Journal of Business and Psychology (2019) 34:337–356
https://doi.org/10.1007/s10869-018-9544-y

ORIGINAL PAPER

The Effects of Empirical Keying of Personality Measures on Faking and Criterion-Related Validity

Jeffrey M. Cucina[1,2] · Nicholas L. Vasilopoulos[3] · Chihwei Su[2] · Henry H. Busciglio[2] · Irina Cozma[4] · Arwen H. DeCostanza[5] · Nicholas R. Martin[6] · Megan N. Shaw[7]

Published online: 7 May 2018

Abstract

We investigated the effects of empirical keying on scoring personality measures. To our knowledge, this is the first published study to investigate the use of empirical keying for personality in a selection context. We hypothesized that empirical keying maximizes use of the information provided in responses to personality items. We also hypothesized that it reduces faking since the relationship between response options and performance is not obvious to respondents. Four studies were used to test the hypotheses. In Study 1, the criterion-related validity of empirically keyed personality measures was investigated using applicant data from a law enforcement officer predictive validation study. A combination of training and job performance measures was used as criteria. In Study 2, two empirical keys were created for long and short measures of the five factors. The criterion-related validities of the empirical keys were investigated using Freshman GPA (FGPA) as a criterion. In Study 3, one set of the empirical keys from Study 2 was applied to experimental data to examine the effects of empirical keying on applicant faking and on the relationship of personality scores and cognitive ability. In Study 4, we examined the generalizability of empirical keying across different organizations. Across the studies, option- and item-level empirical keying increased criterion-related validities for academic, training, and job performance. Empirical keying also reduced the effects of faking. Thus, both hypotheses were supported. We recommend that psychologists using personality measures to predict performance should consider the use of empirical keying as it enhanced validity and reduced faking.

Keywords Personality · Faking · Empirical keying · Validity · Training performance · Job performance · Academic performance · Freshman grade point average · Five-factor model · Impression management

Portions of this paper are based on the doctoral dissertation that Jeffrey Cucina completed while at the George Washington University; he is now at U.S. Customs and Border Protection. In addition, portions of this paper were previously presented at the 2010, 2011, and 2012 meetings of the Society for Industrial and Organizational Psychology and the 2007 of the International Public Management Association for Human Resources Assessment Council. The views expressed in this paper are those of the authors and do not necessarily reflect the views of any agency of the U.S. Federal Government, Development Dimensions International, Aon, or Amazon.

Electronic supplementary material The online version of this article (https://doi.org/10.1007/s10869-018-9544-y) contains supplementary material, which is available to authorized users.

✉ Jeffrey M. Cucina
jcucina@gmail.com; jeffrey.cucina@cbp.dhs.gov

[1] George Washington University, Washington, DC, USA

[2] U.S. Customs and Border Protection, 1400 L ST NW, 7S39, Washington, DC 20229-1145, USA

[3] National Security Agency, Fort Meade, MD, USA

[4] Development Dimensions International, Bridgeville, PA, USA

[5] U.S. Army Research Laboratory, Aberdeen Proving Ground, MD, USA

[6] Aon, Austin, TX, USA

[7] Amazon, Seattle, WA, USA

Introduction

Meta-analytic research has shown that the five factors of personality predict job performance, training performance, and academic performance (Barrick & Mount, 1991; McAbee & Oswald, 2013). However, the relationship for job performance might be overestimated by 30% (Kepes & McDaniel, 2015) and factors other than conscientiousness have low validities (less than .10) for academic performance (McAbee & Oswald, 2013). This is somewhat discouraging given interest in looking beyond cognitive tests (see Soares, 2012, description of the SAT optional movement) and the need to identify other predictors.

Additionally, concerns over the susceptibility of self-report personality measures to faking persist (Morgeson et al., 2007; Kuncel & Hezlett, 2010). One study suggests that the typically reported estimates of the effect size of faking ($d = .83$) underestimate the true effect size ($d = 2$) by not incorporating all of the factors related to faking (Tett, Freund, Christiansen, Fox, & Coaster, 2012). Some psychologists have suggested the use of forced-choice measures of personality (Stark et al., 2014) or warnings (Fan et al., 2012) to reduce faking. Although these approaches reduce faking, they can introduce a correlation with cognitive ability (Christiansen, Burns, & Montgomery, 2005; Vasilopoulos, Cucina, & McElreath, 2005; Vasilopoulos, Cucina, Dyomina, Morewitz, & Reilly, 2006). That said, this side effect does not always occur (Converse et al., 2008; O'Neill et al., 2017) and more research appears to be underway (Van Geert et al., 2016). Nevertheless, the side effect would serve to diminish both incremental validity (over cognitive ability) and construct validity while potentially increasing adverse impact. Furthermore, observer ratings are also susceptible to faking (Konig, Steiner Thommen, Wittwer, & Kleinmann, 2017).

The purpose of this research is to address the concerns about low validity and faking by investigating whether changes to the approach for scoring personality measures, specifically item and option empirical keying, can enhance the criterion-related validity of personality measures used in high-stakes settings, while also reducing faking. *Item empirical keying* involves differentially weighting (e.g., via regression or correlation coefficients) each item in a personality inventory based on its relationship with the criterion. *Option empirical keying* involves determining the relationship between each response option and the criterion and assigning weights for endorsement of each response option accordingly. These approaches can be contrasted with *rational keying* which involves unit weighting each item and using Likert scoring (i.e., 1 = strongly disagree, 2 = disagree, 3 = neither agree nor disagree, 4 = agree, 5 = strongly agree) for response options.

We make the general hypothesis that an applicant's responses to individual personality items provide information not only on their standing on the big five factors and facets but also smaller aspects of personality. This additional personality information can be used, in conjunction with empirical keying, to predict performance. The resulting empirical key is not as obvious as a traditional rational personality scoring key. Thus, identifying which response options maximize scores is more difficult for applicants, causing a reduced faking effect.

We aim to make a contribution to practice by determining whether it behooves practitioners to consider the use of empirically keyed personality measures for selection. Ultimately, the goal of a selection system is to predict future performance as much as possible. If empirical keying enhances prediction, its use would benefit organizations. Empirical keying has the added possibility of improving validity without increasing applicant burden (e.g., testing time) and providing practitioners with a better use of existing data captured from applicants (i.e., item responses). If empirical keying reduces faking, it can help organizations mitigate the effects of faking on hiring decisions. Faking is of concern not only for its possible effects on validity and score inflation but also for perceptions of fairness and the integrity of the hiring process. If we find that empirical keying does not enhance validity or reduce faking, then practitioners can avoid spending time and resources developing empirical keys and they can be prepared to explain to stakeholders that traditional scoring maximizes validity. We hope that providing a systematic investigation of empirical keying will help practitioners better decide how to focus their time and resources when developing selection systems.

We aim to make a contribution to the literature by systematically evaluating the efficacy of empirically keyed personality measures. A substantial body of research exists on the criterion-related validity of personality. Yet, we were unable to locate any published research on the effects of empirical keying on the criterion-related validity of personality for academic, job, or training performance. One exception was Davis (1997) who found that empirical keying improved the prediction of absenteeism and/or turnover (compared to rationally keyed items) in one of three samples considered. Our study adds to the literature by using the more often-studied criteria of academic, training, and job performance, which were not included in Davis's work.

The item and option information that could be used for empirical keying are collected from participants and applicants, but are often ignored or discarded. Our study aims to shed light on the validity of this information. Whether or not empirical keying increases criterion-related validity is an important fact to know for applied psychologists. Practitioners would no doubt want to maximize the criterion-related validity of their assessments for selecting applicants. Academics no doubt would be interested in which additional personality-related individual differences are related to performance. In the next sections, we describe three conceptual rationales, borrowed from Davis's (1997) work, to support our hypothesis that empirical keying increases criterion-related validity. We then describe two additional rationales we developed.

Latent Versus Emergent Models of Personality

The first conceptual rationale from Davis (1997) involves prior work distinguishing latent and emergent models of personality. Bollen and Lennox (1991) described two types of structural models that can occur in situations where there is an unmeasured factor and measured indicators of that factor. In an *emergent model*, causal indicators (e.g., personality items and facets) are viewed as causing the factors (e.g., the big five). In a *latent model*, effect indicators (e.g., personality items and facets) are viewed as being caused by latent factors (e.g., the big five).

Ozer and Reise (1994) maintain that most personality constructs are more properly viewed using an emergent framework, not a latent one. They argue that in an emergent framework, instruments should be evaluated using an external criterion, as opposed to an internal criterion (e.g., internal consistency), which would be more appropriate for latent models. Under an emergent viewpoint, items contain reliable specific variance that should be differentially weighted to predict an external criterion. Thus, the emergent viewpoint predicts that empirical keying will enhance criterion-related validity. Under a latent viewpoint of personality, all of the information at the item level comes from the factor, leaving no reliable specific variance for use in empirical keying. Thus, under a latent viewpoint, empirical keying will not increase validity.

The Bandwidth-Fidelity (BWF) Dilemma

Davis' (1997) second rationale involves the BWF dilemma, which concerns whether narrow measures of a construct provide a better prediction of performance than broad measures. Broad measures focus on the common variance that is shared among related traits, whereas narrow measures contain more specific variance for an individual trait. Many personality instruments follow a two-stratum model whereby items are grouped into facets, which are in turn, grouped into factors. Most researchers have applied the BWF dilemma to the factor-facet interface by comparing the validity of factors and differentially weighted facets. Moon, Hollenbeck, Marinova, and Humphrey (2008) reported that facets of extraversion predicted organizational citizenship behavior in opposite directions (even canceling out when aggregated to an extraversion factor). However, it is also possible to apply the BWF dilemma to the facet-item interface. Personality items tap individual differences in behavior in different situations. It is possible that behavior in certain situations might better predict a criterion than behavior in other situations, even if both situations are linked to items in the same facet.

There is emerging evidence for the presence of personality traits at the item level. Mottus, Kandler, Bleidorn, Riemann, and McCrae (2017) recently investigated the presence of personality *nuances*, which are microdimensions of personality that exist at the level below facets. Nuances often focus on individual differences in response to certain situations and exist at the item level or with small clusters of items (e.g., two to three items). Mottus et al. (2017) reported that self and other personality item scores correlated even after partialling out facet-level variance. This indicates that personality items have unique reliable variance that is not accounted for by the facets. Using longitudinal twin study data, they reported evidence of the temporal stability of the unique reliable item-level variance. They also correlated the unique item-level variance with various criteria (i.e., self-reported conservatism, life satisfaction, interests, body mass index) and found significant correlations when controlling for the facets.

Earlier, Goldberg (1993) argued that in a sample of sufficient size, prediction is maximized at the item level rather than at the factor or facet level. This is due to the presence of reliable specific variance at the item level, which can be used for prediction purposes (although a large sample is needed to reduce capitalization on chance). For somewhat smaller samples, prediction would be maximized at the facet level and for even smaller samples, prediction is only maximized at the factor level. Taking a narrow-trait viewpoint (and following Goldberg's arguments), we hypothesize that item-level scoring (i.e., empirical keying) will increase validity.

Similarity of Biodata and Personality

Davis (1997) also suggested that since biodata and personality exhibit similarities, the fact that empirical keying enhances biodata's validity suggests it will do the same for personality. There is conceptual and empirical evidence that empirical keying increases the criterion-related validity of biodata inventories (Cucina, Caputo, Thibodeaux, & MacLane, 2012; Guion, 1965; Mitchell & Klimonski, 1982). Extending the use of empirical keying from biodata instruments to personality inventories is possible, but it should be mentioned that the similarity between personality and biodata is more evident for soft biodata items than hard biodata items. Indeed, several researchers have noted the similarity in item content between personality and soft biodata inventories (e.g., Davis, 1997; Mael, 1991; Stricker & Rock, 1998). If the two types of instruments have similar item content, then it can be hypothesized that empirical keying (i.e., item and option-level scoring) will also increase the criterion-related validity of personality.

Two Additional Conceptual Rationales

The unpublished work by Davis (1997) provides some basis for the hypothesis that empirical keying will increase the criterion-related validity of personality measures. However, we felt that additional rationales could be developed to compliment and expand upon Davis' (1997) work. We propose two of our own rationales that underlie our hypothesis regarding the increased validity associated with empirically keyed personality measures: Peak-Point Personality Response Option Framework (PPPROF) and Multidimensional Personality Item Framework (MPIF). Both of these frameworks suggest that by collapsing personality items into scale scores, some criterion-relevant variance is obscured. This occurs because the response options in a personality item may have differing curvilinear relationships with a criterion and because the item itself may tap more than one personality/behavioral dimension.

Peak-Point Personality Response Option Framework

Our first framework, Peak-Point Personality Response Option Framework (PPPROF), proposes that each item will have a validity peak, with respect to the criterion score, across the range of response options. The location of the peak depends on the content and constructs of the item and the item's relationship with the criterion. The rational keying method of scoring personality items places the peak at either the high or the low end of the response option scale. PPPROF suggests that some personality items have their peak validity points at other response options. There are a number of recent primary research articles demonstrating the curvilinearity of the relationship between scale-level personality and academic, training, and job performance (e.g., Cucina & Vasilopoulos, 2005; Vasilopoulos, Cucina, & Hunter, 2007; Le et al., 2011).

We suggest that many personality dimensions could exhibit curvilinear relationships with performance, especially at the extreme high and low ends of a trait. For example, an individual with very high conscientiousness might exhibit behavior that, although not at the threshold of obsessive-compulsive disorder, might inhibit performance. Additionally, some personality scales may have asymptotic relationships with performance (e.g., a certain amount of extraversion might be required for a particular job, with further amounts not increasing performance). Furthermore, impression management (including both outright faking and self-deception) can introduce curvilinearity by moving individuals with a true value of a trait (e.g., conscientiousness) to a higher observed score value than they deserve.

Past research examining curvilinearity has largely been focused at the scale level rather than the item level. There are two exceptions whereby researchers have looked at non-linear relationships between personality items and both faking and latent personality factors, but not performance criteria. Kuncel and Tellengen (2009) found that faking can have a non-linear effect on personality item responses. Instead of always maximizing an item score (e.g., by selecting strongly agree on a 5-point positively keyed item), fakers often select other responses options (e.g., agree) because they view those other response options as being more socially desirable. Kuncel and Borneman (2007) looked at the differences in response option endorsement for fakers and honest respondents and developed a faking-resistant personality key. Essentially, they empirically keyed a personality inventory, at the option level, to predict faking. On a related note, ideal-point/unfolding item response theory (IRT) models can improve the measurement of personality (Carter et al., 2014). These models allow for the relationship between personality item responses and the underlying latent trait to be not only non-linear but also non-monotonic. Although neither lines of research demonstrated that personality items have non-linear relationships with performance criteria, both do demonstrate that personality items can exhibit non-linear relationships with other outcomes.

It is possible for different items to capture different aspects of a curvilinear relationship between personality and performance. Consider the hypothetical inverted U-shaped regression curve between a personality dimension and performance shown in Fig. 1; curves similar to this have been shown at the scale level in curvilinearity literature (e.g., Cucina & Vasilopoulos, 2005). In this curve, scores at the low end of the personality dimension are associated with negative performance, scores at the middle are associated with maximal performance, and the relationship begins to bend downward at the high end of the personality dimension. We propose that personality items have differing abilities to measure the curvilinearity of the relationship, mainly due to how they cover the full range of a personality dimension. For example, item A in Fig. 1 is at the low end of the personality dimension and depicts a positive linear relationship across its response options. Item B in Fig. 1 is at the middle of the personality dimension and depicts a curvilinear relationship. Item C, at the high end, has a negative linear relationship.

We propose that there is a peak point across a personality item's response options at which endorsement of that response option is associated with higher performance on the criterion. For example, item A in Fig. 1 has a peak point at response option 5, item B has a peak point at response option 3, and item C has a peak point at response option 1. When personality inventories are rationally keyed, all of the items receive the same weights for each response option and the differences in the relationship between the criterion and the response options are lost. Therefore, by conducting option-level empirical keying, each item is allowed to have a unique curvilinear (or linear) relationship with the criterion.

Multidimensional PersonalityItem Framework

The second framework that we propose, Multidimensional Personality Item Framework (MPIF), reflects the fact that personality items are often related to more than one aspect of

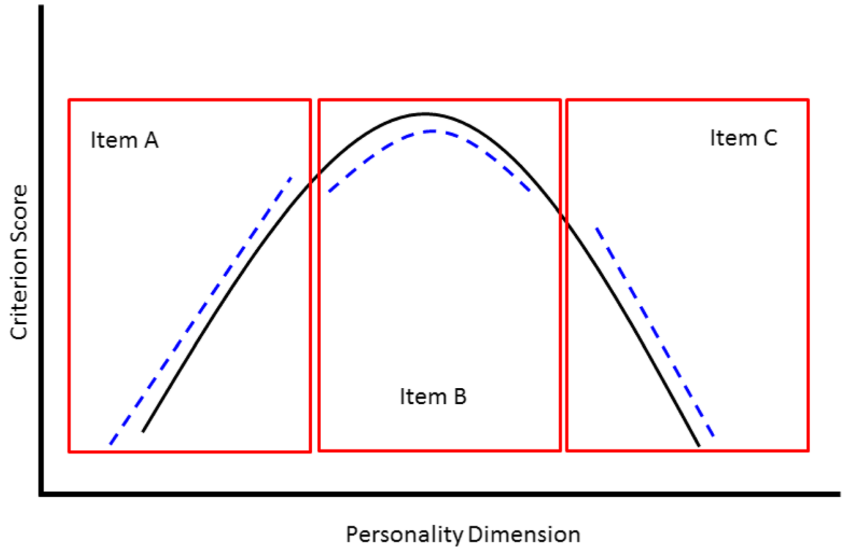

Fig. 1 A hypothetical depiction of the coverage of the latent personality dimension and its relationship with the criterion and its measurement by three different items

personality. To explain, personality items are often written as behaviors or cognitions and several different psychological constructs may underlie a behavior or cognition. Consider the fact that the median test-retest reliability of a conscientiousness personality item is .76 (using publicly available data from Pozzebon et al., 2013, for the Mature Personality and Tidiness scales), yet the median variance explained by item-total correlations (for the two scales) is only .38. A variance decomposition reveals that there is a median loading of .60 of each item on another unique construct, which has the potential to be predictive of performance.

To demonstrate the potential implication of masking unique variance, we use items that could be included on measures of the orderliness and achievement-striving facets of conscientiousness. Using traditional methods, the items included on a single facet scale are written with the intent of achieving construct validity; however, it is unlikely that their criterion-related validities are uniform. For example, construct-valid scales created to assess orderliness typically include items asking if an individual is neat/messy, such as "leave a mess in my room" and organized/disorganized, such as "work according to a plan." It is unlikely that the item "leave a mess in my room" will correlate with job or academic performance because the conceptual alignment between having a messy room and performance in a work or academic setting is nebulous at best. In contrast, the item "work according to a plan" is more likely to correlate with job and academic performance because, in most situations, planning work increases the chance that a project or assignment is completed on time and successfully. In fact, we know that criterion-related validity is slightly increased when personality items refer specifically to a work or academic situation (Kepes & McDaniel, 2015; Shaffer & Postlethwaite, 2012). Thus, a low correlation between the item "leave a mess in my room" and performance could occur because it is directly contextualized to the home, whereas the item "work according to a plan" is contextualized to both the home and work.

Unique aspects of two personality items may also have opposite relationships with performance. For example, in contrast to the orderliness item "work according to a plan," the item "want everything to be just right" may have a negative relationship with performance because individuals who want things to be perfect may spend too much time on a task at the expense of completing other, equally consequential tasks. When these items are aggregated to construct a scale, the negative unique relationship associated with the first item cancels out with the positive relationship associated with the second item causing the overall score to be uncorrelated with the criterion.

In some situations, the unique aspects of two personality items may have a positive relationship with performance, yet with different magnitudes. For example, the achievement-striving items, "work hard" and "demand quality" both should have a positive relationship with performance, although the work hard item can be hypothesized to have a stronger relationship with performance given that it directly assesses the quantity and quality of time spent by an individual on work-related tasks. The demand quality item should have a positive relationship with performance as individuals who demand quality would likely want to see quality in their own work in order to avoid cognitive dissonance (Festinger, 1957). However, the nebulousness of this item (e.g., does it refer to the quality of goods and services that an individual is purchasing or the quality of work that an individual performs) weakens the hypothesized relationship to performance.

The possibility that unique aspects of personality items relate differently to performance is also apparent by examining the content of personality items. For example, consider the following IPIP NEO-PI-R items for Gregariousness: "Love large parties," "Talk to a lot of different people at parties," "Don't like crowded events," and "Seek quiet." We suggest that in addition to measuring extraversion, these items might be related to positive study habits, academic performance, and conscientiousness in academic settings. An individual may not spend time attending parties either due to their low extraversion, or due to their choice to spend their time studying. As a consequence, an individual's responses to these items may indirectly reflect studying behavior, and other aspects of conscientiousness or the intellect portion of openness to experience, in addition to extraversion. Also, note that other IPIP NEO-PI-R items for extraversion may actually have a negative relationship with academic performance. The items "Keep in the background," "Don't like to draw attention to myself," "Have little to say," and "Hold back my opinions," may be negatively related with classroom participation which is included in the grading criteria for some classes. It is also possible that items may be multidimensional with respect to non-personality constructs. For example, consider the IPIP NEO-PI-R Modesty item "Know the answers to many questions." Although the agreeableness factor might not correlate with academic performance, this item might. Knowing "the answers to many questions" could be indicative of crystallized intelligence, which would predict academic performance (via its relationship to *g*). Thus, individuals with higher *g* may endorse this item, causing it to measure *g* (and predict academic performance) in addition to agreeableness.

Items in the morality facet of agreeableness can also be considered. These items could have a U-shaped relationship with academic performance. Students on the high end of this facet may be less likely to engage in counterproductive behaviors and may choose to spend more time studying and attending classes (because they perceive it as the "right" thing to do) and less time engaging in behaviors that could impede academic performance (e.g., drug use, underage drinking, attending parties, skipping classes). However, it is also possible that individuals at the very low end of this trait may also have high criterion scores due to cheating. Individuals low on morality may be more likely to cheat on exams (e.g., by obtaining copies of past year's exams, using imposters to take their exams, copying other students' answers) and other graded assignments. In fact, one of the items in this facet even reads "cheat to get ahead." By ignoring the multidimensionality of items and the differing relationships between items and the criterion, validity may not be maximized. Thus, it could be helpful to individually weight items and response options, using empirical keying.

Connection Between Davis' (1997) Rationales and Both PPPROF and MPIF

There are some connections between the PPPROF and MPIF rationales and Davis' (1997) work. PPPROF suggests that validity is maximized when response options are allowed to have a non-linear relationship with performance. This type of non-linearity is often modeled in biodata instruments via empirical keying and Davis (1997) noted the similarity between biodata and personality measures. Therefore, empirical keying may increase the criterion-related validity of personality measures just as it does for biodata. Davis (1997) also used the BWF dilemma line of research as a rationale for empirical keying of personality measures. MPIF can be viewed as an item-level extension of the BWF dilemma (which typically refers to the facet-level of personality) in that it recognizes that many different aspects of personality exist and that these aspects can have different criterion-related validities. Rather than examining the aspects of personality at the facet level, MPIF extends this distinction to the item level and recognizes that items have multiple sources of variance with differing validities.

Faking and Empirical Keying

In addition to enhancing criterion-related validity, empirical keying may also reduce the effects of faking. Empirical keying has been hypothesized to make the scoring key for self-report measures more subtle (i.e., less obvious) than rational keying (Mumford & Stokes, 1992). There is support for the notion that more subtle items are more resistant to faking (Mumford & Stokes, 1992). Kluger, Reilly, and Russell (1991) found direct empirical support for reduced fakability using empirical keying procedures with a biodata instrument. As discussed above, personality items may have differing types of relationships with performance. Ignoring these effects and focusing just on factor scores makes the scoring key much more transparent and obvious to applicants. Thus, in contrast to rational keying (where items contribute equally to the factor score), item-level empirical keying can make it more difficult for applicants to "fake-good" because the items contributing most to the score are not easily identifiable. Option-level empirical keying may make "faking-good" even more difficult because the response option that yields the maximal score for an item is less obvious to applicants. The most socially desirable response option may be scored lower than one of the less desirable options, making it difficult for applicants to successfully fake their responses. Although the use of empirical keying to reduce faking for biodata is documented, we could locate no published research that examined the effects of empirical keying on score inflation for personality measures. We address this research gap using experimental data. Failing to examine this hypothesis leaves open the possibility that faking is not mitigated in the best way possible.

Relationship to Cognitive Ability

Previous research has shown that under faking conditions, personality measures correlate with cognitive ability test scores when forced-choice items (Christiansen et al., 2005;

Vasilopoulos & Cucina, 2006) and warnings of response verification are used (Vasilopoulos et al., 2005). Vasilopoulos and Cucina (2006) hypothesized that a similar finding could occur for empirically keyed non-cognitive measures (e.g., personality). There are a few reasons why empirical keying may introduce a cognitive load. First, empirical keying involves assigning more weight to items and response options that better correlate with the criterion. It is possible that higher cognitive ability individuals can better identify which items and options correlate with the criterion and inflate their scores accordingly, in comparison to lower cognitive ability individuals. This has been seen at the trait level in the forced-choice literature. Christiansen et al. (2005) reported evidence that higher cognitive ability participants have more accurate implicit job theories about which personality traits are related to job performance. Using that information allowed high cognitive ability participants to inflate their scores on a forced-choice instrument by selecting response options that measured those traits that relate to performance. It is possible that this effect could be extended to the item and response option levels for empirically keyed personality measures.

Second, it is possible that even in the absence of faking, there might be a relationship between cognitive ability and both item-level scores and response option endorsement. Empirical keying has the potential to capture that relationship and harness it to predict a criterion. This has been seen in the meta-analytic literature for biodata, which is typically empirically keyed. According to Schmidt and Hunter's (1998) review, biodata scores correlate with cognitive ability ($r = .50$) and much of their criterion-related validity appears to be derived from cognitive ability as evidenced by low incremental validity ($\Delta R = .01$) and a smaller β-weight (.13) compared to cognitive ability (.45).

Nevertheless, the conceptual underpinnings for a possible relationship between empirically keyed personality scores and cognitive ability are not as robust as those for the other topics covered in this paper. Therefore, we do not make an explicit hypothesis concerning the relationship between empirically keyed personality scores and cognitive ability. However, we believe this is an important topic to explore.

Overview of the Research and Hypotheses

In this paper, we describe four studies that examined the use of empirically keyed personality measures. The primary focus of the analyses is on potential improvements in criterion-related validity and reductions in faking. In the first study, we examine the effects of empirical keying on criterion-related validity for job and training performance using applicant data from a field criterion-related validation study. In the second study, we examine the effects of empirical keying on criterion-related validity for academic performance using a laboratory sample. In the third study, we examine the effects of empirical keying on faking and the relationship between personality scale scores and cognitive ability using a laboratory sample. The purpose of the fourth study is to determine if the generalizability of an empirically keyed personality measure differs from a rationally keyed personality measure. This study addresses the situational specificity and cross-validity of empirical keys for personality measures.

Based on the discussion above, we make the following hypotheses:

> Hypothesis 1: Empirical keys of personality scales will have higher criterion-related validities than rational keys.
> Hypothesis 2: Empirical keying will reduce the extent of personality score inflation in an applicant setting. When using an empirical key, mean scores on the Big Five personality scales will be similar in the applicant and honest response conditions.

We also conduct two supplementary analyses. The first investigates the possibility that empirical keying introduces a correlation with cognitive ability. The second examines the effects of hybridization on empirical keying of personality measures, which has been advanced as a compromise between the purely empirical and purely rational approaches. It is purported to enhance (a) cross-validity by removing weights that are inconsistent with rationality and conceptual models and (b) legal defensibility (Cucina et al., 2012). Large-scale biodata inventories that employ empirical keying often include hybridization (Gandy, Dye, & MacLane, 1994). Thus, we include hybridization to increase the ecological validity of our study to operational practice.

Study 1

This study examined the criterion-related validity of an empirically keyed personality measure using data from a predictive validation study in a work setting.[1] The dataset included personality data obtained from applicants and it was validated against criteria measured in training and on the job several years after the applicants were hired. The study also examined the effects of hybridizing an empirical key on criterion-related validity using applicant data.

Method

Participants

The sample consisted of 854 gun-carrying federal law enforcement officers (LEOs) who took the personality measures as applicants, were subsequently hired, and later

[1] Data on other non-personality predictors from this study have appeared elsewhere (Cucina, Busciglio, & Vaughn, 2013; Cucina, Su, Busciglio, & Peyton, 2015; Cucina, Su, Busciglio, Thomas, & Peyton, 2015).

participated in a criterion-related validation study. Note that the criteria were collected after the applicants were hired and the personality measure was completed as part of the hiring process; thus, the study uses a predictive validation design. Simulation research suggests that the sample size of this archival dataset is adequate for empirical keying (Cucina et al., 2012). As is typical with LEOs, the sample was predominantly male (84%). In terms of race and national origin, 38% of the sample was White, 36% was Hispanic, 11% was African American, 14% was Asian/Pacific Islander, and the remaining participants were either Native American or missing.

Personality Measure

Two of the Big Five dimensions of personality were measured in this study: conscientiousness and emotional stability (a reverse-coded scale of neuroticism), consisting of 9 and 10 items, respectively. These dimensions were selected based on job-analysis results. The items were custom developed to mimic the NEO-PI-R's (Costa & McCrae, 1992) conscientiousness and neuroticism scales that were relevant to specific Federal LEO positions. Vasilopoulos et al. (2005) used a similar measure and reported evidence of construct validity. Note that items not measuring conscientiousness and neuroticism were excluded from this study. In addition, items from the initiative and dependability scales from Vasilopoulos et al. were combined into a single conscientiousness scale in the present study. The instructions asked applicants to use a 5-point scale ranging from 1 (*strongly disagree*) to 5 (*strongly agree*) to indicate the extent to which the behavior depicted in the stem described them.

Rational Key This is the traditional method of scoring personality instruments at the five-factor level. Items were scored using Likert scoring (i.e., an item response of "strongly agree" was scored a 5, or a 1 if the stem was negatively worded). Each item's score was unit weighted and summed to generate scores on each of the Big Five factors.

Item-Level Correlational Scores In this method, items were scored using Likert scoring. The bivariate correlation between each item and the composite criterion was obtained. Items were differentially weighted using this correlation coefficient and were summed to create scale scores. For example, if the bivariate correlation between item 1 and the criterion was 0.15, then the Likert scores on item 1 were multiplied by 0.15.

Item-Level Stepwise Regression Scores In this method, items were scored using Likert scoring. The Likert-scored items were then entered into a stepwise regression procedure to predict the composite criterion. The resulting regression equation was used to compute the unstandardized predicted values of the criterion for both the developmental and holdout portions of the dataset (which are described below). The unstandardized predicted values were then correlated with the criterion in order to estimate the criterion-related validity.

Option-Level Point Biserial Scores The option-level point biserial method was used to empirically key the personality items. Individual response options were coded as 0s or 1s in a manner similar to dummy coding (except that there were k recoded variables rather than $k-1$ recoded variables as in dummy coding). Next, the point biserial correlation between each of the dummy coded response option variables and the criterion was obtained. These correlations formed the response option weights. For example, if the point biserial correlation between endorsement of option A and the criterion was 0.08, then respondents choosing option A received 0.08 points for the item.

Empirical keying capitalizes on chance, necessitating the use of cross-validation. Therefore, the pure empirical key was cross-validated using triple (or threefold) cross-validation, whereby the dataset is divided into thirds and three keys were developed (Brown, 1994). The dataset was randomly divided into the three parts (statistical tests were used to confirm the equivalence of the three parts). Two-thirds of the data (i.e., the developmental sample) were used to develop the key and one-third (i.e., the holdout or cross-validation sample) was used to cross-validate the key. Note that it is possible to arrange the three parts of the data into three different arrangements of two-thirds and one-third. In triple cross-validation, a cross-validity estimate was obtained for each possible arrangement. The average of the three cross-validities serves as an estimate of the true validity.

Hybrid Key To create the hybrid key, two personnel research psychologists reviewed each response option weight from the empirical keying, with respect to personality traits underlying each item and their hypothetical relationship with the criterion, frequency of response option endorsement (a proxy for stability of the obtained empirical weight), and job analytic information. Nonsensical weights were either adjusted (e.g., by collapsing with adjacent options) or were used as grounds for discarding the item in question. Next, a panel of six personnel research psychologists (four of which had peer-reviewed publications and research experience on non-cognitive assessments, and two of which were supervisors/managers) reviewed each response option weight and finalized the key (as part of a second-level review). Due to the time-consuming nature of the development of the hybrid key, it was not feasible to develop a hybrid key for all three cross-validations. Therefore, we developed a hybrid key for a single cross-validation. We chose the cross-validation run that had the median cross-validity (among the three cross-validities from the triple cross-validation). We obtained the developmental sample empirical weights for that cross-validation run and developed a hybrid key. The validity of the hybrid key was determined using the holdout sample for that run. Note that since it is common to develop a final empirical key using the entire sample for implementation purposes, we also developed a hybrid key for the entire sample. We did not estimate the validity of this key since we lacked a holdout sample.

Criteria

Three criteria, developed using SMEs, were used, including training academy course grades, research-based supervisory performance ratings, and scores on a task-based, paper-and-pencil work sample. The work sample was somewhat multidimensional in nature as it included items measuring performance on different duty areas and enforcing different sections of law. A composite criterion score was also computed by summing the standardized scores on the three individual criteria, using unit weighting. The empirical and hybrid keys were created using the composite criterion; however, we report cross-validities for all three criteria and the composite.

Procedure

All applicants completed the personality test items in a proctored setting using paper-and-pencil tests as part of a larger assessment battery that was completed during the application process. The raw response options on the personality items were retained and later used in a predictive validation study. A sample of 854 incumbents who took the personality tests as applicants participated in the validation study. All incumbents had successfully completed a multi-week training academy. The training academy scores were added to the dataset for use as the training performance criterion. The incumbents then completed a multiple-choice paper-and-pencil work sample that was proctored in groups of approximately 1–50 by personnel research psychologists carrying out the validation study. The personnel research psychologists also administered the research-based performance appraisal to supervisors, in groups of approximately 1–10, after conducting a brief training on use of the instrument.

Results

Table 1 presents the criterion-related validity coefficients for the rational keys, the pure empirical keys (which were triple cross-validated), and the hybrid keys (which were single cross-validated). Differences in validity coefficients were identified using Meng, Rosenthal, and Rubin’s (1992) Z test for comparing dependent correlation coefficients and Cohen’s (1992) q statistic for the effect size of the difference between two correlation coefficients. According to Cohen, the values of .10, .30, and .50 correspond to small, medium, and large effect sizes for the q statistic. A summary of these results is available in the Electronic Supplementary Material.

The rational key validities were all non-significant; however, the empirical key validities were statistically significant and were in the mid-teens for the composite criterion (see Table 1). Similar results were found for the training and work sample criteria. Empirical keying did not increase validity when supervisory ratings were used as the criterion. Thus, we only found support for hypothesis 1 for three of the four criteria. We also found that using empirical and hybrid keying reduced the internal consistency reliabilities of the scales (shown in the last column of Table 1). Regarding the effects of hybridization, the cross-validities of the hybrid key were only significantly higher than those for the pure empirical key in one of the eight comparisons. Thus, hybridization did not lead to enhanced criterion-related validity.

An anonymous reviewer inquired about the generalizability of empirically keyed personality measures. We examined the generalizability of empirical keys across different criteria using data from Study 1. Six separate point biserial empirical keys were created to predict training performance, the work sample scores, and supervisory ratings using either the conscientiousness or emotional stability items. Triple cross-validation was used. The supervisory ratings did not correlate with any of the keys. The empirical keys created using training performance as the criterion predicted training performance for both the conscientiousness ($r = .17, p < .001$) and emotional stability ($r = .17, p < .001$) items. These keys also predicted scores on the work sample ($r = .16$, $p < .001$ and $r = .15$, $p < .001$, respectively). Similarly, empirical keys created using the work sample as the criterion predicted not only the work sample ($r = .12$, $p < .001$ and $r = .14$, $p < .001$, respectively) but also training performance ($r = .22$, $p < .001$ and $r = .23$, $p < .001$, respectively).[2] Thus, there is some evidence of the cross-criterion generalizability of empirically keyed personality measures.

Discussion

Perhaps the best way to summarize the findings is to compute the average validities for each method across all criteria and both personality scales. The average observed validity for rational keying was − .01, for hybrid keying was .10, for option-level keying was .12, for item-level correlational keying was .17, and for item-level stepwise regression keying was .12.

[2] It is interesting that the empirical keys created using the work sample as a criterion had higher cross-validity for a criterion they were not developed to predict (i.e., training performance). We did notice that before cross-validation, these empirical keys had similar validities for the two criteria. Shrinkage occurred only for the work sample criterion and led to the keys having higher criterion-related validity for training performance than the work sample after cross-validation. Regardless, these results suggest that empirically keyed personality scales are not necessarily criterion specific in terms of cross-validity.

Table 1 Criterion-related validity coefficients for Study 1 and 300-item IPIP for Study 2

Study 1	Sup. Apr.	Training	Work sample	Overall.	Reliability
Conscientiousness					
Rational key/factor scale scores[a]	.01	− .03	− .02	− .02	.83
Empirical key item correlational scores[b]	.01	.22*[ef]	.17**[efg]	.19**[ef]	− .48
Empirical key item stepwise regression scores[b]	< .01	.14**[ef]	.13**[ef]	.12**[ef]	.04
Empirical key option point biserial-empirical[b]	− .01	.19**[eg]	.14**[eh]	.15**[e]	.39
Empirical key option point biserial-hybrid[c]	− .02	.23**[eg]	.04[gh]	.13**[e]	.54
Emotional stability					
Rational key/factor scale scores[a]	− .03	< .01	.02	− .01	.70
Empirical key item correlational scores[b]	.03	.26**[efgh]	.23**[efgh]	.25**[efgh]	.52
Empirical key item stepwise regression scores[b]	.02	.20**[ef]	.19**[efi]	.19**[ef]	.09[d]
Empirical key option point biserial-empirical[b]	.02	.18**[eg]	.14**[egj]	.16**[egi]	.60
Empirical key option point biserial-hybrid[c]	− .02	.18**[eh]	.10**[hij]	.13**[ehi]	.54
Study 2	(Cross-)validity		*p*		Reliability
Conscientiousness					
Rational key/factor scale scores[a]	.20(.23)		< .001		.91
Empirical key item correlational scores[b]	.26(.29)[ef]		< .001		.92
Empirical key item stepwise regression scores[b]	.31(.35)[efgh]		< .001		.25[d]
Empirical key option point biserial-empirical[b]	.26(.29)[eg]		< .001		.90
Empirical key option point biserial-hybrid[b]	.26(.29)[eh]		< .001		.91
Openness to experience					
Rational key/factor scale scores[a]	.08(.09)		.020		.86
Empirical key item correlational scores[b]	.17(.19)[e]		< .001		.71
Empirical key item stepwise regression scores[b]	.19(.22)[e]		< .001		.08[d]
Empirical key option point biserial-empirical[b]	.14(.16)[ef]		< .001		.68
Empirical key option point biserial-hybrid[b]	.17(.20)[ef]		< .001		.67
Neuroticism					
Rational key/factor scale scores[a]	.07(.08)		.044		.91
Empirical key item correlational scores[b]	.18(.21)[e]		< .001		.74
Empirical key item stepwise regression scores[b]	.19(.22)[e]		< .001		.18[d]
Empirical key option point biserial-empirical[b]	.19(.22)[e]		< .001		.69
Empirical key option point biserial-hybrid[b]	.19(.22)[e]		< .001		.79
Extraversion					
Rational key/factor scale scores[a]	− .06(− .06)		.117		.92
Empirical key item correlational scores[b]	.18(.20)[e]		< .001		.77
empirical key item stepwise regression scores[b]	.17(.19)[e]		< .001		.04[d]
Empirical key option point biserial-empirical[b]	.20(.23)[e]		< .001		.68
Empirical key option point biserial-hybrid[b]	.20(.23)[e]		< .001		.71
Agreeableness					
Rational key/factor scale scores[a]	− .03(− .03)		.402		.87
Empirical key item correlational scores[b]	.16(.18)[e]		< .001		.54
Empirical key item stepwise regression scores[b]	.14(.16)[e]		< .001		− .01[d]
Empirical key option point biserial-empirical[b]	.11(.13)[e]		.002		.67
Empirical key option point biserial-hybrid[b]	.13(.14)[e]		< .001		.73

Study 1—$n = 854$, except for the cross-validity of the hybrid key, which had a sample size of 285. Study 2—$n = 783$; validity coefficients were corrected for unreliability in the criterion (i.e., FGPA). The corrected coefficients appear in parentheses

Sup. Apr. supervisory appraisal

*$p < .05$, **$p < .01$

[a] The rational key/factor scores did not capitalize on chance and thus did not need cross-validation. These validity coefficients were computed using the entire sample

[b] Cross-validities for these keys were obtained using triple cross-validation

[c] Cross-validities for these keys were obtained using a single cross-validation on one-third of the dataset

[d] This is the reliability of the linear composite of positively and negatively weighted items (for which coefficient alpha was computed separately). All other reliabilities were obtained using coefficient alpha

[e] The difference between these validity coefficients and that of the rational key were statistically significant

[f, g, h, i, j] The difference between these pairs of validity coefficients was statistically significant

Thus, we found support for the hypothesis that empirical keying enhances criterion-related validity. Although hybrid keying is often advanced as a compromise between empiricism and rationality, our findings seemed to indicate it was somewhat closer in nature to empirical keying than to a midpoint between empiricism and rationality. The hybrid keys also correlated more strongly with the empirical keys than the rational and tended to have lower validities than the empirical keys.

Regarding practical implications, we found that using a rational key with applicant data resulted in a lack of validity. Thus, organizations may wish to be cautious about using rationally keyed personality measures to select applicants as there is limited applicant data supporting their use. Of course, incumbent studies (e.g., studies 2 and 3 in this paper) can yield valuable findings. However, it could be possible that personality constructs do correlate with job performance but are not adequately measured in applicant settings using rationally keyed instruments. We did find that empirical keying resulted in significant validity. Therefore, organizations considering the use of personality measures for selection could attempt to develop empirical keys. We also found that the time-consuming process of hybridization did not enhance validity; thus, practitioners may want to avoid spending significant amounts of time on hybridization.

One limitation is that the conscientiousness and emotional stability scales were originally created to predict counterproductive work behaviors (CWBs). Although the items do correlate highly with the NEO-PI-R, they were not originally intended to predict overall job performance. While it could be the case that rational keys predict CWBs, due to methodological issues (e.g., low base rates, inaccessibility of data, coding issues), CWBs could not be included here.

Study 2

The second study was a laboratory-based, concurrent, criterion-related validation study of a personality inventory in an undergraduate setting. The study was conducted to test hypothesis 1 (i.e., empirical keying increases criterion-related validity) and to create an empirical key for use in Study 3 for testing hypothesis 2 (i.e., empirical keying will reduce score inflation due to faking). We used two types of personality measures in Study 2, one containing 300 items (60 for each of the 5 factors) and the other containing 100 items (20 for each of the 5 factors). All participants took a combined inventory of 367 items (consisting of the items in the 300- and 100-item measures, of which 33 were common). Since many organizations use short measures of personality, the use of the second measure enhances the ecological validity of our findings.

In the present study, conscientiousness and openness to experience are the primary factors under investigation (however, post hoc analyses will be conducted using the remaining factors). Previous research has demonstrated that conscientiousness is a consistent predictor of academic performance (McAbee & Oswald, 2013), which is why we focused on it. Regarding openness to experience, previous research has found that it is a significant predictor of academic performance when scored curvilinearly (Cucina & Vasilopoulos, 2005) and there are conceptual linkages between the traits it measures (e.g., imagination, intellect) and learning.

Method

Participants

The sample consisted of 783 undergraduates enrolled in psychology courses at a medium-sized, urban university in the eastern USA. The participants completed the inventory either as part of a department-wide subject pool, or as part of a class exercise. The sample was 68% female (32% male), 76% non-Hispanic White, 10% Asian/Pacific islander, 4% African American, 3% Hispanic, and 7% "other." The median age of the participants was 19 years and 43% were freshman, 34% were sophomores, 14% were juniors, and 10% were seniors.

Personality Inventory

The Big Five dimensions of personality were measured using two Preliminary International Personality Item Pool (IPIP; Goldberg, 1999) measures designed to measure constructs similar to the NEO-PI-R (Costa & McCrae, 1992) facets and Big Five factors. The measure was an analogue of the NEO-PI-R, which is one of the most commonly cited and used measures of the FFM and consists of 300 items. The second measure was the 100-item IPIP key that was developed as an analogue of the NEO. The two measures have 33 items in common and these common items were administered only once in Study 2. Thus, a total of 367 personality items were administered to the participants.

Academic Performance

Freshman grade point average (FGPA) is the most commonly used measure of undergraduate academic performance in test validation research and served as the criterion in this study. FGPA was obtained from undergraduate transcripts.

Scoring Methods

The same scoring methods used in Study 1 (i.e., rational key scores, item-level correlational scores, item-level stepwise regression scores, and option-level point biserial scores) were used here for both the 300- and 100-item measures. The hybridization procedure was modified in that the first author hybridized the key separately for each iteration of the triple cross-validation and again for the total sample. Note that all of the empirical keys were subjected to triple cross-validation, which was explained above.

Final Operational Empirical Keys for 100-Item Measure The empirical keys for the 100-item measure were designed for use in the next study, Study 3, with a slight modification. Recall that three keys were generated for each factor and were used to create a triple cross-validity estimate. Study 3 required the use of a final operational key, rather than three separate empirical keys. Typically, when creating an empirically keyed biodata inventory, the final operational key is created using the entire dataset (to maximize stability of the empirical key) but is not cross-validated (since there is no holdout sample on which to cross-validate). This practice, of creating an operational empirical key using the entire sample, is recommended by Hogan (1994). Therefore, the final operational empirical keys were created using the full sample; hybridization was again used for the option-level empirical keying. After the final operational empirical and rational keys were created, the means and standard deviations for each key were obtained and subsequently used to create *T* scores (i.e., *M* = 50; *SD* = 10) for all personality scales.

Results and Discussion

300-Item Measure

Table 1 presents the triple cross-validities for the different scoring methods. Similar to Study 1, statistical tests were conducted to compare the cross-validities of the scoring methods; these results are available in the Electronic Supplementary Material. Although many of the test statistics were statistically significant, only a few demonstrated practical significance. For conscientiousness, validity was maximized when scoring was conducted at the item level using stepwise regression. This method had a cross-validity of .31 (.35 corrected for criterion unreliability), which was significantly higher (based on statistical significance and effect size) than the rational key scores (*r* = .20, .23 corrected). In addition, the cross-validity of the option scoring methods did not surpass that of the item-level stepwise regression procedure. For openness to experience and FGPA, prediction was maximized using the item-level stepwise regression procedure. The cross-validity coefficient of .19 (.22 corrected) was significantly higher than that of the rational key (*r* = .08, .09 corrected). Empirical keying also enhanced the criterion-related validity of neuroticism, extraversion, and agreeableness, as shown in Table 1.

It is notable that empirical keying enhanced the validity of neuroticism, extraversion, and agreeableness. We found that the rational keys had validities that were similar to those reported in McAbee and Oswald's (2013) recent meta-analysis which were .00, − .02, and .06, respectively. When empirical keying was used, the validities for these factors increased to .19 (.22 corrected), .20 (.23 corrected), and .16 (.18, corrected).

We also examined the effects of hybridizing the empirical key on criterion-related validity. As mentioned earlier, hybridizing involves manually reviewing the option-level empirical key and making rationally based modifications (e.g., combining adjacent response options) to provide a compromise between pure empiricism and pure rationality. In biodata research, hybridization is thought to increase the cross-validity of the empirical key by removing sample-specific fluctuations in the weights that do not make conceptual sense or that are due to a low rate of response option endorsement (Cucina et al., 2012). As shown in Table 1, the cross-validities for the keys using hybridized and non-hybridized weights were identical for conscientiousness (.26 vs. .26), neuroticism (.19 vs. .19) and extraversion (.20 vs. .20) and nearly identical for agreeableness (.13 vs. .11). Overall, these results are consistent with those of Cucina et al. (2012) who found that hybridization did not lead to increases in criterion-related validity over pure empirical keys for a biodata inventory. That said, hybridization did increase the cross-validity of the option-level empirical key slightly for openness to experience (.17 vs. .14).

Table 1 also presents reliability estimates for each of the scoring methods; reliability was measured using coefficient alpha, which is a measure of internal consistency. Empirical keying yielded similar (and in one instance slightly higher) coefficient alphas to the rational key, with one exception. The coefficient alpha for the item stepwise empirical key for conscientiousness was negative. Further examination of this key for conscientiousness indicated that it can be divided into two sets of items, one containing positively weighted items (with an alpha of .695) the other negatively weighted items (with an alpha of .601). Using the formula for the reliability of a linear composite, we computed a value of .25 for conscientiousness. Since the situation was similar for the other four factors, we used this approach to compute reliabilities for all five factors and provide the results in Table 1.[3] Notably, for all five factors and across all scoring methods, the item-level stepwise regression method also had the lowest correlations with the rational key. Thus, in some instances, empirical keying has the benefit of increasing criterion-related validity and the side effect of decreasing internal consistency and convergent validity. This finding suggests that this scoring method is more focused on the specific variance at the item level, rather than the broader factors that extend across multiple items.

In some cases, the validity coefficients in Table 1 were higher than the corresponding reliabilities. This can be possible for two reasons. First, the reliability coefficient can be an underestimate of the true reliability of the instrument, especially, if a measure of internal consistency is used for a multidimensional instrument. We believe this is the case with empirically keyed instruments and ideally test-retest reliability estimates should be obtained (but were not feasible in our

[3] As a post hoc analysis, we also applied this approach to the two negative reliability coefficients in Study 1. The reliability of the emotional stability empirical key using item stepwise regression changed sign to .09 (this value is shown in Table 1). However, the reliability of the conscientiousness empirical key using the item correlational method became more negative (the original value appears in Table 1).

study). Second, reliability is the proportion of variance in an observed score accounted for by the true score. The square root of a reliability coefficient is the reliability index, which is the correlation of the observed score with the true score. The reliability index is also the maximum possible value of the criterion-related validity coefficient.

100-Item Measure

The results for the 100-item measure are presented in the first three columns of Table 2. Overall, empirical keying increased criterion-related validity for all five factors. These results suggest that empirical keying is a viable scoring method for shorter measures of the five factors.

Implications

In terms of practical implications, undergraduate admissions decisions have traditionally relied heavily on the SAT and high school GPA. In recent years, a number of institutions became "SAT optional" and have explored the use of other admissions criteria (e.g., the ACT, writing samples). Other potential predictors of academic performance have traditionally been overlooked due to concerns of low validity and fakability (which Study 3 addresses). Here, we found that empirical keying enhances criterion-related validity, resulting in an operational validity of .35 for conscientiousness and statistically significant validities for other personality variables. Thus, colleges and universities could consider adding empirically keyed personality measures to their admissions process. That said, if empirically keyed personality measures are fakable, test coaching becomes an obstacle to the use of this approach.

Study 3

In this study, we examined the effects of faking on empirically keyed personality measures, compared to rationally keyed personality measures. This study was conducted to test hypothesis 2, which predicted that empirical keying would reduce the effects of faking. If empirically keyed personality measures were fakable, it may be of less interest to organizations as faking could reduce the quality of selected applicants, especially in settings where test coaching is prominent (e.g., undergraduate admissions). A second aim of this study was to examine the possibility that empirical keying introduces a cognitive load under faking conditions, causing personality scores to be correlated with those on a cognitive ability test. If empirical keying introduces a cognitive load, organizations might want to avoid its use due to (a) lessened incremental validity over existing cognitive tests, (b) increased chances of legal challenges due to adverse impact, and (c) decreased diversity of selected applicants due to group differences.

Method

We obtained the dataset from Vasilopoulos et al. (2006) for Study 3. Their experiment randomly assigned 327 participants to one of four between-subjects conditions using a 2 (honest vs. faking instructions) × 2 (single-stimulus vs. forced-choice personality measure) design. We removed the participants who completed the forced-choice personality measure from the dataset and focused on the 162 participants who completed the single-stimulus personality measure, of which 81 were in the honest condition and 81 were in the faking condition. We then applied the 100-item measure scoring keys from Study 2 to the single-stimulus personality measure and treated the type of key as a within-subjects independent variable. We provide brief information on the methodology below; more information can be found in Vasilopoulos et al. (2006).

Participants

Participants were 162 undergraduates enrolled in psychology courses at a mid-sized university in the eastern USA. Seventy-four percent were White, 8% were Asian/Pacific Islander, 8% were Hispanic, 3% were African American, and 7% selected "Other" for their race/national origin. Sixty-nine percent were female and 31% were male. The median age was 19 years.

Personality Measure

The empirical keys and *T* score standardization formulas that were developed in Study 2 were applied to the separate dataset for Study 3. Participants in Study 3 completed the 100-item IPIP (Goldberg, 1999) key designed to measure constructs similar to the NEO-PI-R (Costa & McCrae, 1992) Big Five factors. The instructions asked participants to use a 5-point scale ranging from 1 (very inaccurate) to 5 (very accurate) to indicate the extent to which the behavior depicted in the stem described them.

Cognitive Ability Measure

The Wonderlic Personnel Test (WPT; Wonderlic, Inc., 1999) was used to measure cognitive ability. The WPT includes 50 items ordered in terms of increasing difficulty with a time limit of 12 min.

Procedure

Data were collected in group sessions consisting of 10 to 50 participants. Participants completed a background sheet and then were administered the WPT. They were then randomly assigned to one of two experimental conditions (honest vs. applicant). Participants in the honest response condition were instructed to answer honestly, whereas the participants in the

Table 2 Criterion-related validity of 100-item IPIP, effects of faking, and correlation with cognitive ability

	(Cross-)validity			Faking			Relationship with cognitive ability					
				GLM		*t* test	Honest		Faking		Moderated multiple regression	
	r_{obs}	ρ_{ov}	*p*	*p*	*d*	*p*	$r_{\text{pers,g}}$	*p*	$r_{\text{pers,g}}$	*p*	ΔR^2	*p*
Neuroticism												
Rational key/factor scale scores	.02	(.02)	.609		− .69	< .001	− .20	.077	.02	.837	.014	.110
Empirical key item correlational scores	*.18*	(*.21*)	< .001	.003	− .21	.193	.03	.769	.11	.350	.002	.551
Empirical key item stepwise Regr. scores	*.22*	(*.24*)	< .001	< .001	.17	.271	*.23*	*.036*	.12	.288	.019	.073
Empirical key option Pt. Bis.-empirical	*.17*	(*.20*)	< .001	.013	− .42	.008	.01	.914	.07	.515	.001	.747
Extraversion												
Rational key/factor scale scores	.02	(.02)	.599		.53	.001	.03	.813	.01	.961	.001	.758
Empirical key item correlational scores	*.18*	(*.20*)	< .001	.086	.23	.148	*.24*	.028	− .06	.577	.012	.158
Empirical key item stepwise Regr. scores	*.09*	(*.10*)	.015	.010	.02	.893	*.25*	.026	− .08	.509	.015	.125
Empirical key option Pt. Bis.-empirical	*.10*	(*.12*)	.004	.074	.13	.404	.22	.053	− .08	.498	.007	.303
Openness												
Rational key/factor scale scores	.03	(.04)	.376		.64	< .001	.18	.108	< .01	.990	.009	.209
Empirical key item correlational scores	*.17*	(*.19*)	< .001	.010	.19	.219	.14	.217	.16	.160	< .001	.893
Empirical key item stepwise Regr. scores	*.20*	(*.23*)	< .001	.010	.11	.484	.12	.275	*.23*	.043	.003	.504
Empirical key option Pt. Bis.-empirical	*.16*	(*.18*)	< .001	.159	.39	.013	.03	.803	.06	.627	.001	.750
Agreeableness												
Rational key/factor scale scores	< .01	(< .01)	.903		.63	< .001	− .09	.450	.02	.843	.010	.193
Empirical key item correlational scores	*.16*	(*.18*)	< .001	.283	.40	.012	.08	.500	− .02	.846	.005	.375
Empirical key item stepwise Regr. scores	*.10*	(*.11*)	.005	.019	.15	.347	.16	.163	.07	.529	.003	.456
Empirical key option Pt. Bis.-empirical	*.10*	(*.11*)	.007	.192	.37	.019	.09	.446	− .05	.655	.009	.215
Conscientiousness												
Rational key/factor scale scores	*.17*	(*.19*)	< .001		.65	< .001	.02	.853	− .03	.794	.008	.235
Empirical key item correlational scores	*.26*	(*.29*)	< .001	.038	.56	.001	.03	.826	− .05	.689	.008	.248
Empirical key item stepwise Regr. scores	*.24*	(*.27*)	< .001	.003	.20	.196	.08	.493	.01	.934	.003	.504
Empirical key option Pt. Bis.-empirical	*.22*	(*.25*)	< .001	.652	.59	< .001	.02	.854	− .05	.655	.003	.498

The values in the GLM column are the *p* values for the interaction terms (i.e., response instructions [i.e., honest vs. faking] × type of key [i.e., empirical vs. rational]), which serve as a test for the reduction in faking. A significant interaction term and an effect size that is closer to zero than that for the rational key means that empirical keying reduced faking. Moderated multiple regression: the scores for each key were regressed onto the response instructions and cognitive ability scores in step 1. The interaction between response instructions and cognitive ability was entered in step 2. The values in the table represent the ΔR^2 and *p* values for step 2. A statistically significant result would mean that the relationship between the personality keys and cognitive ability differs for honest and faking conditions. Correlations that are statistically significant are shown in italic font

r_{obs} observed correlation, ρ_{ov}: operational validity (corrected for criterion unreliability), $r_{pers,g}$ correlation of personality and cognitive ability (*g*), *GLM* general linear model

applicant response condition were instructed to respond as if they were completing the test as part of the admissions process for a college that they really wanted to attend.

Results and Discussion

Study 3's results are presented in Table 2. As previously shown in Vasilopoulos et al. (2006), there was a significant effect for faking using the rational key. Comparing the scores from the honest and faking conditions, the absolute value of Cohen's (1992) *d* statistic was in the .50 and .60 s and all of the *t* test were significant. Hypothesis 2 (i.e., there would be no faking when empirical keying was used) was also tested using the effect sizes and significant values for a *t* test shown in Table 2. Overall, we found support for hypothesis 2 for many of the scales, as the majority (i.e., 9 out of the 15 *t* tests) of the empirically keyed personality scales revealed no statistically significant differences between the honest and applicant conditions (using *t* tests). At least one of the empirical keying methods eliminated the statistical significance of faking for each of the five factors. In many cases, empirical keying (especially the item-level stepwise regression method) reduced the effects of faking from the *d* values that were in the .50 and .60 s to *d* values that were in the teens or low .20 s. That said, empirical keying was slightly less effective in eliminating the effects of faking for conscientiousness compared to the other keys.

We also tested hypothesis 2 using a general linear model analysis that included the faking condition as a between-subjects main effect, a pairing of rational and empirical keys as a within-subjects main effect, and the interaction between the two as a test for the reduction in faking. A significant interaction would provide support for hypothesis 2 (i.e., empirical keying reduces faking), provided that the effect size was in the correct direction (which as shown in Table 2, it was). We conducted this analysis separately for each empirical keying method and we provide the *p* values for the interaction terms in Table 2. In general, we found further support for hypothesis 2 when examining the statistical significance of the interaction terms.

In Table 2, we provide the correlations between scores on each of the personality scales and cognitive ability for the honest and faking conditions. None of the rational keys had a significant correlation with cognitive ability. Furthermore, under faking conditions, only one of the empirical keys (the item-level stepwise regression scale for openness) had a statistically significant correlation with cognitive ability. Thus, empirical keying does not appear to introduce a correlation between cognitive ability and personality scale scores.

Moderated multiple regression analyses were also conducted to test the possibility that empirical keying introduces a correlation with cognitive ability. In these analyses, the rational key and the empirical key separately served as criteria, with the response instructions (i.e., honest vs. faking) and cognitive ability scores entered in step 1 and the interaction between cognitive ability and response instructions entered in step 2. If the interaction term is statistically significant, this indicates that the relationship between the personality keys and cognitive ability differs for honest and faking conditions. Overall, none of the interaction terms was statistically significant, suggesting that empirical keying does not lead to a correlation with cognitive ability.

In terms of practical implications, the reduction in faking should make the use of empirically keyed personality measures more attractive to organizations than rationally keyed measures. Faking leads to a different rank ordering of applicants since fakers rise to the top of the distribution of scores (Rosse, Stecher, Levin, & Miller, 1998) resulting in lower performance for the group of selected applicants. We found that empirical keying partially reversed this trend, which would allow organizations to measure personality without fears of substantial faking. The lack of a cognitive load suggests that organizations do not need to be concerned about potential adverse impact or diminished incremental validity over existing cognitive measures.

Study 4

An anonymous reviewer raised the possibility that empirical keying capitalizes on chance characteristics within a sample and will not cross-validate or generalize to another setting. Our use of triple cross-validation should eliminate capitalization on chance within the populations we studied. However, the reviewer suggested that this approach would not necessarily yield an empirical key that would apply to another population or setting. The reviewer also mentioned that practitioners might have difficulty implementing empirical keying if the organizations they work with have small populations. To address these concerns, we obtained a dataset with an empirically keyed personality measure developed for use in multiple organizations. This study illustrates the possibility of using a consortium of organizations to increase the sample sizes needed for empirical keying. Additionally, this study allows for an investigation of the generalizability of empirical versus rational keys across organizations.

Method

Participants

The sample consisted of incumbents in a mid-level leader job family (i.e., supervisory jobs between the first-line supervisor and executive leadership levels). A total of 2360 incumbents from 13 organizations were in the validation sample, which contained responses to a personality measure and job performance data. The incumbents were randomly divided into a keying sample of 1581 cases and a holdout sample of 779 cases.

Measure

The personality measure consisted of 84 items divided into 12 scales that covered a range of the 5 factors and facets using a 5-point Likert scale. The scales were based on job analyses and focus on specific facets of personality as opposed to overall dimensions. Thus, this inventory is an overall assessment of the job-related aspects of individuals' personalities. A rational key was created by summing the Likert-scored items. Two empirical keys were created, one using the point biserial method described above and another using the mean criterion (or average performance) approach. Under this approach, the criterion is first standardized. Next, for the first item, the participants who selected the first response option are selected and the average standardized criterion score is computed. This process is repeated for the remaining response options and then for the remaining items. The scoring weight for a particular response option is the average standardized criterion score for incumbents who selected that option.

Results and Discussion

Empirical keying increased criterion-related validity from .14 to .16 for the point biserial method and to .19 for the mean criterion method. This finding partially supports hypothesis 1; however, the main purpose of this study was to determine if empirical keying can generate scoring keys that have similar generalizability as rational keys. Therefore, we obtained the validity coefficients for each of the 13 organizations in the cross-validation sample (which consisted of 779 cases). We then computed meta-analytic (i.e., sample-weighted) criterion-related validity coefficients and determined the percent of variance in the validities accounted for by sampling error. The rational key had a criterion-related validity of .156 and 49.9% of the variance was accounted for by sampling error. The point biserial key had a validity of .169 and 47% of the variance was accounted for by sampling error. The mean criterion key had a validity of .201 and 63.7% of the variance was accounted for by sampling error. These results suggest that empirical keys can have the same (if not slightly higher in the case of the mean criterion method) generalizability of validity coefficients across different organizations. Additionally, this study provides an example of the use of a consortium of organizations for empirical keying.

Our results do not necessarily establish that empirically keyed personality measures have validity generalization since there is still substantial variance that is unaccounted for by sampling error. Other factors, such as organizational culture, differing response styles across organizations, and social requirements for different jobs (see MacLane & Cucina, 2015), could be serving as moderators. More research could be conducted on the validity generalization of empirically keyed personality measures. There is evidence from the biodata literature suggesting that empirical keying generalizes across different settings. Rothstein, Schmidt, Erwin, Owens, and Sparks (1990) reported evidence of validity generalization of empirically keyed biodata for supervisory positions across 79 organizations. Gandy et al. (1994) found evidence of validity generalization of an empirically keyed biodata across 105 occupations and 28 federal government agencies. A review by Schmidt and Rothstein (1994) concluded that empirically keyed biodata can have validity generalization and are not situationally specific, despite earlier suggestions of the contrary. Another study found only minor differences in the validity of a biodata instrument used in 14 different countries (Caputo, Cucina, & Sacco, 2010).

General Discussion

Two common criticisms of personality measures are their relatively low criterion-related validity, and the effects of faking. In order for the use of personality measures to be viable for organizations, progress needs to be made to determine how to maximize the criterion-related validity of personality measures and how to mitigate the effects of faking. Overall, we found evidence that empirical keying can increase the criterion-related validity of personality measures and reduce the effects of faking. Although empirical keying does not raise the criterion-related validity of personality measures to the level of general mental ability, it does explain additional criterion-relevant variance without increasing test administration time. In our experience, many organizations are worried (perhaps unnecessarily; Hardy, Gibson, Sloan, & Carr, 2017) about the burden that assessments place on applicants (in terms of time and effort). There are also calls for harvesting existing information on applicants to supplement or supplant assessments in selection (see Oswald's, 2014, review of the non-I-O psychology literature on big data and personnel selection). Empirical keying addresses these concerns by making better use of existing item-level information that applicants provide when responding to personality measures. Furthermore, empirical keying reduces the effects of faking without introducing a cognitive load into the predictor, which would have had implications for adverse impact and incremental validity.

Although our findings do suggest that psychologists may benefit from using empirical keying to score personality measures, we admit that a tradeoff does exist between using empirical keying and rational keying. In general, empirical keying yields scores with higher criterion-related validity, less susceptibility to faking, lower internal consistency, and less construct validity. In contrast, rational keying generates scores with higher construct validity and internal consistency, but lower criterion-related validity. Of course, empirical keying, with good reason, has never been advocated as a way to enhance construct validity; however, it does enhance criterion-related validity. In research and applied settings where the aim is to maximize criterion-related validity, empirical keying is worthy of serious consideration.

Regarding reliability, some of the empirical keys had low internal consistencies, which may be due to the inherent multidimensionality of empirical keys. Ideally, test-retest reliability coefficients would have been computed; however, these data were unavailable to us. That said, we did locate a pair of datasets from Pozzebon et al. (2013) which contained personality item responses, a criterion with some relevance to applied psychology (i.e., traffic risk), and test-retest data. We created an item stepwise regression empirical key for the conscientiousness items and found a similar pattern of weights that were observed in Study 2. The traffic risk empirical key had a triple cross-validity of .41 (compared to a rational key validity of |.23|) and a test-retest reliability of .63. This suggests that personality instruments can be empirically keyed using item stepwise regression while also being reliable. However, more research is needed on this topic.

At first glance, the finding that the rational keys from Study 1 did not have significant criterion-related validity might seem surprising. However, these were applicant data and a recent meta-analysis by Kepes and McDaniel (2015) reported very little operational data from applicants examining the criterion-related validity of conscientiousness. They only identified four studies and concluded that the "distribution is too small to reach definite conclusions regarding the robustness of the meta-analytic mean estimate" (p. 12). The lack of criterion-related validity with supervisory ratings might also seem surprising. However, supervisors often have a low opportunity to observe performance in law enforcement settings (Schmidt, 2002).

We encourage replication of our findings for other occupations (especially non-law enforcement positions) and investigation of the impact of empirical keying on applicant reactions. We see arguments for empirical keying increasing applicant reactions for honest respondents (by reducing the effects of impression management and dishonesty among fakers with whom an applicant is competing). We also see arguments for the opposite relationship: item subtlety has been linked to low face validity (e.g., Mael, 1991; Bornstein, Rossner, Hill, & Stepanian, 1994) and it seems reasonable that key subtlety could have the same type of relationship. Additionally, empirical keying could be applied to contextualized personality measures, which aim to focus the context of personality in responding to an academic or work setting. This is accomplished by instructing respondents to answer items with a school or work mindset, adding the phrase "at school" or "at work" to the end of each item, or rewriting the items to fully contextualize them. There is evidence that full contextualization enhances criterion-related validity (Holtrop, Born, de Vries, & de Vries, 2014). As suggested by an anonymous reviewer, empirical keying might have less added benefit in enhancing criterion-related validity for contextualized items as these items may be less multidimensional in nature.

Implications for Previous Conceptual Research

We proposed five conceptual rationales that predict empirical keying will maximize the criterion-related validity of personality measures. In terms of the bandwidth-fidelity (BWF) dilemma, our results are consistent with the arguments made by proponents of narrow traits as opposed to those made by proponents of broad traits. As demonstrated in Study 1, empirical keying increased the criterion-related validity of two personality scales from levels that were statistically non-significant to levels that demonstrated statistical and practical significance. In Study 2, the criterion-related validities of the five factors were higher when scoring was conducted at the item and option levels using empirical keying. Our results are consistent with the existence of nuances (i.e., the microdimensions of personality studied by Mottus et al., 2017). Harnessing information contained in personality items via empirical keying improves validity.

In terms of the distinction between latent and emergent models of personality, our results are consistent with an emergent viewpoint of personality (Ozer & Reise, 1994; Bollen & Lennox, 1991), rather than a latent viewpoint. In contrast to the latent viewpoint, the emergent viewpoint predicts that empirical keying will yield higher criterion-related validities than rational keying. Our results are also consistent with previous conjecture that soft biodata items are quite similar to personality items and previous research demonstrating that empirical keying maximizes the criterion-related validity of biodata items. Since option-level keying did not always increase criterion-related validity over and above item-level keying, our results are less consistent with the peak-point-personality response option framework (PPPROF). Although there is clearer evidence that personality item responses have a peak-point response option relationship with faking and latent traits (through the evidence provided by Carter et al., 2014; Kuncel & Tellengen, 2009; Kuncel & Borneman, 2007), the evidence is less clear when performance is used as the criterion. In terms of our multidimensional personality item framework (MPIF), our results are clearly consistent with the notion that personality items measure more than one type of personality characteristic and can vary in terms of criterion-related validity. In the "Introduction" section, we posited that reliable and valid information about an individual's personality is discarded when personality items are summed into scale scores for factors. Item and option empirical keying uses this additional information to predict performance. Thus, the finding that empirical keying increases validity is consistent with our suggestion that broadening the definition of personality beyond the five factors increases the variance in performance that can be accounted for by personality and that items have differing relationships with performance.

Regarding the latter statement, using Study 2, we provide some examples of the differing relationships of items with academic performance. In the "Introduction" section, we described how different items comprising the orderliness facet can have different types of relationships with criteria. Three of these items were used in Study 2 and the validities matched what was suggested. The orderliness item, "leave a mess in my room," did in fact lack empirical alignment to the criterion of academic performance in Study 2 as it had a near zero correlation of .03 ($p = .381$) with FGPA. Additionally, the "work hard" item had a criterion-related validity that was over twice as high ($r = .25$, $p < .001$) as the "demand quality" item ($r = .11$, $p < .002$).

Although there is strong evidence for the existence of the big five factors, it appears that there are more meaningful aspects of personality than can be revealed by summing items into scores for factors and linearly relating them to other life outcomes. Additionally, the finding that empirical keying improves the criterion-related validity of personality measures in some instances suggests that the meta-analytic validity estimates, which are based on rational keys, for personality might underestimate the relationship between personality and performance. This could be a consequence of using personality test items that were written to measure the broad range of individual differences related to personality without fully considering the relationship of the items to performance. Empirical keying provides a path for future personality test developers to hypothesize and test the relationship of proposed personality items to performance.

Practical Implications

Practitioners using personality instruments would be well served to attempt empirical keying *when criterion data are available* and especially when training or academic performance is the criterion of interest. Regarding data requirements, research in the biodata literature suggests that 200 cases will provide 90% power for statistical significance when empirical keying is used (Cucina et al., 2012). That said, empirical keying adds additional work for practitioners. However, in our experience a *purely empirical* (i.e., *non-hybridized*) key can be developed in 1–2 days once the data have been collected. Our results show that empirical keying often explained additional criterion-relevant variance without increasing testing time and often reduced the effects of faking without introducing a cognitive load. These are significant findings for practitioners. The use of empirical keying would be especially helpful for situations where a practitioner is attempting to maximize prediction of a criterion (such as in personnel selection). Additionally, the finding that empirical keying improves the validity of personality measures suggests that the meta-analytic validity estimates, which are based on rational keys, might be underestimates of the relationship between personality dimensions and performance.

We also note that some organizations (especially colleges and universities with admissions programs) are reluctant to use personality measures due to faking. Our finding that empirical keying can partially mitigate the effects of faking might increase the viability of this approach in high-stakes settings. The results of this study support the use of empirical keying as a technique for reducing the impact of faking on personality tests. An important implication of Study 3 is that, unlike other proactive attempts to reducing faking, empirical keying can be implemented without being readily apparent to applicants (just as it is for biodata instruments).

Limitations and Future Research

We alluded to some limitations and areas for future research above. Another limitation is our use of data from the USA; as organizations become more global, it is important for future researchers to replicate our findings in other cultures. Additionally, empirical keying largely did not increase the criterion-related validity for supervisory ratings of job performance, which may be due to a lack of opportunity to observe performance in Study 1. Furthermore, we were unable to examine the incremental validity of empirically keyed personality over high school GPA; this should be investigated by future researchers. Study 2 could have been bolstered by using applicant data. Study 3 could have exaggerated the prevalence of faking as applicants might not engage in faking as much as research participants who are instructed to do so. Nevertheless, our findings might present a worst-case scenario of how bad faking could be and they also allow for a comparison of honest versus faked responses. We also do not know the effects of empirical keying on face validity. One possibility is that applicant's perceptions of face validity are higher after being told that the scoring key for a personality measure depends on the empirical relationship between each response option and performance. There is also the question of the face validity of the empirical key itself when viewed by those that have access to it. Non-psychologists reviewing an empirical key may question the "fairness" of the weights for different items and options, especially in a high-stakes selection setting. As an anonymous reviewer suggested, future researchers could examine the effects of faking on the validity of empirically keyed personality measures in lab studies. Finally, direct comparisons of the latent and emergent models could be conducted.

Conclusion

Overall, we found that empirical keying enhanced the criterion-related validity of personality measures and reduced the effects of faking while maintaining low correlations with cognitive ability. Some drawbacks of empirical keying include reduced internal consistency and potential impacts on construct and face validity. We encourage researchers and practitioners to consider empirically keying personality measures in situations in which prediction is the goal.

References

Barrick, M. R., & Mount, M. K. (1991). The Big Five personality dimensions and job performance: A meta-analysis. *Personnel Psychology, 44*(1), 1–26.

Bollen, K., & Lennox, R. (1991). Conventional wisdom on measurement: A structural equation perspective. *Psychological Bulletin, 110*(2), 305–314.

Bornstein, R. F., Rossner, S. C., Hill, E. L., & Stepanian, M. L. (1994). Face validity and fakability of objective and projective measures of dependency. *Journal of Personality Assessment, 63*(1), 363–386.

Brown, S. H. (1994). Validating biodata. In G. S. Stokes, M. D. Mumford, & W. A. Owens (Eds.), *Biodata handbook: Theory, research, and use of biographical information in selection and performance prediction*. Palo Alto: CPP Books.

Caputo, P. M., Cucina, J. M., & Sacco, J. M. (2010). *Approaches to empirical keying of international biodata instruments*. Poster presented at the 25th meeting of the Society for Industrial and Organizational Psychology, Atlanta, GA.

Carter, N. T., Dalal, D. K., Boyce, A. S., O'Connell, M. S., Kung, M. C., & Delgado, K. M. (2014). Uncovering curvilinear relationships between conscientiousness and job performance: How theoretically appropriate measurement makes an empirical difference. *Journal of Applied Psychology, 99*(4), 564–586.

Christiansen, N. D., Burns, G. N., & Montgomery, G. E. (2005). Reconsidering forced-choice item formats for applicant personality assessment. *Human Performance, 18*, 267–307.

Cohen, J. (1992). A power primer. *Psychological Bulletin, 112*(1), 155–159.

Converse, P. D., Oswald, F. L., Imus, A., Hedricks, C., Roy, R., & Butera, H. (2008). Comparing personality test formats and warnings: Effects on criterion-related validity and test-taker reactions. *International Journal of Selection and Assessment, 16*(2), 155–169.

Costa, P. T., & McCrae, R. R. (1992). *Revised NEO Personality Inventory (NEO PI-R) and NEO Five-Factor Inventory (NEO-FFI) professional manual*. Odessa, FL: Psychological Assessment Resources.

Cucina, J. M., Busciglio, H. H., & Vaughn, K. (2013). Category ratings and assessments: Impact on validity, utility, veterans' preference, and managerial choice. *Applied HRM Research, 13*(1), 51-68.

Cucina, J. M., Caputo, P. M., Thibodeaux, H. F., & MacLane, C. N. (2012). Unlocking the key to biodata scoring: A comparison of empirical, rational, and hybrid approaches at different sample sizes. *Personnel Psychology, 65*, 385–428.

Cucina, J. M., Su, C., Busciglio, H. H., & Peyton, S. T. (2015). Something more than g: Meaningful Memory uniquely predicts training performance. *Intelligence, 49*, 192-206.

Cucina, J. M., Su, C., Busciglio, H. H., Thomas, P. H., & Peyton, S. T. (2015). Video-based testing: A high-fidelity job simulation that demonstrates reliability, validity, and utility. *International Journal of Selection and Assessment, 23*(3), 197-209.

Cucina, J. M., & Vasilopoulos, N. L. (2005). Nonlinear personality-performance relationships and the spurious moderating effects of traitedness. *Journal of Personality, 73*(1), 227–260.

Davis, B. W. (1997). *An integration of biographical data and personality research through Sherwood Forest empiricism: Robbing from personality to give to biodata*. Unpublished Doctoral Dissertation, Louisiana State University.

Fan, J., Gao, D., Carroll, S. A., Lopez, F. J., Tian, T. S., & Meng, H. (2012). Testing a new procedure for reducing faking on personality tests within selection contexts. *Journal of Applied Psychology, 97*, 866–880.

Festinger, L. (1957). *A theory of cognitive dissonance*. Stanford, CA: Stanford University Press.

Gandy, J. A., Dye, D. A., & MacLane, C. N. (1994). Federal government selection: The Individual Achievement Record. In G. S. Stokes, M. D. Mumford, & W. A. Owens (Eds.), *Biodata handbook: Theory, research, and use of biographical information in selection and performance prediction* (pp. 275–309). Palo Alto, CA: CPP Books.

Goldberg, L. R. (1993). The structure of personality traits: Vertical and horizontal aspects. In D. C. Funder, R. D. Parke, C. Tomlinson-Keasey, & K. Widaman (Eds.), *Studying lives through time: Personality and development*. Washington, DC: APA.

Goldberg, L. R. (1999). A broad-bandwidth, public-domain, personality inventory measuring the lower-level facets of several five-factor models. In I. Mervielde, I. J. Deary, F. De Fruyt, & F. Ostendorf (Eds.), *Personality psychology in Europe* (Vol. 7, pp. 7–28). Tilburg: Tilburg University Press.

Guion, R. M. (1965). *Personnel testing*. New York: McGraw-Hill Book Company.

Hardy III, J. H., Gibson, C., Sloan, M., & Carr, A. (2017). Are applicants more likely to quit longer assessments? Examining the effect of assessment length on applicant attrition behavior. *Journal of Applied Psychology, 102*(7), 1148–1158.

Hogan, J. R. (1994). Empirical keying of background data measures. In G. S. Stokes, M. D. Mumford, & W. A. Owens (Eds.), *Biodata handbook: Theory, research, and use of biographical information in selection and performance prediction*. Palo Alto, CA: Consulting Psychologists Press, Inc..

Holtrop, D., Born, M. P., de Vries, A., & de Vries, R. E. (2014). A matter of context: A comparison of two types of contextualized personality measures. *Personality and Individual Differences, 68*, 234–240.

Kepes, S., & McDaniel, M. A. (2015). The validity of conscientiousness is overestimated in the prediction of job performance. *PLoS One, 10*(10), e0141468.

Kluger, A. N., Reilly, R. R., & Russell, C. (1991). Faking biodata tests: Are option-keyed instruments more resistant? *Journal of Applied Psychology, 76*, 889–896.

Konig, C. J., Steiner Thommen, L. A., Wittwer, A.-M., & Kleinmann, M. (2017). Are observer ratings of applicants' personality also faked? Yes, but less than self-reports. *International Journal of Selection and Assessment, 25*, 183–192.

Kuncel, N. R., & Borneman, M. J. (2007). Toward a new method of detecting deliberately faked personality tests: The use of idiosyncratic item responses. *International Journal of Selection and Assessment, 15*(2), 220–231.

Kuncel, N. R., & Hezlett, S. A. (2010). Fact and fiction in cognitive ability testing for admissions and hiring decisions. *Current Directions in Psychological Science, 19*(6), 339–345.

Kuncel, N. R., & Tellengen, A. (2009). A conceptual and empirical re-examination of the measurement of the social desirability of items: Implications for detecting desirable response style and scale development. *Personnel Psychology, 62*, 201–228.

Le, H., Oh, I., Robbins, S. B., Ilies, R., Holland, E., & Westrick, P. (2011). Too much of a good thing: Curvilinear relationships between personality traits and job performance. *Journal of Applied Psychology, 96*(1), 113–133.

MacLane, C. N., & Cucina, J. M. (2015). Generalization of cognitive and noncognitive validities across personality-based job families. *International Journal of Selection and Assessment, 23*(4), 316–331.

Mael, F. A. (1991). A conceptual rationale for the domain and attributes of biodata items. *Personnel Psychology, 44*, 763–792.

McAbee, S. T., & Oswald, F. L. (2013). The criterion-related validity of personality measures for predicting GPA: A meta-analytic validity competition. *Psychological Assessment, 25*(2), 532–544.

Meng, X. L., Rosenthal, R., & Rubin, D. B. (1992). Comparing correlated correlation coefficients. *Psychological Bulletin, 111*(1), 172–175.

Mitchell, T. W., & Klimonski, R. J. (1982). Is it rational to be empirical? A test of methods for scoring biographical data. *Journal of Applied Psychology, 67*(4), 411–418.

Moon, H., Hollenbeck, J. R., Marinova, S., & Humphrey, S. E. (2008). Beneath the surface: Uncovering the relationship between extraversion and organizational citizenship behavior through a facet approach. *International Journal of Selection and Assessment, 16*, 143–154.

Morgeson, F. P., Campion, M. A., Dipboye, R. L., Hollenbeck, J. R., Murphy, K., & Schmitt, N. (2007). Reconsidering the use of personality tests in personnel selection contexts. *Personnel Psychology, 60*(3), 683–729.

Mottus, R., Kandler, C., Bleidorn, W., Riemann, R., & McCrae, R. R. (2017). Personality traits below facets: The consensual validity, longitudinal stability, heritability, and utility of personality nuances. *Journal of Personality and Social Psychology, 112*(3), 474–490.

Mumford, M. D., & Stokes, G. S. (1992). Developmental determinants of individual action. In M. D. Dunnette & L. M. Hough (Eds.), *Handbook of industrial and organizational psychology* (Vol. 3). Palo Alto, CA: Consulting Psychologists Press.

O'Neill, T. A., Lewis, R. J., Law, S. J., Larson, N., Hancock, S., Radan, J., Lee, N., & Carswell, J. J. (2017). Forced-choice pre-employment personality assessment: Construct validity and resistance to faking. *Personality and Individual Differences, 115*, 120–127.

Oswald, F. (2014). *Under the Hood of Big Data in Personnel Selection.* Presentation for the November 7, 2014 meeting of the Personnel Testing Council of Southern California.

Ozer, D. J., & Reise, S. P. (1994). Personality assessment. *Annual Review of Psychology, 45*, 357–388.

Pozzebon, J., Damian, R. I., Hill, P. L., Lin, Y., Lapham, S., & Roberts, B. W. (2013). Establishing the validity and reliability of the Project Talent Personality Inventory. *Frontiers in Psychology, 4*, 968.

Rosse, J. G., Stecher, M. D., Levin, R. A., & Miller, J. L. (1998). The impact of response distortion on preemployment personality testing and hiring decisions. *Journal of Applied Psychology, 83*, 634–644.

Rothstein, H. R., Schmidt, F. L., Erwin, F. W., Owens, W. A., & Sparks, C. P. (1990). Biographical data in employment selection: Can validities be generalizable? *Journal of Applied Psychology, 75*(2), 175–184.

Schmidt, F. L. (2002). The role of general cognitive ability and job performance: Why there cannot be a debate. *Human Performance, 15*(1/2), 187–210.

Schmidt, F. L., & Hunter, J. E. (1998). The validity and utility of selection methods in personnel psychology: Practical and theoretical implications of 85 years of research findings. *Psychological Bulletin, 124*(2), 262–274.

Schmidt, F. L., & Rothstein, H. R. (1994). Application of validity generalization to biodata scales in employment selection. In G. S. Stokes, M. D. Mumford, & W. A. Owens (Eds.), *Biodata handbook: Theory, research, and use of biographical information in selection and performance prediction* (pp. 237–260). Palo Alto: CA CPP Books.

Shaffer, J. A., & Postlethwaite, B. E. (2012). A matter of context: A meta-analytic investigation of the relative validity of contextualized and noncontextualized personality measures. *Personnel Psychology, 65*(3), 445–494.

Soares, J. A. (Ed.). (2012). *SAT wars: The case for test-optional college admissions*. New York, NY: Teachers College Press.

Stark, S., Chernyshenko, O. S., Drasgow, F., White, L. A., Heffner, T., Nye, C. D., & Farmer, W. L. (2014). From ABLE to TAPAS: A new generation of personality tests to support military selection and classification decisions. *Military Psychology, 26*, 153–164.

Stricker, L. J., & Rock, D. A. (1998). Assessing leadership potential with a biographical measure of personality traits. *International Journal of Selection and Assessment, 6*(3), 164–184.

Tett, R. P., Freund, K. A., Christiansen, N. D., Fox, K. E., & Coaster, J. (2012). Faking on self-report emotional intelligence and personality tests: Effects of faking opportunity, cognitive ability, and job type. *Personality and Individual Differences, 52*, 195–201.

van Geert, E., Orhon, A., Cioca, J. A., Mamede, R., Golusin, S., Hubena, B., & Morillo, D. (2016). Study protocol on intentional distortion in personality assessment: Relationship with test format, culture, and cognitive ability. *Frontiers in Psychology, 7*, 933.

Vasilopoulos, N. L., & Cucina, J. M. (2006). Faking on non-cognitive measures: The interaction of cognitive ability and test characteristics. In R. Griffith (Ed.), *A closer examination of applicant faking behavior*. Greenwich, CT: Information Age Publishing, Inc..

Vasilopoulos, N. L., Cucina, J. M., Dyomina, N. V., Morewitz, C. L., & Reilly, R. R. (2006). Forced-choice personality tests: A measure of personality and cognitive ability? *Human Performance, 19*(3), 175–199.

Vasilopoulos, N. L., Cucina, J. M., & Hunter, A. E. (2007). Personality and training proficiency: Issues of validity, curvilinearity, and bandwidth-fidelity. *Journal of Organizational and Occupational Psychology, 80*, 109–131.

Vasilopoulos, N. L., Cucina, J. M., & McElreath, J. M. (2005). Do warnings of response verification moderate the relationship between personality and cognitive ability? *Journal of Applied Psychology, 90*(2), 306–322.

Wonderlic, Inc. (1999). *Wonderlic personnel test and scholastic level exam*. Libertyville, IL: Wonderlic Personnel Test, Inc..