Journal of Applied Testing Technology, Vol 22(2), 1-20, 2021

Assessing Competencies in the Workplace: Developing a Modular Measure with Universal Applicability

Yin Lin[1]*, Alexandra Livesey[1] and Kathy Tuzinski[2]

[1]SHL, Thames Ditton, United Kingdom; Yin.Lin@shl.com
[2]Human Measures, Greater Minneapolis - St. Paul Area, United States

Abstract

Competencies have been a common tool for talent management operations for decades. In an attempt to standardize and streamline competency modelling and assessment in the varied and evolving workplace, this paper presents a measurement architecture consisting of a modular but comprehensive construct framework and a technology-enabled assessment tailoring workflow utilizing automated test assembly and quality-check technology. The resulting tailored competency assessments demonstrated promising construct and criterion validity in a series of empirical studies. While research is still ongoing, we review the initial findings and reflect on the implications and future potentials arising from this research.

Keywords: Automated Test Assembly, Assessment Tailoring, Competency Modelling, Competency Assessment, Personnel Selection, Measurement System

1. Introduction

Competency-based Assessments in the Changing Workplace

Following their introduction in 1973 (McClelland, 1973), competencies have become a common tool for talent management. A survey of over 1,400 human resource (HR) representatives found that a majority of Global Fortune 500 companies rely on competencies for describing, measuring, and rewarding the job performance of their employees (Kantrowitz, 2014). A more recent survey of global HR professionals revealed that 30% of the surveyed organizations considered the creation and implementation of competency models to be a top talent priority, with the emphasis being more prominent in emerging economies where competency modelling was growing more rapidly (Kantrowitz, Tuzinski & Raines, 2018). The adoption of competencies in talent management operations allowed complex job performance to be decoded into objective and transparent goal-related behaviors, enabling a criterion-centric approach to modelling, measuring and communicating job requirements and expectations, ultimately leading to better business outcomes (Bartram, 2005). As a result, competencies have utility throughout the employee lifecycle: from the description of job requirements in a job analysis, to the comparison of candidates in personnel selection, to the structuring of a performance review process, to the development of current employees and the planning of future successors (e.g., Kasemsap, 2016; Mansfield, 1996; Spencer & Spencer, 1993). In order to support the use of competencies in talent management operations, competency-based assessments are widely used. About a third (32%) of surveyed HR professionals reported that their organizations were using assessments to measure competency models, while a further 30% indicated that they were planning to use assessments for the same purpose in the near future (Kantrowitz, Tuzinski

*Author for correspondence

& Raines, 2018). Clearly, competency modelling and competency assessments have an important role to play in the workplace today.

The current practices of competency modelling and competency assessment vary in the specificity of the constructs selected for the competency model. At the most generic level, there are competency models with broad constructs that are relevant to any job role, e.g., the Great Eight model of competencies (Bartram, Robertson, & Callinan, 2002; Kurz & Bartram, 2002). Such generic competency models and the accompanying assessments have wide applicability across situations and settings, and thus are valuable for applications involving multiple job roles – for example, in organizational-wide appraisals or cross-cultural research. Competency assessments for such widely-applicable constructs therefore tend to have more opportunities for empirical data collection and careful content revisions over a good number of operational years, thus likely leading to more robust and transferable scientific validation evidence for the measures, as well as a larger variety of language versions, norms and benchmarks. However, when attempting to capture and contrast the specific requirements of different job roles, broad and generic competency constructs tend to struggle, capturing only the obvious differences but providing no fine differentiations between roles. As a consequence, measurements for such broad and generic constructs tend to lack customizability for job-specific applications. On the other end of the spectrum, at the most specific level, competency models can be constructed solely for a given application targeting a particular job role within a single organization, allowing the intricacies of the role to be captured in great detail. Custom assessments based on such a tailor-made competency model have the greatest construct flexibility, context specificity and face validity. But such assessments don't immediately come with validity evidence or norms, and also lack transferability across settings and may have limited reusability if the target roles evolve over time.

The struggle to find a balance between these two extremes is known as the bandwidth-fidelity dilemma (BFD) of psychological testing (Cronbach & Gleser, 1965). Practitioners often have to weigh up the pros and cons of using broad and generic measures versus using narrow and specific ones, and there is evidence to support both sides of the argument. For example, Ones and Viswesvaran (1996) found the use of broad measures leads to better predictions. On the contrary, other authors (e.g., Driskell, Hogan, Salas & Hoskin, 1994; Hogan & Roberts, 1996) have found narrow and specific constructs to capture more situation and trait-specific variance.

In practice, many competency assessments target a common job function or industry (e.g., Bish, Newton, Browning, O'Connor & Anibaldi, 2014; Le Bon, 2015; Skorková, 2016; Testa & Sipe, 2012), thereby striking a balance in the middle of the BFD spectrum – allowing some role specificity in the model while also maintaining cross-organizational generalizability. While this approach allows for industry-level specificity, it doesn't take into account organizational-level context and variations. Organizations seldom find competency models and assessments that are completely aligned to their needs, and the adoption of a one-size-fits-all approach will lead to redundancies (i.e., spending time on measuring unimportant constructs) as well as deficits (i.e., missing out on important constructs). Some organizations may thus wish to adopt a fully customized model. However, the complexity and cost (in human, time, financial and other resources) of such an approach can be prohibitive. A fully customized competency model requires a dedicated team of psychologists to design and determine the constructs, and the accompanying assessment requires weeks if not months of development from assessment content creation to pre-testing and validation by psychometricians. While such an investment may be worthwhile for repeated use in large organizations, it ceases to be economically viable for smaller organizations. Therefore, most organizations are still limited to off-the-shelf competency models and assessments, even in situations where they would prefer a solution tailored specifically to their needs. Another practical consideration is that an organization may wish to compare its different functions. If different competency models and assessments were used to cover different functions within the same organization, such a comparison becomes operationally complex or even infeasible.

While trade-offs seem unavoidable on the BFD spectrum when taking a traditional approach to competency modelling and competency assessment, the customizability-generalizability trade-off and the inhibiting cost of customization are not insolvable problems. This paper presents the key ideas underlying a measurement architecture for systematic and automated production of tailored competency models

and assessments. First, we describe a comprehensive competency framework with generic building blocks that addresses the customizability-generalizability trade-off of competency constructs. Second, we present a streamlined assessment tailoring workflow featuring automated assessment assembly and quality-check technology to drastically reduce the cost of assessment customization. Third, we present initial empirical studies exploring the construct and criterion-related validities of assessments created using this new assessment architecture. Finally, we review current findings and remaining challenges, and discuss the utilities, implications, and future possibilities arising from this research.

2. The Universal Competency Framework

A competency model is typically a list of competency constructs that cover important behaviors for some purpose or context. When a competency model has a more defined psychometric structure and hierarchy of how the competencies relate to each other and to external constructs, it can be considered a competency *framework* (Bartram, 2004). In the design of our measurement architecture for tailored competency assessments, we adopted the Universal Competency Framework (UCF; Bartram, 2005; SHL, 2003, 2019b) and associated underlying item contents (SHL, 2019a).

The latest iteration of the UCF (SHL, 2019b) consists of 96 competency components, which are nested within 20 dimensions, which are again nested within 8 factors. The broad factors and dimensions are generic enough for big-picture research comparing countries, industries and organizations within the same framework (e.g., Bartram, 2005). The measurement architecture described in this paper, however, utilizes the 96 competency components. Competency components are narrow building blocks which, when mixed in different combinations and proportions, provide enough specificity and differentiation to describe the unique competency requirements for a large variety of settings and contexts in the workplace, all under the same common competency framework (e.g., Mansfield, 1996; McLagan, 1988).

The UCF components and associated underlying item contents are suitable building blocks for competency modelling and measurement for a number of reasons (SHL, 2019a). First, the components are based on **context-free** behavioral expressions that are objective, observable and goal-related, making them universally meaningful regardless of settings. Second, the components are **narrow and distinct**, each addressing one and only one unique behavior, enabling precise and flexible competency modelling as well as psychometrically-robust competency measurement. Finally, the collection of competency components provides **comprehensive** coverage of the large variety of important behavioral requirements across most settings. The UCF can thus be used to profile the behavioral job requirements in very different settings. For example, a requirement to be "good at critical evaluation" may be interpreted differently depending on the job role – for a researcher it may be a combination of "Gathers Information", "Analyzes Information", "Critically Evaluates" and "Makes Rational Judgements", but for an airport security officer it may instead be a combination of "Understands Others", "Complies with Rules and Regulations", "Shows Safety Awareness", and "Critically Evaluates". The modular but universal UCF components thus provide the ingredients for addressing the customizability-generalizability trade-off. The next challenge, then, is to combine the modular assessment content in order to meet standards for reliability and validity in a customized assessment.

3. Assessment Tailoring Workflow

While it was common practice to develop a single, authoritative assessment, and then mold the application and usage around it, this traditional assessment-centric approach is gradually losing its appeal compared to an application-centric approach. Indeed, assessment users and test regulations are placing increasing demands on the direct relevance of assessments to the application scenario – the assessment must be face valid for its use, the assessment content must be candidate-friendly, the scores must have been shown to predict key outcomes, and the reporting of results must reflect the needs of the assessment scenario. Nowadays, the assessment needs to mold around the application scenario.

The traditional customized assessment development process largely depends on human experts. In order to make this process more repeatable and scalable, we enhanced this human-centric process with technology-

enabled automated test assembly and quality-check techniques. A streamlined human-system integrated workflow resulted, enabling systematic and efficient assessment tailoring to suit a large variety of application scenarios. This assessment tailoring workflow starts with competency modelling, with the assessment users conducting a job analysis using the UCF to structure and express the job requirements as a combination of competency components. An assessment design process then followed, using Item Response Theory (IRT) based automated test assembly technology to create and optimize a tailored competency assessment from the modular UCF item bank. Then, the assessment system automatically conducts a response simulation study to estimate the psychometric quality of the tailored assessment. Finally, the resulting assessment content and psychometric properties are reviewed by the user, allowing the user to decide whether to accept the assessment as-is, or refine its design further. This section illustrates the logical design of each step of the workflow through an empirical study. While some of the assessment requirements and features in the illustrative example are unique, the process is general and can be re-applied to different usage scenarios.

3.1 Competency Modelling

A company in the telecommunication industry was hiring for Social Media Specialists (SMS) – agents that provide support and services through interactions with customers on social media platforms. This role only came about following the rise of social media in the digital age. As such, there was limited understanding of the job requirements, and no tailored assessments designed for hiring into this role.

A job analysis was conducted to decode and express the SMS job requirements in terms of UCF competencies. Trained I/O psychologists gathered and reviewed available job information, including the current job description, selection processes, and performance metrics, then conceptualized the requirements as underlying UCF behavioral components. The UCF not only provided the structure for this conceptual mapping, but also formed the basis of a Job Analysis Questionnaire (JAQ). The JAQ was then administered to a small sample (N=30) of high-performing incumbents in the SMS role, to confirm and quantify the importance of each of the identified competency behaviors for performing their job effectively. Following the JAQ analysis, a profile of 20 UCF components were selected to represent the competency requirements for the SMS role (Table 1). Up until this point, the workflow largely resembled existing processes, with the only exception being the adoption of the UCF as the construct framework, which provided the necessary link to modular assessment content in the next stage of the assessment tailoring workflow.

3.2 Assessment Design

The UCF item bank (SHL, 2019a) formed the content basis for assessment tailoring. First, the item bank was filtered so that only items from the 20 essential competency components were considered. These items were further filtered based on the expected educational level of the job applicants – items with too high a Flesch-Kincaid reading grade (Kincaid, Fishburne, Rogers & Chissom, 1975) for the typical job applicant were excluded. Then, using the filtered item bank as the final content pool, an IRT-based automated test assembly (ATA) algorithm was employed to construct a tailored assessment.

This ATA algorithm constructs questions where three statements from different competencies are presented simultaneously and ranked by the respondent (Figure 1). This Multidimensional Forced-Choice (MFC) response format is chosen due to its greater resistance to faking and impression management, making it more robust than traditional Likert-like rating scales especially in personnel selections and other high stakes situations (e.g., Cao, 2016; Christiansen, Burns & Montgomery, 2005; Jackson, Wroblewski & Ashton, 2000; Martin, Bowen & Hunt, 2002).

Please select one statement that is most true or typical of you, and another statement that is least like you:	Most	Least
I remain polite even with people I do not get on well with		
I occasionally break promises I have made		
I often volunteer to be the one accountable for seeing issues through to their resolution		

Figure 1. Example multidimensional forced-choice question.

With such a question format, a specialized IRT model is utilized to model the responses. The Thurstonian IRT model is specially designed for modelling ipsative forced-choice data in order to recover normative trait scores (Brown & Maydeu-Olivares, 2011, 2013). In this model, ranking responses to a MFC block of three items, {i, j, k}, is decoded into three dichotomous pairwise comparisons, {i, j}, {i, k} and {j, k}. Then, the response probability for each pairwise comparison is modelled by the Thurstonian IRT Item Characteristic Function (Brown & Maydeu-Olivares, 2011):

$$Pr\left(y_{\{i,k\}}=1\mid\eta_a,\eta_b\right)=\Phi\left(\frac{\mu_i-\mu_k+\lambda_i\eta_a-\lambda_k\eta_b}{\sqrt{\psi_i^2+\psi_k^2}}\right) \quad (1)$$

In this expression, $y_{\{i,k\}}$ denotes the dichotomous response to pairwise comparison *{i,k}* and takes value 1 if item *i* is preferred and 0 if item *k* is preferred; η_a and η_b denote the person's standing on the traits measured by items *i* and *k* respectively; μ_i, λ_i and ψ^2_i denote the item mean, loading, and uniqueness IRT parameters for item *i* (and likewise for item *k*); and Φ denotes the cumulative standard normal distribution function. Finally, the information contributions of each pairwise comparison to the two associated traits are modelled by the Item Information Functions (Brown & Maydeu-Olivares, 2011):

$$I^a_{\{i,k\}}(\eta_a,\eta_b)=\frac{\left[\beta_i-\beta_k corr(\eta_a,\eta_b)\right]^2\left[\varphi\left(\alpha_{\{i,k\}}+\beta_i\eta_a-\beta_k\eta_b\right)\right]^2}{Pr\left(y_{\{i,k\}}=1\mid\eta_a,\eta_b\right)\left[1-Pr\left(y_{\{i,k\}}=1\mid\eta_a,\eta_b\right)\right]} \quad (2)$$

$$I^b_{\{i,k\}}(\eta_a,\eta_b)=\frac{\left[\beta_i corr(\eta_a,\eta_b)-\beta_k\right]^2\left[\varphi\left(\alpha_{\{i,k\}}+\beta_i\eta_a-\beta_k\eta_b\right)\right]^2}{Pr\left(y_{\{i,k\}}=1\mid\eta_a,\eta_b\right)\left[1-Pr\left(y_{\{i,k\}}=1\mid\eta_a,\eta_b\right)\right]} \quad (3)$$

In these expressions, $\alpha_{\{i,k\}}=(\mu_i-\mu_k)/\sqrt{\psi_i^2+\psi_k^2}$, $\beta_i=\lambda_i/\sqrt{\psi_i^2+\psi_k^2}$, $\beta_k=\lambda_k/\sqrt{\psi_i^2+\psi_k^2}$, and φ denotes the standard normal density function. The interested reader is referred to Brown and Maydeu-Olivares (2011, 2013) for a more detailed account of the Thurstonian IRT model.

The Thurstonian IRT model quantifies the measurement information gain from any potential MFC block. Utilizing this information, the ATA algorithm takes the filtered item bank, considers possible combinations of three items into MFC triplets, and selects triplets that maximize information gain as indicated by the Thurstonian IRT Item Information Functions, while also taking into account content rules to ensure coverage of all selected competencies, and to balance social desirability of items within the same triplet to enhance fake-resistance. The ATA algorithm automatically builds triplets one by one until the desired assessment length is reached. For the SMS role, the resulting assessment contained 84 triplets measuring all 20 identified competencies (typically taking 20-30 minutes to complete). Responses to the resulting assessment can then be scored using the Thurstonian IRT model.

This stage of the workflow presents the greatest deviation from existing processes, which usually require a psychometrician to manually put the competency assessment together. With the help of an ATA algorithm, once the assessment requirements are determined, the assessment assembly process is automated and only takes a few seconds or minutes (depending on the complexity of the request). Through capturing job requirements as different combinations of UCF competencies, the ATA algorithm can create tailored competency assessments for different roles. This unparalleled flexibility represents major advancement beyond the typical ATA algorithm, which is usually fine-tuned towards creating a very specific type of assessment for a pre-defined assessment scenario, with many of the assessment design decisions (e.g., item pool, content coverage and design rules, test length) optimized for that scenario and hard-coded into the system. Our ATA algorithm does not hard-code assessment design decisions. Rather, it presents a menu of design options and allows the user to freely define what is needed – from which UCF competency to include in the assessment, to the reading level of the item content, to how long the assessment needs to be, and all the way to highly technical settings such as information optimization rules. And in order to reduce the technical complexity of the ATA algorithm for non-technical users, only the project-specific design decisions are mandatory (i.e., competencies to measure, job level, reading grade, assessment length). All other design options have pre-populated recommended defaults (some of which auto-adjust according to user-entered project-specific design decisions). This way, a non-technical user can access project-specific assessment design options and create an assessment independently, only involving

psychometric experts for trouble-shooting, enhancement and maintenance. This ATA algorithm thus enables flexible and efficient assessment tailoring on a large scale with limited dependency on the availability of psychometric experts.

3.3 Assessment Quality Check

Despite the flexibility provided by the many assessment design options in the ATA algorithm, not all combinations of settings will result in an assessment that is psychometrically-sound. Unrealistic inputs may be entered (e.g., attempting to measure 20 competencies with five triplets), or there may be a lack of appropriate item content (i.e., the available items for the selected competencies are insufficient for making a good assessment in the MFC format that satisfies all content rules). Unfortunately, given the numerous assessment design possibilities, and the psychometric complexities associated with within-question multidimensionality of the MFC response format in assessment creation and scoring, it is unrealistic to completely pre-define what would work and what would fail. However, unreliable assessments must be prevented, and it is essential to check the psychometric quality of any newly-created competency assessments.

A newly-created assessment doesn't have any empirical response data yet. The psychometric checks therefore rely on response simulation studies that make use of population characteristics and item properties, as established in the form of multivariate-normal score distribution statistics from past trial samples and item IRT parameters derived from past administrations of the items in calibration trials. Based on the established population characteristics and item properties, a sample of responses to the newly-created assessment is simulated according to Thurstonian IRT, and the resulting response patterns are then scored by Thurstonian IRT with a maximum a posterior estimator. The reliability of the score for each competency, as measured by the newly-assembled assessment, can then be calculated as the square of the correlation between true and estimated scores on that scale, giving an indication of the accuracy of the newly-assembled assessment in recovering candidates' true score profiles.

The response simulation study runs automatically and immediately following the assembly of an assessment, requiring no additional user input. A user can then review the assembled assessment's questions and psychometric properties, and decide whether the tailored assessment is adequate for the intended use. If the user is not satisfied with any aspect of the assessment, it can be regenerated with adjusted design settings. For example, if the reading level of the assessment was deemed too high, the item reading level threshold may be adjusted; if the measurement properties of the assessment were unsatisfactory, the assessment may be lengthened to boost reliability. With automated assessment assembly and simulation checks, the time needed for assessment design and refinement is greatly reduced and the reliance on experts is minimized (although they are still consulted). The automation of assessment quality checks thus further streamlines the assessment tailoring workflow, making the assessment tailoring process more repeatable and scalable.

For the example study, simulation showed that the tailored assessment's 20 competency scores had reliabilities ranging from .72 to .81 (mean = .77). The assessment and the reliability estimates were reviewed by an I/O psychologist, and it was considered adequate for use. This decision completed the assessment tailoring workflow, and the resulting assessment was released for data collection in order to gather empirical validation evidence.

As part of a concurrent validation study, the tailored SMS assessment was administered to a sample of 203 incumbents. The participating employees were roughly balanced in gender (37% male, 46% female, 16% not disclosed). The majority (79%) of the employees were less than 40 years old (8% aged 40 or above, 13% not disclosed). An overall score was created for each employee by averaging their assessment scores across all 20 competencies. At the same time, the employees' managers provided information in a job performance rating (JPR) survey, indicating the employee's performance on each of the 20 selected competencies (the competency definitions and key behavioral indicators were provided) on a five-point rating scale ("below average", "average", "above average", "well above average", "one of the best"). A manager could also choose to select an "unable to rate" response if they felt they didn't have sufficient information to rate an employee on a particular competency, so not all 20 competencies were rated for all employees. If an

Table 1. Observed intercorrelations between SMS assessment scores (above diagonal) and between performance ratings (below diagonal)

Competency		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
1	Accepts Others		.30	.23	.11	.17	.42	.57	.31	.14	.16	.27	.40	.27	.47	.18	.26	.26	.16	.03	.21
2	Acts Ethically	.54		.18	.09	.40	.07	.19	.54	.36	.28	.33	.46	.42	.28	.17	.27	.15	.30	.44	.14
3	Adapts to Change	.62	.37		.10	−.01	.26	.01	.13	.22	.68	.12	.15	.22	.31	.25	.67	.24	.22	.06	.08
4	Attends Work Reliably	.38	.40	.32		.35	.26	.06	.20	.10	.10	.04	.09	.07	.19	.33	.09	.60	.19	.22	.16
5	Complies with Rules and Regulations	.56	.64	.54	.51		.23	.19	.26	.12	.06	.05	.27	.29	.25	.25	.11	.25	.13	.28	.06
6	Controls Emotions	.56	.36	.48	.33	.49		.10	.29	.17	.15	−.11	.13	.03	.77	.21	.53	.22	.12	.05	.19
7	Demonstrates Empathy	.46	.26	.45	.21	.40	.30		.17	.01	−.01	.52	.37	.28	.25	.04	.00	.07	−.05	.05	.19
8	Earns Trust	.45	.56	.43	.32	.58	.48	.37		.19	.15	.14	.33	.24	.30	.21	.29	.25	.24	.19	.06
9	Gathers Information	.37	.35	.49	.39	.43	.40	.27	.54		.40	.32	.27	.15	.24	.21	.30	.21	.24	.60	.15
10	Learns Quickly	.41	.36	.58	.30	.41	.33	.26	.50	.67		.24	.05	.16	.12	.19	.52	.29	.36	.30	.28
11	Offers Help	.40	.34	.43	.31	.28	.31	.34	.40	.43	.49		.32	.30	.02	−.05	.03	.03	.00	.22	.11
12	Puts Customers First	.47	.41	.51	.29	.49	.34	.44	.60	.51	.55	.44		.61	.43	.33	.25	.11	.24	.40	.11
13	Shows Commitment to the Organization	.49	.51	.49	.37	.52	.36	.41	.39	.43	.46	.57	.46		.19	.25	.20	.11	.16	.23	.00
14	Shows Courtesy	.60	.45	.40	.24	.46	.47	.48	.45	.31	.31	.24	.38	.43		.33	.59	.18	.16	.17	.20
15	Stays Focused	.35	.47	.35	.34	.42	.35	.13	.36	.45	.46	.29	.38	.25	.35		.34	.40	.43	.28	.25
16	Thrives Under Pressure	.39	.32	.49	.26	.36	.52	.20	.43	.54	.55	.43	.41	.43	.29	.43		.28	.38	.20	.19
17	Uses Time Efficiently	.30	.38	.50	.38	.40	.29	.25	.47	.53	.55	.46	.45	.41	.21	.52	.47		.32	.22	.12
18	Works Autonomously	.30	.37	.42	.30	.30	.36	.09	.44	.54	.64	.39	.45	.34	.26	.46	.51	.56		.35	.27
19	Works to High Quality Standards	.39	.51	.57	.30	.54	.34	.33	.53	.57	.57	.36	.65	.51	.39	.49	.43	.57	.56		.30
20	Writes with Clarity	.20	.37	.40	.25	.33	.30	.14	.46	.61	.57	.45	.54	.41	.17	.47	.45	.53	.61	.62	

Table 2. Observed correlations between SMS assessment scores (rows) and JPR ratings (columns)

Competency		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21
1	Accepts Others	**.07**	.20	.03	.06	.08	−.07	−.02	−.13	−.03	.03	.02	.10	−.02	.08	.05	−.09	−.05	.01	.04	−.07	**.06**
2	Acts Ethically	.02	**.11**	.02	.14	.05	.01	.02	.14	.15	.16	.17	.24	.12	.03	.07	−.04	.12	.16	.17	.08	**.16**
3	Adapts to Change	.10	.02	**.24**	.02	.06	.13	.09	−.06	.11	.16	.12	.08	.12	−.04	.07	.16	.07	.07	.05	.02	**.15**
4	Attends Work Reliably	.17	.11	.11	**.24**	.11	.07	.01	.02	.19	.12	.08	.03	.01	.09	.08	.08	.16	.04	−.04	−.04	**.14**
5	Complies with Rules and Regulations	.06	.04	.01	.01	**.03**	.02	.10	.14	−.05	−.06	.00	.06	−.02	.02	−.19	−.19	−.01	−.11	−.04	−.08	**.00**
6	Controls Emotions	.23	.09	.10	.02	.09	**.21**	.01	.02	−.06	.02	−.06	.00	−.07	.17	.03	.13	.00	−.01	−.04	−.10	**.08**
7	Demonstrates Empathy	.06	.09	−.07	.01	−.06	−.05	**.05**	−.09	−.07	−.04	.01	.15	.03	.03	−.01	−.10	−.13	−.06	.05	−.10	−.04
8	Earns Trust	.06	.00	.02	.12	−.08	.00	.02	**.01**	.05	.10	.03	.09	−.10	.02	−.08	.01	.01	.02	−.01	−.06	**.04**
9	Gathers Information	.03	−.03	−.02	.04	.00	.04	−.06	.06	**.09**	.10	.14	.13	.03	−.03	.10	.02	−.02	.02	.05	.04	**.11**
10	Learns Quickly	.01	−.01	.12	.01	−.08	.10	−.04	.01	.07	**.19**	.09	.17	.07	−.10	.08	.10	.09	.17	.05	.08	**.07**
11	Offers Help	.01	.00	−.04	.06	−.02	−.05	−.01	.05	.13	.18	**.26**	.24	.16	−.08	.03	−.05	.01	.08	.16	.11	**.10**
12	Puts Customers First	.07	.16	.06	.00	.09	−.06	.07	.04	.02	.05	.08	**.15**	.06	.03	.04	−.07	.00	−.02	.13	−.07	**.11**
13	Shows Commitment to the Organization	−.05	−.04	−.01	−.13	−.04	−.15	.14	−.13	−.02	.00	.10	.13	**.07**	−.08	−.11	−.16	.00	−.09	−.03	−.12	**−.03**
14	Shows Courtesy	.14	.13	.08	.03	.10	.14	.02	.06	−.03	.02	−.03	.03	−.03	**.14**	.03	.08	−.03	−.01	.02	−.17	**.10**
15	Stays Focused	−.03	.01	.04	.01	.03	−.02	.02	−.03	.05	−.02	−.06	−.03	.00	.01	**.01**	.05	.02	−.07	−.07	−.04	**.04**
16	Thrives Under Pressure	.08	−.02	.20	.00	−.03	.11	.03	.02	.04	.14	.05	.09	.06	−.03	−.02	**.16**	.00	.08	.04	−.02	**.10**
17	Uses Time Efficiently	.09	.10	.09	.25	.07	.04	.00	−.02	.14	.04	.06	.07	.07	.07	.09	.11	**.18**	.01	.05	.02	**.14**
18	Works Autonomously	.06	.08	.16	.11	.00	.04	.00	.11	.14	.22	.08	.21	.11	.05	.13	.16	.16	**.29**	.09	.10	**.21**
19	Works to High Quality Standards	.10	.07	.07	.05	.07	.11	−.02	.20	.12	.09	.10	.13	.02	.08	.13	.00	.11	.12	**.14**	.10	**.16**
20	Writes with Clarity	−.01	.09	−.09	−.07	−.11	.00	−.07	−.11	−.03	.01	−.06	−.04	−.04	.00	.07	.03	−.02	.02	−.04	**.06**	**−.05**
21	Overall	.13	.12	.11	.09	.03	.07	.04	.02	.10	.15	.12	.20	.07	.05	.06	.04	.06	.08	.07	−.02	**.17**

Table 3. Summary of concurrent validation results for the tailored SMS assessment

	Validity against Matching Competency			
Competency	**N**	**Observed Correlation**	**Observed Correlation Significance**	**Estimated Operational Validity***
Accepts Others	169	.07	.39	.09
Acts Ethically	159	.11	.17	.14
Adapts to Change	180	.24	.00	.31
Attends Work Reliably	170	.24	.00	.30
Complies with Rules and Regulations	113	.03	.78	.03
Controls Emotions	169	.21	.01	.27
Demonstrates Empathy	168	.05	.53	.06
Earns Trust	135	.01	.88	.02
Gathers Information	159	.09	.25	.12
Learns Quickly	165	.19	.01	.25
Offers Help	173	.26	.00	.33
Puts Customers First	145	.15	.06	.20
Shows Commitment to the Organization	160	.07	.40	.09
Shows Courtesy	160	.14	.07	.18
Stays Focused	158	.01	.92	.01
Thrives Under Pressure	179	.16	.03	.21
Uses Time Efficiently	161	.18	.02	.24
Works Autonomously	169	.29	.00	.38
Works to High Quality Standards	184	.14	.06	.18
Writes with Clarity	146	.06	.49	.07
Overall	**152**	**.17**	**.04**	**.21**

*Operational validities were statistically corrected for criterion unreliability using a conservative reliability estimate of .60.

employee's manager rated at least 75% (i.e., 15 out of 20) of the competencies, an overall performance measure was created by averaging the performance ratings across all rated competencies. Table 1 displays the intercorrelations between competency assessment scores (above diagonal) and between JPR ratings (below diagonal) across the 20 competencies. The forced-choice assessment scores had weaker intercorrelations compared to the Likert performance ratings, as the latter were likely affected by halo effects as well as other method factors associated with Likert ratings.

The observed correlations between SMS assessment scores and JPR ratings are shown in Table 2, while Table 3 summarizes the convergent validity of SMS assessment scores against manager ratings of the same competency. The uncorrected validities ranged from .01 to .29 with an average of .13. Results were significant for nine competencies at the $p<.05$ level, and twelve competencies at the $p<.10$ level. These results helped to identify the

competencies that related most strongly to performance in the SMS role, which informed the design of a shortened assessment with fewer but validated scales, created using the same assessment tailoring workflow. Alternatively, as one reviewer rightly suggested, there is merit in using the full assessment despite some competencies not reaching statistical significance. The non-significant competencies were considered important for the role from job analysis results, and the CRV study might fail to pick up a relationship due to the competencies being less observable by the managers (e.g., "Stays Focused") or having limited variance in the incumbent sample (e.g., "Complies with Rules and Regulations"). Indeed, the overall score from all 20 competencies still had an uncorrected validity of .17 (p=.04) against the overall performance measure. Whether to proceed with a full or shortened assessment will depend on the available assessment time, and the validity evidence needed to meet legal requirements for the purpose of the assessment.

As illustrated by this example, while psychometricians are still relied upon for content update, system maintenance, general troubleshooting and advice, the workflow is very accessible to non-technical users and repeatable across applications, thus reducing the time and resources needed for customization. And because competency models and assessments make use of generic component building blocks from the UCF model, generalizability across roles is maintained at the component level. This streamlined assessment tailoring workflow combined with the UCF model therefore has the potential to resolve the customizability-generalizability trade-off, as well as producing targeted competency assessments en masse at lower cost. Anecdotally, at the beginning of the 2020 global pandemic when workers were forced to work from home, this measurement architecture enabled the swift creation of a Remote Working Questionnaire (SHL, 2020), providing timely diagnostic feedback to workers as they adjusted to the new way of working. If a traditional competency assessment development approach was adopted instead, the competency modelling would usually take at least a week to finalize, followed by weeks of assessment content writing, trialing and analysis. However, using this new measurement architecture, the competency modelling and the competency assessment design took less than a week in total.

4. Empirical Validity

A series of empirical studies were conducted in order to establish the validity of competency assessments created using the new approach. To date, most competency scales have been included in construct validation studies, with additional studies planned for the remaining scales. However, the greatest challenge comes from criterion-related validity (CRV) data collection. An attempt to validate 96 components simultaneously in a CRV study is unrealistic, thus most validation studies focused on about 10-20 relevant scales for a specific job setting. Therefore, in order to cover all UCF components, a large amount of data from many organizations spanning multiple job roles and industries is needed, and the recruiting of organizations to partake in this effort is a considerable undertaking that will likely take many years.

While the empirical trialing effort is still ongoing, here we present the initial results for the subset of scales that have sufficient data for both criterion and construct validation analysis. At the time of writing, 24 scales have been included in construct validation studies against an established instrument (N between 584 to 2,112 per scale), as well as trialed in at least five CRV studies (combined N across studies between 508 to 1,712 per scale). This paper focuses solely on these 24 scales, where the amount of available data and the variety of CRV studies give rise to stable results for in-depth discussions. It is important to note that the availability of data is the only criteria used to choose the 24 scales for presentation in this paper, so we're hopeful that the pattern of results will be repeated for the remaining scales as more studies emerge.

4.1 Construct Validity

Construct validity of the UCF assessment content was evaluated against the Occupation Personality Questionnaire (OPQ32r, with normative scoring enabled by IRT; Bartram, Brown, Fleck, Inceoglu, & Ward, 2006; SHL, 2014). Studies conducted by the publisher of the OPQ showed construct validity against other popular personality measures (Bartram, et al., 2006; SHL, 2014). The OPQ has also been shown to have a strong relationship with job performance (Robertson & Kinder, 1993; Saville, Sik, Nyfield, Hackston, & MacIver, 1996). Moreover, while the OPQ does not measure competency behaviors directly, it is conceptually linked to the UCF through a

Table 4. Correspondence between OPQ and UCF conceptual ratings and empirical correlations

Conceptual Rating Interval	Count	Mean Observed Empirical Correlation
[−4, −3)	0	
[−3, −2)	5	−.24
[−2, −1)	6	−.10
[−1, 0)	91	−.08
[0, 0]	331	−.01
(0, 1]	252	.06
(1, 2]	45	.16
(2, 3]	28	.24
(3, 4]	10	.47

series of empirically derived equations for measuring the personality aspect of competencies. Finally, the OPQ offered a comprehensive measure of personality through its 32 scales. A construct validation study was thus conducted to explore the relationships between scores from the UCF competency assessments (giving direct measures of competency behaviors) and scores from the OPQ (giving indirect measures of competency potential from personality).

An online sample of 3,415 respondents was recruited from a website offering practice for pre-employment assessments and personalized feedback for professional development purposes (there were no monetary incentives). Each participant completed a UCF competency assessment alongside the OPQ32r. To avoid fatigue, the competency assessment for each participant contained only a subset of 25 to 27 UCF components, requiring approximately 30 minutes to complete. The sample was balanced in terms of gender (50.9% male, 49.1% female), and all working ages along with a wide range of geographical regions were represented (with the majority coming from English-speaking countries). Independently, a group of I/O psychologists rated the strength and direction of the conceptual links between OPQ scales and UCF competencies on a scale of −4 (strong negative conceptual link) to 4 (strong positive conceptual link).

The correlation between the conceptual ratings and empirical correlations was .59, suggesting the pattern of empirical correlations between OPQ and UCF scores largely matched theoretical expectations. Table 4 summarizes the correspondence between conceptual ratings and empirical correlations. Table 5 reports the three OPQ scales showing the strongest empirical correlations (either positive or negative) with each UCF scale. Most of the strongest-correlated personality traits were in line with expectation. For example, the Competitive personality trait correlated strongest (r = .69) with the competency behavior of Thrives on Competition (conceptual rating = 3.8), and the Caring personality trait correlated strongest (r = .62) with the competency behavior of Demonstrates Empathy (conceptual rating = 3.8). Some observed empirical relationships were unexpected initially. For example, Independent Minded personality correlated negatively to a moderate degree (r = −.32) with the competency behavior of Shows Courtesy (conceptual rating = 0.4). However, upon reviewing the item content, we saw how this relationship could result from independent-minded individuals asserting their own thinking when disagreement arises, possibly escalating into acting in a disrespectful and impolite manner. Some competencies showed only weak correlations with personality. For example, Puts Customers First was only weakly related to personality traits (conceptual ratings from −0.2 to 1.2). The Puts Customers First competency is indicated by prioritizing meeting customer needs in one's actions, and the lack of relationship with personality traits might indicate that there are more than a single type of person who could prioritize customer needs – people with different personality profiles can have different strategies and approaches for doing so. The weak correlations for Puts Customers First therefore

Table 5. UCF competency scores and the three strongest related OPQ personality traits

UCF Component	OPQ Traits and Correlations
Acts Ethically	Rule Following (.26), Caring (.14), Emotionally Controlled (–.13)
Adapts to Change	Worrying (–.33), Socially Confident (.31), Achieving (.26)
Attends to Multiple Tasks	Vigorous (.31), Achieving (.24), Controlling (.22)
Attends Work Reliably	Rule Following (.33), Conventional (.29), Detail Conscious (.27)
Builds Rapport	Socially Confident (.69), Outgoing (.56), Persuasive (.37)
Complies with Rules and Regulations	Rule Following (.74), Conventional (.50), Variety Seeking (–.38)
Controls Emotions	Tough Minded (.34), Emotionally Controlled (.33), Outspoken (–.29)
Copes with Setbacks and Criticism	Tough Minded (.43), Worrying (–.34), Persuasive (.28)
Demonstrates Empathy	Caring (.62), Affiliative (.32), Data Rational (–.32)
Earns Trust	Rule Following (.28), Adaptable (–.16), Independent Minded (–.15)
Focuses on Self-Development	Achieving (.18), Democratic (.17), Forward Thinking (.13)
Learns Quickly	Innovative (.33), Achieving (.32), Data Rational (.29)
Offers Help	Caring (.36), Trusting (.22), Democratic (.22)
Offers Practical Solutions	Innovative (.52), Evaluative (.42), Achieving (.41)
Plans Ahead	Detail Conscious (.32), Forward Thinking (.30), Rule Following (.26)
Puts Customers First	Caring (.18), Persuasive (.18), Evaluative (–.17)
Shows Commitment to the Organisation	Achieving (.40), Forward Thinking (.37), Persuasive (.36)
Shows Courtesy	Independent Minded (–.32), Caring (.28), Tough Minded (.22)
Strives to Achieve	Achieving (.52), Forward Thinking (.28), Competitive (.28)
Takes Action	Vigorous (.57), Evaluative (–.26), Controlling (–.17)
Thrives on Competition	Competitive (.69), Achieving (.29), Modest (–.27)
Thrives Under Pressure	Worrying (–.31), Relaxed (.29), Controlling (.27)
Works Autonomously	Democratic (–.44), Independent Minded (.28), Behavioural (–.25)
Works to High Quality Standards	Detail Conscious (.44), Rule Following (.40), Conscientious (.32)

Table 6. Observed correlations between UCF competency scores and OPQ personality traits

OPQ Trait	Acts Ethically	Adapts to Change	Attends to Multiple Tasks	Attends Work Reliably	Builds Rapport	Complies with Rules and Regulations	Controls Emotions	Copes with Setbacks and Criticism	Demonstrates Empathy	Earns Trust	Focuses on Self-Development	Learns Quickly	Offers Help	Offers Practical Solutions	Plans Ahead	Puts Customers First	Shows Commitment to the Organization	Shows Courtesy	Strives to Achieve	Takes Action	Thrives on Competition	Thrives Under Pressure	Works Autonomously	Works to High Quality Standards
Persuasive	.05	.19	.13	−.10	.37	−.23	−.17	.28	−.09	−.02	−.01	.19	.04	.32	.01	.18	.36	.01	.19	−.13	.22	.20	−.20	−.13
Contro-lling	.07	.23	.22	−.15	.20	−.22	−.21	.26	−.21	−.03	−.07	.23	.12	.37	−.02	−.01	.27	−.13	.27	−.17	.16	.27	−.17	−.21
Outspoken	.02	.10	.10	−.09	.17	−.18	−.29	.18	−.11	−.07	.00	.16	.06	.21	−.07	−.04	.03	−.21	.10	−.16	.13	.10	−.10	−.12
Indepen-dent Minded	−.12	−.12	.04	−.13	−.05	−.32	−.08	−.05	−.12	−.15	−.11	.13	−.11	.14	−.17	−.14	−.12	−.32	.09	−.08	.15	.01	**.28**	−.17
Outgoing	.00	.24	.05	−.04	**.56**	−.16	−.12	.12	.21	−.04	−.05	.09	.12	.05	−.07	.16	.11	.00	.11	.04	.15	.08	−.12	−.09
Affiliative	.00	.10	−.04	−.05	.29	−.10	.00	−.07	.32	−.04	.04	−.04	.18	.00	.02	.11	.03	.12	.01	.10	−.05	.00	−.14	−.07
Socially Confident	.06	.31	.09	−.03	**.69**	−.11	.05	.20	.11	.06	.03	.11	.13	.12	−.04	.15	.20	.16	.10	−.09	.07	.24	−.09	−.04
Modest	−.03	−.09	−.12	.05	−.18	.08	.24	−.11	.08	.08	−.10	−.09	−.10	−.20	−.10	−.03	−.19	.13	−.17	.04	−.27	−.05	.10	−.01
Demo-cratic	.09	−.02	−.02	−.12	.07	−.08	−.07	.05	.18	.01	.17	−.09	.22	−.01	.11	.01	.13	**.10**	−.09	−.11	−.22	−.09	**−.44**	−.06
Caring	.14	.02	−.11	−.07	.13	.00	.01	−.11	**.62**	.15	.09	−.11	.36	−.09	−.02	.18	.02	**.28**	−.06	.05	−.24	−.01	−.19	−.09
Data Rational	.00	.12	.09	−.09	−.15	.02	−.11	.14	−.32	−.01	.06	.29	−.04	.30	.06	−.11	.10	−.06	.17	−.08	.06	.11	.10	.14
Evaluative	.06	.04	.16	−.15	−.06	−.22	−.10	.12	−.30	−.05	.02	.26	−.05	.42	−.05	−.17	.09	−.19	.13	−.26	−.05	.11	.00	−.01
Behavi-oural	.02	−.02	−.07	−.19	.14	−.32	−.05	−.07	.29	−.05	.06	−.06	.13	.06	−.12	.03	−.02	**.05**	−.10	−.17	−.13	−.07	−.25	−.23
Conven-tional	.03	**−.13**	−.06	.29	−.11	**.50**	.02	−.06	−.01	.11	.03	−.19	−.07	−.25	.26	.05	−.02	.11	−.09	.05	.00	−.07	.07	.30

Concep-tual	.04	−.03	.11	−.16	.04	−.25	.01	.09	−.09	−.11	.08	.20	.00	.28	−.11	−.07	.02	−.10	.02	−.11	−.04	.03	−.02	−.08
Innovative	.05	.17	.15	−.18	.13	−.22	−.14	.19	−.15	−.07	.05	.33	.04	.52	−.02	.00	.23	−.12	.21	−.08	.05	.17	−.07	−.16
Variety Seeking	−.02	**.04**	.09	−.24	.03	−.38	−.04	.06	.00	−.11	−.01	.11	.08	.19	−.20	−.10	−.03	−.20	.06	−.09	−.03	−.03	−.02	−.30
Adaptable	−.11	**−.03**	−.06	−.07	**.03**	−.23	−.12	−.12	.11	−.16	−.06	−.09	.01	−.08	−.10	−.02	−.05	−.07	−.11	−.12	.02	−.12	−.06	−.16
Forward Thinking	.12	.16	.15	−.06	.05	−.12	−.10	.19	−.17	−.03	.13	.16	−.03	.29	**.30**	.00	.37	−.06	.28	−.10	.09	.12	−.11	.01
Detail Conscious	.06	.07	.18	.27	−.02	.27	.01	.08	−.10	.08	.10	.08	−.01	.05	.32	.12	.17	.09	.15	.05	−.05	.11	.19	**.44**
Conscie-ntious	.08	.18	.21	**.25**	−.05	.23	−.05	.09	−.12	.12	.09	.09	−.02	.12	**.18**	.16	.22	.07	**.25**	**.09**	−.01	.17	.13	**.32**
Rule Following	**.26**	.01	−.01	**.33**	−.09	**.74**	.01	.02	.02	**.28**	.04	−.08	.02	−.13	.26	.16	.19	.19	.03	.09	.00	.06	.06	.40
Relaxed	−.04	.22	.08	−.01	.19	.06	**.14**	**.24**	−.06	.08	−.04	.05	−.03	−.03	−.06	.02	.06	.20	−.03	−.08	−.02	**.29**	.02	−.04
Worrying	−.06	−.33	−.17	.12	−.28	.08	**.00**	**−.34**	.21	−.06	−.02	−.22	−.04	−.27	.02	−.01	−.23	−.06	−.17	.17	−.14	**−.31**	.09	.08
Tough Minded	.01	.10	.06	−.08	.14	−.06	**.34**	**.43**	−.16	.08	.03	.04	.02	.08	−.16	−.04	.05	.22	−.04	−.09	−.06	.22	−.03	−.11
Optimistic	.08	.25	.09	−.12	.20	−.02	−.06	**.23**	.02	.03	.07	.04	.11	.07	.09	.08	.25	.09	.14	−.10	.02	.16	−.16	−.10
Trusting	.07	.03	−.02	−.12	.05	−.02	.07	.05	.19	.11	.02	−.09	.22	−.03	−.05	.05	.11	.16	−.08	.02	−.25	.01	−.20	−.17
Emotio-nally Controlled	−.13	−.13	−.05	.04	−.19	.04	**.33**	−.03	−.10	.00	−.12	−.08	−.12	−.14	−.12	−.12	−.15	.09	−.14	.00	−.03	.01	.15	−.05
Vigorous	−.04	.06	**.31**	.16	−.01	.11	−.06	.03	−.01	−.01	.09	.11	.03	.08	.05	.09	.12	.11	.24	**.57**	−.10	**.16**	.18	.20
Compe-titive	−.04	.16	.02	.01	.04	−.14	−.23	.09	−.23	−.10	**−.02**	.12	−.12	.18	.07	.03	.19	−.19	**.28**	−.09	**.69**	.09	.01	−.02
Achieving	.07	.26	.24	−.04	.07	−.13	−.14	.24	−.27	−.08	**.18**	.32	.03	.41	.16	.04	.40	−.08	**.52**	**.03**	.29	.25	.03	.07
Decisive	−.05	.23	.20	−.20	.06	−.20	−.18	.10	−.20	−.04	−.13	.24	−.02	.33	−.14	−.04	.03	−.13	.12	−.12	.07	.23	.04	−.27

* Bold numbers indicate theoretical conceptual mappings of at least medium strength.

did not indicate a lack of convergent construct validity, but rather indicate that the OPQ and other personality assessments may not be the most suited assessment for construct-validating this particular competency – in this case an alternative measure centered around customer service would be better suited. Table 6 reports the full correlation matrix between UCF competency scores and OPQ personality traits, where bolded numbers highlight where strong conceptual links (with magnitudes greater than 2) were expected.

The correlations between the two measures were good but never very strong. This is in line with expectations, as a very strong relationship may imply that the assessments are measuring the same constructs. The UCF assessment and OPQ are measuring conceptually linked but different constructs (i.e., direct competency measures and indirect predictions of competency potential). Overall, the empirical relationships with the OPQ provided strong construct validity evidence for the UCF assessments created using the new approach.

4.2 Criterion-related Validity

The CRV of UCF competency assessments created using the new approach was examined by establishing the relationships between self-rated assessment results against manager-rated job performance, through a series of concurrent CRV studies (including the aforementioned example study). The studies spanned a number of job roles and industries, including healthcare, travel, banking, telecommunications, finance, automotive, utilities, and retail. Employees in each study completed a UCF competency assessment tailored to reflect essential competencies for their role (as identified by UCF competency modelling in a job analysis). Employees' managers provided ratings of their observed performance on the same competencies through a JPR survey. Through utilizing the UCF as a unified language for describing job requirements and performance, the JPR survey focused on observable workplace behaviors, prompting managers to rate actual behavior rather than general perceptions of job performance, thus giving a more reliable reflection of job performance than generic job performance ratings.

Positive correlations between self-rated competency assessment scores and manager-rated job performance observations would provide CRV evidence of the new assessment content and approach. It is important to note that this setup is different from inter-rater agreement, where a very strong relationship is expected between two independent raters rating the same constructs about the same person using the same measure. In the case of our CRV studies, we have two raters rating the same conceptual constructs but differentially – the employee rated how they viewed themselves, and the manager rated actual observed behaviors. Moreover, different measures were needed for self and manager ratings, as most of the time it was impractical to expect managers to complete a JPR survey as detailed as the self-rated UCF assessment for multiple subordinates. In many studies, the JPR survey had to rely on ratings on a single item to capture the performance associated with each UCF competency, thus greatly reducing the reliability of the performance measure and its ability to differentiate between high and low performers. Given this setup, we interpret our CRV correlations against the benchmarks established by Bosco, Aguinis, Singh, Field and Pierce (2015). Bosco et al. (2015) summarized the expected correlations when predicting job performance from psychological characteristics (including competencies): correlations below .10 are considered low, correlations between .10 and .23 are considered medium, and correlations above .23 are considered high. Moreover, in order to provide estimated operational CRVs, the observed correlations were corrected for criterion unreliability using a conservative reliability estimate of .60. Salgado and Moscoso (2019) estimated ratings of behaviors and competencies (i.e., "task performance") in research settings to have reliabilities around .52. In order to avoid over-correction and thus overestimating the utility of the assessments, a slightly higher reliability estimate of .60 was adopted to arrive at more conservative operational validity estimates.

A meta-analysis (Hunter & Schmidt, 2004) was conducted in order to summarize the results across multiple CRV studies. Results for the 24 competencies with at least five studies are reported in Table 7. All observed CRV correlations were positive, ranging from .03 to .28 with an average of .14. The majority of scales (19 out of 24) had medium to strong CRV according to Bosco et al. (2015) and the remaining five scales had low CRV. Most observed correlations (20 out of 24) were significant at the $p<.05$ level. After correction for criterion unreliability, the operational validities ranged from .04

Table 7. Criterion-related validity of UCF competency scores

UCF Component	No. of studies (k)	Combined Sample Size (N)	Observed Correlation (r)	Estimated Operational Validity (ρ)
Acts Ethically	7	862	.06	.08
Adapts to Change	12	1712	.10	.13
Attends to Multiple Tasks	5	508	.22	.28
Attends Work Reliably	6	730	.13	.17
Builds Rapport	7	925	.14	.19
Complies with Rules and Regulations	5	542	.14	.18
Controls Emotions	8	1201	.20	.26
Copes with Setbacks and Criticism	7	833	.11	.14
Demonstrates Empathy	7	822	.15	.19
Earns Trust	5	570	.07	.09
Focuses on Self-Development	6	763	.11	.14
Learns Quickly	7	819	.18	.23
Offers Help	8	1022	.14	.18
Offers Practical Solutions	5	687	.12	.15
Plans Ahead	5	756	.03	.04
Puts Customers First	8	983	.07	.09
Shows Commitment to the Organization	5	649	.05	.06
Shows Courtesy	7	899	.19	.25
Strives to Achieve	8	1070	.21	.27
Takes Action	7	1090	.16	.21
Thrives on Competition	6	764	.28	.36
Thrives Under Pressure	8	1010	.15	.20
Works Autonomously	9	1335	.13	.17
Works to High Quality Standards	9	1045	.18	.24

*Operational validities were statistically corrected for criterion unreliability using a conservative reliability estimate of .60.

to .36, with an average of .18. Thus, the meta-analytic results provide strong support for the CRV of the new competency assessment content and approach.

5. Discussion

This paper described the development of a novel way of addressing the bandwidth-fidelity dilemma. The balance between customizability and generalizability is navigated though the adoption of a modular but generic UCF model and associated item bank, and the power of this approach is illustrated through a streamlined and technology-enabled workflow for competency modelling and assessment tailoring. Initial empirical data for a subset of scales were analyzed, showing promising construct and criterion validity evidence. Findings from this study provide confidence in the quality and utility of tailored competency assessments created using the new methodology.

It is acknowledged that there are still many unanswered questions and operational challenges around this new methodology. Further data collection and analysis are needed to cover all competency components of the UCF across multiple job roles, levels, industries, regions, cultures, languages, and other important operational variables. Further item content development is needed to broaden and deepen the item bank, in order enhance the assessment tailoring power and measurement accuracy. Further system fine-tuning and enhancement is needed to ensure the stability and optimality of future assessment creations, especially as the usage scenarios expand. Further field applications need to quantify any operational requirements and risks, for instance those around assessment certification and legal defensibility when using a computer-generated assessment (i.e., current test review guidelines tend to focus on a single assessment design, making a measurement architecture that can create an almost unlimited number of tailored assessments challenging to certify, thus creating legal risks for certain geographical regions).

Despite the remaining challenges, this research demonstrated empirically that automated creation of tailored competency assessments utilizing a modular framework of behavioral competencies can be both practically viable and operationally valid. This capability has extensive potential in field applications. For a specific job role, the modular UCF content provides sufficient granularity for very detailed competency modelling and assessment, as highlighted in this paper. For an organization, the context-independent nature of UCF content allowed multiple job functions to be described under the same competency framework, removing the operational complexity of function-specific models, while also improving transparency and transferability across functions for career pathing and succession planning. For a common job function across multiple organizations, common competency components can be identified to create a function-specific competency profile and measure, giving a quick off-the-shelf solution for talent operations that also provides a common basis for industry benchmarking. For general research across functions, organizations, and even countries, the modular UCF content enabled aggregation of data from different tailored assessments for general description, comparison, and more advanced statistical modelling with other variables. Apart from these more traditional and stable application areas, the UCF also enabled the analysis and assessment of trending concepts, for example, "Digital Readiness" and "Resilience". Many such concepts can be decomposed into UCF components, and a targeted measure can be created following the streamlined assessment tailoring workflow, thereby reducing assessment development time while also gaining validity support for the new measure from established assessment content. This way, assessment offerings can keep up with ever-changing workplace concepts. To sum up, the modular competency modelling and automated assessment tailoring capability described in this paper have utilities not only cross-sectionally (i.e., useful for many types of applications), but also longitudinally (i.e., capturing trends and changes over time), making it an exciting modular measure with universal applicability.

While this paper focuses solely on self-report forced-choice competency assessments for individuals in the workplace, the general methodology may be applicable to other constructs and/or types of assessments. For instance, there may be utilities in modelling and measuring other psychological constructs (e.g., values, motivation, preferences) of individuals in a similar manner. Even further, there may be possibilities in assessing not individuals, but organizations, cultures, product features, etc., using a similar systematic approach. Enormous potential has been unlocked through a modular, component approach in physical

engineering – from toy models to flat pack furniture to bespoke computer building. Why shouldn't complex psychological measurement benefit from adopting a similar approach, with standardized components but flexible configurations? If achieved, it would be analogous to moving psychometric assessment creation from individual craftsmanship to industrialized production. A similar strategy is being developed in the educational assessment space with efforts focused on accelerating the creation of items through automatic item generation (Gierl & Haladyna, 2013). The psychometric "industrial revolution" is on the horizon, with methodologies as described here building the machinery that will help enable it.

6. Conclusion

This paper presented a new and exciting measurement architecture for mass production of tailored competency models and assessments. For competency modelling, a comprehensive and empirically refined framework of constructs was described to address the bandwidth-fidelity dilemma and customizability-generalizability trade-off, and also ensured long-term re-usability and flexibility of the associated assessment content. For competency assessment, a streamlined assessment tailoring workflow illustrated how automated assessment assembly and quality-check technology can reduce the time and effort required for assessment customization, thus making bespoke assessments more affordable and scalable. The new approach has been trialed and tested in a number of empirical studies, establishing promising initial validity evidence. We truly believe in the utility and future possibilities of this approach in addressing the needs of competency modelling and assessment in the workplace, in a standardized, systematic, efficient and repeatable manner.

7. Conflict of Interest Disclosure

The authors are/were employed by SHL, the organization sponsoring this research. A patent (U.S. Patent No. 10,460,617) has been granted to SHL for the measurement architecture described in this paper, with the first author being one of the inventors. SHL currently offers assessments created using this measurement architecture under the product name of Apta™.

8. References

Bartram, D. (2004). Assessment in organisations. *Applied Psychology, 53*(2), 237-259. https://doi.org/10.1111/j.1464-0597.2004.00170.x

Bartram, D. (2005). The great eight competencies: A criterion-centric approach to validation. *Journal of Applied Psychology, 90*(6), 1185-1203. https://doi.org/10.1037/0021-9010.90.6.1185. PMid:16316273

Bartram, D., Brown, A., Fleck, S., Inceoglu, I. and Ward, K. (2006). *OPQ32 technical manual*. Thames Ditton, UK: SHL.

Bartram, D., Robertson, I. T. and Callinan, M. (2002). Introduction: A framework for examining organizational effectiveness. In I. T. Robertson, M. Callinan & D. Bartram (Eds.), *Organizational effectiveness: The role of psychology* (pp. 1-10). Chichester, UK: Wiley. https://doi.org/10.1002/9780470696736.ch

Bish, A. J., Newton, C. J., Browning, V., O'Connor, P. and Anibaldi, R. (2014). An exploration of the professional competencies required in engineering asset management. *European Journal of Engineering Education, 39*(4), 432-447. https://doi.org/10.1080/03043797.2014.895701

Bosco, F. A., Aguinis, H., Singh, K., Field, J. G. and Pierce, C. A. (2015). Correlational effect size benchmarks. *Journal of Applied Psychology, 100*(2), 431-449. https://doi.org/10.1037/a0038047. PMid:25314367

Brown, A. and Maydeu-Olivares, A. (2011). Item response modeling of forced-choice questionnaires. *Educational & Psychological Measurement, 71*(3), 460-502. https://doi.org/10.1177/0013164410375112

Brown, A. and Maydeu-Olivares, A. (2013). How IRT can solve problems of ipsative data in forced-choice questionnaires. *Psychological Methods, 18*(1), 36-52. https://doi.org/10.1037/a0030641. PMid:23148475

Cao, M. (2016). *Examining the fakability of forced-choice individual differences measures* (Ph.D.). Retrieved from https://www.ideals.illinois.edu/handle/2142/93064

Christiansen, N. D., Burns, G. N. and Montgomery, G. E. (2005). Reconsidering forced-choice item formats for applicant personality assessment. *Human Performance, 18*(3), 267-307. https://doi.org/10.1207/s15327043hup1803_4

Cronbach, L. J. and Gleser, G. C. (1965). The bandwidth-fidelity dilemma. *Psychological Tests and Personnel Decisions*, 97-107.

Driskell, J.E., Hogan, J., Salas, E. and Hoskin, B. (1994). Cognitive and personality predictors of training performance. *Military Psychology*, 6(1), 31-46. https://doi.org/10.1207/s15327876mp0601_2

Gierl, M. A. and Haladyna, T. M. (Eds.). (2013). *Automatic item generation: Theory and practice*. New York, NY: Routledge. https://doi.org/10.4324/9780203803912 PMCid:PMC3336114

Hogan, J. and Roberts, B.W. (1996). Issues and non-issues in the fidelity–bandwidth trade-off. *Journal of Organizational Behavior,* 17(6), 627-637. https://doi.org/10.1002/(SICI)1099-1379(199611)17:6<627::AID-JOB2828>3.0.CO;2-F

Hunter, J. E. and Schmidt, F. L. (2004). *Methods of meta-analysis: Correcting error and bias in research findings* (2nd ed.). Thousand Oaks, CA: Sage. https://doi.org/10.1207/S15327043HUP1304_3

Jackson, D. N., Wroblewski, V. R. and Ashton, M. C. (2000). The impact of faking on employment tests: Does forced choice offer a solution? *Human Performance, 13*(4), 371-388.

Kantrowitz, T. M. (2014). *Global assessment trends report*. Thames Ditton, UK: SHL.

Kantrowitz, T. M., Tuzinski, K. A. and Raines, J. M. (2018). *Global assessment trends report*. Thames Ditton, UK: SHL.

Kasemsap, K. (2016). Analyzing the roles of human capital and competency in global business. In Management Association (Ed.), *Project management: Concepts, methodologies, tools, and applications* (pp. 2190-2218). Hershey, PA: IGI Global. https://doi.org/10.4018/978-1-5225-0196-1.ch109

Kincaid, J. P., Fishburne, R. P., Rogers, R. L. and Chissom, B. S. (1975). *Derivation of new readability formulas (automated readability index, fog count, and flesch reading ease formula) for navy enlisted personnel.* (Research Branch Report No.8–75). Naval Air Station Memphis: Chief of Naval Technical Training. https://doi.org/10.21236/ADA006655. PMCid:PMC432753

Kurz, R. and Bartram, D. (2002). Competency and individual performance: Modeling the world of work. In I. T. Robertson, M. Callinan & D. Bartram (Eds.), *Organizational effectiveness: The role of psychology* (pp. 227-255). Chichester, UK: Wiley. https://doi.org/10.1002/9780470696736.ch10

Le Bon, J. (2015). Training and qualification: Developing a competency model to assess sales leaders' equity. In M. Zeuch (Ed.), *Handbook of human resources management* (pp. 1-16). Berlin, Heidelberg: Springer Berlin Heidelberg.. https://doi.org/10.1007/978-3-642-40933-2_144-1

Mansfield, R. S. (1996). Building competency models: Approaches for HR professionals. *Human Resource Management, 35*(1), 7-18. https://doi.org/10.1002/(SICI)1099-050X(199621)35:1<7::AID-HRM1>3.0.CO;2-2

Martin, B. A., Bowen, C. C. and Hunt, S. T. (2002). How effective are people at faking on personality questionnaires? *Personality and Individual Differences, 32*(2), 247-256. https://doi.org/10.1016/S0191-8869(01)00021-6

McClelland, D. C. (1973). Testing for competence rather than for "intelligence". *American Psychologist, 28*(1), 1-14. https://doi.org/10.1037/h0034092 PMid:4684069

McLagan, P. A. (1988). Flexible job models: A productivity strategy for the information age. In J. P. Campbell, R. J. Campbell & Associates (Eds.), *Productivity in organizations: New perspectives from industrial organizational psychology*. San Francisco, CA: Jossey-Bass.

Ones, D. S. and Viswesvaran, C. (1996). Bandwidth–fidelity dilemma in personality measurement for personnel selection. *Journal of Organizational Behavior,* 17(6), 609-626. https://doi.org/10.1002/(SICI)1099-1379(199611)17:6<609::AID-JOB1828>3.0.CO;2-K

Robertson, I. T. and Kinder, A. (1993). Personality and job competences: The criterion-related validity of some personality variables. *Journal of Occupational and Organizational Psychology, 66*(3), 225-244. https://doi.org/10.1111/j.2044-8325.1993.tb00534.x

Salgado, J. F. and Moscoso, S. (2019). Meta-Analysis of Interrater Reliability of Supervisory Performance Ratings: Effects of Appraisal Purpose, Scale Type, and Range Restriction. Frontiers in Psychology, 10, 2281. https://doi.org/10.3389/fpsyg.2019.02281 PMid:31681087 PMCid:PMC6813221

Saville, P., Sik, G., Nyfield, G., Hackston, J. and Maclver, R. (1996). A demonstration of the validity of the

occupational personality questionnaire (OPQ) in the measurement of job competencies across time and in separate organisations. *Applied Psychology, 45*(3), 243-262. https://doi.org/10.1111/j.1464-0597.1996.tb00767.x

SHL. (2003). *The SHL competency framework: Technical reference manual.* Thames Ditton, UK: SHL.

SHL. (2014). *OPQ32r™ Technical manual*. Thames Ditton, UK: SHL.

SHL. (2019a). *Apta™ Architecture Technical manual*. Thames Ditton, UK: SHL.

SHL. (2019b). *Updated Universal Competency Framework Technical Manual.* Thames Ditton, UK: SHL.

SHL. (2020). *RemoteWorkQ Technical manual*. Thames Ditton, UK: SHL.

Skorková, Z. (2016). *Competency models in public sector.* https://doi.org/10.1016/j.sbspro.2016.09.029

Spencer, L. M., and Spencer, S. M. (1993). *Competence at work*. New York: Wiley.

Testa, M. R., and Sipe, L. (2012). *Service-leadership competencies for hospitality and tourism management* https://doi.org/10.1016/j.ijhm.2011.08.009

Viswesvaran, C., Ones, D. S., and Schmidt, F. L. (1996). Comparative analysis of the reliability of job performance ratings. *Journal of Applied Psychology*, 81(5), 557–574. https://doi.org/10.1037/0021-9010.81.5.557