of in-baskets, and to researchers who evaluate the validity of in-baskets in various settings. We discuss each of these moderators in turn.

1.1.1. Job-specific versus generic content of the in-basket

Several studies have used in-baskets designed for generic management jobs (e.g., the Bureau of Business in-basket; Frederiksen, 1962). Several test vendors offer online in-baskets that assess general managerial competencies across jobs and organizations. Although job analysis is used to justify their use in organizations, these in-baskets are not tailored to specific jobs in specific organizations. Thus, these in-baskets can be characterized as generic. On the other hand, in-baskets can be designed specifically for a particular job or organization (Atkins & Wood, 2002; Bentz, 1968). Job analysis is used to identify essential tasks to be simulated using the in-basket. Job-specific knowledge, skills, and abilities can be assessed using task-related documents that increase the fidelity of the in-basket. Thus, these in-baskets can be characterized as job-specific.

This distinction is analogous to the content categories of employment interviews studied by McDaniel, Whetzel, Schmidt, and Maurer (1994). They differentiated between job-related and psychological interviews. Job-related interviews were those conducted by a personnel officer or hiring official and the questions assessed past behaviors and job-related information of relevance to a specific job in an organization. Psychological interviews were typically conducted by a psychologist and the questions assessed general personal traits, such as conscientiousness. Because their content was relatively generic, psychological interviews applied to multiple jobs. McDaniel et al. (1994) found that job-related (specific) interviews were more valid than psychological (generic) interviews (.39 vs. .29). Thus, we offer:

Hypothesis 2: In-baskets with job-specific content will yield higher operational validity estimates than in-baskets with generic content.

1.1.2. Objective versus subjective scoring of the in-basket

Objectivity, in this discussion, refers to the use of a scoring key and the degree of human judgment involved in deriving a score. On one extreme, multiple-choice tests can be used to score in-baskets objectively. In such cases, the examinee selects an option that reflects what he/she would do in response to each item or rates the effectiveness of response options identified by the developer. A similar approach involves a list on which raters check off whether the examinee exhibited a given behavior (Felker, Curtin, & Rose, 2007). Objective methods are often used when in-baskets are administered and scored by computer. An advantage of an objective approach is likely increased reliability because of scoring standardization. Because reliability places constraints on the maximum levels of validity, we begin with a brief review of reliability evidence of work sample tests and then review the reliability of job performance ratings.

The joint services job performance measurement project conducted extensive studies of the reliability of work samples. They found increased reliability by focusing on observable behaviors and scoring task performance on each checklist item as a dichotomy (go/no go; Wigdor & Green, 1986) using a behavioral checklist. Using this approach, high levels of reliability were found for Project A (Knapp & Campbell, 1993). Similarly, Carey (1990) and Felker et al. (1988) reported agreements exceeding 90% between test scorers and 'shadow' scorers for a variety of Marine Corps job sample tests. Hedge, Lipscomb, and Teachout (1988) reported pairwise agreements ranging from almost 75% to 90% across three teams of test administrators and three Air Force occupations. Possible causes of these high levels of reliability include the rigorous approaches to developing, administering, and scoring work sample tests, including the thorough training of assessors and careful delineation of work steps used in the checklists. Thus, carefully developed work samples that focused on observable behaviors and were scored objectively (e.g., go/no go) likely contributed to these levels of interrater agreement.

In contrast, a subjective approach involves the use of rating scales (e.g., behaviorally anchored rating scales or graphic rating scales) to measure competencies assessed by the in-basket. Using a behaviorally based approach, a scale may include behavioral examples of responses a person could make at various levels of proficiency; raters then compare an examinee's response to examples provided in the scale. Studies have investigated the reliability of subjective performance ratings. Viswesvaran, Ones, and Schmidt (1996) compared reliability estimates of ratings from various sources. They found that supervisory ratings appeared to have higher reliability than peer ratings, and that the mean interrater reliability of supervisory ratings was .52 (using 40 reliability coefficients and a total sample of 14,650 participants) for overall job performance. Salgado, Moscoso, and Lado (2003) found levels of test–retest reliability of job performance ratings across performance dimensions (.57) that were comparable with Viswesvaran and colleagues.

Although we present the objective/subjective distinction as dichotomous, one could view the differences between behavioral checklists and behaviorally anchored rating scales as differing only in degree of objectivity. Thus, degree of objectivity could be considered a continuum. For this study, we made a distinction between objective and subjective methods of scoring to inform

developers of potential differences in the validity of these scoring options. Given the apparent higher level of reliability of objective scoring methods over subjective, we offer the following hypothesis:

Hypothesis 3: In-baskets that are objectively scored will have higher operational validity estimates than those that are subjectively scored.

1.1.3. Published versus unpublished data sources

We also investigated whether the source of the study was from published literature or from unpublished documents. There is a great deal of controversy, particularly in the medical field (e.g., Curfman, Morrissey, & Drazen, 2006), about the selective suppression of research results. In the typical case of publication bias, studies with small samples and statistically insignificant results are unavailable in the published literature (Chan, Hróbjartsson, Haahr, Gøtzsche, & Altman, 2004; Dickersin, 2005; McDaniel, Rothstein, & Whetzel, 2006; Pigott, 2009; Song et al., 2010). In the medical community, publication bias is a problem when the medical efficacy of a drug or treatment is substantially overestimated because of suppression of small effect results. In the fields of psychological, educational, and behavioral treatment, Lipsey and Wilson (1993) showed that published studies tend to yield higher effects sizes than unpublished studies (.53 vs. .39; table 3). Closer to the field of industrial/organizational (I/O) psychology, Kepes, McDaniel, Banks, and Whetzel (2012) found that a comparison of published with unpublished distributions for structured interviews indicated that published samples have larger effect size estimates. They attributed this result to a suppression of small effect size samples in the published literature. The higher effect sizes found in published studies may reflect better quality research or a preference of journals to report statistically significant results. Given the consistency of these results, we offer the following hypothesis:

Hypothesis 4: Validity studies that have been published (e.g., journals and book chapters) will yield higher validities than studies that have not been published (e.g., technical reports and conference presentations).

1.1.4. Concurrent versus predictive study design

We coded whether each study used a concurrent or predictive research design. Studies that use a predictive design involve administering the in-basket at one point in time (e.g., at application for a job or promotion) and then assessing job performance after the person is in the job for a period of time. Concurrent studies typically involve administering the test and a criterion measure at the same time. Given that criterion data are available, the participants in such studies are usually job incumbents. For tests of cognitive ability, results show that these two designs do not yield significantly different results (Barrett, Phillips, & Alexander, 1981; Schmitt, Gooding, Noe, & Kirsch, 1984). However, meta-analytic research has shown that concurrent study designs yield higher validity estimates than predictive study designs for several selection methods. For SJTs, the estimated operational validity using predictive studies was .18 ($k = 6$; $N = 346$), whereas the estimated operational validity using concurrent studies was .35 ($k = 96$; $N = 10{,}294$) (McDaniel, Morgeson, Finnegan, Campion, & Braverman, 2001). Huffcutt, Conway, Roth, and Klehe (2004) investigated the effect of study design on the validity of the interview. They found that studies using a predictive design yielded lower estimated population validities (.38 in the overall sample) than studies using a concurrent design (.48 in the overall sample; .53 in the adjusted sample). These findings are aligned with other meta-analytic results (e.g., Hough, 1998; Ones, Viswesvaran, & Schmidt, 1993). Thus, we offer:

Hypothesis 5: Studies using a concurrent design will yield higher operational validities than studies using a predictive design.

1.2. Cognitive loading of in-baskets

To the extent that in-baskets place demands on cognitive skills, such as reading, reasoning, and problem solving, it is reasonable to suggest that in-baskets are correlated with cognitive ability. This is useful information because it has implications for the potential incremental validity of in-baskets over *g* and the potential for subgroup differences. Simulations of varying levels of fidelity have been found to covary with cognitive ability. For example, McDaniel et al. (2007) found that SJTs were correlated with cognitive ability (mean observed validity = .29; $k = 95$; $N = 30{,}859$). Similarly, Roth et al. (2005) found a mean observed correlation of .32 ($k = 43$, $N = 17{,}563$) between work samples and cognitive ability. Thus, the correlations tend to establish a low to moderate relationship with cognitive ability but not so high as to rule out the possibility that in-baskets may provide incremental validity over cognitive ability. Given the consistency of the correlations between other types of simulations and cognitive ability, we offer the following hypothesis:

Hypothesis 6: Because of the cognitive requirements of in-baskets, they will be positively correlated with *g*.

1.3. Sensitivity analyses

Because meta-analytic reviews typically summarize all known studies on a particular topic, they tend to be cited frequently and have a large influence on the literature. Therefore, it is important to investigate the extent

to which meta-analytic results are robust to the influences of potential bias (Kepes et al., 2012; Kepes, McDaniel, Brannick, & Banks, 2013). Publication bias is defined as the tendency to publish studies depending on the magnitude, direction, or statistical significance of the results (Dickersin, 2005). Thus, to enhance the accuracy of meta-analytic reviews, both the samples (i.e., their effect sizes) included and those left out of the analysis must be considered.

Several methods for conducting sensitivity analyses were used for the current study. First, we identified a possible outlier (i.e., a study with large numbers of subjects that yielded atypical results) and conducted a separate meta-analysis without it to determine if the study's omission changed the results. Second, we conducted publication bias analyses consisting of trim and fill analyses (Duval & Tweedie, 2000a, 2000b) and two correlational and regression-based publication bias methods (Begg & Mazumdar, 1994; Egger, Smith, Schneider, & Minder, 1997).

1.4. Unique contribution of this study to science and applied practice

The primary purpose of this study was to evaluate the criterion-related validity of the in-basket exercise. Schppmann, Prien, and Katz (1990) conducted an exhaustive narrative review on the psychometric characteristics of the in-basket. They concluded that, in spite of its face validity, the compiled studies provided only modest evidence of criterion-related validity. Several meta-analyses have investigated the criterion-related validity of assessment centers, which included exercises such as in-baskets, role–plays, and leaderless group discussions. As such, one cannot determine the validity of the in-basket independent of the other exercises. Because of their fidelity to the job, in-baskets can be considered work samples. The Roth et al. (2005) study is an excellent review of the work sample literature; however, they included only one in-basket study (Meyer, 1970). Thus, although it substantially informs our knowledge of work samples in general, the Roth et al. study contributes little to our knowledge of the in-basket in particular.

The data compiled for this meta-analysis of in-basket validity is the most comprehensive validity collection to date. Thus, in addition to estimating criterion-related validity, we investigate plausible moderators related to features of the in-basket and features of the validation studies. These moderators will inform developers and users regarding decisions made during the creation of in-basket materials and their scoring. We also conduct sensitivity analyses to assess the extent to which our findings are robust to potential biases. As such, this article is an important contribution beyond Schippmann, Prien, and Katz's (1990) narrative review, the assessment center meta-analyses, and Roth et al.'s (2005) meta-analysis of work samples.

2. Method

2.1. Literature search

We consulted a variety of sources to gather available literature on the psychometric characteristics of in-baskets. First, we searched *PsycINFO*. Keywords for computerized searches included in-basket, in-box, in-tray, assessment center, administrative performance, managerial performance, validity, reliability, decision making, planning, organizing, delegation, and organizational skills. Once we obtained relevant articles, their reference lists were reviewed for additional studies. Second, we contacted test publishers to obtain technical documentation describing the validity of in-baskets. Third, we issued a call for papers via the Society for Industrial and Organizational Psychology (SIOP) website (http://www.SIOP.org). Fourth, we requested articles using http://www.LinkedIn.com listserv groups, including SIOP, International Personnel Assessment Council, Personnel Testing Council (PTC) of Metropolitan Washington, PTC Northern California, PTC Southern California, I/O Practitioners Network. We also requested articles through the HR listserv of the Academy of Management, and the Academy of International Business listserv. Note that out of the 31 coefficients used in this study, 23 (74%) came from published literature, four (13%) came from two consulting organizations, two (6%) came from a call for papers through LinkedIn, and two (6%) came via personal communication to the first author.

2.2. Decision rules

Several decision rules were used for inclusion of studies in this meta-analysis. We classified the decision rules into categories of how the studies were reported, the nature of the criteria, and scoring of the in-basket. Regarding the nature of the studies, because the purpose of this meta-analysis was to determine the criterion-related validity of the in-basket technique as a predictor of job performance, only studies that used the in-basket in an employment setting were included in the analysis. Further, many of the original assessment center studies that used in-baskets only provided validities using an overall assessment rating (OAR) or exercise dimension ratings. Because we could not isolate in-basket validities, these studies were not included (e.g., Dunnette, 1971; Huck & Bray, 1976; Wilson & Tatge, 1973). One study provided validity results for two in-basket scales (e.g., productivity and content) as well as a multiple *R* with a criterion and in that case, we coded the multiple *R* (Hakstian, Woolsey, & Schroeder, 1986). When studies

reported a range of numbers of participants in a study, we coded the smaller number to provide a more conservative estimate (e.g., Melchers & Kleinmann, 2009; Turnage & Muchinsky, 1984). Two samples were omitted because they reported only significant correlations of the in-basket with the criterion (i.e., Ginsburg & Silverman, 1972; Meyer, 1970 cross-validation sample). Inclusion of such studies or samples would tend to result in overestimating the operational validity. Data from independent samples of participants were entered into the meta-analysis as separate correlation estimates. Thus, when samples were identical across multiple studies, the most current edition served as the data source for that sample (e.g., Bentz, 1962, 1968). When two in-baskets were administered within the same study (Cross, 1969), one prior to employment and the other after being hired, we coded the validity from the first in-basket administration because it could be used to predict future job performance. In the Cross (1969) study, the in-basket was scored in more than one manner. The validity estimate generated using the method that resembled traditionally scored in-baskets (scoring on categories of 'stylistic' behavior such as 'discusses with subordinates' and 'recognizes good work') was chosen over the nontraditionally scored in-basket (scoring on semantic differential scale, such as urbane-rough and forceful-tentative). Studies that reported validity coefficients that had been corrected for artifacts, such as range restriction and criterion unreliability, were transformed back to the observed values. When the necessary information was not provided to make these adjustments, those coefficients were not included (e.g., Hakstian & Harlos, 1993).

Regarding the nature of the criterion, studies that did not include correlations of the in-basket with a criterion were not included (e.g., research that provided only multi-trait multi-method matrices and/or factor analyses [e.g., Donahue, Truxillo, Cornwell, & Gerrity, 1997] or research that did not include a criterion [e.g., Craik et al., 2002]). In addition, criteria such as personal temperament and interests (Frederiksen, 1966) and starting salary (Ward, 1960) were deemed unrepresentative of the job performance construct. Although an increase in managerial responsibility could be considered a useful proxy for job performance, Wollowick and McNamara (1969) was the only study found that used this criterion. Thus, it was excluded from the present analysis. Promotions, however, were considered a proxy for job performance (e.g., Melchers & Kleinmann, 2009). Note that with the exception of Melchers and Kleinmann, all criterion measures were supervisor ratings of job performance or supervisor ratings of promotion potential.

Per the Meta-Analysis Reporting Standards specified in the *Publication Manual of the American Psychological Association* (2010), we have listed the study source, *r*, *N*, and the coding of moderators in the Appendix A.

2.3. Inter-coder agreement

The first two authors of this study, one an experienced researcher with over 20 years of applied assessment experience, the other a faculty member with research and teaching experience in the area of employee selection, both with doctorate degrees in industrial/organizational psychology, coded all validity studies. For each study, we compared five data points: (a) *N*; (b) *r*; (c) objective versus subjective; (d) concurrent versus predictive; and (e) job specific versus generic. The published versus unpublished moderator was not part of this analysis because it was considered too obvious and would artificially inflate the estimate of interrater agreement.

To determine the level of inter-coder agreement, we identified the number of data points and the number of disagreements. For this analysis, there were 190 data points and 18 disagreements, yielding a 91% level of agreement. Disagreements were resolved by referring back to the decision rules (or creating new ones) and discussion between the two researchers. This high level of agreement replicates the findings of Whetzel and McDaniel (1988) concerning inter-coder agreement for validity generalization databases.

2.4. Meta-analytic techniques

Psychometric meta-analytic procedures (Hunter & Schmidt, 2004) were used in this study. Hunter and Schmidt's most recent meta-analytic program was used to analyze the data (Schmidt & Le, 2005). We used the Comprehensive Meta-Analysis (CMA) software (Borenstein, Hedges, Higgins, & Rothstein, 2005) to conduct the publication bias analyses.

2.4.1. One correlation per sample

Only one criterion-related coefficient was coded per sample and only one correlation with cognitive ability was coded per sample. When multiple correlation coefficients per sample were provided, we calculated a composite correlation when possible and used a mean correlation otherwise (Ghiselli, Campbell, & Zedeck, 1981, p. 163). We computed a composite under the following two conditions. The first condition was when multiple in-basket predictor dimensions were correlated separately with a criterion and either the (a) correlations among the predictors were provided or (b) mean correlation among predictors could be computed (e.g., Brass & Oldham [1976] reported correlations between six in-basket dimension scores). The second condition was when a single in-basket score was correlated with multiple criteria (e.g., one correlation for each of several supervisor ratings) and the correlations among the scales were provided (e.g., Kesselman, Lopez, & Lopez, 1982 who reported in-basket scores correlated with two performance criteria: Form A [traits ratings] and

Form B [behavioral observation scale ratings]). When the correlations among the variables were not provided, we averaged the validities for the sample. When overall performance was provided as one of the criteria, we coded it rather than computing a composite among remaining scales.

2.4.2. Corrections for statistical artifacts

Because 29 of the 31 validity coefficients in the job performance distribution were based on supervisor ratings, we adjusted the mean validity coefficient for unreliability (two coefficients were supervisor ratings of promotion potential). We relied on Pearlman, Schmidt, & Hunter's (1980) assumed distribution of proficiency criterion reliabilities, in which the expected value is .60 (Pearlman et al.'s table 1). This may provide an overestimate of reliability as later research estimated supervisor ratings reliability to be approximately .57 or .52 (Salgado et al., 2003; Viswesvaran et al., 1996). Thus, our estimates of the operational validity are likely conservative (i.e., downwardly biased). Our tables include both the 'bare bones' meta-analysis results with no corrections, as well as the fully corrected estimates. We also computed the operational validity estimates with and without correcting for range restriction. Because we did not have range restriction data available in our studies, we used range restriction estimates provided by Pearlman et al. (1980) and Schmidt and Hunter (1977). This will likely underestimate the validity of in-baskets because we corrected for direct range restriction only and because range restriction is more severe for more complex jobs (Schmidt, Shaffer, & Oh, 2008). When we conducted the meta-analysis of correlations between *g* and the in-basket, we assumed that *g* had a reliability of .90 (Berry, Sackett, & Landers, 2007; Huffcutt, Roth, & McDaniel, 1996).

2.5. Publication bias

Confidence in the validity and robustness of meta-analytic results depends on the extent to which publication bias influences research results (Kepes et al., 2012; Rothstein, Sutton, & Borenstein, 2005). CMA (Borenstein et al., 2005) was used to complete the following analyses: (a) a meta-analysis of observed correlations using a random-effects model; (b) Duval and Tweedie's (2000a, b) trim and fill analysis on the random-effects model (using the symmetry of a distribution of correlations to detect bias and imputing the 'missing' correlations to reestimate the overall effect size); (c) Begg and Mazumdar's (1994) rank correlation test (using the correlation between the effect size and precision [similar to *N*]); (d) Egger et al.'s (1997) regression intercept (using precision to correct the standardized effect size, which is the effect size divided by the

Table 1. Meta-analysis of in-basket operational validity and publication bias results using four plausible moderators

Criterion/moderator			Bare bones meta-analysis		Corrected for criterion unreliability only			Corrected for criterion unreliability and range restriction			Publication bias analyses				
					Operational validity and SD	80% credibility interval		Operational validity and SD	80% credibility interval		Trim and fill				I^2
	N	*k*	$\bar{r}$	SD_r	ρ	SD_ρ	Lower	Upper	ρ	SD_ρ	Lower	Upper	Studies imputed	$\Delta\bar{r}_o$	$\bar{r}_o$	%
Job performance	3958	31	.16	.10	.21	.12	.06	.36	.42	.13	.25	.58	7	.03	.13	57
No outliers	3325	30	.18	.10	.23	.12	.07	.39	.45	.12	.29	.61	4	.02	.16	56
Content																
Job-specific	2050	19	.17	.12	.22	.15	.03	.41	.42	.19	.18	.67	5	.04	.13	63
Generic	1908	12	.16	.06	.20	.07	.11	.30	.41	0	.41	.41	1	.01	.15	47
Scoring																
Objective	1089	12	.15	.09	.20	.11	.06	.34	.39	.13	.23	.55	2	.04	.11	49
Subjective	2238	15	.15	.11	.20	.14	.02	.37	.39	.19	.15	.62	5	.05	.10	68
Data source																
Published	2385	16	.16	.08	.21	.10	.08	.34	.42	.08	.32	.52	6	.05	.11	56
Unpublished	1573	15	.17	.11	.22	.14	.04	.39	.41	.18	.19	.64	0	0	.17	61
Study design																
Predictive	771	8	.02	.10	.03	.13	−.13	.19	.06	.25	−.26	.39	0	0	.02	57
Concurrent	3187	23	.20	.06	.26	.06	.18	.33	.50	0	.50	.50	9	.05	.15	36

Note: N = total sample size across studies; k = number of coefficients; $\bar{r}$ = mean correlation of the observed distribution; SD_r = standard deviation of the observed distribution; ρ = estimated operational mean validity; SD_ρ = estimated operational standard deviation; Studies imputed is the number of studies that would need to be imputed to achieve a symmetrical distribution; $\Delta\bar{r}_o$ is the difference in validity after studies are imputed and $\bar{r}_o$ is the resulting validity estimate; I^2 is the amount of variance remaining after artifactual sources have been removed. Statistical significance of rank correlation (Begg & Mazumdar, 1994) and intercept tests (Egger et al., 1997) are cited in the text.

standard error); and (e) Higgins, Thompson, Deeks, and Altman (2003) I^2 (using an estimate of the proportion of variance not due to random sampling error). Higgins et al.'s criteria were used to evaluate the I^2: 25% is low, 45% is moderate, and 75% is large. For the trim and fill estimates, Kepes et al. (2012) classified publication bias as essentially absent or negligible when the difference between the meta-analytic mean and the trim and fill adjusted mean estimate is less than 20%. If the difference between two mean estimates is between 20% and 40%, Kepes et al. classified the presence of publication bias as moderate. Finally, publication bias was denoted as severe when the difference is 40% or greater. Note that the publication bias analyses were conducted on observed, not corrected, correlations. We also note that the rank correlation and the intercept tests are relatively low-power tests because their sample size is the number of studies. Therefore, when the results were not statistically significant, we reserved judgment on the existence of publication bias, consistent with recommended practice (Borenstein, Hedges, Higgins, & Rothstein, 2009).

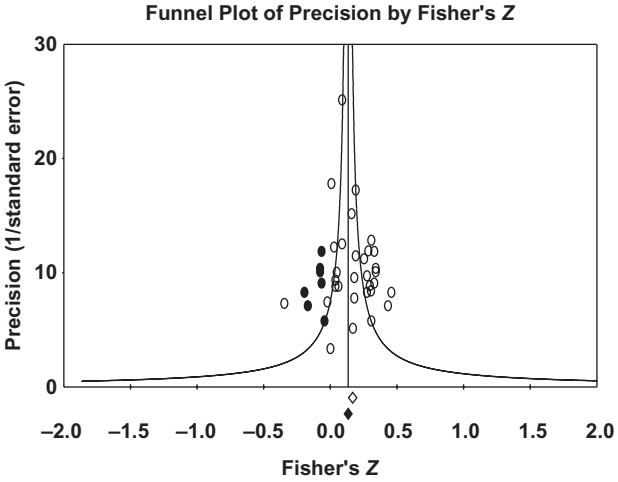

Figure 1. Distribution of observed validities for the job performance criterion (seven studies imputed; change in $r = .04$). *Y*-axis is an indicator of precision (1/standard error), which is highly correlated with sample size. *X*-axis is estimate of validity (expressed as Fisher's Z) Open circles represent observed studies; darkened circles represent studies imputed using trim and fill. Open diamond represents original validity estimate; darkened diamond represents validity estimate after studies are imputed.

3. Results

3.1. Reliability

The estimated mean reliability of the in-basket based on interrater reliability was .76 ($k = 24$; $N = 2{,}325$). The literature also reports numerous internal consistency (i.e., alpha) reliability estimates for the in-basket. We argue that these are inappropriate estimates given that in-baskets can and likely do measure multiple constructs. Thus, we offer .76 as the most accurate estimate of the in-basket's reliability based on cumulative research.

3.2. Criterion-related validity

Our operational criterion-related validity estimates[1] and publication bias results for the in-basket are shown in Table 1. We also provide the validity as moderated by characteristics of the in-basket (scoring and content) and study characteristics (predictive vs. concurrent and published vs. unpublished).

3.2.1. Hypothesis 1: Operational validity estimate

The mean operational validity estimate of in-baskets ($k = 31$; $N = 3{,}958$) was .42 (observed mean validity = .16), which supports our main hypothesis regarding the extent to which the validity of in-baskets for predicting job performance is comparable with the validity of other simulations. Our first sensitivity analysis for this distribution involved computing the operational validity without an outlier study (Goldstein, Yusko, Braverman, Smith, & Chung, 1998). The study was deemed a potential outlier because of its comparatively large sample size ($N = 633$). As shown in Table 1, the difference between this analysis and that using the full set of studies was negligible (operational mean estimates of .42 for the full set of studies vs. .45 for the reduced set of studies). Consequently, subsequent analyses were performed using all studies available.

Our second sensitivity analysis involved computing indices of publication bias. Consistent with literature (Kepes et al., 2012), these estimates are conducted on observed correlations. Our results showed moderate evidence of publication bias, as seen in Table 1. Regarding the trim and fill analysis for the overall sample, imputing seven studies would result in an observed validity estimate that is .03 lower than the observed correlation (.16) (see Figure 1 for a graphic illustration). Because the resulting estimate (.13) is 19% lower, this is considered negligible evidence of publication bias (Kepes et al., 2012). We note, however, that a trim and fill analysis is not robust to violation of the assumption that all variance is due to sampling error (Terrin, Schmid, Lau, & Olkin, 2003). Thus, publication bias is best evaluated in moderator subgroups where heterogeneous variance due to moderators is restricted. Note also that in sub-

[1]Consistent with literature, we define operational criterion-related validity as the estimated mean validity obtained from criterion reliability corrections and range restriction corrections, but not from predictor reliability corrections. In Table 1, we also present an operational validity estimate without the range restriction correction. To simplify presentation, we do not discuss the estimate of the operational validity without the range restriction correction.

sequent analyses, we analyzed the coefficients resulting from the addition of the 'missing' studies in order to assess the effect of each moderator. This is not to suggest that these are the better estimates of operational validity, but merely a way to establish a common metric for making comparisons and evaluating hypotheses. Neither the rank correlation test nor the test of the intercept was statistically significant. We note that both tests have limited power (Borenstein et al., 2009), and that their results may not be statistically significant if the number of samples is small, even if publication bias is present. Note that Higgins I^2 showed that 57% of variance is due to nonrandom sampling error heterogeneity, suggesting the presence of moderators. According to Higgins et al. (2003), this is in the moderate range. Because of the width of the credibility interval and the I^2 suggest moderate variance remaining, we evaluated four moderators of in-basket validity.

3.2.2. Hypothesis 2: In-basket content (generic vs. job specific)

In-baskets with job-specific and generic content yielded nearly equal operational validity estimates (.42 and .41, respectively). Trim and fill publication bias results suggested that the mean observed validity of job-specific in-baskets was reduced by .04 (from the mean observed correlation of .17) by adding five studies to the distribution, yielding a mean observed validity of .13. As such, the adjusted validity is 24% lower than the original validity estimate, providing moderate evidence of publication bias. The validity of in-baskets developed for multiple jobs (generic) suggests that, after adding one study, the mean observed validity decreased by .01 (from the mean observed correlation of .16), yielding a mean observed validity of .15. Because the resulting trim and fill estimate was less than 20% lower than the observed estimate, publication bias is considered to be absent or negligible. Neither the rank correlation test nor the intercept test yielded significant results for either job-specific or generic validity estimate distributions. The Higgins I^2 test shows a moderate amount variance remaining in both distributions that is not attributable to random sampling error. Given the similarity of the validity estimates both before and after addressing publication bias, Hypothesis 2 is not supported. Specifically, there is no meaningful difference in the validity of generic versus job-specific in-baskets.

3.2.3. Hypothesis 3: In-basket scoring (objective vs. subjective)

Both objectively and subjectively scored in-baskets yielded equal operational validity estimates (.39), as shown in Table 1. The trim and fill publication bias analyses suggested that the mean observed validity of objectively scored in-baskets decreased by .04 (from the mean observed correlation of .15) when two studies were imputed, yielding a mean observed validity of .11. This is indicative of moderate publication bias because the adjusted estimate is 27% lower than the original observed estimate. Similarly, the mean observed validity of subjectively scored in-baskets decreased by .05 (from the mean observed correlation of .15) when five studies were added, yielding a mean observed validity of .10 (a 33% difference), indicative of moderate publication bias. Neither the rank correlation test nor the intercept test was statistically significant for the objectively or subjectively scored in-basket validity distributions. The Higgins I^2 test for both objectively and subjectively scored in-baskets suggests that there is a moderate amount of unexplained variance beyond random sampling error. Because of the nearly equal levels of validity in the observed and publication bias adjusted results, Hypothesis 3 was not supported. Specifically, there is no meaningful difference in validity scored with objective approaches versus subjective approaches.

3.2.4. Hypothesis 4: Source of study (published vs. unpublished)

Published and unpublished studies yielded nearly equal operational validity estimates (.42 and .41, respectively), as shown in Table 1. However, the trim and fill publication bias analysis suggests that the mean observed validity of in-baskets as documented in published studies was reduced by .05 (from the mean observed correlation of .16) after adding six studies, making the mean adjusted observed validity .11. For unpublished studies, no studies needed to be imputed to achieve symmetry in the distribution, so the mean observed validity of .17 remained the same. These results suggest that the distribution of published studies is moderately affected by publication bias because the adjusted estimate is 31% lower than the observed estimate. However, contrary to Hypothesis 4, the unpublished studies exhibited higher levels of validity than the published studies. The rank correlation test did not yield statistically significant results for either distribution; the intercept test yielded significant results for the published distribution. As with the other distributions, the amount of variance remaining after accounting for random sampling error was moderate. Thus, Hypothesis 4 was not supported even though greater evidence of publication bias was evident in the published distribution than in the unpublished distribution.

3.2.5. Hypothesis 5: Study design (predictive vs. concurrent)

We also investigated the extent to which study design affects in-basket validity results, as shown in Table 1. Predictive studies yielded lower operational mean validity estimates than concurrent studies (.06 and .50, respectively) providing support for Hypothesis 5.

Publication bias analyses suggest that no validity estimates were imputed for the distribution of predictive validity estimates. For concurrent studies, publication bias analysis showed that after adding nine studies, the mean observed validity decreased by .05 (from the mean observed correlation of .20), yielding an observed validity of .15. The effect of publication bias was moderate because the trim and fill adjusted estimate was 25% lower than the original estimate. The rank correlation test yielded significant results for neither distribution; the intercept test yielded significant results for the concurrent distribution. The Higgins I^2 test showed a moderate amount of variance remaining after accounting for random sampling error. Both the meta-analytic estimates and the publication bias results suggest that the mean validity estimates from predictive studies are lower than those for concurrent studies, thus providing support for Hypothesis 5.

However, we offer two caveats to this conclusion. First, there were relatively few validity coefficients ($k = 8$) from predictive studies, which makes the results more likely to be distorted by second-order sampling error (Hunter & Schmidt, 2004). Second, publication bias analyses may be suspect for distributions with fewer than 10 effect sizes (Kepes et al., 2012). Still, the magnitude of the difference between predictive and concurrent studies is sufficiently large such that we conclude that validities obtained from current samples are larger than those obtained from predictive samples.

3.2.6. Hypothesis 6: Cognitive loading of in-baskets

Regarding the correlation between the in-basket and *g*, the observed estimate is .26, which suggests that the in-basket is somewhat *g*-loaded. Trim and fill publication bias analysis showed that one study was needed to achieve a symmetrical distribution, but the validity remained unchanged. Neither the rank order correlation test nor the intercept test was statistically significant. The Higgins I^2 test showed a moderate amount of variance remaining after accounting for random sampling error (Table 2).

4. Discussion

For approximately 50 years, the in-basket has been used to assess training and developmental needs and to predict performance for a wide variety of occupations. The constructs measured by the in-basket are likely very different than those measured by other simulation measures often included in assessment centers. In general, the constructs measured by the in-basket are more cognitive (e.g., procedural and declarative job knowledge and managerial/administrative capabilities) than those measured by other simulation methods, such as role-plays and leaderless group discussions, that measure interpersonal skills. Its widespread use, both online and in paper and pencil form, is likely due to its face validity and ease of use. In-baskets are easier to administer than many other assessment center exercises, such as role-plays and leaderless group discussions, because they do not require human interaction. In fact, because many vendors have created in-baskets that are administered and scored online, their use will likely increase in the future. Until this article, there has been no comprehensive empirical review of the reliability and validity of in-baskets.

Our results show that in-baskets can be scored reliably with our best estimate of .76, based on interrater reliability. This estimate is comparable to or slightly higher than those found for other work samples (e.g., Roth et al., 2005). The estimated mean operational criterion-related validity was .42 (observed mean = .16), which is comparable to (and somewhat higher than) other meta-analytic reviews of work samples and simulations (e.g., Arthur et al., 2003; Gaugler et al., 1987; Hermelin et al., 2007; McDaniel et al., 2007; Roth et al., 2005). There was no moderator effect for in-basket content, which suggests that generic in-baskets, such as those administered online, are likely just as valid as those developed for a particular job or organization. A plausible explanation is that in-baskets generally measure administrative or managerial skills (e.g., planning, organizing, and prioritizing) and it may not matter

Table 2. Mean correlation between *g* and the in-basket

			Bare bones meta-analysis		Corrected for unreliability in g only				Publication bias analyses			
					Operational validity and SD		80% credibility interval		Trim and fill			I^2
	N	k	$\bar{r}$	SD_r	ρ	SD_ρ	Lower	Upper	Studies imputed	$\Delta\bar{r}$	$\bar{r}_c$	%
g	3259	19	.26	.08	.30	.08	.20	.40	1	.00	.26	51

Note: N = total sample size across studies; k = number of coefficients; $\bar{r}$ = mean correlation of the observed distribution; SD_r = standard deviation of the observed distribution; ρ = estimated operational mean validity; SD_ρ = estimated operational standard deviation; Studies imputed is the number of studies that would need to be imputed to achieve a symmetrical distribution; $\Delta\bar{r}_c$ is the difference in validity after studies are imputed and $\bar{r}_c$ is the resulting validity estimate; I^2 is the amount of variance remaining after artifactual sources have been removed. Statistical significance of rank correlation (Begg & Mazumdar, 1994) and intercept tests (Egger et al., (1997) are cited in the text.

whether the in-basket is job-specific or developed for any number of jobs. The operational validity estimates of subjectively and objectively scored in-baskets were nearly equal, even after adjustments for publication bias, possibly because there is a continuum of objectivity with checklists being somewhat more objective than rating scales. Data source did act as a moderator after adjusting for publication bias; just not as we expected. Published and unpublished studies yielded approximately equal validity estimates after adjusting for publication bias (publication bias was moderate for the published studies and nonexistent for the unpublished studies). This is consistent with Ferguson and Brannick's (2012) findings demonstrating that publication bias is relatively common (41% of the studies displayed evidence of publication bias). Consistent with our hypothesis, predictive studies yielded lower validity estimates than concurrent studies, and this result was robust to the influence of publication bias. This finding is likely due to predictor contamination (a rater's knowledge about a candidate's job performance may bias the rater's assessment of the candidate's in-basket results). Another possible explanation is that incumbents' experience and knowledge of the organization may influence in-basket performance as well as their job performance. The opportunity for incumbents to use their experience and knowledge from their present positions when answering questions may magnify the correlation with job performance (Cascio, 1998; Huffcutt et al., 2004).

We also found that cognitive ability had a moderate observed mean correlation with the in-basket (.26), which suggests that in-baskets are somewhat *g*-loaded. This is not surprising given the reading and reasoning requirements of in-baskets. Given the correlation with cognitive ability and the cognitive demands of in-baskets, there may be mean subgroup differences on in-basket scores. To provide a more complete answer to this question Roth, Bobko, McFarland, and Buster (2008) found that Black–White ethnic group differences on work samples were .73 favoring White job applicants. They concluded that work sample measures have more adverse impact than previously believed. Additionally, Goldstein et al. (1998) found that the in-basket produced a .35 standard deviation difference favoring Whites over Blacks. Regardless of which study provides the more accurate estimate, we conclude that use of in-baskets may result in subgroup score differences, on average.

4.1. Limitations of this study

The first limitation of this study is the relatively small number of studies in each distribution. That said, this study provides the largest known database of in-basket validities. We recommend that as more data become available, these meta-analyses be rerun.

Similarly, we did not have sufficient information on range restriction or criterion reliability to correct studies individually. Thus, we relied on Pearlman et al.'s (1980) distributions. For criterion reliability, Pearlman et al.'s estimate had an average of .60. We note that Viswesvaran et al. (1996) identified .52 as a more accurate estimate and Salgado et al. (2003) obtained a sample size weighted mean reliability of .57 across performance dimensions. We used the Pearlman et al.'s distribution because: (a) they provided a distribution of criterion reliabilities that enabled both mean and variance corrections to be made and (b) the average of .60 would yield a more conservative estimate of validity than .52 or .57. Thus, it is likely that our results underestimate the mean validity of the in-basket. Because of the lack of predictor reliability data in studies that provided validity data, we did not correct the variance in validities for predictor reliability. This likely resulted in overestimating the variance of the operational validity distribution and artificially reducing the lower end of the 80% credibility interval. We recommend that when conducting validity studies, the reliability of the measure be investigated and reported.

4.2. Recommendations for practitioners

This article provides important information to practitioners regarding the reliability and criterion-related validity of in-baskets. Regarding reliability, we note that the correlation between two raters is considered the reliability of a single rater. Thus, we used the Spearman–Brown formula to estimate the reliability of in-baskets, assuming we are computing a composite of two raters. The resulting reliability estimate is .86. Thus, we recommend that more than one rater provide in-basket scores in practice.

To obtain interrater agreement at the level suggested by this study, raters must be trained on how to use the rating scales. If the in-basket responses are provided orally, raters need to be trained on how to observe behavior and take notes. If the in-basket responses are provided in written form, raters need to be trained on how to score the responses independent of the potential confound of writing ability. Whether responses are provided orally or in writing, raters using behaviorally based rating scales or checklists need to be able to compare the in-basket responses to the examples provided and decide on a score based on the behavior that most closely matches the behavior they observe or read.

Because our results show that both job-specific and generic in-baskets have relatively high levels of validity, assessment designers can tailor in-baskets to various jobs and to specific competencies. For example, an in-basket for a HR professional would be different from that developed for a law enforcement officer/

supervisor. On the other hand, the results of this study inform the use of in-baskets, regardless of technical job content. Whether in-baskets are designed to be job-specific or generic, we recommend that their use be justified by an assessment of the extent to which the competencies measured in the in-basket match those required by the job. Because of the cognitive demands of the in-basket, we recommend that the jobs for which in-baskets are used include those that take place, at least in part, in an office environment, and that they require some amount of reasoning and decision making. For example, performance in jobs requiring mostly physical skills outside an office environment (e.g., drywall installers or brick masons) may not be well predicted using an in-basket.

4.3. Recommendations for future research

Many researchers have developed and studied in-baskets intended to measure numerous constructs. These studies suggest that the in-basket technique measures various administrative skills linked to managerial success. We note that the possible constructs measured by the in-basket are a likely source of unexplained variance in validity because some constructs may have greater validity than others. For our study, there was an insufficient number of studies of any one skill or construct to conduct a credible meta-analysis of a single skill or construct. Thus, we recommend that future studies describe the constructs in sufficient detail to permit the development of a taxonomy or coding scheme that would apply across in-basket exercises. Also, most studies only assert attributes measured by the in-basket. The literature would be substantially improved if authors provided empirical evidence for constructs measured by a given in-basket.

Given the multidimensional nature of in-baskets (similar to SJTs and interviews), we recommend computing and reporting interrater reliability estimates rather than coefficient alpha or split half reliability. Interrater reliability is likely more appropriate given that in-baskets are measurement methods that assess multiple constructs. The literature also needs far greater reporting of range restriction data.

One moderator we were unable to investigate because of lack of data is the use of online versus paper and pencil in-baskets. As test vendors provide online in-baskets, making them available to anyone with access to the Internet, their use likely will proliferate. Thus, we recommend that research be conducted to study the similarity of constructs assessed using both administration modes. There are a host of other moderators we could have studied had the data been available and we recommend these as avenues of future research. These include validity differences between stand-alone in-baskets versus those administered as part of an assessment center and consensual scoring versus nonconsensual scoring. In consensual scoring, raters discuss their individual ratings and the raters come to a consensus on the score. In nonconsensual scoring, ratings by different raters are combined with discussion, usually by averaging ratings.

We recommend better reporting of study design and features of in-baskets. There were several studies where it was difficult to determine how the in-baskets were developed and scored. Also, we urge comparison of objective versus subjective scoring of in-baskets controlling for content. For example, using the same stimulus materials, one could create two scoring keys, one objective and one subjective, and compare their validities.

Although we were unable to investigate the effect of job complexity on the validity of the in-basket, Huffcutt et al. (2004) investigated job complexity as a moderator of the validity of interviews. The mean validity of situational interviews was lower for high complexity positions than for lower-complexity positions. Their explanation was that it is difficult to write questions that capture the intricacies and subtleties often found in high-level positions. Similarly, it is difficult to develop rating scale anchors that enable the differentiation among complex, sophisticated responses. The same may be true for in-baskets. That is, developing in-basket items and scales that capture the nuances required for higher-level positions can be more complex for high-level positions, thus possibly reducing the validity of the measure. One also needs to consider the restriction of range in cognitive complexity that one will encounter when using in-baskets for higher-level positions. Unlike interviews, which are used when screening for jobs with a wide range of cognitive complexity, in-baskets are more likely to be used for higher-complexity jobs than for lower-complexity jobs.

Finally, we reiterate our call for more predictive validity studies of the in-basket. In this study, based on only eight validities, we report much lower validities for predictive studies than for concurrent studies. The precision of our estimated validity for predictive validity studies is thus limited. Our analyses should be redone as more predictive validity data accumulate.

5. Conclusion

This study provides the most comprehensive evaluation of the criterion-related validity of in-baskets to date. Our best point estimate of the validity of in-baskets for predicting job performance is .42, which is comparable to other research on work samples and simulations. However, the extant in-basket literature has several deficiencies and we provide recommendations for future research to redress these deficiencies. We also

demonstrate that in-baskets can be scored reliably and are *g*-loaded. To the extent that in-baskets are available online, and because they have a high degree of face validity, their popularity as a selection device will likely increase. Thus, this article makes a substantial contribution to practice as well as to science.

References

*Study was used in validity analyses.

†Study was used in reliability analyses.

‡ Study was used for correlation with *g*.

American Psychological Association. (2010). *Publication manual of the American Psychological Association* (6th ed.) Washington, DC: Author.

Arthur, W., Day, E. A., McNelly, T. L., & Edens, P. S. (2003). A meta-analysis of the criterion related validity of assessment center dimensions. *Personnel Psychology*, 56, 125–154.

Arthur, W., & Villado, A. J. (2008). The importance of distinguishing between constructs and methods when comparing predictors in personnel selection research and practice. *Journal of Applied Psychology*, 93, 435–442.

*†Atkins, P. W. B. & Wood, R. E. (2002). Self- versus others' ratings as predictors of assessment center ratings: Validation evidence for 360-degree feedback programs. *Personnel Psychology*, 55, 871–904.

‡Avolio, B. J., O'Connell, M. S., Martiz, D., & Kennedy, C. (1999). *Assessment of transformation leadership potential*. Atlanta, GA: Symposium presented at the14th annual meeting of the Society for Industrial and Organizational Psychology.

†Bader, G. E. (1987). Understanding individual political behaviors in organizations: Instrument development and validation. *Dissertation Abstracts International*, 47(11-A), 3915. University of San Diego.

Barrett, G. V., Phillips, J. S., & Alexander, R. A. (1981). Concurrent and predictive validity designs: A critical reanalysis. *Journal of Applied Psychology*, 38, 41–56.

Begg, C. B., & Mazumdar, M. (1994). Operating characteristics of a rank correlation test for publication bias. *Biometrics*, 50, 1088–1101.

Bentz, V. J. (1962). *The Sears experience in the investigation, description, and prediction of executive behavior. Executive study conference*. Princeton, NJ: Educational Testing Service.

*†‡Bentz, V. J. (1968). The Sears experience in the investigation, description and prediction of executive behavior. In J. A. Meyer (Ed.), *Predicting managerial success* (pp. 59–152). Ann Arbor, MI: Foundation for Research on Human Behavior.

*Bernthal, P., Schmidt, D., Stehura, A. M. (2010). *Technical summary from Manager Ready (RM)*. Bridgeville, PA: Development Dimensions International, Inc.

Berry, C. M., Sackett, P. R., & Landers, R. N. (2007). Revisiting interview-cognitive ability relationships: Attending to specific range restriction mechanisms in meta-analysis. *Personnel Psychology*, 60, 837–874.

Blum, M. L. (1943). Selection of sewing machines operators. *Journal of Applied Psychology*, 27, 35–40.

*†Bobrow, W. & Leonards, J. S. (1997). Development and validation of an assessment center during organizational change. *Journal of Social Behavior and Personality*, 12, 217–236.

Borenstein, M., Hedges, L., Higgins, J., & Rothstein, H. R. (2005). *Comprehensive meta-analysis. Version 2*. Englewood, NJ: Biostat.

Borenstein, M., Hedges, L. V., Higgins, J. P., & Rothstein, H. R. (2009). *Introduction to meta-analysis*. Chichester, West Sussex, UK: Wiley.

†Bourgeois, R. P. & Slivinski, L. W. (1974). The inter-rater reliability of The Consolidated Fund in-basket. *Studies in Personnel Psychology*, 6, 47–52.

†Boyd, M. K. (1990). A comparison of two methods of scoring in-baskets. Unpublished Master's Thesis. University of South Florida.

*†Brass, D. J., & Oldham, G. R. (1976). Validating in-basket test using an alternative set of leadership scoring dimensions. *Journal of Applied Psychology*, 61, 652–657.

†Bray, D. W., & Grant, D. L. (1966). The assessment center in the measurement of potential for business management. *Psychological Monographs: General and Applied*, 80, 1–27.

†Brostoff, M. & Meyer, H. H. (1984). The effects of coaching on in-basket performance. *Journal of Assessment Center Technology*, 7, 17–21.

Carey, N. B. (1990). *An assessment of surrogates for hands-on tests: Selection standards and training needs* (CRM 90-47). Alexandria, VA: Center for Naval Analyses.

Cascio, W. F. (1998). *Applied psychology in human resources management* (5th ed.) Upper Saddle River, NJ: Prentice Hall.

Chan, A.-W., Hróbjartsson, A., Haahr, M. T., Gøtzsche, P. C., & Altman, D. G. (2004). Empirical evidence for selective reporting of outcomes in randomized trials: Comparison of protocols to published articles. *Journal of the American Medical Association*, 291, 2457–2465. doi: 10.1001/jama.291.20.2457.

Craik, K. H., Ware, W. P., Kamp, J., O'Reilly, C., Staw, B., & Zedeck, S. (2002). Explorations of construct validity in a combined managerial and personality assessment programme. *Journal of Occupational and Organizational Psychology*, 75, 171–193.

*†Cross, W. R. (1969). Relationship between elementary school principals' in-basket performance and their on-the-job behavior. *The Journal of Educational Research*, 63, 26–30.

Curfman, G. D., Morrissey, S., & Drazen, J. M. (2006). Expression of concern reaffirmed. *New England Journal of Medicine*, 354, 1193–1193. doi: 10.1056/NEJMe068054.

†‡Denning, D. L. (1980). An examination of in-basket scores. *Dissertation Abstracts International*, 41(10-B), 3924. University of Georgia.

Dickersin, K. (2005). Publication bias: Recognizing the problem, understanding its origins and scope, and preventing harm. In H. R. Rothstein, A. J. Sutton, & M. Borenstein (Eds.), *Publication bias in meta-analysis: Prevention, assessment and adjustments* (pp. 11–34). Chichester, UK: Wiley.

Donahue, L. M., Truxillo, D. M., Cornwell, J. M., & Gerrity, M. J. (1997). Assessment center construct validity and behavioral checklists: Some additional findings. In Riggio, R.E. & Mayes, B.T. (Eds.). Assessment centers: Research and applications [Special issue]. *Journal of Social Behavior and Personality*, 12, 85–108.

Dunnette, M. D. (1971). The assessment of organizational talent. In P. McReynolds (Ed.), *Advances in psychological assessment* 2, (pp. 79–108). Palo Alto, CA: Science and Behavior Books.

Duval, S. J., & Tweedie, R. L. (2000a). A non-parametric 'trim and fill' method of accounting for publication bias in meta-analysis. *Journal of the American Statistical Association*, 95, 89–98.

Duval, S. J., & Tweedie, R. L. (2000b). Trim and fill: A simple funnel plot-based method of testing and adjusting for publication bias in meta-analysis. *Biometrics*, 56, 276–284.

Egger, M., Smith, G. D., Schneider, M., & Minder, C. (1997). Bias in meta-analysis detected by a simple, graphical test. *British Medical Journal*, 315, 629–634.

Engel, J. D. (1970). *Development of a work sample criterion for general vehicle mechanic*. Fort Knox, KY: HumRRO for Research and Development Department of the Army (Report 70-11).

†Fedorko, M. G. (1992). The effects of training and feedback format on reactions to in-basket feedback and on in-basket performance. *Dissertation Abstracts International*, 52(9-B), 5005. Old Dominion University.

Felker, D. B., Crafts, J. L., Rose, A. M., Harnest, C. W., Edwards, D. S., Bowler, E. C. *et al.* (1988). *Developing job performance tests for the United States Marine Corps infantry occupational field* (AIR-47500-9/88-FR). Washington, DC: American Institutes for Research.

Felker, D. B., Curtin, P. J., & Rose, A. M. (2007). Tests of job performance. In D. L. Whetzel & G. R. Wheaton (Eds.), *Applied measurement: Industrial psychology in human resources management* (pp. 319–348). Mahwah, NJ: Lawrence Erlbaum Associates.

Ferguson, C. J., & Brannick, M. T. (2012). Publication bias in psychological science: Prevalence, methods for identifying and controlling, and implications for the use of meta-analyses. *Psychological Methods*, 17, 120–128.

†Frederiksen, N. (1962). Factors in in-basket performance. *Psychological Monographs: General and Applied*, 76, 1–25.

‡Frederiksen, N. (1966). Validation of a simulation technique. *Organizational Behavior and Human Performance*, 1, 87–109.

*†Frederiksen, N., Jenson, O. & Beaton, A. (1972). *Prediction of organizational behavior*. New York: Pergammon.

†‡Frederiksen, N., Saunders, R., & Wand, B. (1957). The in-basket test. *Psychological Monographs: General and Applied*, 71, 1–28.

Gaugler, B. B., Rosenthal, D. B., Thornton, G. C., & Bentson, C. (1987). Meta-analysis of assessment center validity. *Journal of Applied Psychology*, 72, 493–511.

Ghiselli, E. E., Campbell, J. P., & Zedeck, S. (1981). *Measurement theory for the behavioral sciences*. San Francisco, CA: Freeman.

†Gill, R. W. T. (1979). The in-tray (in-basket) exercise as a measure of management potential. *Journal of Occupational Psychology*, 52, 185–197.

Ginsburg, L. R., & Silverman, A. (1972). The leaders of tomorrow: Their identification and development. *Personnel Journal*, 51, 662–666.

*†‡Goldstein, H. W., Yusko, K. P., Braverman, E. P., Smith, D. B., & Chung, B. (1998). The role of cognitive ability in the subgroup differences and incremental validity of assessment center exercises. *Personnel Psychology*, 51, 357–374.

Hakstian, A. R., & Harlos, K. P. (1993). Assessment of in-basket performance by quickly scored methods: Development and psychometric evaluation. *International Journal of Selection and Assessment*, 1, 135–142.

*†Hakstian, A. R., & Scratchley, L. S. (1997). In-basket assessment by fully objective methods: Development and evaluation of a self-report system. *Educational and Psychological Measurement*, 57, 607–630.

*†‡Hakstian, A. R., Woolsey, L. K., & Schroeder, M. L. (1986). Development and application of a quickly scored in-basket exercise in an organizational setting. *Educational and Psychological Measurement*, 46, 385–396.

*Hardoin, M. M., Waugh, G., Keenan, P. A., & O'Shea, G. (2010). *Development and implementation of the Supervisory Veterans Service Representative (SVSR) skills certification test.* (FR-10-24). Alexandria, VA: Human Resources Research Organization.

Hedge, J. W., Lipscomb, M. S., & Teachout, M. S. (1988). Work sample testing in the Air Force job performance measurement project. In M. S. Lipscomb & J. W. Hedge (Eds.), *Job performance measurement: Topics in the performance measurement of Air Force enlisted personnel* (AFHRL-TP-87-58). Brooks Air Force Base, TX: Air Force Human Resources Laboratory.

*†‡Hemphill, J. K., Griffiths, D. E., & Frederiksen, N. (1962). *Administrative performance and personality: A study of the principal in a simulated elementary school*. New York: Teachers College, Bureau of Publications.

Hermelin, E., Lievens, F., & Robertson, I. T. (2007). The validity of assessment centres for the prediction of supervisory performance ratings: A meta-analysis. *International Journal of Selection and Assessment*, 15, 405–411.

Higgins, J., Thompson, S. G., Deeks, J. J., & Altman, D. G. (2003). Measuring inconsistency in meta-analyses. *British Medical Journal*, 327, 557–560.

Hough, L. M. (1998). Personality at work: Issues and evidence. In M. D. Hakel (Ed.), *Beyond multiple choice: Evaluating alternatives to traditional testing for selection* (pp. 131–166). Mahwah, NJ: Lawrence Erlbaum.

Huck, J. R., & Bray, D. W. (1976). Management assessment center evaluations and subsequent job performance of white and black females. *Personnel Psychology*, 29, 13–30.

Huffcutt, A. I., Conway, J. M., Roth, P. L., & Klehe, U. (2004). The impact of job complexity and study design on situational and behavior description interview validity. *International Journal of Selection and Assessment*, 12, 262–273.

Huffcutt, A. I., Roth, P. L., & McDaniel, M. A. (1996). A meta-analytic investigation of cognitive ability in employment interview evaluations: Moderating characteristics and implications for incremental validity. *Journal of Applied Psychology*, 81, 459–473.

Hunter, J. E., & Schmidt, F. L. (2004). *Methods of meta-analysis: Correcting error and bias in research findings* (2nd ed.) Newbury Park, CA: Sage.

†Jaffee, C. L., & Michaels, C. E. (1978). Is in-basket performance subject to coaching effects. *Journal of Assessment Center Technology*, 1, 13–17.

Kepes, S., McDaniel, M. A., Banks, G., & Whetzel, D. L. (2012). Publication bias in the organizational sciences. *Organizational Research Methods*, 15, 624–662.

Kepes, S., McDaniel, M. A., Brannick, M. T., & Banks, G. C. (2013). Meta-analytic reviews in the organizational sciences: Two meta-analytic schools on the way to MARS (the Meta-Analytic Reporting Standards). *Journal of Business and Psychology*, 28, 123–143.

*†Kesselman, G. A., Lopez, F. M., & Lopez, F. E. (1982). The development and validation of a self-report scored in-basket test in an assessment center setting. *Public Personnel Management Journal*, 11, 228–238.

†Kim, J. S. (1980). Relationships of personality to perceptual and behavioral responses in stimulating and nonstimulating tasks. *Academy of Management Journal*, 23, 307–319.

Knapp, D. J., & Campbell, J. P. (1993). *Building a joint-service classification research roadmap: Criterion-related issues* (AL/HR-TP-1993-0028). Brooks Air Force Base, TX: Armstrong Laboratory, Manpower and Personnel Research Division.

‡Lance, C. E., Newbolt, W. H., Gatewood, R. D., Foster, M. R., French, N. R., Smith, D. E. (2000).). Assessment center exercise factors represent cross-situational specificity, not method bias. *Human Performance*, 13, 323–353.

*Lilienthal, R. A., & Mack, M. J. (1992). *Criterion development revisited: Heartache to heartbreak*. Paper presented at the 16th Annual Conference of the International Personnel Management Association Assessment Council.

Lipsey, M. W., & Wilson, D. B. (1993). The efficacy of psychological, educational, and behavioral treatment. Confirmation from meta-analysis. *American Psychologist*, 48, 1181–1209.

*†‡Lopez, F. M. (1966). *Evaluating executive decision making: The in-basket technique* (AMA Research Study No. 75). New York: American Management Association.

†Mayes, B. T., Belloli, C. A., Riggio, R. E., & Aguirre, M. (1997). Assessment centers for course evaluations: A demonstration. In R. E. Riggio & B. T. Mayes (Eds.), Assessment centers: Research and application. *Journal of Social Behavior and Personality [special issue]*, 12, 303–320.

McDaniel, M. A., Hartman, N. S., Whetzel, D. L., & Grubb, W. L. (2007). Situational judgment tests, response instructions, and validity: A meta-analysis. *Personnel Psychology*, 60, 63–91.

McDaniel, M. A., Morgeson, F. P., Finnegan, E. B., Campion, M. A., & Braverman, E. P. (2001). Predicting job performance using situational judgment tests: A clarification of the literature. *Journal of Applied Psychology*, 86, 730–740.

McDaniel, M. A., Rothstein, H. R., & Whetzel, D. L. (2006). Publication bias: A case study of four test vendors. *Personnel Psychology*, 59, 927–953.

McDaniel, M. A., Whetzel, D. L., Schmidt, F. L., & Maurer, S. D. (1994). The validity of employment interviews: A comprehensive review and meta-analysis. *Journal of Applied Psychology*, 79, 599–616. doi: 10.1037/0021-9010.79.4.599.

*Melchers, K. G., & Kleinmann, M. (2009). *Occupational fidelity as a moderator of AC validity*. Paper presented at the 24th annual conference of the Society for Industrial and Organizational Psychology, New Orleans, LA.

*‡Meyer, H. H. (1970). The validity of the in-basket test as a measure of managerial performance. *Personnel Psychology*, 23, 297–307.

Motowidlo, S. J., Dunnette, M. D., & Carter, G. W. (1990). An alternative selection procedure: The low-fidelity simulation. *Journal of Applied Psychology*, 75, 640–647. doi:10.1037/0021-9010.75.6.640.

*†Nowack, K. M. (1997). Congruence between self-other ratings and assessment center performance. In R. E. Riggio & B. T. Mayes (Eds.), Assessment centers: Research and application. *Journal of Social Behavior and Personality [special issue]*, 12, 145–166.

*‡O'Connell, M. S., Kung, M. C., & Lawrence, A. (2010). *Normative summary of group leader assessment system at (company withheld)*. *Technical Report*. Pittsburgh, PA: Select International, Inc.

*‡O'Connell, M. S., Wolf, D., & Kato, M. (2001). *Predictive validation report for leadership and professional level employees at (company withheld)*. *Technical Report*. Pittsburgh, PA: Select International, Inc.

*‡O'Connell, M. S., Wolf, D., & Klinvex, K. (1999). *Concurrent validation report for leadership and professional level employees at (company withheld)*. *Technical Report*. Pittsburgh, PA: Select International, Inc.

Ones, D. S., Viswesvaran, C., & Schmidt, F. L. (1993). Comprehensive meta-analysis of integrity test validities: Findings and implications for personnel selection and theories of job performance. *Journal of Applied Psychology*, 78, 679–703.

Pearlman, K., Schmidt, F. L., & Hunter, J. E. (1980). Validity generalization results for tests used to predict job proficiency and training success in clerical occupations. *Journal of Applied Psychology*, 65, 373–406.

Pigott, T. D. (2009). Handling missing data. In H. Cooper, L. V. Hedges, & J. C. Valentine (Eds.), *The handbook of research synthesis and meta-analysis* (2nd ed.) (pp. 399–416). New York, NY: Russell Sage Foundation.

Ployhart, R. E., Schneider, B., & Schmitt, N. (2006). *Staffing organizations: Contemporary practice and research*. Mahwah, NJ: Erlbaum.

†‡Reed, N. S. (1981). An examination of the trainability of assessment center dimension performance on the in-basket: An exploratory study. *Dissertation Abstracts International*, 42(10-B), 4234. University of Georgia, 1981, Order # DA8206304.

Roth, P. L., Bobko, P., & McFarland, L. (2005). A meta-analysis of work sample test validity: Updating and integrating some classic literature. *Personnel Psychology*, 58, 1009–1037.

Roth, P. L., Bobko, P., McFarland, L., & Buster, M. (2008). Work sample tests in personnel selection: A meta-analysis of black-white differences in overall an exercise scores. *Personnel Psychology*, 61, 637–662.

Rothstein, H. R., Sutton, A. J., & Borenstein, M. (2005). Publication bias in meta-analysis. In H. R. Rothstein, A. J. Sutton, & M. Borenstein (Eds.), *Publication bias in meta-analysis: Prevention, assessment and adjustments* (pp. 1–7). West Sussex, UK: Wiley.

Salgado, J. F., Moscoso, S., & Lado, M. (2003). Test–retest reliability of ratings of job performance dimensions in managers. *International Journal of Selection and Assessment*, 11, 98–101.

Schippmann, J. S., Prien, E. P., & Katz, J. A. (1990). Reliability and validity of in-basket performance measures. *Personnel Psychology*, 43, 837–859.

Schmidt, F. L., & Hunter, J. E. (1977). Development of a general solution to the problem of validity generalization. *Journal of Applied Psychology*, 62, 529–540.

Schmidt, F. L., & Le, H. (2005). Hunter and Schmidt meta-analysis programs (V 1.1).

Schmidt, F. L., Shaffer, J. A., & Oh, I. S. (2008). Increased accuracy for range restriction corrections: Implications for the role of personality and general mental ability in job and training performance. *Personnel Psychology*, 61, 827–868.

Schmitt, N., Gooding, R. Z., Noe, R. A., & Kirsch, M. (1984). Meta analyses of validity studies published between 1964 and 1982 and the investigation of study characteristics. *Personnel Psychology*, 37, 407–422.

†Shapira, Z., & Dunbar, R. L. M. (1980). Testing Mintzberg's managerial roles classification using an in-basket simulation. *Journal of Applied Psychology*, 65, 87–95.

†Silverman, A. I. (1972). The effects of preferred-perceived environmental compatibility and self-esteem on managerial in-basket behavior. Unpublished doctoral dissertation.

Song, F. Parekh S., Hooper L., Loke Y. K., Ryder J., Sutton A. J. *et al*. (2010). Dissemination and publication of research findings: An updated review of related biases. *Health Technology Assessment*, 14, 1–220. doi: 10.3310/hta14080.

Terrin, N., Schmid, C. H., Lau, J., & Olkin, I. (2003). Adjusting for publication bias in the presence of heterogeneity. *Statistics in Medicine*, 22, 2113–2125.

†Tett, R. P., & Jackson, D. N. (1990). Organization and personality correlates of participative behaviors using an in-basket exercise. *Journal of Occupational Psychology*, 63, 175–188.

Thornton, G. C., & Byham, W. C. (1982). *Assessment centers and managerial practice*. New York, NY: Academic Press.

Tsacoumis, S. (2007). Assessment centers. In D. L. Whetzel & G. R. Wheaton (Eds.), *Applied measurement: Industrial psychology in human resources management* (pp. 259–293). Mahwah, NJ: Lawrence Erlbaum Associates.

*Turnage, J. T., & Muchinsky, P. M. (1984). A comparison of the predictive validity of assessment center evaluations versus traditional measures in forecasting supervisory performance: Interpretive implications of criterion distortion for the assessment paradigm. *Journal of Applied Psychology*, 69, 595–602.

‡Tziner, A., & Dolan, S. (1982). Validity of an assessment center for identifying future female officers in the military. *Journal of Applied Psychology*, 67, 728–736.

*Van Hooft, E. A. J., van der Flier, H., Minne, M. R. (2006). Construct validity of multi-source performance ratings: An examination of the relationship of self-, supervisor-, and peer-ratings with cognitive and personality measures. *International Journal of Selection and Assessment*, 14, 67–81.

Viswesvaran, C., Ones, D. S., & Schmidt, F. L. (1996). Comparative analysis of the reliability of job performance ratings. *Journal of Applied Psychology*, 81, 557–574.

Ward, L. B. (1960). The performance of business school students on an in-basket test. In J. Hemphill (Ed.), *Proceedings of the conference on the executive study: The in-basket technique* (pp. 91–105). Princeton, NJ: Educational Testing Service.

West, L., & Bolanovich, D. J. (1963). Evaluation of typewriting proficiency: Preliminary test development. *Journal of Applied Psychology*, 47, 403–407.

Whetzel, D. L., & McDaniel, M. A. (1988). Reliability of validity generalization databases. *Psychological Reports*, 63, 131–134.

Wigdor, A. K., & Green, B. F. (Eds.). (1986). *Assessing the performance of enlisted personnel: Evaluation of a joint-service research project*. Washington, DC: National Academy Press.

Wilson, J. E., & Tatge, W. A. (1973). Assessment centers – further assessment needed. *Personnel Journal*, 172–178.

Wollowick, H. B., & McNamara, W. J. (1969). Relationship of the components of an assessment center to management success. *Journal of Applied Psychology*, 53, 348–352.

Appendix A

Table A1. Performance distribution

Study ID	*r*	*N*	Time lag	Predictor used	Predictor content	Source
Atkins and Wood (2002)	.18	63	0	*Assessor ratings*	Job-specific	Published
Avolio et al. (1999)	.063	80	0	*Assessor ratings*	Generic	Unpublished
Bentz (1968)	–.33	56	5 years	*Assessor ratings*	Job-specific	Unpublished
Bernthal, Schmidt, and Stehura (2010)	.3	167	0	*Assessor ratings*	Generic	Unpublished
Bobrow and Leonards (1997) concurrent sample	.27	71	0	*Assessor ratings*	Job-specific	Published
Bobrow and Leonards (1997) predictive sample	.17	29	1 year	*Assessor ratings*	Job-specific	Published
Brass and Oldham (1976)	.43	71	0	Behavior checklist	Job-specific	Published
Cross (1969)	.003	14	2–5 years	Behavior checklist	Generic	Published
Frederiksen, Jensen, and Beaton (1972) BB in-basket	.05	103	0	Behavior checklist	Generic	Published
Goldstein et al. (1998)	.09	633	0	*Assessor ratings*	Generic	Published
Hakstian and Scratchley (1997)	.32	143	0	*Assessor ratings*	Generic	Published
Hakstian et al. (1986) women	.33	110	0	Mixture	Job-specific	Published
Hakstian et al. (1986) men	.25	128	0	Mixture	Job-specific	Published
Hardoin, Waugh, Keenan, and O'Shea (2010)	.33	104	0	*Assessor ratings*	Job-specific	Unpublished
Hemphill, Griffiths, and Frederiksen (1962)	.16	232	0	*Assessor ratings*	Job-specific	Published
Kesselman et al. (1982)	.32	85	0	Behavior checklist	Job-specific	Published
Lilienthal and Mack (1992) First level super	.19	299	0	Unknown	Generic	Unpublished
Lilienthal and Mack (1992) Second level super +	.18	94	0	Unknown	Generic	Unpublished
Lopez (1966) Administrative Services	–.02	58	Predictive Unknown	Behavior checklist	Job-specific	Unpublished
Lopez (1966) Facility Management	.27	97	0	Behavior checklist	Job-specific	Unpublished
Lopez (1966) Secretarial (Port Authority)	.3	36	0	Behavior checklist	Job-specific	Unpublished
Lopez (1966) Lieutenant	.04	80	0	Behavior checklist	Job-specific	Unpublished
Melchers and Kleinmann (2009) Sales	.03	152	5 years	Behavior checklist	Job-specific	Unpublished
Melchers and Kleinmann (2009) Nonsales	.04	90	5 years	Behavior checklist	Generic	Unpublished
Meyer (1970)	.29	81	0	*Assessor ratings*	Job-specific	Published
Nowack (1997)	.28	144	0	Behavior checklist	Job-specific	Published
O'Connell et al. (1999)	.298	73	0	*Assessor ratings*	Generic	Unpublished
O'Connell et al. (2001)	.409	53	Predictive Unknown	*Assessor ratings*	Generic	Unpublished
T. Kiger (personal communication, March 25, 2013)	.19	134	0	*Assessor ratings*	Job-specific	Unpublished
Turnage and Muchinsky (1984)	.01	319	4 years	*Assessor ratings*	Job-specific	Published
van Hooft, van der Flier, and Minne (2006)	.09	159	0	Electronic scoring	Generic	Published

Note: For *Time lag*, all those with 0 are concurrent; for *Predictor used*, those considered subjective are italicized.

Table A2. In-basket reliability distribution

Study ID	r	*N*	Reliability
Bader (1986) sample 1	.71	13	Interrater
Bader (1986) sample 2	.78	20	Interrater
Boyd (1990) objective scoring	.86	99	Interrater
Boyd (1990) subjective scoring	.60	99	Interrater
Bobrow and Leonards (1997)	.78	169	Interrater
Bourgeois & Slivinski (1974)	.83	38	Interrater
Bray and Grant (1966)	.92	355	Interrater
Brostoff and Meyer (1984)	.87	32	Interrater
Denning (1980)	.38	25	Interrater
Fedorko (1992)	.97	60	Interrater
Frederiksen, Jensen, and Beaton (1972)	.46	258	Interrater
Frederiksen, Saunders, and Wand (1957) CO Problems	.91	112	Interrater
Frederiksen et al. (1957) D/M Problems	.74	110	Interrater
Frederiksen et al. (1957) D/O Problems	.86	106	Interrater
Frederiksen et al. (1957) D/P Problems	.80	102	Interrater
Gill (1979)	.48	15	Interrater
Jaffee and Michaels (1978)	.91	39	Interrater
Kim (1980)	.71	96	Interrater
Lopez (1966)	.81	80	Interrater
Mayes, Belloli, Riggio, and Aguirre (1997)	.28	62	Interrater
Nowack (1997)	.93	144	Interrater
Reed (1981)	.59	154	Interrater
Shapira and Dunbar (1980) rater 1 and rater 2	.93	112	Interrater
Silverman (1972)	.76	25	Interrater

Table A3. Distribution of correlations with cognitive ability

Study ID	*r*	*N*	Reliability	Test of *g*
Avolio, O'Connell, Martiz, and Kennedy (1999)	.08	156	.68	Reasoning (15 items) + critical thinking (15 items)
Bentz (1968)	.27	56	.85	Unknown
Denning (1980)	.11	79	.38	Kit of factor-referenced cognitive tests
Frederiksen (1966) four tests	.05	115	.70	Average across all four cog tests with all in-basket ratings
Frederiksen, Saunders and Wand (1957)	.25	92	.70	Unknown
Goldstein et al. (1998)	.29	633	.87	Wesman Personnel Classification Test – verbal and numerical
Hakstian et al. (1986) females	.28	110	.95	Wonderlic and Cattell Culture Fair Intelligence Test
Hakstian et al. (1986) males	.24	128	.94	Wonderlic and Cattell Culture Fair Intelligence Test
Hemphill, Griffiths, and Frederiksen (1962)	.37	232	.61	Kit of selected tests for reference aptitude and achievement factors
Lance et al. (2000)	.24	353	.78	Written cognitive ability test: math, writing fluency, reading comp, following directions
Lopez (1966) Admin Services	.34	58	.70	School and College Ability Test (SCAT)
Lopez (1966) Facility Management	.44	97	.70	School and College Ability Test (SCAT)
Lopez (1966) Lieutenant	.27	80	.81	Composite Alpha 9 and ACE
Meyer (1970)	.42	126	.72	Wonderlic
O'Connell et al. (1999)	.11	95	.68	Nonverbal reasoning (15 items) + critical thinking (15 items)
O'Connell et al. (2010)	.16	449	.68	Nonverbal reasoning (15 items) + critical thinking (15 items)
O'Connell et al. (2001)	.40	53	.68	Nonverbal reasoning (15 items) + critical thinking (15 items)
Reed (1981)	.37	154	.59	Shipley total math and verbal
Tziner and Dolan (1982)	.34	193	.70	Composite among verbal, nonverbal and inductive intelligence